

REVUE DE STATISTIQUE APPLIQUÉE

GISÈLE MERSCH

Tables de contingence 2×2 pour les petites échantillons

Revue de statistique appliquée, tome 22, n° 3 (1974), p. 69-81

http://www.numdam.org/item?id=RSA_1974__22_3_69_0

© Société française de statistique, 1974, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

TABLES DE CONTINGENCE 2x2 POUR LES PETITES ÉCHANTILLONS ⁽¹⁾

Gisèle MERSCH (*)

Dans le présent article nous proposons une méthode de programmation sur ordinateur de l'épreuve d'homogénéité dans une table de contingence 2x2 dont les effectifs sont petits. Cette méthode est basée sur la similitude qui existe entre cette épreuve d'homogénéité et le problème de Wilcoxon des deux échantillons avec ex-aequo. Notre but est de permettre à l'utilisateur de traiter un grand nombre de problèmes successivement avec un minimum d'efforts.

1 – INTRODUCTION

Soient deux dichotomies π_A et π_B de proportions respectives p_A et p_B . Pour éprouver l'hypothèse

$$p_A = p_B$$

à partir de deux échantillons simplement fortuits extraits des populations π_A et π_B et d'effectifs m et n respectivement, on effectue généralement une épreuve chi-carré d'homogénéité sur la table de contingence 2x2 suivante :

	Succès	Echecs	
A	a ₁	a ₂	m
B	b ₁	b ₂	n
	t ₁	t ₂	N

(*) Assistante, chaire de probabilité et statistique, Université de Liège.

(1) Article remis le 9/2/73, révisé le 25/4/73

Cependant l'épreuve chi-carré n'est qu'une approximation valable uniquement pour les grands échantillons. En 1934 F. Yates calcule la probabilité d'observer une table de contingence 2 x 2 donnée conditionnellement aux totaux marginaux m, n, t_1 et t_2 . Cette probabilité est donnée par la loi hypergéométrique, à savoir :

$$\Pr [a_1 = a ; a_2 = b ; b_1 = c ; b_2 = d | m, n, t_1, t_2] = \frac{\binom{t_1}{a} \binom{t_2}{b}}{\binom{N}{m}} \quad (1)$$

La fonction de répartition de la variable aléatoire a_1 est donc donnée par la relation

$$\Pr [a_1 \leq a] = \frac{1}{\binom{N}{m}} \sum_{s=\max(0, t_1 - n)}^a \binom{t_1}{s} \binom{t_2}{m-s} ; \max(0, t_1 - n) \leq a \leq \min(t_1, m) \quad (2)$$

Yates compare les résultats obtenus au moyen de la distribution exacte et de l'approximation chi-carré, constate que cette dernière sous-estime la probabilité de dépassement, et propose des corrections de continuité à appliquer à l'épreuve chi-carré lorsque les effectifs sont petits. Il donne les conditions d'application de la correction de continuité et conseille d'utiliser cette dernière si la plus petite répétition théorique est inférieure à 500.

Cependant, même corrigée, l'approximation chi-carré ne donne pas toujours une probabilité de dépassement très exacte, surtout lorsque les effectifs m et n sont très différents.

Depuis cette époque des tables de la distribution exacte des tables de contingence 2 x 2 ont été établies. Parmi celles-ci citons par exemple celles de R. Latscha [voir 3], B.J. Finney, R. Latscha, B.M. Bennett et P. Hsu [voir 1], et E.S. Pearson et H.O. Hartley [voir 5] dont les étendues sont respectivement :

$$\begin{aligned} 3 &\leq \max(m, n) \leq 20 ; \\ 3 &\leq \max(m, n) \leq 40 ; \\ \text{et} \quad 3 &\leq \max(m, n) \leq 15. \end{aligned}$$

Néanmoins ces tables deviennent rapidement encombrantes et sont encore insuffisantes pour qu'on puisse se contenter de l'approximation chi-carré lorsque les effectifs m et n dépassent leur étendue.

Aussi avons-nous pensé qu'un algorithme de calcul permettant une construction rapide d'un nombre quelconque de distributions exactes pourrait rendre service à de nombreux utilisateurs en leur évitant des manipulations ennuyeuses.

2 – RELATION EXISTANT ENTRE L'ÉPREUVE DE WILCOXON DES DEUX ÉCHANTILLONS AVEC EX-AEQUO ET L'ÉPREUVE D'HOMOGÉNÉITÉ.

Conditionnellement aux totaux marginaux, les variables aléatoires a_1 , a_2 , b_1 et b_2 peuvent toutes s'exprimer linéairement en fonction de l'une d'entre elles. Par exemple, en fonction de a_1 , on a :

$$a_2 = m - a_1 \quad ;$$

$$b_1 = t_1 - a_1 \quad ;$$

$$b_2 = n - t_1 + a_1$$

Etablir la distribution exacte d'une table de contingence 2 x 2 revient donc à établir celle de la variable aléatoire a_1 .

En 1956 C. Van Eeden [voir 7] a montré que le problème des deux échantillons avec ex-aequo est identique à l'épreuve d'homogénéité. Si l'on note x_1, \dots, x_m et y_1, \dots, y_n les échantillons issus des populations π_A et π_B respectivement et (*)

$$D = \neq \{(x_i, y_j) : x_i = y_j ; i = 1, \dots, m ; j = 1, \dots, n\},$$

$$E = \neq (x_i, y_j) : x_i > y_j ; i = 1, \dots, m ; j = 1, \dots, n\}$$

$$W = 2E + D,$$

on voit aisément qu'entre les variables aléatoires W et a_1 on a la relation :

$$W = mn - (Na_1 - mt_1). \quad (3)$$

Par exemple, dans la table suivante

	succès	échecs	
A	5	5	10
B	10	5	15
	15	10	25

on a

$$D = 5 \times 5 + 5 \times 10 = 75$$

$$E = 5 \times 10 = 50$$

$$W = 2 \times 50 + 75$$

$$= 175$$

$$= 10 \times 15 - (25 \times 5 - 10 \times 15)$$

$$= 175$$

(*) [$\neq A$ désignant le nombre d'éléments de l'ensemble A]

Il est donc équivalent d'établir la distribution de W ou de a_1 . Or, en 1956 également L.J. Smid [voir 6] donne des formules de récurrence permettant d'établir la distribution des variables aléatoires W ou S , entre lesquelles on a la relation :

$$S = m n - W \quad (4)$$

ou encore

$$S = N a_1 - m t_1.$$

Appliquées au cas de deux dichotomies, ces formules se simplifient en la relation (2).

3 – PROBABILITES DE DEPASSEMENT UNILATERALES ET BILATERALES.

L'épreuve chi-carré est essentiellement bilatérale, c'est-à-dire qu'elle ne permet que d'éprouver l'hypothèse

$$H_0 : a_1 = a$$

vis-à-vis de l'hypothèse contradictoire

$$H_1 : a_1 \neq a.$$

Cependant, en pratique, il est parfois intéressant de pouvoir considérer des hypothèses contradictoires unilatérales. En effet, l'hypothèse sous épreuve peut être fautive parce que le nombre a de succès observés dans l'échantillon extrait de la population A est soit trop grand, soit trop petit par rapport au nombre total t_1 de succès. Les probabilités de dépassement correspondant à ces épreuves d'hypothèses unilatérales ne peuvent être calculées que si la distribution exacte de a_1 est connue.

Elles sont respectivement égales à

$$PD_d = \frac{1}{\binom{N}{m}} \sum_{s=a}^{\min(t_1, m)} \binom{t_1}{s} \binom{t_2}{m-s} \quad (5)$$

et

$$PD_g = \frac{1}{\binom{N}{m}} \sum_{s=\max(0, t_1-n)}^a \binom{t_1}{s} \binom{t_2}{m-s} \quad (6)$$

D'autre part, dans certains cas pratiques, il est impossible de préciser si ce sont les petites ou les grandes valeurs de a qui sont défavorables à l'hypothèse sous épreuve. Il convient alors de calculer une probabilité de dépassement bilatérale.

Si la distribution de a_1 est symétrique, c'est-à-dire

$$\text{si } m = n \text{ ou si } t_1 = t_2,$$

la probabilité de dépassement correspondant à l'épreuve bilatérale est :

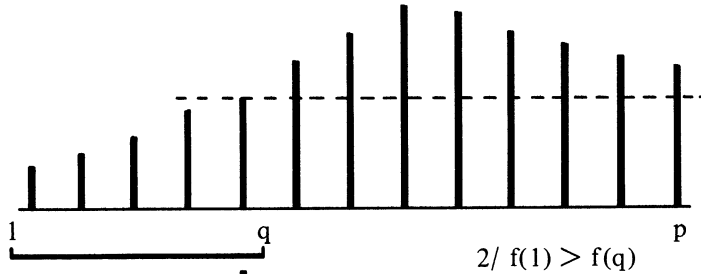
$$PD_b = 2 \min (PD_g, PD_d).$$

Si, par contre, la distribution de a_1 est dissymétrique, l'ensemble R des valeurs de a_1 au moins aussi défavorables à l'hypothèse sous épreuve que la valeur observée a , peut être défini de diverses manières. Mais, pour obtenir les plus grandes régions critiques bilatérales à un niveau d'incertitude donné, il faut définir R comme étant l'ensemble de toutes les valeurs de a_1 dont la probabilité est inférieure ou égale à celle de a .

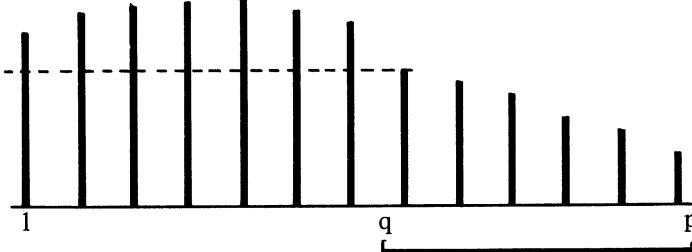
Soient donc $a_1(1), \dots, a_1(p)$ les p valeurs possibles de a_1 rangées par ordre de grandeur croissante et soient $f(1), \dots, f(p)$ les probabilités correspondantes. La valeur observée de a_1 , a , est évidemment l'une des valeurs $a_1(i)$, disons : $a_1(q)$

Trois cas peuvent se présenter (Fig. 1)

1/ $f(p) > f(q)$



2/ $f(1) > f(q)$



3/ $f(1) < f(q), f(p) < f(q)$

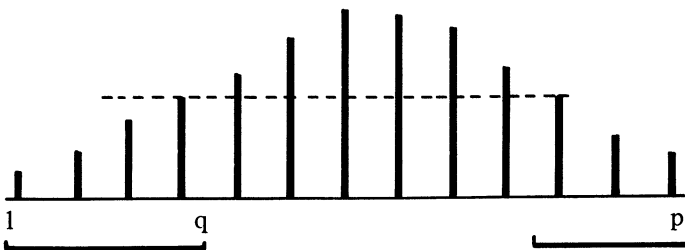


figure 1

Constitution de la zone critique R

1) La probabilité $f(p)$ de la plus grande valeur, $a_1(p)$, de a_1 est strictement supérieure à chacune des probabilités $f(1), \dots, f(q)$ correspondant aux q premières valeurs de a_1 . Dans ce cas, on a

$$R = \{a_1(1), a_1(2), \dots, a_1(q)\}.$$

2) La probabilité $f(1)$ de la plus petite valeur de a_1 est strictement supérieure à chacune des probabilités $f(q), f(q+1), \dots, f(p)$ relatives aux $p - q + 1$ plus grandes valeurs de a_1 .

Dans ce cas, on a

$$R = \{a_1(q), a_1(q+1), \dots, a_1(p)\}.$$

3) Il existe aux deux extrémités de la distribution de a_1 , des valeurs discrètes dont la probabilité est inférieure ou égale à $f(q)$. Dans chaque cas, il y a des éléments de R à chaque extrémité de la distribution de a_1 .

Soient v_1, v_2, \dots, v_l les indices des l valeurs de a_1 dont l'ensemble R est constitué.

La probabilité de dépassement bilatérale est définie par

$$PD_b = \sum_{j=1}^l f(v_j). \quad (7)$$

4 – EXEMPLES NUMERIQUES ET PROGRAMMATION.

Le seul algorithme de calcul pratique de la distribution de la variable aléatoire W qui ait été publié est, à notre connaissance, celui de J. Klotz (1966) [voir 2] qui donne un procédé permettant d'énumérer directement les valeurs discrètes de W . Cet algorithme, différent des formules de récurrence de L.J. Smid, peut être utilisé pour les dichotomies.

Cependant, ayant programmé la construction des distributions de W suivant les méthodes de L.J. Smid et de J. Klotz, nous nous sommes aperçus que la première était beaucoup plus efficace que la seconde, qui est de toute façon trop lourde pour un simple problème de table de contingence 2×2 .

De plus, la version originale de la seconde méthode contenait une erreur (corrigée l'année suivante par C. Odoroff [voir 4]), qui a dû contribuer à son manque de diffusion.

Aussi avons-nous écrit un programme en langage Fortran IV, composé d'un *programme principal* qui calcule les probabilités de dépassement exactes unilatérales et bilatérales ainsi que l'approximation chi-carré avec la correction de F. Yates, d'un *sous-programme* DIST qui construit la distribution de la variable aléatoire S conditionnellement aux totaux marginaux, et d'une *fonction* COMB qui calcule les combinaisons $\binom{m}{s}$.

Notons que le sous-programme DIST peut servir de première étape dans la construction générale des distributions de la variable aléatoire W relative au problème des deux échantillons. Nous nous proposons d'ailleurs de reprendre ce dernier dans un article ultérieur.

Signalons enfin que le programme proposé ici (et dont l'état imprimé est repris en annexe) est très efficace puisqu'il permet de traiter 40 problèmes successivement en 3 secondes sur ordinateur IBM 360/65.

Nous donnons ci-dessous trois exemples numériques, un d'une distribution symétrique et deux de distributions dissymétriques.

Premier exemple

	SUCCES	ECHECS	
A	1	11	12
B	7	5	12
	8	16	24

DISTRIBUTION DE A1

A1	S	R	F*10EXP+4	F CUMULEES
0	-96	1820.000000	6.7303809	*10EXP+4
1	-72	34944.000000	129.22331	6.7303809
2	-48	224224.000000	829.18293	135.95369
3	-24	640640.000000	2369.0941	965.13663
4	0	900900.000000	3331.5386	3334.2307
5	24	640640.000000	2369.0941	6665.7693
6	48	224224.000000	829.18293	9034.8634
7	72	34944.000000	129.22331	9864.0463
8	96	1820.000000	6.7303809	9993.2696
				10000.000

NOMBRE DE CAS POSSIBLES= 2704156.00000000

A1= 1 S= -72

ZONE CRITIQUE=PETITES VALEURS DE A1 .. PD= 135.9537*10EXP-4
 GRANDES VALEURS DE A1 .. PD= 9993.2696*10EXP-4

PD BILATERALE= 271.9074*10EXP-4

APPROXIMATION CHI-CARRE

PD BILATERALE= 303.8282*10EXP-4

Deuxième exemple

	SUCCES	ECHECS	
A	17	10	27
B	2	8	10
	19	18	37

DISTRIBUTION DE A1

A1	S	R	F*10EXP+4	F CUMULEES
				*10EXP+4
9	-180	92378.00000	2.6520243	2.6520243
10	-143	1662804.000	47.736438	50.388463
11	-106	11564046.00	331.98523	382.37369
12	-69	41116608.00	1180.3919	1562.7656
13	-32	83023920.00	2383.4837	3946.2493
14	5	99628704.00	2860.1804	6806.4297
15	42	71954064.00	2065.6859	8872.1156
16	79	30837456.00	885.29394	9757.4096
17	116	7482618.000	214.81397	9972.2235
18	153	923780.0000	26.520243	9998.7438
19	190	43758.00000	1.2562221	10000.000

NOMBRE DE CAS POSSIBLES= 348330136.000000

A1= 17 S= 116

ZONE CRITIQUE=PETITES VALEURS DE A1 .. PD= 9972.2235*10EXP-4
 GRANDES VALEURS DE A1 .. PD= 242.5904*10EXP-4
 PD BILATERALE= 292.9789*10EXP-4
 APPROXIMATION CHI-CARRE PD BILATERALE= 509.7518*10EXP-4

Troisième exemple

	SUCCES	ECHECS	
A	5	15	20
B	3	69	72
	8	84	92

DISTRIBUTION DE A1

A1	S	R	F*10EXP+4	F CUMULEES
				*10EXP+4
0	-160	0.10735998890 20	1285.8726	1285.8726
1	-68	0.26427074190 20	3165.2249	4451.0975
2	24	0.26627279300 20	3189.2039	7640.3014
3	116	0.14307194850 20	1713.6021	9353.9034
4	208	0.44709983900 19	535.50065	9889.4041
5	300	0.82940259990 18	99.339251	9988.7433
6	392	0.88864564280 17	10.643491	9999.3868
7	484	0.50064543260 16	0.59963330	9999.9865
8	576	0.11299289280 15	0.135333910-01	10000.000

NOMBRE DE CAS POSSIBLES= 0.834919322388326D 20

A1= 5 S= 300

ZONE CRITIQUE=PETITES VALEURS DE A1 .. PD= 9988.7433*10EXP-4
 GRANDES VALEURS DE A1 .. PD= 110.5959*10EXP-4
 PC BILATERALE= 110.5959*10EXP-4
 APPROXIMATION CHI-CARRE PD BILATERALE= 132.6280*10EXP-4

Les deux premiers exemples traités nous ont été proposés par un laboratoire de botanique étudiant la floraison de *Sinapis alba* L. soumis à diverses conditions de culture. En se basant sur son expérience, le botaniste détermine à priori certains traitements susceptibles soit de réduire, soit d'augmenter la proportion de floraisons survenant lorsque les plantes sont, soit placées dans des conditions naturelles, soit soumises à un traitement dont les effets sont bien connus.

Les hypothèses à éprouver sont donc de type unilatéral.

Par convention, appelons succès une floraison.

Dans le premier exemple repris ci-dessus, l'hypothèse à éprouver était : le traitement A diminue la proportion de floraison par rapport au traitement B. Etant donné que sur 24 plantes le nombre total de floraisons observées est égal à 8, la probabilité d'obtenir 0 ou 1 floraison dans l'échantillon A d'effectif 12, est égal à 0,0136, si elle est calculée au moyen de la distribution exacte de a_1 , et à 0,0304 si elle est calculée par l'approximation chi-carré avec correction de F. Yates.

Si le niveau d'incertitude choisi a priori est 2 %, nous voyons donc que si l'expérimentateur utilise l'épreuve exacte de Fisher, il ne rejettera pas l'hypothèse que les conditions expérimentales qu'il a imposées à ses cultures entraînent une diminution de la proportion de floraisons.

Par contre, l'usage de l'approximation chi-carré l'amènera à conclure que son expérience n'a pas atteint l'objectif recherché.

Le deuxième exemple est semblable au premier ; il n'en diffère que par le fait que le traitement A était destiné à augmenter la proportion de floraisons par rapport au traitement B.

Sachant que sur 37 plantes 19 ont fleuri et que l'effectif de l'échantillon A est égal à 27, les probabilités exactes et approximatives d'observer au moins 17 floraisons sont respectivement 0,0243 et 0,0510.

Si le niveau d'incertitude choisi est 5 %, les conclusions à formuler seront identiques à celles du premier exemple.

Quant au troisième exemple, il n'est autre que celui donné par F. Yates et sort de la portée des tables.

BIBLIOGRAPHIE

- [1] B.J. FINNEY, R. LATSCHA, B.M. BENNETT, P. HSU — Tables for testing significance in a two-by-two contingency table, 1963, Cambridge, University Press.
- [2] J. KLOTZ — "The Wilcoxon, ties and the computer", *Journal of the Am. Stat. Ass.*, vol. 61, n° 315, 1966 pp. 772 à 788.

- [3] R. LATSCHA – Tests of significance in a 2×2 contingency table : extension of Finney's tables", *Biometrika* 40, 1953, pp. 74 à 86.
- [4] C. ODOROFF – Corrigenda du *Journal of the Am. Stat. Ass.*, vol. 62, n° 320, 1967, p. 1520.
- [5] E.S. PEARSON H.O. HARTLEY – "Biometrika tables for statisticians, vol. 1, 1958, Cambridge University Press.
- [6] L.J. SMID – "On the distribution of the test statistics of Kendall and Wilcoxon's two sample test when ties are present" *Statistica Neerlandica*, jaargang 10, n° 3/4, 1956, pp. 205, à 214.
- [7] C. VAN EEDEN – "Verdelingsvrije toetsen voor twee steekproeven en de methode der 2×2 table" *Statistica Neerlandica*, jaargang 10, n°3/4, 1956 pp. 157 à 162.
- [8] F. YATES, B.A. – "Contingency tables involving small numbers and the X^2 test" *Suppl. to the Journal of the Royal Stat. Soc.*, vol. 1 n.° 2, 1934, pp. 217 à 235.

ANNEXE

Etats imprimés des programmes en langage Fortran IV

p. 79 - 80 : Programme principal
 p. 81 { : Fonction COMB
 { : Sous-programme DIST

(voir ci-dessus, parag. 4)


```

      IND=0
      PDG=0.
      DO 8 I=1,NCAS
      IND=IND+1
      PDG=PDG+R(I)
      IF(IS(I).EQ.ISO) GO TO 20
8 CONTINUE
20 PDD=0.
      DO 9 I=IND,NCAS
9 PDD=PDD+R(I)
      PDG=1.E+4*PDG/PN
      PDD=1.E+4*PDD/PN
C
C   PROBABILITE DE DEPASSEMENT BILATERALE EXACTE
C
      PDB=0.
      I=1
      J=NCAS
2 IF(IS(I).GT.ISO.OR.IS(J).LT.ISO) GO TO 3
      IF(DABS(R(I)-R(J)).LT.1.D-1) GO TO 6
      IF(R(I)-R(J))5,6,7
5 PDB=PDB+R(I)
      I=I+1
      IF(I-NCAS)2,2,3
6 PDB=PDB+R(J)+R(I)
      I=I+1
      J=J-1
      IF(I.GT.NCAS.OR.J.LT.1) GO TO 3
      GO TO 2
7 PDB=PDB+R(J)
      J=J-1
      IF(1-J)2,2,3
3 PDB=1.E+4*PDB/PN
      IF(PDB.GE.1.E+4) PDB=1.E+4
C
C   PROBABILITE DE DEPASSEMENT BILATERALE APPROCHEE
C
      FNN=DFLOAT(NN)
      RT1=DFLOAT(IT1*M)/FNN
      RT2=DFLOAT(IT2*M)/FNN
      RT3=DFLOAT(IT1*N)/FNN
      RT4=DFLOAT(IT2*N)/FNN
      DA=DABS(IA1-RT1)-0.5
      DB=DABS(IA2-RT2)-0.5
      DC=DABS(IB1-RT3)-0.5
      DD=DABS(IB2-RT4)-0.5
      CHICA=DA*DA/RT1+DB*DB/RT2+DC*DC/RT3+DD*DD/RT4
      GAUSS=DSQRT(CHICA/2.D+0)
      PDCH=DERFC(GAUSS)*1.E+4
      WRITE(6,34)PDG
      WRITE(6,35)PDD
      WRITE(6,36)PDB
      WRITE(6,37)PDCH
      GO TO 19
      END

```

```

FUNCTION COMB(M, KK)
IMPLICIT REAL*8(A, B, C, D, E, F, G, H, O, P, Q, R, S, T, U, V, W, X, Y, Z)
NK=M-KK
K=MINO(KK, NK)
IF(K.EQ.0) GO TO 1
COMB=1.
IK=K+1
DO2 I=2, IK
2 COMB=COMB*DFLOAT(M+2-I)/DFLOAT(I-1)
RETURN
1 COMB=1.
RETURN
END

```

```

SUBROUTINE DIST( IT1, IT2, M, N, NN, R, IS, NCAS, COMB)
C
C
C
CALCUL DE LA DISTRIBUTION DE LA VARIABLE ALEATOIRE S
IMPLICIT REAL*8(A, B, C, D, E, F, G, H, O, P, Q, R, S, T, U, V, W, X, Y, Z)
DIMENSION IS(1), R(1)
NN=IT1+IT2
N=NN-M
JMAX=MINO(IT2, M)
NU=IT2-N
JMIN=MAXO(0, NU)
NCAS=JMAX-JMIN+1
IS(1)=IT2*M-JMAX*NN
NT2=IT2-JMAX
MJ=M-JMAX
NJ=N-IT2+JMAX
R(1)=COMB(IT2, JMAX)*COMB(IT1, MJ)
IF(NCAS.EQ.0) RETURN
DO 2 K=2, NCAS
IS(K)=IS(K-1)+NN
KJ2=K-2
KJ1=K-1
DNUM=DFLOAT(JMAX-KJ2)*DFLOAT(NJ-KJ2)
DEN=DFLOAT(NT2+KJ1)*DFLOAT(MJ+KJ1)
2 R(K)=R(K-1)*DNUM/DEN
RETURN
END

```