Notes sur le test Q de Cochran

F.-G. Carpentier - 20 mai 2009

1) Le test Q de Cochran

On a évalué une variable dichotomique X sur un échantillon de n individus statistiques o_1, o_2, …, o_n observés dans k conditions différentes c_1, c_2, .., c_k. Les observations constituent donc un tableau de taille (n, k) :

	c_1	c_2	…	c_k
o_1	x_{11}	x_{12}	…	x_{1k}
o_2	x_{21}	x_{22}	…	x_{2k}
…	…	…	…	…
o_n	x_{n1}	x_{n2}	…	x_{nk}

où x_{ij} prend ses valeurs dans {0,1}.

On souhaite étudier si les chances de succès sont les mêmes dans toutes les conditions, autrement dit, on désigne par π_1, π_2, ..., π_k les fréquences de succès correspondant aux k conditions dans la population parente et on souhaite tester l'hypothèse :

$$H_0 : \pi_1 = \pi_2 = ... = \pi_k$$

contre :

$$H_1 : \text{non}(H_0)$$

La statistique de test est alors :

$$Q = \frac{(k-1)\sum_{j=1}^{k} G_j^2 - G^2}{k\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2} = \frac{k(k-1)\sum_{j=1}^{k}\left(G_j - \overline{G}\right)^2}{k\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2}$$

où G_j désigne la somme sur la colonne j $\left(G_j = \sum_{i=1}^{n} x_{ij}\right)$, L_i désigne la somme sur la ligne i $\left(L_i = \sum_{j=1}^{k} x_{ij}\right)$, G désigne la somme de toutes les valeurs du tableau $\left(G = \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij} = \sum_{j=1}^{k} G_j = \sum_{i=1}^{n} L_i\right)$ et $\overline{G}$ désigne la moyenne des valeurs G_j $\left(\overline{G} = \frac{\sum_{j=1}^{k} G_j}{k}\right)$.

On remarque que cette valeur n'est pas affectée par la suppression éventuelle des lignes "sans variation", c'est-à-dire composées uniquement de 1 ou uniquement de 0.
Par conséquent, dans toute la suite, **nous supposerons que les lignes ne sont pas composées uniquement de 1 ou uniquement de 0**, c'est-à-dire que nous supposerons que :

$$\forall i = 1,2,..,n, \quad 1 \le L_i \le k-1.$$

Dans ces conditions, il est généralement indiqué que, pour $k \ge 4$ et/ou $nk \ge 24$, la statistique Q suit une loi du khi-2 à (k-1) degrés de liberté [2].

2) Diverses remarques sur le test Q de Cochran

Sous une apparente simplicité, le test de Cochran cache quelques difficultés, ce qui explique par exemple, qu'on ne trouve pas facilement des tables relatives à ce test pour les petites valeurs de n et k.

1) Cas où k=2

Lorsque le nombre de groupes *k* est égal à 2, la valeur de la statistique *Q* de Cochran est exactement celle du khi-2 de Mac Nemar, sans correction de Yates. On sait par ailleurs que le khi-2 de Mac Nemar est aussi le test des signes, lorsque les deux variables comparées sont dichotomiques.

2) Remarques faites par M.W. Tate et S.M. Brown [3]

Ces deux auteurs indiquent que le test est en général mal décrit dans les publications postérieures à celle de Cochran [1]. Les auteurs soulignent notamment deux points :

- Ces publications ne soulignent pas suffisamment que les lignes "sans variation", c'est-à-dire composées uniquement de 1 ou composées uniquement de 0 n'ont pas d'effet sur le calcul de *Q*. Il importe donc d'éliminer ces lignes dans toutes les considérations sur les effectifs.

- Certaines publications semblent indiquer que les protocoles pris en compte pour calculer la p-value d'une valeur *Q* donnée sont obtenus en distribuant au hasard les 1 dans les lignes et colonnes de la matrice des observations. Or, Cochran indique clairement que le nombre de succès (de 1) de chaque ligne est considéré comme fixe, et que les protocoles pris en compte sont obtenus en permutant, dans chaque ligne i, les valeurs de la suite (x_{ij}) pour j= 1, 2, …, k.

D'éventuelles tables pour les petites valeurs des paramètres doivent donc prendre en compte non seulement les valeurs de *n* (nombre d'individus observés) et *k* (nombre de variables ou conditions différentes), mais aussi le nombre de lignes dont la somme vaut 1, 2, …, *k*-1 (on rappelle ici que les lignes composées uniquement de 1 (somme = *k*) ou de 0 (somme = 0) sont éliminées). De telles tables ont été élaborées, par exemple par ces mêmes auteurs dans une autre publication, mais ne sont pas largement diffusées dans les ouvrages spécialisés.

3) Approximation de la distribution de *Q* pour les petites valeurs de *n* et *k*

On peut obtenir une valeur approchée de la fréquence des différentes valeurs prises par *Q* lorsque *n*, *k* et les différentes sommes des lignes sont fixées, par une méthode de bootstrap : à partir d'un protocole donné remplissant ces conditions, on génère un (grand) nombre de protocoles obtenus en effectuant des permutations des valeurs sur les lignes, et pour chacun d'eux, on calcule la valeur de *Q* qui en résulte.

Ceci peut être fait à l'aide du programme R suivant. La fonction cochranq.test est reprise d'un programme diffusé par M. Schwartz sur un forum d'utilisateurs de R (http://tolstoy.newcastle.edu.au/) :

```
# Calcul de la statistique Q de Cochran et évaluation du niveau
# de significativité à l'aide de l'approximation par le khi-2

cochranq.test <- function(mat)
{ k <- ncol(mat)
  C <- sum(colSums(mat) ^ 2)
  R <- sum(rowSums(mat) ^ 2)
```

```
  T <- sum(rowSums(mat))

  num <- (k - 1) * ((k * C) - (T ^ 2))
  den <- (k * T) - R
  Q <- num / den

  df <- k - 1
  names(df) <- "df"
  names(Q) <- "Cochran's Q"

  p.val <- pchisq(Q, df, lower = FALSE)
  QVAL <- list(statistic = Q, parameter = df, p.value = p.val,
               method = "Cochran's Q Test for Dependent Samples",
               data.name = deparse(substitute(mat)))
class(QVAL) <- "htest"
  return(QVAL) }

cochranq.exact.test <- function(X)
{ library(e1071)
  n <- dim(X)[1] # nombre de lignes
  L <- dim(X)[2] # nombre de colonnes
  perm <- permutations(L) # tableau des L! permutations sur 1..L
  qobs <- cochranq.test(X)$statistic
  freq <- 0

  X1 <- X
  pval <-c()
  nbiter <- 10000 #Nombre de tirages - à modifier au besoin
  for (boot in 1:nbiter) {
    for (j in 1:n) {
        k <- 1+ as.integer(runif(1)*gamma(L+1))
        X1[j,] <- X[j,perm[k,]]}
    res <- cochranq.test(X1)$statistic
    if (res >= qobs) {freq <- freq + 1}}

    p.val <- freq / nbiter
    QVAL <- list(statistic = qobs, p.value = p.val,
               method = "Cochran's Q Test for Dependent Samples (approximation)",
               data.name = deparse(substitute(X)))

  class(QVAL) <- "htest"
  return(QVAL)
}
```

Pour un petit nombre de colonnes ($k < 10$ par exemple), la distribution (approximative) de la statistique Q pour l'ensemble des protocoles correspondant à un protocole donné pourra être donnée par :

```
cochranq.exact.dist <- function(X)
{ library(e1071)
  n <- dim(X)[1] # nombre de lignes
  L <- dim(X)[2] # nombre de colonnes
  perm <- permutations(L) # tableau des L! permutations sur 1..L
  qobs <- cochranq.test(X)$statistic
  freq <- 0

  X1 <- X
  pval <-c()
  nbiter <- 10000 #Nombre de tirages - à modifier au besoin
  for (boot in 1:nbiter) {
    for (j in 1:n) {
        k <- 1+ as.integer(runif(1)*gamma(L+1))
        X1[j,] <- X[j,perm[k,]]}
    res <- cochranq.test(X1)
    pval <- c(pval,res$statistic)
    }
    ta <-table(pval) / nbiter
    return(ta) }
```

et la distribution cumulative par :

```
cochranq.exact.cumdist <- function(X)
{ library(e1071)
  n <- dim(X)[1] # nombre de lignes
  L <- dim(X)[2] # nombre de colonnes
  perm <- permutations(L) # tableau des L! permutations sur 1..L
  qobs <- cochranq.test(X)$statistic
  freq <- 0

  X1 <- X
  pval <-c()
  nbiter <- 20000 #Nombre de tirages - à modifier au besoin
  for (boot in 1:nbiter) {
    for (j in 1:n) {
        k <- 1+ as.integer(runif(1)*gamma(L+1))
        X1[j,] <- X[j,perm[k,]]}
    res <- cochranq.test(X1)
    pval <- c(pval,res$statistic)
    }
    ta <- matrix(c(sort(unique(pval)),table(pval)),byrow=TRUE,nrow=2)
    ta[2,] <-ta[2,] / nbiter
    ll <- ncol(ta)
    lll <- ll - 1
    tc <- ta
    tc[1,] <- ta[1,]
    for (i in 1: lll ) { tc[2,ll-i] <- ta[2,ll-i] + tc[2,ll-i+1]}
    return(tc) }
```

4) Tables de la statistique Q pour n et k petits

Des tables détaillées doivent prendre en compte non seulement les paramètres k et n , mais aussi les nombres de lignes correspondant à chaque somme possible (de 1 à n-1). Par exemple, pour k=3 et n=6, on obtiendrait la table suivante :

		Nombre de lignes ayant pour somme		Valeurs à 5%				Valeurs à 1%			
k	n	1	2	Q non sig	p-val	Q sig	p-value	Q non sig	p-val	Q sig	p-value
3	6	6	0	7,00	0,0551	12,00	0,0036	7,00	0,0551	12,00	0,0036
3	6	5	1	5,33	0,0914	6,33	0,0496	6,33	0,0496	10,33	0,0100
3	6	4	2	5,33	0,0751	6,33	0,0496	8,33	0,0161	9,33	0,0085
3	6	3	3	7,00	0,0543	9,00	0,0091	7,00	0,0543	9,00	0,0091
3	6	2	4	5,33	0,0751	6,33	0,0496	8,33	0,0161	9,33	0,0085
3	6	1	5	5,33	0,0914	6,33	0,0496	6,33	0,0496	10,33	0,0100
3	6	0	6	7,00	0,0551	12,00	0,0036	7,00	0,0551	12,00	0,0036

Pour chaque configuration des sommes de lignes, cette table donne, aux seuils de 5% et de 1%, la plus grande valeur non significative ainsi que la plus petite valeur significative. La p-value (c'est-à-dire la probabilité d'observer une valeur de Q supérieure ou égale à la valeur indiquée) est également mentionnée.

Au seuil de 1% par exemple, on constate que toute valeur de Q supérieure ou égale à 9 est significative, tandis que toute valeur de Q inférieure ou égale à 8,33 est non significative.

En revanche, au seuil de 5%, la valeur 6,33 par exemple, est significative lorsque le protocole comporte 5 lignes de somme 1 et 1 ligne de somme 2, alors que même 7 est une valeur est non significative lorsque le protocole comporte 3 lignes de somme 1 et 3 lignes de somme 2. L'ensemble de ces lignes ne peut donc être résumé de manière simple en une seule ligne de table. En revanche, il serait possible de faire des tables simplifiées en indiquant que :

- d'une part toute valeur strictement supérieure à 7 est significative au seuil de 5% ;
- d'autre part toute valeur strictement inférieure à 6,33 est non significative au seuil de 5% ;
- et enfin, pour les valeurs comprises, au sens large, entre 6,33 et 7, il convient de se référer à des tables détaillées.

De telles tables pourraient être disposées comme suit :

		Seuil 5%		
k	n	Sig, assurée si Q > à	Non-sig assurée si Q < à	Zone d'incertitude
3	4	6,50	8,00	
3	5	5,80	6,40	
3	6	7,00	6,33	6,33 - 7,00

Dans cette table, les valeurs de la troisième colonne sont obtenues comme maximum des plus grandes valeurs non significatives de la table détaillée, les valeurs de la quatrième colonne sont obtenues comme minimum des plus petites valeurs significatives de la table détaillée.

5) Tests post hoc - Comparaisons par paires

Lorsque le test Q de Cochran conduit à une valeur Q_{obs} significative au seuil fixé, cela indique seulement qu'il existe au moins deux conditions différentes entre elles. Des tests par paires permettent ensuite de déterminer les paires de conditions entre lesquelles les différences ne sont pas l'effet du hasard.

On peut par exemple utiliser une procédure dérivée de la méthode de Bonferroni-Dunn.

Pour faire des tests par paires bilatéraux, on calcule la différence minimale entre les proportions p_a et p_b observées dans les conditions *a* et *b*, notée CD_C à l'aide de la formule :

$$CD_C = z_{adj}\sqrt{2\ \frac{k\sum_{i=1}^{n} L_i - \sum_{i=1}^{n} L_i^2}{n^2 k(k-1)}}$$

où z_{adj} est la valeur de la loi normale centrée réduite correspondant au seuil $\frac{2\alpha}{k(k-1)}$, autrement dit la valeur t vérifiant $P(|Z| > t) = \frac{2\alpha}{k(k-1)}$ ou encore $P(Z > t) = \frac{\alpha}{k(k-1)}$.

On compare ensuite les $\frac{k(k-1)}{2}$ valeurs absolues des proportions $|p_a - p_b|$ à la valeur CD_C.

6) Exemple

On dispose des observations suivantes :

	c_1	c_2	c_3	L_i	L_i^2
O_1	1	1	0	2	4
O_2	0	1	0	1	1
O_3	1	1	1	3	9

O_4	0	1	0	1	1
O_5	0	1	0	1	1
O_6	0	1	1	2	4
O_7	0	0	0	0	0
O_8	0	1	0	1	1
O_9	1	1	0	2	2
O_{10}	0	1	0	1	1
O_{11}	0	0	0	0	0
O_{12}	0	0	1	1	1
G_j	3	9	3	15	27
p_j	0,25	0,75	0,25		

Ici, le nombre de lignes de somme différente de 0 ou 3 est n=9.

La valeur de Q est : $Q = \frac{2\left(3\left(3^2 + 9^2 + 3^2\right) - 15^2\right)}{3 \times 15 - 27} = 8.$

De même, on peut ne considérer que les lignes "utiles" du tableau :
On dispose des observations suivantes :

	c_1	c_2	c_3	L_i	L_i^2
O_1	1	1	0	2	4
O_2	0	1	0	1	1
O_4	0	1	0	1	1
O_5	0	1	0	1	1
O_6	0	1	1	2	4
O_8	0	1	0	1	1
O_9	1	1	0	2	4
O_{10}	0	1	0	1	1
O_{12}	0	0	1	1	1
G_j	2	8	2	12	18
p_j	0,22	0,89	0,22		

La valeur de Q est : $Q = \frac{2\left(3\left(2^2 + 8^2 + 2^2\right) - 12^2\right)}{3 \times 12 - 18} = 8.$

Pour k=3 et n=9, avec 6 lignes égales à 1 et 3 lignes égales à 3, la table de la statistique Q donne :

		Nombre de lignes ayant pour somme		Valeurs à 5%				Valeurs à 1%			
k	n	1	2	Q non sig	p-val	Q sig	p-value	Q non sig	p-val	Q sig	p-value
3	9	6	3	6,00	0,0540	8,00	0,0266	8,67	0,0144	10,67	0,0023

Le test est donc significatif au seuil de 5%.

Il y a ici 3 comparaisons possibles. Le seuil utilisé pour les comparaisons par paires est donc :
$\alpha_{PC} = \frac{\alpha}{3} = \frac{0,05}{3} = 0,0167$, et, pour un test bilatéral, z_{adj} = 2,39. Sheskin [2] mène alors le calcul de CD_C en utilisant les 12 observations. Il semble cependant préférable d'éliminer ici les lignes "sans variation" et de prendre n=9.
On obtient alors CD_C=0,65, d'où une différence significative entre la condition 2 et chacune des deux autres conditions, car dans chacun des cas on a : $|p_a - p_b| = 0,89 - 0,22 = 0,67.$

Bibliographie

[1] Cochran, W.G., "The Comparison of Percentages in Matched Samples", Biometrika, 37 (1950), 256-66.

[2] Sheskin, D.J., Handbook of Parametric and Nonparametric Statistical Procedures, 3rd Edition, Chapman and Hall, CRC, 2004.

[3] Tate, M.W. and Brown, S.M., "Note on the Cochran's Q test", Journal of the American Statistical Association, 65 (March 1970), 155-60.