Test d'équirépartition, quel risque d'erreur ?[1]

Louis-Marie Bonneval

Le test d'équirépartition est une notion au programme de terminale S et ES. Un aspect délicat est l'interprétation du résultat observé.
On lit dans certains énoncés[2] : *Avec un risque d'erreur de 10 %, doit-on accepter ou refuser l'hypothèse d'équirépartition ?*
Ce qui suit voudrait montrer que cette formulation n'est pas correcte, car le risque d'erreur n'est pas le même selon qu'on accepte ou qu'on rejette l'hypothèse[3].

On va considérer d'abord deux exemples : une épreuve à deux issues, puis une épreuve à trois issues. Ensuite on traitera le cas général d'une épreuve à k issues.

I. Un premier exemple : garçon ou fille ?

Posons-nous la question : quand on attend la naissance d'un enfant, y a-t-il autant de chances de voir naître un garçon qu'une fille ?
Avant de tenter une réponse, notons que cette réponse ne pourra pas être simplement oui ou non. En effet l'équiprobabilité n'existe pas dans la Nature[4] : c'est une notion mathématique, autrement dit un *modèle*. La question signifie en fait : « Le modèle d'équiprobabilité est-il adapté pour le sexe d'un enfant à naître ? » C'est pourquoi elle ne peut attendre qu'une réponse de physicien : « oui, à un certain degré de précision ; non, à un degré de précision plus fin »[5].

(1) Merci à Jean-François Kentzel, qui par son travail approfondi, ses recherches et ses commentaires, a fortement contribué à cet article.
(2) Voir le bac ES de juin 2003 en métropole (exercice 1, question 4), téléchargeable sur : http://www.maths.ac-aix-marseille.fr/annales/bac/annales03.htm.
et les « sujets zéro » diffusés par l'inspection générale en novembre 2003 (exercice commun pour les séries S et ES « le meunier », troisième question) téléchargeable sur : http://www.eduscol.education.fr/D0056/exmathsES-S.htm.
(3) Une simple remarque permet d'ailleurs de s'en rendre compte : si la démarche préconisée en terminale conduit à accepter l'hypothèse d'équiprobabilité « au seuil de 10 % » (c'est-à-dire en utilisant le 9^{e} décile), la même démarche conduirait à l'accepter aussi « au seuil de 1 % » (c'est-à-dire en utilisant le 99^{e} centile, qui est plus grand). Cela montre bien que ce 10 % ne peut pas être le risque d'erreur quand on accepte l'hypothèse d'équiprobabilité.
(4) On pourrait discuter sur la différence entre équiprobabilité et équirépartition. À strictement parler, le deuxième mot peut traduire une réalité objective : dans une population où il y a des individus de plusieurs catégories, ils peuvent être équirépartis ou non. Mais dans les situations utiles où s'applique le test, il s'agit d'équirépartition approchée, et c'est bien le modèle d'équiprobabilité qui est en cause. En fait, une bonne image mentale de l'équiprobabilité est celle d'une urne où les boules de différentes couleurs sont équiréparties.
(5) On pourrait faire une analogie avec la géométrie : à la question « ma chambre est-elle carrée ? » il se peut que je réponde oui ou non selon que je mesure les longueurs au décimètre près ou au centimètre près.

L'outil de mesure est bien entendu statistique : on observe sur un grand nombre de naissances les proportions de garçons et de filles ; si elles sont voisines de $\frac{1}{2}$, on répondra oui, sinon on répondra non. C'est ce qu'on appelle *faire un test*.

Cette démarche très naturelle soulève au moins trois difficultés :
• Que veut dire un grand nombre de naissances ?

• Que veut dire voisin de $\frac{1}{2}$?

• Que faire, si sur certains registres de naissances, les proportions sont voisines de $\frac{1}{2}$ et sur d'autres non ?

Il nous faut donc préciser les choses :

Notons e l'épreuve aléatoire qui consiste à observer le sexe d'un enfant à naître, p la probabilité (inconnue) de naissance d'un garçon[6]. L'épreuve e étant une alternative, la répartition de probabilité[7] sur son univers {G, F} est entièrement déterminée par le seul nombre p.
Observer n naissances, c'est répéter n fois l'épreuve e. Notons e^n cette épreuve. Une issue de e^n, c'est-à-dire un élément de $\{G, F\}^n$, est un *échantillon*[8] *de taille n* de l'épreuve e.
Notons F_1 la variable aléatoire qui à chaque échantillon possible associe la proportion[9] de garçons. Comme nF_1 suit la loi binomiale $B(n, p)$, F_1 a pour espérance p et pour écart-type $\sqrt{\frac{p(1-p)}{n}}$: cela indique d'une part que F_1 fluctue autour de p, d'autre part que sa dispersion diminue comme $\frac{1}{\sqrt{n}}$ quand la taille n de l'échantillon augmente.
La valeur de F_1 obtenue après observation d'un échantillon sera notée f_1.

1) Tester un modèle

Pour tester l'hypothèse d'équiprobabilité $p = \frac{1}{2}$, il faut comparer f_1 à $\frac{1}{2}$: si f_1 s'écarte *significativement* de $\frac{1}{2}$, autrement dit si la valeur de $\left|f_1 - \frac{1}{2}\right|$ est trop élevée, on

(6) d'une fille si on préfère…
(7) On dit aussi *distribution de probabilité*. Je suggère de réserver l'expression *loi de probabilité* à une variable aléatoire : on peut l'utiliser (comme le fait le programme de première) si on code numériquement les issues, par exemple ici par 1 et 0.
(8) Ce mot évoque les tirages dans une urne. De fait, on peut se représenter l'épreuve e comme le tirage d'une boule dans une urne comportant une proportion p de boules marquées G ; l'épreuve e^n consiste alors à tirer n fois de suite une boule avec remise.
(9) F est bien sûr l'initiale de « fréquence ».

rejettera l'hypothèse d'équiprobabilité.
Mais que signifie une valeur trop élevée ?
C'est à l'appréciation de l'utilisateur :
• Il peut décider qu'elle est trop élevée si, sur l'ensemble des échantillons possibles, elle fait partie des 1 % les plus grandes[10]. Autrement dit si elle dépasse le *quantile* d'ordre 0,99 de $\left|F_1 - \frac{1}{2}\right|$.
• S'il est plus exigeant, il peut décider qu'une valeur trop élevée est une valeur faisant partie des 5 % les plus grandes. Autrement dit si elle dépasse le quantile d'ordre 0,95 de $\left|F_1 - \frac{1}{2}\right|$. Ce *niveau de confiance*, ou *niveau de signification*, de 95 % est souvent choisi par les statisticiens.
• S'il est encore plus exigeant, il peut considérer que les valeurs trop élevées sont les 10 % plus grandes, autrement dit celles qui dépassent le quantile d'ordre 0,90 de $\left|F_1 - \frac{1}{2}\right|$. Ce niveau de confiance de 90 % est celui que nous retiendrons dans la suite, conformément à ce que suggère le programme de Terminale.

Appelons r ce quantile d'ordre 0,9 de $\left|F_1 - \frac{1}{2}\right|$. Il est défini[11] par

$$P\left(\left|F_1 - \frac{1}{2}\right| \leq r\right) = 0{,}9.$$

La lettre P désigne la répartition de probabilité sur l'univers $\{G, F\}^n$, déduite de la répartition de probabilité $\left(\frac{1}{2}, \frac{1}{2}\right)$ sur l'univers $\{G, F\}$, autrement dit l'équiprobabilité sur $\{G, F\}^n$.
La question de la détermination pratique de r est traitée au paragraphe IV.

On rejettera donc le modèle d'équiprobabilité si $\left|f_1 - \frac{1}{2}\right| > r$. On juge en effet que l'écart entre f_1 et $\frac{1}{2}$ est *significatif (au niveau de 90 %)*, c'est-à-dire trop important pour être dû au hasard du choix de l'échantillon.

(10) La plus grande valeur possible de $\left|f_1 - \frac{1}{2}\right|$ est $\frac{1}{2}$: elle est obtenue quand f_1 prend la valeur 0 ou la valeur 1, c'est-à-dire quand l'échantillon ne comporte que des filles ou que des garçons. Cet événement est possible, mais très peu probable : sous l'hypothèse d'équiprobabilité, sa probabilité est $\frac{2}{2^n} = \frac{1}{2^{n-1}}$.

(11) $\left|F_1 - \frac{1}{2}\right|$ étant une variable aléatoire discrète, l'existence et l'unicité de r appelleraient quelques commentaires, plus théoriques que pratiques, que nous n'aborderons pas ici.

Quelle décision prendre si $\left|f_1 - \frac{1}{2}\right| \leq r$? On accepte alors l'hypothèse d'équiprobabilité. Mais il serait abusif d'en déduire que l'équiprobabilité est le seul modèle valable : ce modèle n'a pas été contredit par le résultat du test, mais d'autres modèles peuvent parfaitement être valables aussi. La valeur la plus naturelle pour p serait d'ailleurs f_1 (on démontre que c'est la meilleure *estimation ponctuelle* de p). Les valeurs de p acceptables[12] au niveau de confiance de 90 % se répartissent autour de f_1 et constituent la *fourchette de sondage*[13] $[f_1 - r\ ;\ f_1 + r]$. Tester l'hypothèse $p = \frac{1}{2}$ consiste à observer si cette fourchette contient $\frac{1}{2}$; ou, ce qui est équivalent, si *l'intervalle de dispersion* $\left[\frac{1}{2} - r; \frac{1}{2} + r\right]$ contient f_1.

On dit que $\left|\mathrm{F}_1 - \frac{1}{2}\right|$ est la *variable de décision*, et r la *valeur critique* (c'est-à-dire celle qui sépare les deux cas). On peut dire aussi que F_1 est la variable de décision[14], l'intervalle $\left[\frac{1}{2} - r; \frac{1}{2} + r\right]$ la *zone d'acceptation* de l'hypothèse $p = \frac{1}{2}$, son complémentaire dans [0 ; 1] la *zone de rejet* de cette hypothèse.

La décision de rejeter l'hypothèse d'équiprobabilité comporte un risque d'erreur : celui de rejeter ce modèle alors qu'il est pertinent. D'une urne où les boules sont équiréparties, il est en effet possible de tirer un échantillon fortement dissymétrique. Ce risque est $\mathrm{P}\left(\left|\mathrm{F}_1 - \frac{1}{2}\right| > r\right)$: il est égal ici à 10 %.

Inversement, il est tentant de se demander quel est le risque d'accepter le modèle d'équiprobabilité alors qu'il n'est pas adapté. Posée sans plus de précision, cette question *n'a pas de réponse* : en effet si on rejette l'hypothèse $p = \frac{1}{2}$ sans proposer une autre valeur de p, on ne dispose plus d'un modèle pour l'épreuve e, on ne peut donc plus définir de répartition de probabilité sur l'univers $\{\mathrm{G}, \mathrm{F}\}^n$. L'événement $\left\{\left|\mathrm{F}_1 - \frac{1}{2}\right| \leq r\right\}$ est défini, puisque l'univers est toujours $\{\mathrm{G}, \mathrm{F}\}^n$, mais sa probabilité ne l'est pas. Comme si on tirait un échantillon dans une urne de composition non précisée.

La situation serait tout autre si on *opposait deux modèles*.

(12) Le modèle étant caractérisé par le seul nombre p, la théorie des tests rejoint ici la théorie de l'estimation.

(13) L'*intervalle de confiance* quant à lui est l'intervalle aléatoire $[\mathrm{F}_1 - r\ ;\ \mathrm{F}_1 + r]$; sa valeur observée sur l'échantillon prélevé est donc la fourchette de sondage.

(14) On pourrait aussi prendre comme variable de décision l'effectif de garçons, ce qui conduit à multiplier par n les bornes des intervalles.

2) Choisir entre deux modèles

Posons-nous maintenant la question[15] : le modèle défini par $p = 0{,}51$ est-il meilleur que le modèle défini par $p = 0{,}5$?

La variable de décision étant toujours F_1, la zone d'acceptation de l'hypothèse $p = 0{,}5$ ne peut plus être un intervalle centré en 0,5.
Une idée naturelle serait de choisir 0,505 comme valeur critique : si $f_1 > 0{,}505$, on rejettera l'hypothèse $p = 0{,}5$; si $f_1 \leq 0{,}505$, on acceptera l'hypothèse $p = 0{,}5$.

• Le risque de rejeter l'hypothèse $p = 0{,}5$ alors qu'elle est pertinente est alors $P(F_1 > 0{,}505)$, où P désigne comme plus haut la répartition de probabilité sur $\{G, F\}^n$ induite par $p = 0{,}5$. On l'appelle le *risque de première espèce*, et on le note traditionnellement α.

• Le risque d'accepter l'hypothèse $p = 0{,}5$ alors que l'autre hypothèse $p = 0{,}51$ est pertinente est $P'(F_1 \leq 0{,}505)$, où P′ désigne la répartition de probabilité sur $\{G, F\}^n$ induite par $p = 0{,}51$. On l'appelle le *risque de deuxième* espèce, et on le note traditionnellement β.

Bien entendu, α et β dépendent tous les deux de n. On trouve par exemple[16] :

taille n de l'échantillon	risque de première espèce α	risque de deuxième espèce β
1 000	0,36	0,39
10 000	0,16	0,16
16 000	0,10	0,10
25 000	0,06	0,06
100 000	0,0008	0,0008

On constate qu'ici α et β sont très voisins : cela s'explique par le fait que pour n assez grand les lois de F_1 engendrées par les probabilités P et P′ sont pratiquement symétriques par rapport à 0,505. Ce tableau montre que les deux risques d'erreur diminuent quand la taille de l'échantillon augmente : ainsi il faut observer au moins 25 000 naissances pour pouvoir trancher sans trop de risque entre les deux modèles.

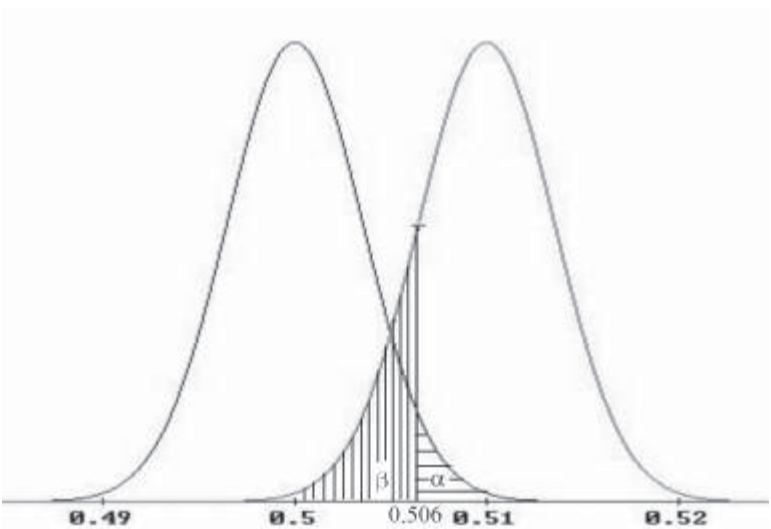
Mais la démarche qui vient d'être exposée, où l'on a pris le nombre 0,505 comme valeur critique, n'est pas celle qui est utilisée couramment. En effet elle symétrise les deux risques, ce qui en général n'est pas souhaitable d'un point de vue pratique : on veut surtout savoir si p peut être choisi égal à 0,5. C'est donc le risque de rejeter à tort cette hypothèse qu'on veut minimiser. L'usage est donc de se fixer *a priori* le risque de première espèce α, et d'en déduire la valeur critique comme quantile d'ordre $1 - \alpha$ de $F_1 - 0{,}5$ (dans le modèle d'équiprobabilité). On calcule ensuite le risque de deuxième espèce β, essentiellement différent de α.

Supposons par exemple qu'on se fixe $\alpha = 0{,}1$.

(15) Les démographes proposent souvent cette valeur 0,51 (voire la valeur 0,512) comme probabilité de naissance d'un garçon.
(16) Voir les calculs en [13]. Pour n = 1 000 on a utilisé les loi binomiales $B(n, 0{,}5)$ et $B(n, 0{,}51)$. Pour les autres valeurs de n on a utilisé leurs approximations normales.

Pour $n = 10\,000$, le quantile d'ordre 0,9 de $F_1 - 0,5$ dans le modèle d'équiprobabilité est 0,0064.
Donc :
• si f_1 dépasse 0,5064, on rejette l'hypothèse d'équiprobabilité, avec un risque d'erreur $\alpha = 0,1$.
• si f_1 ne dépasse pas 0,5064, on accepte l'hypothèse, avec un risque d'erreur[17] $\beta \approx 0,24$.

Pour $n = 25\,000$ le quantile d'ordre 0,9 de $F_1 - 0,5$ dans le modèle d'équiprobabilité est 0,004 0.
Donc :
• si f_1 dépasse 0,504, on rejette l'hypothèse d'équiprobabilité, avec un risque d'erreur $\alpha = 0,1$.
• si f_1 ne dépasse pas 0,504, on accepte l'hypothèse, avec un risque d'erreur[18] $\beta \approx 0,03$.



La courbe ci-dessus et celle de la page suivante représentent les densités des lois normales qui approximent respectivement les lois binomiales B(0,5 ; n) et B(0,51 ; n). L'échelle verticale est telle que l'aire sous chaque courbe soit égale à 1. La « largeur » de chaque courbe (écart entre les deux points d'inflexion) est exactement $\frac{1}{\sqrt{n}}$ pour la première, pratiquement $\frac{1}{\sqrt{n}}$ pour la deuxième[19] : c'est pourquoi dans le deuxième schéma les courbes sont plus étroites que dans le premier, et donc plus discriminantes.

(17) Voir les calculs en [13].
(18) Voir le calcul en [13].
(19) Respectivement $2\sqrt{\frac{0,5 \times 0,5}{n}}$ et $2\sqrt{\frac{0,51 \times 0,49}{n}}$.

3) Opposer un modèle à tous les autres

On peut opposer le modèle $p = 0{,}5$ à un autre modèle $p = p'$, où $p' \neq 0{,}5$ n'est pas précisé.

Reprenons alors $\left|F_1 - \frac{1}{2}\right|$ comme variable de décision, et r comme valeur critique : si $\left|f_1 - \frac{1}{2}\right| > r$, on rejettera l'hypothèse $p = 0{,}5$; si $\left|f_1 - \frac{1}{2}\right| \leq r$, on acceptera l'hypothèse $p = 0{,}5$.

Le risque de première espèce est $\alpha = \mathrm{P}\left(\left|F_1 - \frac{1}{2}\right| > r\right)$.

Le risque de deuxième espèce est $P'\left(\left|F_1 - \frac{1}{2}\right| \leq r\right)$, où P′ désigne la mesure de probabilité sur $\{G, F\}^n$ induite par p'. On peut le noter $\beta(p')$, et chercher à le majorer[20] quand p' décrit $[0;1] \setminus \left\{\frac{1}{2}\right\}$.

4) Opposer un modèle à certains autres

Posons-nous maintenant la question : quand on attend la naissance d'un enfant, y a-t-il plus de chances de voir naître un garçon qu'une fille ?
Il s'agit alors d'opposer le modèle $p = 0{,}5$ à tous les autres modèles $p = p'$ pour lesquels $p' > 0{,}5$. Les statisticiens parlent dans ce cas de *test unilatéral* (par opposition au *test bilatéral* ci-dessus).

La variable de décision doit alors être $F_1 - \frac{1}{2}$ (et non plus $\left|F_1 - \frac{1}{2}\right|$) : si $f_1 - \frac{1}{2} \leq r'$, on acceptera l'hypothèse $p = 0{,}5$; si $f_1 - \frac{1}{2} > r'$, on rejettera l'hypothèse $p = 0{,}5$. La valeur critique r' est déterminée par le niveau de confiance adopté : si on l'a choisi égal à 0,9, on prend pour r' le quantile d'ordre 0,9 de $F_1 - \frac{1}{2}$ dans le modèle d'équiprobabilité.

Le risque de première espèce $\alpha = P\left(F_1 - \frac{1}{2} > r'\right)$ est alors égal à 10 %.

Le risque de deuxième espèce $P'\left(F_1 - \frac{1}{2} \leq r'\right)$ dépend de p' : on peut comme ci-dessus le noter $\beta(p')$, et chercher à le majorer[21] quand p' décrit $\left]\frac{1}{2};1\right]$.

II. Un deuxième exemple : liquide, chèque ou carte bancaire ?

Pour ce deuxième exemple on a choisi une épreuve à trois issues. Travailler sur des triplets permet en effet une interprétation géométrique éclairante. La généralisation à k issues sera ensuite plus facile.

La direction d'un hypermarché se pose la question suivante : les trois types de paiement utilisés par les clients (liquide, chèque, carte bancaire) sont-ils équirépartis ?

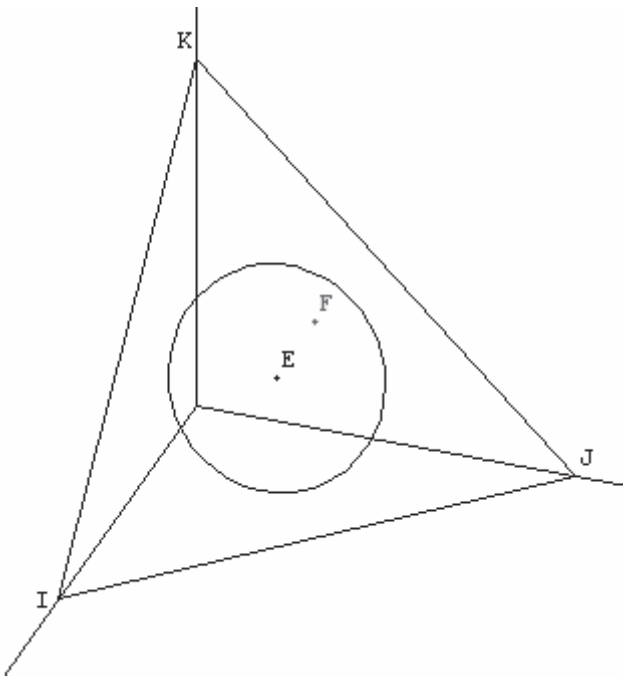
On peut faire les mêmes remarques que plus haut : la réponse ne peut être simplement oui ou non, elle dépendra de la précision que l'on peut atteindre, c'est-à-dire à la fois du *nombre d'observations* dont on dispose et du *niveau de confiance* choisi.

(20) Voir [13].
(21) Voir [13].

Appelons e l'épreuve qui consiste à observer le paiement d'un client à la caisse : elle a trois issues L, C, B, de probabilités (inconnues) p_1, p_2, p_3.
Observer n paiements consiste à répéter n fois l'épreuve e : appelons e^n cette épreuve. Une issue de e^n, c'est-à-dire un élément de $\{L, C, B\}^n$, est *un échantillon de taille n* de l'épreuve e.
À chaque échantillon possible on peut associer les proportions de chaque type de paiement. Appelons F_1, F_2, F_3 les variables aléatoires(22) ainsi définies, et f_1, f_2, f_3 leurs valeurs obtenues sur l'échantillon observé.

Il s'agit de comparer (f_1, f_2, f_3) à $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

Pour cela on peut considérer dans l'espace muni d'un repère orthonormé le point aléatoire F de coordonnées (F_1, F_2, F_3). Comme $F_1 + F_2 + F_3 = 1$, F appartient au plan(23) défini par I (1, 0, 0), J (0, 1, 0), K (0, 0, 1). Comme F_1, F_2, F_3 sont positifs, il appartient à l'intérieur du triangle IJK(24).



(22) Pour chaque i, nF_i suit la loi binomiale $B(n, p_i)$. Le triplet (nF_1, nF_2, nF_3) suit une loi multinomiale.
(23) F décrit un espace de dimension 2 : le triplet (f_1, f_2, f_3) a *deux degrés de liberté*. On peut aussi remarquer que F est le barycentre de I, J, K affectés des coefficients f_1, f_2, f_3.
(24) Il y a un nombre fini de points F possibles : en effet $f_i = \frac{n_i}{n}$, où n_i est le nombre de paiements de type i. Il y a donc autant de points possibles que de triplets (n_1, n_2, n_3) de naturels tels que $n_1 + n_2 + n_3 = n$, c'est-à-dire $\frac{(n+1)(n+2)}{2}$.

Appelons D sa distance au point $E\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$:

$$D=\sqrt{\left(F_1-\frac{1}{3}\right)^2+\left(F_2-\frac{1}{3}\right)^2+\left(F_3-\frac{1}{3}\right)^2}.$$

Sa valeur sur l'échantillon observé est

$$d=\sqrt{\left(f_1-\frac{1}{3}\right)^2+\left(f_2-\frac{1}{3}\right)^2+\left(f_3-\frac{1}{3}\right)^2}.$$

Si d est trop élevée, on rejettera l'hypothèse d'équiprobabilité.
Que signifie une valeur trop élevée ? C'est à l'appréciation de l'utilisateur. Si comme plus haut il a choisi un niveau de confiance de 90 %, il considérera le quantile d'ordre 0,9 de la variable aléatoire D dans le modèle d'équiprobabilité, c'est-à-dire[25] le nombre R tel que $P(D \leq R) = 0{,}9$, et rejettera l'hypothèse si d dépasse R ; autrement dit si F tombe en dehors du disque de centre E et de rayon R.
Cette décision comporte un risque d'erreur : celui de rejeter l'hypothèse d'équiprobabilité alors qu'elle est pertinente. Ce *risque de première espèce* α est $P(D > R)$: il est égal à 10 %.

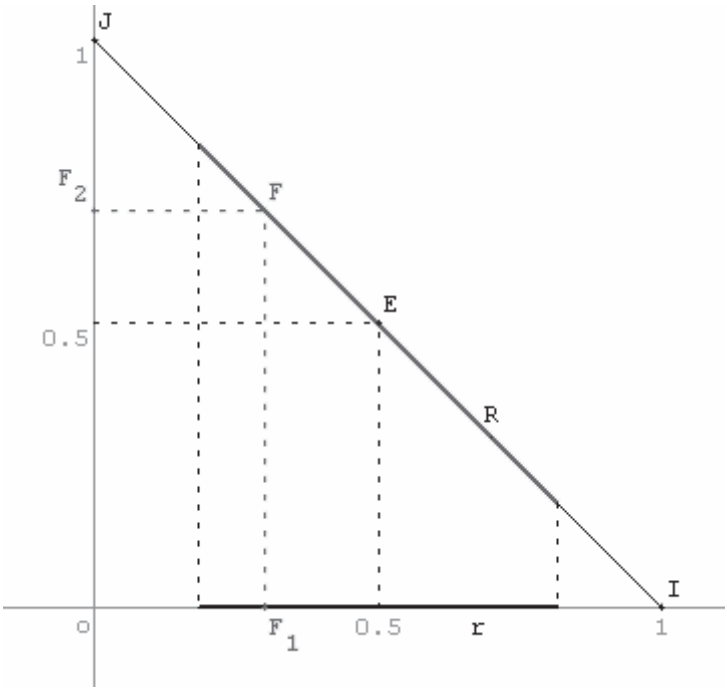
En revanche, si d est inférieure à R, il retiendra l'équiprobabilité comme l'un des modèles acceptables.
Le *risque de deuxième espèce* n'est défini que si on oppose au modèle d'équiprobabilité $\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$ un modèle concurrent (p_1, p_2, p_3). Ce risque β alors égal à $P'(D \leq R)$, où P' désigne la répartition de probabilité induite sur $\{L, C, B\}^n$ par la répartition de probabilité (p_1, p_2, p_3) sur {L, C, B}.

III. Retour sur le premier exemple

L'interprétation géométrique ci-dessus suggère de reprendre l'étude du premier exemple.
Au lieu de raisonner sur la proportion F_1 de garçons, on peut traiter ensemble les proportions F_1 et F_2 de garçons et de filles. Il s'agit de comparer (F_1, F_2) à $\left(\frac{1}{2},\frac{1}{2}\right)$.

Pour cela on peut considérer dans le plan muni d'un repère orthonormé le point aléatoire F de coordonnées (F_1, F_2). Comme $F_1 + F_2 = 1$, il appartient à la droite[26] définie par I (1, 0) et J (0, 1). Comme F_1 et F_2 sont positifs, il appartient même au segment [IJ][27].

(25) La détermination de R est traitée au paragraphe IV dans le cas général.
(26) F décrit un espace de dimension 1 : le couple (F_1, F_2) a *un degré de liberté*.
(27) Il y a $n + 1$ positions possibles de F, puisque $f_1 = \frac{n_1}{n}$, où le nombre de garçons n_1 est un entier compris entre 0 et n.



Notons D sa distance au point E $\left(\frac{1}{2},\frac{1}{2}\right)$:

$$D=\sqrt{\left(F_1-\frac{1}{2}\right)^2+\left(F_2-\frac{1}{2}\right)^2}=\sqrt{2}\left|F_1-\frac{1}{2}\right|.$$

Sa valeur sur l'échantillon observé est *d*.

Choisissons par exemple le niveau de confiance 90 %. Appelons R le quantile d'ordre 0,9 de D dans le modèle d'équiprobabilité. Si *d* dépasse R, on rejettera l'hypothèse d'équiprobabilité, sinon on l'acceptera.

Dans l'étude faite au I.1, on avait pris $\left|F_1-\frac{1}{2}\right|$ comme variable de décision et *r* comme valeur critique. Cela revient à projeter E et F sur l'axe des abscisses. Comme $D=\sqrt{2}\left|F_1-\frac{1}{2}\right|$, on voit que $R=r\sqrt{2}$.

IV. Cas général

On s'intéresse à une épreuve aléatoire e ayant *k* issues. On se demande si ces *k* issues sont équiprobables. Autrement dit si la répartition de probabilité $\left(\frac{1}{k},\frac{1}{k},\ldots,\frac{1}{k}\right)$ est un modèle adapté pour décrire cette épreuve.

Pour tester cette hypothèse, on répète n fois l'épreuve e, et on observe les fréquences $f_1, f_2, \ldots, f_k$ d'apparition des k issues. Pour mesurer l'écart entre $(f_1, f_2, \ldots, f_k)$ et $\left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$, on pose[28] $d = \sqrt{\sum\left(f_i - \frac{1}{k}\right)^2}$. d est la valeur observée de la variable aléatoire $\mathrm{D} = \sqrt{\sum\left(\mathrm{F}_i - \frac{1}{k}\right)^2}$.

On choisit *a priori* le niveau de confiance souhaité (en général 0,95, mais cela peut être 0,90 ou 0,99), et on en déduit par complément à 1 le risque d'erreur de première espèce α qu'on est prêt à accepter (0,05 ou 0,1 ou 0,01).

On cherche alors le quantile d'ordre $1 - \alpha$ (dans le modèle d'équiprobabilité) de D, c'est-à-dire le nombre R tel que $\mathrm{P}(\mathrm{D} \leq \mathrm{R}) = 1 - \alpha$.
Pour cela, on peut soit l'estimer par simulation, soit le calculer par la théorie. Par les deux méthodes on obtient une valeur approchée.

• Estimer R par simulation :

C'est ce qui est proposé par le programme de terminale S et ES : on simule un grand nombre de fois le modèle d'équiprobabilité de l'épreuve e^n, on calcule pour chaque simulation la valeur de D, et on en déduit une estimation (ponctuelle) de R en prenant le quantile d'ordre $1 - \alpha$ de la série statistique obtenue.
Les avantages et inconvénients de cette méthode sont analysés dans [11].

• Calculer R par la théorie :

On a besoin pour cela de la loi de probabilité de D dans le modèle d'équiprobabilité : c'est une loi discrète, qui dépend de k et de n. Elle se déduit de la loi multinomiale du vecteur aléatoire $(n\mathrm{F}_1, n\mathrm{F}_2, \ldots, n\mathrm{F}_k)$. Le calcul est malcommode, même avec un logiciel de calcul formel. Mais on démontre que pour n grand, la loi de $kn\mathrm{D}^2$ peut être approchée par la loi du χ^2 à $k - 1$ degrés de liberté. Appelons alors Q le quantile d'ordre $1 - \alpha$ de cette loi (donné par les tables ou les tableurs) : il dépend de α et de k, mais pas de n (cf. page suivante).

De l'égalité $\mathrm{P}\left(kn\,\mathrm{D}^2 \leq \mathrm{Q}\right) \approx 1 - \alpha$, on déduit $\mathrm{P}\left(\mathrm{D} \leq \sqrt{\frac{\mathrm{Q}}{kn}}\right) \approx 1 - \alpha$, d'où :

$$\mathrm{R} \approx \sqrt{\frac{\mathrm{Q}}{kn}}.$$

(28) Si l'on voulait tester un modèle quelconque $(p_1, p_2, \ldots, p_k)$, la distance de Pearson $\Delta = \sqrt{\sum_1^k \frac{(f_i - p_i)^2}{p_i}}$ serait préférable à la distance euclidienne, car $n\Delta^2$ suit asymptotiquement la loi du χ^2 à $k - 1$ degrés de liberté, ce qui facilite le calcul de la valeur critique. Mais quand les p_i sont égaux à $\frac{1}{k}$, Δ est égal à $\mathrm{D}\sqrt{k}$, d'où $n\Delta^2 = kn\mathrm{D}^2$.

En pratique, plutôt que de comparer d à R, il est plus commode de comparer knd^2 à Q.

Le critère de décision est alors le suivant :

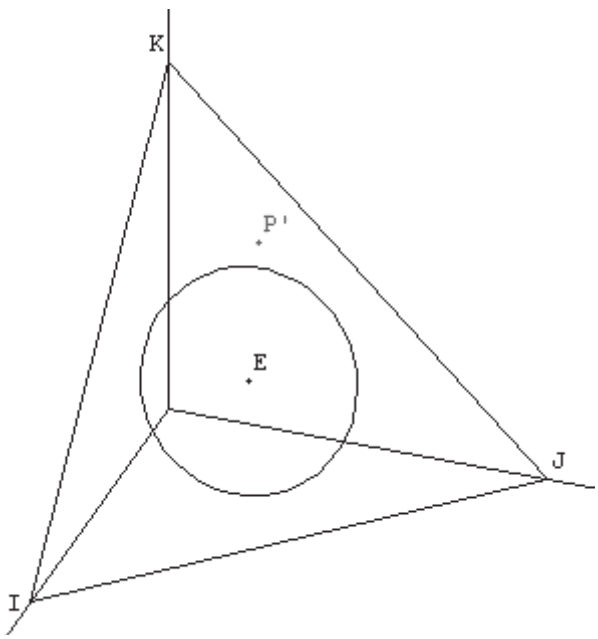
• si d > R, on rejette l'hypothèse d'équiprobabilité. Ce faisant, on prend le risque α de se tromper (risque de première espèce).

• si $d \leq$ R, on accepte l'hypothèse d'équiprobabilité. Le risque β de se tromper (risque de deuxième espèce) dépend alors du modèle concurrent $(p_1, p_2, \ldots, p_k)$ que l'on oppose au modèle d'équiprobabilité. On peut essayer de le majorer en considérant tous les modèles concurrents, ou seulement certains d'entre eux. C'est l'objet du paragraphe suivant.

k	deg lib	quantile 0,9	quantile 0,95	quantile 0,99
2	1	2,7	3,8	6,6
3	2	4,6	6,0	9,2
4	3	6,3	7,8	11,3
5	4	7,8	9,5	13,3
6	5	9,2	11,1	15,1
7	6	10,6	12,6	16,8
8	7	12,0	14,1	18,5
9	8	13,4	15,5	20,1
10	9	14,7	16,9	21,7
11	10	16,0	18,3	23,2
12	11	17,3	19,7	24,7
13	12	18,5	21,0	26,2
14	13	19,8	22,4	27,7
15	14	21,1	23,7	29,1
16	15	22,3	25,0	30,6
17	16	23,5	26,3	32,0
18	17	24,8	27,6	33,4
19	18	26,0	28,9	34,8
20	19	27,2	30,1	36,2

V. Majoration du risque de deuxième espèce

Appelons P′ le k-uplet $(p_1, p_2, \ldots, p_k)$ que l'on oppose au modèle d'équiprobabilité[29], et notons β_n(P′) le risque de deuxième espèce.



(29) Ce k-uplet P′ détermine sur l'univers des échantillons la répartition de probabilité que l'on a aussi notée P′ dans ce qui précède.

Les P′ possibles décrivent un domaine de dimension $k-1$: si $k=3$, c'est l'intérieur du triangle IJK, privé du point central E.
On peut démontrer[30] que le sup de $\beta_n(\text{P}')$ sur ce domaine est $1-\alpha$.
Mais, le risque α ayant été choisi et le rayon R déterminé en conséquence, on peut chercher le sup de $\beta_n(\text{P}')$ sur le complémentaire du disque de centre E et de rayon R. Autrement dit, ne prendre en compte que des modèles P′ qui sont assez éloignés de l'équiprobabilité. En effet si P′ est à l'intérieur de ce disque, en vertu de la loi faible des grands nombres, le point F y est aussi avec une très forte probabilité lorsque n est grand et on ne peut pas obtenir de majoration intéressante de $\beta_n(\text{P}')$.

Mais $\beta_n(\text{P}')$ dépend aussi de n, et on peut réduire le risque de deuxième espèce en augmentant la taille de l'échantillon : on peut ainsi démontrer[31] que

$$\beta_n(\text{P}') \leq \frac{k^2}{4n(\text{EP}'-\text{R})^2},$$

où EP′ désigne la distance euclidienne $\sqrt{\sum_1^n \left(p_i - \frac{1}{k}\right)^2}$.

Bibliographie

[1] ENGEL A., *Les certitudes du hasard* (Aléas), 1990.
[2] SAPORTA G., *Probabilités, analyse des données et statistiques* (Technip), 1990.
[3] SPIEGEL M.-R., *Théorie et applications de la statistique* (Mc-Graw-Hill), 1993 (2e édition).
[4] DROESBEKE J.-J., *Éléments de statistique* (Ellipses), 1997 (3e édition).
[5] VEYSSEYRE R., *Statistique et probabilités pour l'ingénieur* (Dunod), 2000.
[6] LEJEUNE Michel, *Statistique : la théorie et ses applications* (Springer), 2004.
[7] DRESS F., *Probabilités Statistique*, Série TD DEUG sciences (Dunod).
[8] DUTARTE P. et PIEDNOIR J.-L., *Enseigner les statistiques au lycée : des enjeux aux méthodes* (IREM de Paris-Nord).
[9] *Statistique et probabilité au lycée – Carnets de stage*, (IREM de Paris-Nord, brochure 124), 2003.
[10] ROBERT C., *Contes et décomptes de la statistique* (Vuibert).
[11] BONNEVAL L.-M., *Test d'équirépartition : qui a dit khi-deux ?* in Bulletin de l'APMEP nº 441.

Ouèbographie

[12] KENTZEL J.-F., *Quelques précisions sur le test de l'adéquation de données à une loi équirépartie*, sur le site des « classes virtuelles » d'aromath :
http://www.aromath.net/moodle/

(30) C'est graphiquement évident pour $k=2$. Pour k quelconque, voir [12].
(31) Voir [12].

[13] BONNEVAL L.-M., *Test d'équirépartition : quelle erreur ? Annexe : calcul de alpha et beta dans le cas* $k = 2$, sur le même site :
http://www.aromath.net/moodle/
[14] POSS J.-L., *Probabilités et statistiques*, téléchargeable sur :
http://www.aix.ensam.fr/departement/info-math/6G01/poly_math1 complet.pdf
[15] http://www.math-info.univ-paris5.fr/<smel/cours/ts/ts.html
[16] http://moire4.u-strasbg.fr/bouquins/proba/tabmat1.htm
[17] http://www.emse.fr/~messiaen/3MI/probaStats/Contenu_prob_stat.html
[18] http://www.univ-lr.fr/formations/idea/duCultureMath/statistiques/index.htm
[19] http://www.chups.jussieu.fr/polys/biostats/poly/stats.pdf
[20] http://cons-dev.univ-lyon.fr/Enseignement/Stat/St.html
[21] http://www.rfv.insa-lyon.fr/~jolion/STAT/node1.html