sommaire du site

http://alain.pichereau.pages.perso-orange.fr
marc.pichereau@wanadoo.fr

Test du *khi-deux*

1- But du test du *khi-deux*

Le problème est le suivant : on cherche à savoir si oui ou non une variable aléatoire X' (ayant k résultats possibles $x_1,x_2,...,x_k$) a pour loi celle de la variable aléatoire X prenant les k valeurs x_i avec $P(X=x_i)=p_i$ pour i variant de 1 à k.

Par exemple on dispose d'un dé numéroté 1 à 6 et on veut savoir s'il est parfaitement équilibré, c'est-à-dire si la probabilité d'obtenir chaque numéro est effectivement 1/6 : ici X' est la variable aléatoire correpondant au dé possédé (on ne connaît pas les probabilités théoriques d'apparition de chaque numéro) et X est la va définie par P(X=i)=1/6 pour i=1 à 6.

Autre exemple : on veut savoir si un générateur de chiffres au hasard génére vraiment des chiffres au hasard, c'est-à-dire si effectivement la probabilité d'apparition du chiffre i est 1/10 pour i=0 à 9.

Le test du *khi-deux* est un outil statistique qui en fait ne permet pas de répondre exactement à la question posée ci-dessus, mais il va permettre de prendre une décision "pratique" :

soit oui, on peut accepter que la loi de X' ait la même loi que X, cela avec un risque de se tromper de t%

soit non, on ne peut accepter que la loi de X' ait la même loi que X, cela avec un risque de se tromper de t%

Le niveau de risque, t%, peut être choisi arbitrairement ; en pratique il est choisi aux environs de 5%.

2-La méthode va reposer sur les 2 résultats théoriques suivants:

R1

d étant un entier supérieur ou égal à 1, la loi du *khi-deux* à d degrés de liberté est la loi suivie par la variable aléatoire KHI2 égale à la somme des carrés de d variables aléatoires gaussiennes

(=normales) centrées, réduites et indépendantes (mutuellement ou globalement) ; c'est en fait la loi $\Gamma(d/2,1/2)$.

Quelques précisions :

la densité est toujours nulle sur $]-\infty;0[$, tend vers 0 en $+\infty$ et

si d=1 la densité est strictement décroissante sur $]0;+\infty[$

si d=2 la densité est strictement décroissante sur $[0;+\infty[$ (il s'agit dans ce cas de la loi exponentielle de paramètre 2 ou (1/2 selon les auteurs))

si $d\geq 3$, la densité est nulle en 0, puis strictement croissante jusqu'à l'abscisse d-2 où elle atteint son maximum, puis strictement décroissante : sa représentation graphique est donc une sorte de courbe en cloche.

Cette variable aléatoire a pour moyenne d (le nombre de degrés) et pour écart-type $\sqrt{(2\times d)}$.

Pour d=1/2/3/4/5 la médiane est respectivement (environ) 0,45/1,39/2,37/3,36/4,35 et pour d>5 elle est environ d-0,66.

Des tables à double entrées d et p donnent le seuil se tel que P(KHI2>se)=p ; par exemple pour d=5, p=0,05 on a se=11,07 (voir ci-dessous une petite table (§3) et deux outils de calculs permettant d'obtenir se ou p en ligne (§4)).

R2

Le théorème de Pearson :

on considère n variables aléatoires indépendantes (mutuellement ou globalement) $X_1,X_2,.....,X_n$ suivant toutes la même loi : celle de X et pour j=1,2,..,k (k>=2) on appelle Y_j la variable aléatoire égale au nombre de X_i (pour i=1,2,...,n) prenant la valeur x_j

alors la variable aléatoire

$$S=\sum_{j=1}^{k}\frac{(Y_j-n\times p_j)^2}{n\times p_j}$$

tend lorsque n tend vers l'infini vers une variable aléatoire suivant une loi du *khi-deux* à k-1 degrés de liberté.

Il s'agit bien ici de k-1, et non de k, degrés de liberté : cela s'explique (en partie) par le fait que si S est effectivement la somme de k carrés, ces k carrés ne sont pas indépendants puisque que la somme des Y_j est la constante n.

On notera (chose étonnante... pour moi) que *le résultat ne dépend pas* des p_j mais uniquement de k .

Si k=1, S est en fait la variable aléatoire toujours nulle.

Pour certains auteurs la loi limite est atteinte si tous les $n\times p_j$ sont ≥ 5, pour d'autres c'est si $n\geq 100$?

Exercice : montrer, pour j=1,2,...,k (k>=2), que Y_j suit une loi binomiale de paramètres n et p_j.

3-Application des 2 résultats précédents à la réalisation du test

Pour savoir si X' a pour loi celle de X on fait n réalisations (indépendantes) de la variable aléatoire X' (donc celle qu'on veut tester) et on note n_i le nombre de fois où on a obtenu la valeur x_i ; la fréquence d'apparition de x_i est donc $f_i=n_i/n$. On calcule alors

$$s=\sum_{j=1}^{k}\frac{(n_j-n\times p_j)^2}{n\times p_j}$$

qui s'écrit aussi

$$s=n\sum_{j=1}^{k}\frac{(f_j-p_j)^2}{p_j}$$

et dans le cas particulier où X suit une loi équiprobale (tous les p_j égaux à 1/k)

$$s=n\times k\times\sum_{j=1}^{k}(f_j-1/k)^2$$

remarque : dans la deuxième formule le coefficient de n s'interpréte comme une sorte (à cause de la division par les p_j) de carré d'une distance euclidienne entre un point dont les coordonnées seraient les f_j et un autre point dont les coordonnées seraient les p_j ; par contre dans le cas où X suit une loi équiprobable alors le coefficient de n×k est vraiment le carré de la distance euclidienne habituelle : pour cette raison d'ailleurs certains notent

$$d^2=\sum_{j=1}^{k}(f_j-1/k)^2 \text{ et donc } s=n\times k\times d^2$$

Dans tous les cas s sera d'autant plus petit (nul) que les f_j seront proches des (égaux aux) p_j.

Si la loi de X' était effectivement celle de X alors s serait une réalisation de la variable aléatoire S parfaitement connue d'un point de vue théorique : par exemple grâce à des tables on peut déterminer le seuil se tel que P(S>se)=5% : pour k=6 (cas où on veut tester un dé, et donc on a une loi du *khi-deux* à 5 degrés de liberté) on a se=11,07. Donc si la valeur trouvée s est >11,07 il serait étonnant que X' ait la même loi que X, car si c'était le cas il y aurait peu de chances (5%) que l'on obtienne un s>11,07 : comme on a dépassé le seuil se c'est qu'il est "peu probable" que X' ait la même loi que X.

On convient alors de décider :

si s > 11,07 on rejette l'hypothèse que X' ait la même loi que X, en rajoutant le commentaire : au risque de se tromper de 5%

si s <11,07 on accepte l'hypothèse que X' ait la même loi que X, en rajoutant le commentaire : au risque de se tromper de 5%

11,07 est appelé le seuil de confiance à 5%

D'une façon générale le seuil de confiance à 1% correspond au 99ième centile, celui à 5% au 95ième centile et celui à 10% au 9ième décile.

Bien entendu si la loi de X est équiprobable (avec k modalités) alors s>se équivaut à d^2>se/(n×k).

Pour mieux comprendre le principe de décision ci-dessus voici 4 remarques :

Remarque 1 : accepter l'hypothèse que X' ait la même loi que X au risque de se tromper de 5% (par exemple) ne signifie pas que la probabilité que X' ait la même loi que X est 95%.....car en fait l'événement X' a la même loi que X n'est pas véritablement un événement aléatoire : X' a la même loi que X ou pas.

Remarque 2 : accepter l'hypothèse que X' ait la même loi que X ne signifie pas qu'il n'existe pas de loi de probabilité coïncidant mieux que la loi de X avec les observations de la loi testée X', mais seulement on n'a pas de raisons importantes de rejeter cette hypothèse.

Remarque 3 : si on rejette pour X' la loi de X, on ne sait pas pour autant quelle est la vraie loi de X', mais seulement que l'hypothèse X' a pour loi celle de X est peu vraisemblable

Donc, en aucun cas le test du khi2 ne permet d'affirmer avec certitude que X' suit la loi de X ou que X' ne suit pas la loi de X.

Remarque 4 : le 5% de l'exemple ci-dessus est rigoureusement la probabilité conditionnelle de rejeter l'hypothèse que la loi de X' soit celle de X sachant que la loi de X' est effectivement celle de X : c'est la probabilité de l'erreur de 1ère espèce ; 95% étant évidemment la probabilité d'accepter l'hypothèse que la loi de X' soit celle de X sachant que la loi de X' est effectivement celle de X.

Et qui dit 1ère espèce dit 2ième espèce : l'erreur de 2ième espèce est l'erreur qui consiste à accepter pour X' la loi de X sachant que la loi de X' n'est pas celle de X. La probabilité de cette erreur est beaucoup plus délicate à évaluer et il n'y a pas de relation très simple entre les deux probabilités d'erreur de 1ère et de 2ième espèce.

En fait, en toute rigueur la probabilité d'erreur de 2ième espèce dépend du choix (autre que la loi de X) d'une loi pour X' : ca se complique...

Cependant diminuer la probabilité de l'erreur de 1ère espèce va augmenter la probabilité de l'erreur de 2ième espèce : en effet, diminuer la probabilité de l'erreur de 1ère espèce revient à augmenter le seuil se et donc s a "plus de chances" d'être inférieur à se, cela quelque soit la loi de X'.

Ce résultat permet de voir tout de suite qu'augmenter considérablement le seuil se pour avoir un risque de se tromper (=la probabilité de l'erreur de 1ère espèce) quasiment nul est illusoire, car parallèlement on a augmenté la probabilité de l'erreur de 2ième espèce : le choix du seuil est donc en pratique le résultat d'un compromis, l'usage (qui satisfait en pratique les

utilisateurs) étant semble-t-il de faire en sorte que le risque de se tromper soit aux alentours de 1 ou 5%.

D'ailleurs en prenant un seuil "très très grand", on est "presque sûr qu'à tout coup" on acceptera que X' ait pour loi celle de X avec un risque de se tromper quasiment nul....ce qui n'est pas très réaliste.

Remarque 5 : je réponds là à une question posée par un lecteur qui déçu de constater que le test du khi2 ne permettait pas de décider à coup sûr (ou à 99%) que X' suit ou pas la loi de X me proposa la méthode suivante : on estime l'espérance et l'écart-type de X' et on compare avec ceux de X. Certes on peut estimer espérance et écart-type (voir exemple 3), mais le problème c'est que moyenne et écart-type ne suffisent pas à caractériser une loi de variable aléatoire. Par exemple n'importe quelle variable aléatoire de moyenne m et d'écart-type σ a même moyenne et écart-type que la loi normale de moyenne m et ecart-type σ, et pourtant toute variable aléatoire ne suit pas une loi normale.

TABLE des seuils du *khi-deux* pour quelques degrés de liberté (d) et quelques probabilités (p)

par exemple P(KHI2>15,09)=0,01 pour 5 degrés de liberté

$d\|^p$	0,10	0,05	0,01
1	2,71	3,84	6,63
2	4,61	5,99	9,21
3	6,25	7,81	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81
10	15,99	18,31	23,21
12	18,55	21,03	26,22
24	33,2	36,42	42,98

4-Calculs en ligne

Pour ceux souhaitant des valeurs complémentaires (ou souhaitant vérifer les valeurs ci-dessus!) voici deux outils de calculs en ligne sur la loi du *khi-deux* (ou hors-ligne si vous enregistrez cette page!). Ils sont réalisés grâce à 2 programmes écrits en Java-script en provenance du site de John Walker. Je me suis contenté d'expliciter en français les messages apparaissant à l'écran et de

changer la présentation (dans le script original il est écrit " Both the original C code and this JavaScript edition are in the public domain").

Bien entendu il ne faut pas avoir désactivé l'éxécution des programmes Java-script : dans le champ ci-dessous vous allez savoir si oui ou non Java-script est activé dans votre navigateur.

Java-script n'est pas activé dans votre navigateur.

-4.1 Pour un degré de liberté d et un seuil se donnés calcul de P(KHI2>se)

Entrez (mettre un point pour la virgule) le seuil se dans ce champ et le degré d dans celui-là

cliquez pour lancer le calcul

P(KHI2>se)=

-4.2 Pour un degré de liberté d et une probabilité p donnés calcul du seuil se tel que P (KHI2>se)=p

Entrez la probabilité p dans ce champ et le degré d dans celui-là

cliquez pour lancer le calcul

Le seuil se tel que P(KHI2>se)=p est égal à

Pour les curieux de la programmation, le programme Java-Script est constitué de 4 fonctions principales :

pos(z)=probabilité qu'une va normale centrée réduite soit inférieure à z (l'intégration de $\exp(-x^2/2)$ est faite grâce à 2 approximations polynomiales : précision annoncée 10^{-6}).

pochisq(x,df)= probabilité qu'une variable suivant la loi du *khi-deux* à df degrés de libertés soit supérieure à x ; uniquement dans le cas où df est impair cette fonction fait appel à la précédente.

critchi(p,df)= le seuil x tel que pochisq(x,df)=p ; il est obtenu à l'aide d'une dichotomie à partir de la valeur initiale x=df/sqrt(p).

trimfloat(x,d) réalise l'affichage d'un nombre x avec d chiffres après la virgule.

5- Exemples

Exemple 1

On dispose d'une pièce de monnaie et on veut savoir si elle est équilibrée ou pas, c'est-à-dire si la variable aléatoire X' correspondant à un lancer de cette pièce a oui ou non la loi de X : P(X=pile)

=1/2 et P(X=face)=1/2. La variable aléatoire S (voir §2) suit une loi du *khi-deux* à 1 degré de liberté, le seuil de confiance à 5% étant 3,84 (P(KHI2>3,84)=5%).

Après avoir fait n=100 (ou 10 car tous les $10\times p_j$ sont ≥ 5) lancers de cette pièce on calcule s (voir le §3) :

$$d^2=\sum_{j=1}^{2}(f_j-1/2)^2 \text{ puis } s=2\times n\times d^2$$

si s < 3,84 on dira que la pièce n'est pas truquée, au risque de se tromper de 5%

si s > 3,84 on dira que la pièce est truquée, au risque de se tromper de 5%

Remarque 1 : dans ce cas particulier où k=2 on a $(f_2-1/2)^2=(f_1-1/2)^2$ et donc $2\times n\times d^2<3,84$ équivaut à $|f_1-1/2|< \sqrt{3,84}/(2\times\sqrt{n})$ soit environ $1,96/(2\times\sqrt{n})$. On retrouve ainsi (par approximation de la loi de S par une loi du *khi-deux*) le résultat général suivant : la probabilité que la fréquence observée d'un événement A de probabilité p (ici A c'est pile et p=1/2) soit autour de p à $0,98/\sqrt{n}$ près est 95% : c'est le fameux intervalle de confiance à 95% (bien souvent 0,98 est remplacé par 1).

Rappelons la méthode habituelle pour obtenir ce résultat. Soit Z la variable aléatoire égale au nombre de fois où l'événement A de probabilité p s'est réalisé lors de n réalisations (indépendantes) d'une même épreuve aléatoire, alors Z suit une loi binomiale de paramètre n et p, et $(Z-n\times p)/\sqrt{(n\times p\times q)}$ converge en loi vers la loi normale centrée réduite. Donc pour n grand (n>30 et 15/n < p <1-15/n, mais tout le monde ne donne pas les mêmes conditions) on peut écrire

$$P(|Z/n-p|<\frac{k\sqrt{(p\times q)}}{\sqrt{n}})\approx\frac{1}{\sqrt{(2\pi)}}\int_{-k}^{k}\exp(-x^2/2)dx$$

D'où en faisant p=0,5 et k=1,96 on obtient $P(|Z/n-p|<0,98/\sqrt{n})\approx 0,95$, ce qui est bien le résultat obtenu précedemment puisque Z/n n'est autre que la fréquence.

Mais on peut en déduire aussi que pour tout $\varepsilon > 0$ la limite, lorsque n tend vers +infini, de $P(|Z/n-p|\leq\varepsilon)$ est 1, c'est-à-dire la fréquence d'apparition d'un événement converge en probabilité vers la probabilité de cet événement : c'est la loi faible des grands nombres. Laquelle peut aussi s'obtenir par application de l'inégalité $P(|Z/n-p|\leq\varepsilon) > 1-p\times q/(n\times\varepsilon^2)$ de Bienaymé-Tchebitchev.

Remarque 2 : et si on n'a pas sous la main une table du *khi-deux* , ni les merveilleux [:-)] outils de calculs ci-dessus, comment faire pour trouver le seuil voulu, par exemple celui à 5%?

On peut songer à faire une simulation de cette loi du *khi-deux* , c'est-à-dire calculer, par exemple , un millier de s et à chercher le 95 ième centile de la série obtenue. Mais en fait il ne s'agit pas de calculer ces 1000 nombres s en faisant à chaque fois n=100 réalisations de la variable aléatoire X' (ici lancer de la pièce dont on cherche à savoir si elle est truquée ou pas) mais avec n=100 réalisations de X, c'est-à-dire en faisant n lancers d'une pièce non truquée, car S suit une loi du *khi-deux* sous réserve que les X_i suivent la même loi que X (puisque les p_j intervenant dans S sont relatifs à X; en fait il suffit de générer une variable aléatoire X prenant 2 valeurs distinctes, peu importe sa loi de probabilité pourvue qu'elle soit connue, puisque le résultat du §3 ne dépend que de k). Je trouve que cet aspect n'est pas assez signalé dans certains ouvrages.

En outre, quitte à faire 1000×100 réalisations de X (à condition de disposer, pour cet exemple, d'une pièce non truquée!), pourquoi ne pas faire 100000 réalisations de X'? En effet les fréquences observées donneront alors une bonne idée de la loi de probabilité de X' puisque d'après la remarque précédente (et en notant que $1/\sqrt{100000}$ est peu différent de 0,003) pour tout j il y a 95% de chances que p'_j soit dans l'intervalle $[f_j-0,003;f_j+0,003]$: on pourrait aller alors juger de façon immédiate s'il y a adéquation avec la loi de X.

En fait l'intérêt du test du *khi-deux* est, je pense, de pouvoir donner une conclusion relativement pertinente grâce au calcul d'un seul s (avec un n pas considérablement élevé).

Et donc simuler la loi du *khi-deux* pour trouver un seuil, puis faire ensuite le test du *khi-deux*, c'est perdre une grosse partie de l'intérêt de ce test.

Exemple 2

Sous excel la formule ENT(10*ALEA()) génére au hasard les chiffres 0;1;2..;9 : on veut savoir si la probabilté d'obtenir le chiffre i est bien 1/10.

Ici X' est la variable aléatoire correspondant au générateur ENT(10*ALEA()) et X la variable aléatoire définie sur {0;1;2;3;4;5;6;7;8;9} par P(X=i)=1/10. La variable aléatoire S (voir §2) suit une loi du *khi-deux* à 9 degrés de liberté, le seuil de confiance à 5% étant 16,9 (P(KHI2>16,9) =5%). Pour n=100 on calcule s (voir §3) : pour cela, dans une cellule on tape =ENT(10*ALEA()) et on propage 99 fois cette formule, puis grâce à la fonction NB.SI on calcule les n_i puis les f_i et enfin s.

Le premier s que j'ai obtenu a été 10,4 : puisque s<16,9 on peut dire que le générateur X' suit la loi X, au risque de se tromper de 5%.

Par curiosité j'ai ensuite fait une série de 50 calculs de s : parmi eux 6% soit 3 résultats 17,2 / 17,4 / 20,6 ont été supérieurs à 16,9 et la moyenne de ces 50 résultats a été de 9,54 (pour une valeur théorique de 9).

Une autre série de 50 calculs a donné 2% soit un seul résultat (18,8) qui a été supérieur à 16,9 pour une moyenne de 8,5.

Globalement, on serait donc amené à considérer que la formule ENT(10*ALEA()) génére effectivement des nombres au hasard. Mais ce raisonnement est-il vraiment justifié? En effet S suit une loi du *khi-deux* sous réserve que les variables aléatoires X_i sont indépendantes ; et donc les n réalisations de X' (qui servent à calculer s) doivent être n réalisations indépendantes de X', c'est-à-dire ici le résultat donné par un aléa doit être indépendant des résultats donnés par les aléa précédents. Or le test du *khi-deux* ne peut tester cet aspect puisqu'il ne prend pas en compte l'ordre d'apparition des résultats. Donc sur cet exemple on applique une méthode sans être sûr que les hypothéses la justifiant soient vérifiées : c'est pour le moins gênant!

D'ailleurs pour montrer que le test du *khi-deux* ne donne pas toujours un résultat pertinent certains considèrent le générateur de chiffres 0,1,2,...9 un peu particulier suivant : il donne d'abord 0 puis, après tout chiffre n≥0, apparaît toujours le chiffre n+1 si n≤8 sinon c'est 0. La suite des chiffres obtenus est donc de période 10 : 0,1,2,3,4,5,6,7,8,9,0,1,2,3.....

Quelque soit n multiple de 10 (petit ou grand) on aura f_j=1/10, soit s=0 et donc, à n'importe quel risque de se tromper, on acceptera l'hypothèse que ce générateur ...produit des nombres au hasard, ce qui est faux puisque chaque chiffre donné par ce générateur dépend du chiffre précédent par une formule simple (cependant il donne bien des chiffres équirépartis!). En fait la séquence obtenue est entiérement déterminée : le terme de rang i est le reste de la divison de i par 10 diminué de 1 (sauf lorsque le reste est 0, auquel cas le terme est 9). Puisque les chiffres donnés par ce générateur ne correspondent pas à des réalisations indépendantes d'une même variable aléatoire X', il n'est pas justifié de lui appliquer le test du *khi-deux*.

De même il n'est pas justifié, à mon avis, de l'appliquer pour tester l'ALEA d'excel puisque on ne sait pas si les réalisations successives de cet ALEA sont vraiment indépendantes.

On ne peut appliquer le test du *khi-deux* que si l'on est sûr que les n nombres à partir desquels on calcule s correspondent à n réalisations indépendantes d'une même variable aléatoire X'.

(ce qui est bien le cas de l'exemple 1 : lancer n fois une pièce (truquée ou pas), c'est faire n réalisations indépendantes d'une même variable aléatoire).

Exemple 3

En fait ce test du *khi2* peut aussi servir à tester l'adéquation avec une loi continue : uniforme, normale, exponentielle....

Supposons qu'on dispose d'une série de n factures réparties en 8 classes d'effectifs n_j : la fréquence d'appartenance à la classe j est donc $f_j=n_j/n$. On cherche à savoir si cette distribution peut être approximée par une loi normale.

Il faut d'abord estimer la moyenne m et l'écart-type σ de cette loi normale : on prend les estimateurs habituels, c'est-à-dire on prend pour m la moyenne de la série observée et pour σ on prend $\sigma'\times\sqrt{(n/(n-1))}$ où σ' est l'écart-type de la série observée.

On est alors en mesure de calculer la probabilité p_j que la variable aléatoire X suivant la loi normale (m,σ) soit dans la classe j. Et donc pour voir s'il y a adéquation entre la distribution des factures et cette loi normale (m,σ) il suffit de calculer le s du §3 avec les f_j et p_j ci-dessus (k étant égal à 8) et de le comparer au seuil correspondant à la précision voulue.

En espérant ne pas avoir dit trop de bêtises (ne pas hésiter à me les signaler : voir mon adresse à sommaire du site).
Alain Pichereau