Consultation statistique avec le logiciel 

Comment comparer des fréquences très faibles ?

D. Chessel

14 mars 2006

Pour un copain d'un ami, le souvenir d'une consultation pour V. Nigon sur les taux d'anomalies chromosomiques rares.

Table des matières

1 Les copains des amis

Un ami m'a écrit :

> Bonjour Daniel,
> Un copain biologiste me pose la question ci-dessous.
> `Connais-tu un (ou des) test(s) de comparaison de fréquences très faibles au sein de grandes populations dans des cas où le test du Chi2, le test G et la Probabilité exacte de Fisher sont théoriquement invalides ou peu valides.`
> `Exemple 1 : population A : 1 cas sur 106`
> `population B : 14 cas sur 107`
> `Exemple 2 : population A : 0 cas sur 106`
> `population B : 6 cas sur 10788`
> Il veut faire une comparaison entre les pop A et B, indépendamment dans chaque exemple.
> Qu'en penses-tu ? Tu peux lui répondre directement à ...
> Merci et bonne journée

Ma réponse est rapide :

> Je ne réponds pas directement car la réponse est toujours : posez votre question sur le forum adelist, la réponse peut être utile à d'autres (et ensuite je réponds, parce que c'est public).
> Pour toi, c'est différent, tu es tellement exceptionnel.
>
> La réponse est "pas de problème", dans R la fonction chisq.test intègre un paramètre sim = TRUE qui remplace l'approximation qui est douteuse par une similation de Monte-Carlo parfaitement adaptée à la situation. Ci-joint un fichier de consultation pour V. Nigon qui était venu me voir il y a quelques années. C'était EXACTEMENT ton problème.

Cette discussion avec V. Nigon est rapportée ici.

2 La question des anomalies chromosomiques

Le problème est défini par V. Nigon dans une lettre remarquable :

> La vie à haute altitude intègre l'action de plusieurs facteurs, en particulier, une réduction de la pression partielle d'oxygène et une intensité accrue du rayonnement cosmique. Ces facteurs sont susceptibles d'exercer un effet sur les structures chromosomiques et d'y déterminer la formation d'anomalies. Si cet effet existe, il doit comporter une composante de sommation et donc se trouver plus apparent chez des personnes âgées que sur des sujets jeunes. En vue de répondre à ces interrogations du sang a été prélevé, en Bolivie, sur des personnes, âgées de 60 à 80 ans, vivant les unes à La Paz (3500 à 4000 mètres d'altitude), les autres à Santa Cruz (300 mètres d'altitude). Les dites personnes ont fait l'objet d'un examen médical préalable pour s'assurer qu'elles ne sont pas affectées d'une autre pathologie, susceptible d'entraîner la formation d'anomalies chromosomiques. Les sangs recueillis ont été amenés dans des laboratoires, français ou boliviens, dans lesquels ils ont été employés pour réaliser des cultures de lymphocytes. Les préparations de ces lymphocytes ont été montées en frottis et colorées, de façon à permettre l'analyse des métaphases. On a concentré l'attention sur 3 sortes d'anomalies chromosomiques connues pour résulter d'influences exercées sur les phases G0 du cycle cellulaire (=interphases). L'objectif du travail est de déterminer si l'on peut trouver entre les cellules des diverses origines des fréquences d'anomalies présentant des différences statistiquement significatives. Parmi les facteurs susceptibles d'exercer des influences perturbatrices, il faut compter le traitement dans des laboratoires différents. Ainsi, les échantillons cultivés en France ont subi un voyage durant lequel les conditions de conservation, et peut être d'autres facteurs, ont été différentes de celles auxquelles ont été soumises les cultures directement traitées en Bolivie.
>
> A titre de comparaison, on dispose uniquement d'un travail de Lloyd et al. (1992) [1] qui étudie la fréquence des mêmes anomalies sur des sujets âgés d'environ 30 ans vivant en Europe, dans diverses villes de faible altitude. Ces sujets diffèrent des sujets boliviens par

l'âge et, pour les sujets de La Paz, par l'altitude de leur lieu d'existence.

Les résultats sont portés dans les tableaux I, II et III. Le tableau III comporte 3 sous catégories correspondant, respectivement, à deux séries de cultures faites en France et une série de cultures faites en Bolivie. Chaque ligne représente les observations faites sur les cultures issues d'un seul sujet : nombre des métaphases examinées, nombre observés, respectivement, de dicentriques, d'anneaux et d'acentriques + minutes. Dans le tableau III, les lignes 3 et 10 représentent les résultats obtenus à partir d'un même sujet prélevé 2 fois.

Les questions posées sont les suivantes :

1. Les résultat d'une même série peuvent ils être valablement considérés comme des tirages au hasard provenant d'un même échantillon ou les paramètres étudiés seraient distribués selon une loi de Poisson ?
2. Dans le tableau III, peut-on ou non trouver des différences significatives entre les diverses séries ?
3. Peut-on détecter des différences significatives entre les résultats des divers tableaux ?

Il n'est bien sûr pas possible de rendre publiques ces données qui représentent un travail considérable. On s'en tiendra, pour un exercice utile, aux conversations qui avaient précédés cette consultation et qui portaient sur des données en cours d'acquisition.

3 Un exemple

D'autre part, je me permets de vous rappeler le problème de distribution de Poisson dont nous avons parlé et dont la solution présente pour nous une grande importance.

Il faut bien répondre. Dans [1], on apprend que six laboratoires se sont associés pour une étude de grande ampleur. Pour des doses de radiations comprises entre 0 et 19.3, qu'on peut, à la lecture de l'article cité, considérées comme sans effet (témoin), on a trouvé [1, Table 2, p. 338] :

Laboratoire	n	DIC
1	10000	10
2	10054	8
3	10000	26
4	9986	9
5	9547	17
6	10000	13

`n` est le nombre de cellules examinées et `DIC` le nombre de cellules présentant l'aberration de dicentrisme. Des moyens considérables sont mis en œuvre pour obtenir ces résultats. Il n'est pas fait mention d'une éventuelle variation du taux entre donneurs des cellules examinées. L'événement a un taux observé de 83/59587=0.00139. On peut se poser la question de l'identité des probabilités dans les différents échantillons. V. Nigon ajoutait :

La question de la taille de l'échantillon nécessaire pour une comparaison avec ces données de référence est posée. Nous avons trouvé après étude de 11 donneurs les résultats suivants (`n` = nombre de cellules examinées pour une personne, `DIC` nombre de cellules anormales trouvées chez cette personne) :

n	443	113	101	148	145	16	278	100	734	35	232
DIC	5	0	0	0	0	0	0	0	8	0	0

D'où la question :

Est-il possible que ce soient des échantillons de loi de Poisson avec une moyenne constante ? Quelle et la précision de l'estimation du taux dans un cas ou dans l'autre ? Combien faut-il de mesures pour distinguer un taux de 1/100, 5/1000 et 1/1000, pour être sûr qu'il est supérieur à 1/1000 ?

```
rm(list = ls())
tge <- c(10000, 10054, 10000, 9986, 9547, 10000, 10,
    8, 26, 9, 17, 13)
tge <- matrix(tge, nrow = 2, byrow = T)
tge[1, ] <- tge[1, ] - tge[2, ]
chisq.test(tge)
```

```
        Pearson's Chi-squared test

data:  tge
X-squared = 16.9922, df = 5, p-value = 0.004515
```

Les choses commencent mal : les données de référence ne sont pas homogènes. Évidemment, si on en enlève un des échantillons, tout va mieux, mais ceci manque singulièrement d'orthodoxie statistique.

```
chisq.test(tge[, -3])
```

```
        Pearson's Chi-squared test

data:  tge[, -3]
X-squared = 5.3364, df = 4, p-value = 0.2545
```

Le troisième échantillon est différent des autres. Mais ce n'est pas la question posée.

```
ess = c(443, 113, 101, 148, 145, 16, 278, 100, 734, 35,
    232)
suc = c(5, 0, 0, 0, 0, 0, 0, 0, 8, 0, 0)
ech = ess - suc
chisq.test(cbind(suc, ech))
```

```
        Pearson's Chi-squared test

data:  cbind(suc, ech)
X-squared = 12.98, df = 10, p-value = 0.2248
```

```
chisq.test(suc, p = ess/sum(ess))
```

```
        Chi-squared test for given probabilities

data:  suc
X-squared = 12.9081, df = 10, p-value = 0.2289
```

```
Chi-squared approximation may be incorrect in:
    chisq.test(cbind(suc, ech))
Chi-squared approximation may be incorrect in:
    chisq.test(suc, p = ess/sum(ess))
```

Voilà l'objet de la discussion : les conditions d'utilisation du χ^2 peuvent ne pas être remplies. En fait, il s'agit simplement de savoir si la convergence de la statistique utilisée vers une loi χ^2, sous l'hypothèse nulle de tirage aléatoire à taux constant, est acceptable.

Remarquons d'abord qu'il y a plusieurs façons de faire le calcul. Soit m le nombre d'échantillons et pour l'échantillon i on note n_i le nombre d'essais, s_i le nombre de succès, e_i le nombre d'échecs, n le nombre total d'essais, s le nombre total de succès et e le nombre total d'échecs.

On peut dire que e_i est la réalisation d'une variable aléatoire E_i binomiale de paramètres n_i et q estimée par $\hat{q} = \frac{e}{n}$. Les variables E_i sont indépendantes, ce qui conduit à la statistique :

$$X_1 = \sum_{i=1}^{i=m} \frac{(e_i - n_i\hat{q})^2}{n_i\hat{q}(1-\hat{q})}$$

C'est le χ^2 de comparaison de m pourcentages indépendants.

De même, on peut dire que s_i est la réalisation d'une variable aléatoire binomiale de paramètres n_i et p estimée par $\hat{p} = \frac{s}{n}$ ce qui conduit à la statistique :

$$X_2 = \sum_{i=1}^{i=m} \frac{(s_i - n_i\hat{p})^2}{n_i\hat{p}(1-\hat{p})}$$

On peut aussi dire que les $(s_i)_{1\leq i\leq m}$ est la réalisation d'une variable aléatoire hypergéométrique $(S_i)_{1\leq i\leq m}$ de paramètres s et $(n_i)_{1\leq i\leq m}$ ce qui conduit par le biais de la loi multinomiale à la statistique :

$$X_2 = \sum_{i=1}^{i=m} \frac{(s_i - n_i\hat{p})^2}{n_i\hat{p}(1-\hat{p})} = X1$$

C'est alors le χ^2 d'ajustement à une distribution de fréquences entièrement connue.

On peut dire enfin que les $(s_i)_{1\leq i\leq m}, (e_i)_{1\leq i\leq m}$ forment la réalisation d'une variable aléatoire multinomiale de paramètres $(p_i)_{1\leq i\leq m}, (q_i)_{1\leq i\leq m}$ ce qui conduit à la statistique :

$$X_3 = \sum_{i=1}^{i=m} \frac{(s_i - n_i\hat{p})^2}{n_i\hat{p}} + \frac{(e_i - n_i(1-\hat{p}))^2}{n_i(1-\hat{p})}$$

C'est, cette fois, le χ^2 d'indépendance de la table de contingence, indépendance entre l'appartenance à un échantillon et résultat (succès ou échec).

Bien sûr, ces statistiques ne suivent des lois χ^2 sous l'hypothèse nulle

asymptotiquement

Quand les effectifs tendent vers l'infini, la loi tend vers une lois χ^2. Pour ce qui concerne l'utilisateur, il convient de savoir quand les effectifs sont suffisamment grands pour que les approximations soient valides. Beaucoup d'encre a coulé sur le sujet qui n'a plus beaucoup de signification.

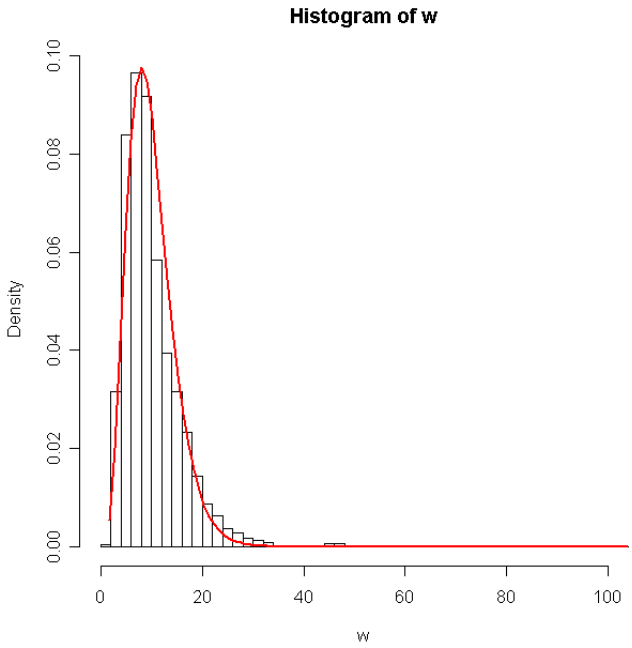
4 Simuler la loi

Si on veut savoir quelle confiance on peut accorder à un théorème d'approximation, c'est maintenant bien facile.

```
w <- rbind(suc, ech)
w
```

```
    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
suc    5    0    0    0    0    0    0    0    8     0     0
ech  438  113  101  148  145   16  278  100  726    35   232
```

```
proba = ess/sum(ess)
f1 <- function(k) {
    ww <- sample(1:11, 13, replace = T, proba)
    ww <- as.array(table(ww))
    wsuc = rep(0, 11)
    names(wsuc) = (1:11)
    wsuc[names(ww)] <- ww
    wech <- ess - wsuc
    chisq.test(rbind(wsuc, wech))$statistic
}
w <- unlist(lapply(1:9999, f1))
hist(w, nclass = 50, proba = T)
x0 <- seq(min(w), max(w), le = 100)
lines(x0, dchisq(x0, 10), lwd = 2, col = "red")
```



On a simplement fait le raisonnement suivant. Le nombre de prélèvement par individus est fixé. Au total on a 2345 cellules examinées. 13 d'entre elles sont dicentriques. Si les taux sont constants, ces 13 succès apparaissent au hasard (`sample`) conformément à la distribution des essais (`proba`). On fait 9999 fois l'expérience. On a une bonne idée de la variabilité d'échantillonnage de la statistique du χ^2.

Or les effectifs théoriques attendus dans les cases succès sont :

```
round(13 * ess/sum(ess), dig = 1)
```

```
 [1] 2.5 0.6 0.6 0.8 0.8 0.1 1.5 0.6 4.1 0.2 1.3
```

C'est réputé insuffisant, et pourtant, on n'est pas très loin d'un χ^2 à 10 degrés de liberté. En tout cas, l'approximation est suffisante pour savoir que la valeur observée pour le résultat expérimental n'a rien d'extraordinaire. C'est ce raisonnement qui est mis en œuvre dans la fonction `chisq.test` qu'il suffit d'utiliser avec l'option `simulate.p.value = TRUE`. L'approximation mathématique sera alors remplacée par une simulation sur – par défaut – 2000 sélections aléatoires.

```
chisq.test(rbind(suc, ess), sim = T)
```

```
        Pearson's Chi-squared test with simulated p-value (based on
        2000 replicates)

data:  rbind(suc, ess)
X-squared = 12.8377, df = NA, p-value = 0.2289
```

5 Écart au taux de référence

Si on admet que les donneurs sont homogènes, on peut réunir les échantillons et dire qu'on a observé 13 aberrations sur 2345 cellules. Est-ce compatible avec 83 aberrations sur 59587 cellules de l'expérience de référence :

```
chisq.test(matrix(c(13, 2345 - 13, 83, 59587 - 83), 2,
    2))
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  matrix(c(13, 2345 - 13, 83, 59587 - 83), 2, 2)
X-squared = 22.5061, df = 1, p-value = 2.095e-06
```

La réponse est non. Ou encore (probabilité de trouver au moins 13 aberrations sur 2345 cellules avec un taux de 1.4 pour mille) :

```
1 - pbinom(13, 2345, 83/59587)
```

```
[1] 8.565272e-06
```

La réponse est non. Ou encore :

```
prop.test(13, 2345)
```

```
        1-sample proportions test with continuity correction

data:  13 out of 2345, null probability 0.5
X-squared = 2291.311, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.003085675 0.009729812
sample estimates:
          p
0.00554371
```

La réponse est encore non.

6 Effectif des donneurs négatifs

On se demande alors si le nombre de 0 n'est pas l'indication d'une hétérogénéité de l'ensemble des donneurs (la question est subsidiaire mais le statisticien est pinailleur) :

```
proba = ess/sum(ess)
f2 <- function(k) {
    ww <- sample(1:11, 13, replace = T, proba)
    ww <- length(unique(ww))
}
w <- unlist(lapply(1:9999, f2))
wobs <- 2
table(w)
```

```
w
   2    3    4    5    6    7    8    9   10
   5   68  545 2029 3334 2787 1047  177    7
```

```
sum(w <= wobs)/10000
```

```
[1] 5e-04
```

Sur 10000 expériences il y en a 5 qui ne donne pas plus de deux donneurs positifs. Ce n'est pas normal. On observe alors que si on diminue le taux théorique, le nombre de donneurs positifs devient acceptable mais qu'il est impossible de trouver un nombre aussi important de cellules anormales. Un réexamen dans ce sens des données de référence (*op. cit.*) serait du plus grand intérêt. Il y a vraiment une faible probabilité que les données acquises soient issues d'une population à taux constant.

Plusieurs mois après cette étude, V. Nigon propose le même jeu de données vérifiées et complétées par des observations supplémentaires. Pour se prémunir d'un biais éventuel l'appartenance des lames observées est inconnue pour l'observateur pendant la dernière série de comptages.

Des enregistrements suspects sont éliminés. Les individus n'ayant fourni que peu de cellules observées sont définitivement exclus de la discussion, logiquement car les petits effectifs de comptage ne peuvent contenir de l'information dans les conditions numériques présentes.

La conclusion sera très simple : L'anomalie est un événement rare équiprobable chez les 9 donneurs. Sa fréquence est comprise entre 4 et 9 pour mille au risque d'erreur de 5%. *La valeur de référence est absolument incompatible avec le taux de référence de l'article cité.*

Références

[1] D.C. Lloyd, A.A. Edwards, A. Leonard, G.L. Deknudt, L. Verschaeve, A.T. Natarajan, F. Darroudi, G. Obe, F. Palitti, C. Tanzarella, and E.J. Tawn. Chromosomal aberrations in human lymphocytes induced in vitro by very low doses of x-rays. *International Journal of Radiation Biology*, 61(3) :335–343, 1992.