# Bulletin de Méthodologie Sociologique

http://bms.sagepub.com


## Introduction a l'utilisation du logiciel d'Inference bayésienne sur les frequences (IBF2XK)


Jean-Marc Bernard and Philippe Cibois



*Bulletin de Méthodologie Sociologique* 1988; 20; 12
DOI: 10.1177/075910638802000104

*Bulletin de méthodologie sociologique*, n. 20, octobre 1988, pp. 12-24

# INTRODUCTION A L'UTILISATION DU LOGICIEL D'INFERENCE BAYESIENNE SUR LES FREQUENCES (IBF2XK)

**par**

Jean-Marc Bernard
(GMP-CNRS, 12 rue Cujas, 75005 Paris; tél 43 29 12 13, p 32 48)
et
Philippe Cibois
(LISH-CNRS, 54 bd Raspail, 75006 Paris; tél. 42 22 83 96)

Summary. **Introduction to the use of the computer program "Bayesian Inference based on Frequencies" (IBF2XK).** With several different examples of frequency distributions, the authors show how to use this microcomputer program of bayesian analysis to answer typical questions of statistical inference that arise in social science research. Examples include inferences from one or two frequency distributions, significance tests, variance from independance, and confidence intervals. **Bayesian analysis, statistical inference, frequency distributions, significance tests, variance from independance, confidence intervals.**

**Résumé:** A travers plusieurs exemples de distributions de fréquences différentes, les auteurs montrent comment utiliser ce logiciel de micro-ordinateur d'analyse bayésienne pour répondre à des questions typiques d'inférence statistique rencontrées dans la recherche en sciences sociales. Les exemples comprennent les inférences sur une ou deux distributions de fréquences, les tests de signifiance, l'écart à l'indépendance, et les intervalles de confiance. **Analyse bayésienne, inférence statistique, distributions de fréquences, tests de signifiance, écart à l'indépendance, intervalles de confiance.**

## Introduction

Le but de cet article est d'aider à voir, à partir d'exemples concrets, comment utiliser le programme IBF2XK, tant du point de vue technique que pour la compréhension des résultats. On trouvera des utilisations plus complexes dans Bernard 1986.

Le but de ce programme IBF2XK, qui signifie **Inférence Bayésienne sur les Fréquences**, pour (au maximum) 2 groupes d'observations catégorisées en K classes, est de tester dans quelle mesure des résultats, obtenus sur des fréquences à partir d'un nombre limité d'individus pris au hasard, peuvent être généralisés à une population plus vaste.

## Exemple simple: inférence sur une fréquence

Pour montrer l'emploi du programme nous partirons de l'exemple suivant. En débarquant à Calais, un Anglais fait les observations suivantes sur les 10 premières personnes qu'il rencontre:

Yeux foncés: 8
Yeux bleus: 2

On est donc ici en présence d'un groupe d'observations de 10 individus qui se répartissent en 2 classes: yeux foncés, yeux bleus.

Etant plus prudent que son compatriote qui voyait toutes les Françaises rousses à partir de 3 observations, il ne pense pas que 80% des Français aient les yeux foncés, mais que cette proportion est certainement plus forte que la moitié, dans la mesure où l'on peut considérer que cet échantillon de 10 personnes est représentatif de la population française en ce qui concerne la couleur des yeux. C'est ce que nous allons mettre à l'épreuve avec IBF2XK.

Pour lancer le programme il suffit, sur un micro-ordinateur compatible disposant de 512K de mémoire, de taper IBF2XK puis RETURN. Le programme répond:

```
INFERENCE BAYESIENNE SUR DES FREQUENCES -  IBF2XK  - VERSION 2.4
*****************************************************************
METHODE: J-M. BERNARD (CNRS-UA1201) PROGRAMME: J. POITEVINEAU

ENTREZ LE NOMBRE DE COLONNES DE VOTRE TERMINAL (80 OU 132 ):
```

Répondre 80 pour les écrans habituels puis RETURN (que l'on supposera toujours dorénavant). Le programme demande la confirmation suivante:

```
NOMBRE DE COLONNES ADOPTE:  80  OK ? (O/N):
```

Répondre O pour oui. Le programme est prêt à enregistrer une demande en affichant les lignes suivantes:

```
- I . B . F -
ENTREZ VOTRE DEMANDE (FIN& POUR FINIR)
```

Les demandes se font en deux temps: dans une première étape, on donne les caractéristiques de l'observation (étape IBF dans le langage du programme); dans une deuxième étape, on fait des demandes d'inférence sur les données entrées (étape appelée INF).

La première étape doit commencer par IBF, puis comporter, séparés par des points-virgules, les renseignements suivants:

- Les effectifs du groupe observé, repérés par le mot-clé G1EF= suivi des effectifs des différentes classes de l'observation, séparés par de simples virgules. Dans le cas présent, la demande commencera donc par: IBF; G1EF=8,2; puisque 8 et 2 sont les effectifs des deux classes (yeux foncés et yeux bleus).

- Les "forces" initiales sur les classes qui traduisent l'information ou l'opinion que l'on a, *a priori*, sur les fréquences vraies (les fréquences dans la population): le principe des méthodes bayésiennes est la modification, ou l'actualisation, par l'observation, de cette information ou opinion *a priori*. Lorsqu'on ne dispose pas d'information initiale, ou qu'on ne veut pas en tenir compte dans l'analyse, on prendra des forces initiales exprimant un état d'ignorance: c'est la "solution standard" qui consiste à répartir une force totale de 1 (qui pèsera peu par rapport à l'effectif de l'échantillon), équitablement entre les classes (pour ne favoriser aucune fréquence particulière).

Cette répartition *a priori* est codée par le mot clé G1IN= (IN signifiant INitiale) avec les forces de chaque classe séparées par des virgules, une force de 0,5 étant codée 0.5 ou même .5

Ici on aurait G1IN=.5,.5.

On va demander le graphique de densité de probabilité en entrant le mot-clé DENS.

Enfin, on va préciser sur quoi va se faire l'inférence bayésienne par le mot-clé COMP (pour COMParaison). Quand cette demande porte sur une classe (ici la première, relative à l'effectif 8), il suffit de donner son numéro d'ordre, ici 1. Le mot-clé est codé de la manière suivante: COMP=1.

L'écriture se fait en format libre sur une ou plusieurs lignes, les mots-clés étant séparés par des points-virgules et la fin de demande étant signalée par un &.

La demande précédente est donc:

IBF; G1EF=8,2 ; G1IN= .5 , .5;DENS ;COMP=1  ;&

La commande est lancée par RETURN et le programme répond:

```
COMMANDES LUES:
IBF;G1EF=8,2;G1IN=.5,.5;DENS;COMP=1;

==========================  I . B . F  ======================
-----------  INFERENCE BAYESIENNE SUR DES FREQUENCES  ----------



(ERREUR SUR PROBA=1.00E-05  MAX ITER = 2000. APPROX BETA >1000.)

INFERENCE BAYESIENNE DANS UN PROTOCOLE S------> U2
COMPARAISON SUR UNE FREQUENCE:  1(G1)

--EFFECTIFS DE G1    :      8.      2.
--DISTR. INIT. DE G1 :   .5000   .5000

--FREQUENCE OBSERVEE =  .800000

***
```

Les commandes sont lues, les erreurs éventuelles détectées (ici il n'y en a pas - "ERREUR SUR PROBA" indique la précision des calculs), les effectifs et les distributions *a priori* affichés, et la fréquence sur laquelle porte l'inférence est calculée (ici .8 qui correspond à la proportion de l'effectif de la première classe à 8 individus).

Les 3 étoiles signifient que l'écran est plein. Pour avoir la suite, il suffit de faire RETURN et on a alors:

```
--Caracteristiques de la distribution BAYESIENNE :
     PHI distribuee BETA (8.5000000 , 2.5000000)
     Moyenne    (PHI)=  .7727273
     Variance   (PHI)= 0.1463499E-01
     Ecart-type(PHI)=  .1209751
     Bornes  inf =  0        sup =  1
```

```
*** TAPEZ "ENTREE" POUR AVOIR LE GRAPHIQUE DE LA DENSITE
```

On nous donne ici les paramètres de la courbe de densité de la fréquence examinée (ici celle de la première catégorie). On remarquera que pour chaque classe les effectifs sont augmentés de la force *a priori*. La moyenne est proche de la fréquence observée .8, ce qui sera le cas général. L'écart-type est assez fort mais cela vient de la faiblesse de notre échantillon. Les bornes données sont celles de la fréquence sur laquelle porte l'inférence. Elle ne peut varier que de 0 à 1.

En tapant "ENTREE" (c'est à dire RETURN), on a:

```
                DENSITE DE BETA (8.5000000 , 2.5000000)

3.3716    +                                  *******
          !                                ***********
          !                             ***************
          !                           ******************
          !                          ********************
          !                        ***********************
          !                      **************************
1.6857    +                    ****************************
          !                  *******************************
          !                ***********************************
          !             *************************************
          !          *****************************************
          !      *********************************************
          !***************************************************
 .0000    ++---------+---------+---------+---------+---------+
         .43951              .65187              .86422
                   .54569              .75804             .97040

---------------------------------------------------------------
```

La lecture de cette courbe se fait d'une manière analogue à celle d'une loi normale; la surface lue sous la courbe nous donne la probabilité recherchée. On a en ordonnée la densité de la probabilité (nombre sans dimension) et en abscisse la valeur de la fréquence. Par exemple, si on cherche la probabilité pour que que la fréquence de la catégorie examinée soit supérieure à .5, on regarde l'aire de la courbe à droite d'une verticale qui a pour abscisse .5. C'est la figure suivante:

```
3.3715    +                                  *******
          !                                ***********
          !                             ***************
          !                           ******************
          !                          ********************
          !                        ***********************
          !                      **************************
1.6857    +                    ****************************
          !                  *******************************
          !                ***********************************
          !   !         *************************************
          !   !      *****************************************
          !   !  *********************************************
          !   !***********************************************
 .0000    ++--!------+---------+---------+---------+---------+
             .5              .65187              .86422
                   .54569              .75804           .97040
```

On voit que l'on prend presque toute la surface sous la courbe, ce qui veut dire qu'on a presque toutes les chances d'avoir raison en affirmant que la proportion des yeux foncés dépasse 50% de la population, au vu d'une observation de 8 individus aux yeux foncés sur 10 observés.

La valeur exacte de cette probabilité va être obtenue en demandant maintenant des inférences. Cette demande est introduite par le mot clé INF suivi par la valeur de la fréquence dont on veut qu'elle serve de limite inférieure. Le mot-clé LIM= introduit cette valeur. Ici la demande est:

INF ; LIM= .5 ; &

On notera que, comme précédemment, les mots-clés sont séparés par des points-virgules et que la demande est terminée par &. La demande est introduite à la suite de l'affichage suivant:

```
- I . B . F -
ENTREZ VOTRE DEMANDE (FIN& POUR FINIR):
```

On introduit la demande et on obtient:

```
COMMANDES LUES:
INF;LIM=.5
;

 ...   .90000 = PROB (PHI >  .60516975)
 ...   .95000 = PROB (PHI >  .54749588)
 ...   .99000 = PROB (PHI >  .43950636)
 ... PROB (PHI >  .50000000) =   .97396
```

La réponse à notre question se trouve à la dernière ligne qui nous dit que la probabilité pour que la fréquence examinée soit, dans la population de référence, supérieure à la moitié est de .97.

Les autres lignes nous donnent (par défaut et donc sans que nous ayons à les demander) les valeurs de la fréquence correspondant aux seuils de confiance standards de .90, .95 et .99. Par exemple, on peut dire que l'on a 9 chances sur 10 d'avoir raison en disant que la proportion des yeux foncés est supérieur à 60% (.605).

Pour terminer le programme, il suffit d'entrer FIN&.

```
- I . B . F -
ENTREZ VOTRE DEMANDE (FIN& POUR FINIR):
COMMANDES LUES:
FIN
FIN DES DEMANDES
STOP
```

## Quelques expériences

Si au lieu de la distribution 8 contre 2, on entre la distribution 7 contre 3. Que se passe-t-il? Tout d'abord, du point de vue du programme, il suffit de relancer une étape IBF en changeant seulement G1EF, car tous les autre paramètres conservent leur valeur antérieure qui est gardée par défaut. La demande se réduit donc à:

IBF ; G1EF=7,3;&

De même pour la demande d'inférence, comme on conserve la même demande que précédemment (LIM=.5), il suffit de relancer INF& pour obtenir le résultat qui est le suivant:

```
PROB (PHI > .5 ) = .89798
```

Ce résultat signifie que l'on a encore pratiquement 9 chances sur 10 d'avoir raison en affirmant que la proportion des yeux foncés est supérieure à la moitié.

Si maintenant, de la même façon, on tente l'expérience avec une distribution de 6 contre 4, la probabilité tombe à .73. En prenant le seuil usuel de 9 sur 10, on ne peut plus avoir la même conclusion que précédemment. La proportion observée a beau être de .6, on ne peut plus conclure que dans la population de référence la vraie valeur ait de bonnes chances d'être supérieure à la moitié.

Par contre, si notre observation était de 60 contre 40, la probabilité remonterait à .977. Ce résultat est fondamental car si la proportion observée est bien la même que précédemment, c'est à dire .6, le pouvoir généralisateur de l'observation est bien plus grand du fait que l'effectif est supérieur et nous autorise ainsi à conclure, avec une bonne garantie, que la fréquence dans la population dépasse .50.

On observe une fréquence de .6. Il est raisonnable de penser que la vraie valeur, dans la population de référence, n'en sera pas tellement loin. Mais pour l'affirmer, il faudra accepter de s'écarter de cette valeur, et ceci d'autant plus que l'échantillon est petit.

Plus l'effectif d'observation est grand, plus son pouvoir généralisateur est fort, moins la "taxe à payer" pour une inférence est grande.

## Inférence sur deux groupes: les écarts à l'indépendance

On examine ici comment utiliser le programme pour tester le pouvoir généralisateur d'un écart à l'indépendance observé dans un tableau croisé. On reprend les exemples cités dans l'annexe technique d'un article consacré à une analyse secondaire d'une enquête sur la natalité (Cibois 1986).

Le premier cas est relatif à l'écart à l'indépendance de 8,7525 individus observé en croisant l'opinion sur les conséquences de la dénatalité avec l'opinion politique. On observe en effet que entre la modalité "la dénatalité conduit au vieillissement" et la modalité "orientation à gauche", il existe un écart positif à l'indépendance. Le tableau observé est le suivant:

| | | Conséquences de la dénatalité | | |
|---|---|---|---|---|
| | | Vieillissement | autre | total |
| Orientation | Gauche | 150 | 275 | 425 |
| politique | Autre | 187 | 402 | 589 |
| | Total | 337 | 677 | 1014 |

Pour la case "Gauche"-"Vieillissement", l'effectif observé est de 150; l'effectif théorique sous l'hypothèse d'indépendance est de 337 x 425 / 1014 = 141,2475 et l'écart positif à l'indépendance de 150 - 141,2475 = 8,7525.

Le problème est de savoir si cet écart est assez fort pour que l'on puisse conclure que dans la population de référence, il existe bien une attraction entre les deux modalités. Pour ce faire, on va définir dans le programme deux groupes ayant chacun deux modalités de réponse et l'on va faire porter le test inférentiel sur la différence entre les proportions.

Chaque groupe correspond à une ligne du tableau. Le premier groupe de distribution est 150 et 275. Le deuxième est de distribution 187 et 402. La proportion de la modalité "vieillissement" dans le premier groupe est de 150/425= .353, la même proportion dans le deuxième groupe est de 187/589= .317. La proportion est plus forte dans le premier groupe, ce qui indique donc une attraction. La force de cette attraction est donnée par la différence des proportions: .353 - .317 = .035.

Pour savoir si cet écart positif des différences est généralisable, il suffira de tester si la différence est positive en donnant LIM=0 dans les demandes d'inférence.

Les données sont présentées dans IBF de la façon suivante:

```
ENTREZ VOTRE DEMANDE (FIN& POUR FINIR) :
COMMANDES LUES:
IBF;G1EF=150,275;G2EF=187,402;G1IN=.25;G2IN=.25,.25;DENS;COMP=1-1;
```

G1EF correspond à la première ligne du tableau et G2EF à la seconde. On doit donner les distributions *a priori* en utilisant G1IN et G2IN, avec un total des forces *a priori* sur les 2 groupes égal à 1 et un partage équitable de cette force totale entre les 2 groupes, puis entre leurs classes respectives, ceci pour manifester notre ignorance. La demande de comparaison se fait par COMP=1-1, ce qui signifie que l'on examine la différence entre les fréquences des premières classes de chacun des deux groupes. Ne pas oublier de terminer la demande par le & commercial, même si celui-ci n'est pas indiqué par le programme quand il affiche la commande lue.

Le programme donne les indications suivantes:

```
INFERENCE BAYESIENNE DANS UN PROTOCOLE S<G2>--> U2
COMPARAISON PHIx - PHIy :  1(G1) - 1(G2)

--EFFECTIFS DE G1     :    150.   275.
--DISTR. INIT. DE G1 :   .2500  .2500

--EFFECTIFS DE G2     :    187.   402.
--DISTR. INIT. DE G2 :   .2500  .2500

--DIFFERENCE OBSERVEE =  .035454
```

On retrouve le tableau croisé et la différence étudiée. C'est cette dernière dont on étudie la distribution.

```
--Caracteristiques de la distribution BAYESIENNE:
     PHIx distribuee BETA (150.25000 , 275.25000)
     PHIy distribuee BETA (187.25000 , 402.25000)
     DELTA = PHIx-PHIy  (PHIx et PHIy independantes)
     Moyenne    (DELTA)= 0.3547191E-01
     Variance   (DELTA)= 0.9026335E-03
     Ecart-type(DELTA)= 0.3004386E-01
     Bornes  inf = -1       sup =  1
--Pour les calculs ulterieurs cette distribution est APPROXIMEE par:
     2 * BETA (572.34438 , 533.13105) - 1

          DENSITE DE 2 * BETA (572.34438 , 533.13105) - 1

13.269    +                       *****
          !                    ***********
          !                   *************
          !                 *****************
          !                *******************
          !               *********************
          !              ************************
6.6347    +            ***************************
          !           *****************************
          !         *********************************
          !        ***********************************
          !      ****************************************
          !   *********************************************
          !***************************************************
 .0000    ++---------+---------+---------+---------+---------+
        -.03448              .02141              .07729
                 -.00654              .04935              .10523
```

L'aire couverte par la courbe qui nous intéresse est celle située à droite de la différence égale à zéro:

```
13.269    +                       *****
          !                    ***********
          !                   *************
          !            !    *****************
          !            !   *******************
          !            !  *********************
          !            ! ************************
6.6347    +            !**************************
          !            !***************************
          !            !*****************************
          !            !******************************
          !            !*********************************
          !            !***********************************
          !            !**************************************
 .0000    ++---------+-!-------+---------+---------+---------+
       -.03448         0     .02141              .07729
                  -.00654              .04935              .10523
```

On voit que cette surface à droite recouvre la plus grande partie de la probabilité. Pour savoir sa valeur exacte, il suffit, dans l'étape INF, de faire la demande suivante:

```
- I . B . F -
ENTREZ VOTRE DEMANDE (FIN& POUR FINIR):
COMMANDES LUES:
INF;LIM=0

 ...   .90000 = PROB (DELTA > -0.3056584E-02) *
 ...   .95000 = PROB (DELTA >  -.0139863) *
 ...   .99000 = PROB (DELTA >  -.0344789) *
 ... PROB (DELTA >  .00000000) =  .88099 *
```

La dernière ligne nous indique que la surface, qui est à droite de la valeur zéro de la différence, recouvre 88% de la probabilité. On peut donc conclure de manière inférentielle que l'on a environ 9 chances sur 10 d'avoir raison en affirmant qu'il y a une attraction entre crainte du vieillissement et opinion politique de gauche dans la population de référence.

Refaisons l'expérience avec l'opinion politique croisée cette fois avec une autre modalité de la perception des conséquences de la dénatalité: la crainte pour la retraite. Il y a encore attraction entre cette modalité et la gauche, mais avec un écart plus faible. Le tableau est le suivant:

| | | Conséquences de la dénatalité | | |
|---|---|---|---|---|
| | | Peur pour retraite | autre | total |
| Orientation | Gauche | 196 | 229 | 425 |
| politique | Autre | 267 | 322 | 589 |
| | Total | 463 | 551 | 1014 |

L'écart à l'indépendance est de 1,9418 individu, ce qui correspond à une différence de proportion entre les deux groupes, pour la première modalité, de .0078, soit moins de 1%.

Pour entrer les données, il suffit de modifier les paramètres G1EF et G2EF. Tous les autres paramètres restent, par défaut, les mêmes.

```
- I . B . F -
ENTREZ VOTRE DEMANDE (FIN& POUR FINIR):
COMMANDES LUES:
IBF;G1EF=196,229;G2EF=267,322;
```

On a les résultats suivants:

```
INFERENCE BAYESIENNE DANS UN PROTOCOLE S<G2>--> U2
COMPARAISON PHIx - PHIy :  1(G1) - 1(G2)

--EFFECTIFS DE G1      :    196.    229.
--DISTR. INIT. DE G1 :    .2500   .2500

--EFFECTIFS DE G2      :    267.    322.
--DISTR. INIT. DE G2 :    .2500   .2500

--DIFFERENCE OBSERVEE =   .007866

--Caracteristiques de la distribution BAYESIENNE:
     PHIx distribuee BETA (196.25000 , 229.25000)
     PHIy distribuee BETA (267.25000 , 322.25000)
     DELTA = PHIx-PHIy   (PHIx et PHIy independantes)
     Moyenne    (DELTA)= 0.7871795E-02
     Variance   (DELTA)= 0.1002325E-02
     Ecart-type(DELTA)= 0.3165952E-01
     Bornes  inf = -1        sup =  1
--Pour les calculs ulterieurs cette distribution est
     APPROXIMEE par :
     2 * BETA (502.21166 , 494.38649) - 1

          DENSITE DE 2 * BETA (502.23166 , 494.38649) - 1

12.591     +                      *****
           !                   ***********
           !                  *************
           !                ******************
           !               *******************
           !              *********************
           !            *************************
           !           ***************************
           !          *****************************
           !        *********************************
           !       ***********************************
           !     ***************************************
           !  *********************************************
6.2957     +
 .0000     ++---------+---------+---------+---------+---------+
        -.06576              -.00687               .05202
                 -.03632               .02257               .08146
```

On voit sur cette courbe que la valeur 0 pour la différence se trouve pratiquement au centre, ce qui signifie que la probabilité à droite ne va pas être très forte. Si on demande cette limite de LIM=0 dans une étape INF, la valeur répondue est de .59. Elle signifie que l'écart observé pour notre échantillon n'est pas généralisable puisque cette attraction ne se retrouve pas avec une probabilité suffisamment grande. On conclura donc en disant que cet écart n'est pas significatif et que l'on ne peut pas affirmer qu'être de gauche rend plus craintif face aux retraites.

On peut cependant aller plus loin et essayer d'affirmer que l'écart observé est "négligeable". Il suffit pour cela de se mettre d'accord sur ce qu'on appelle une différence de proportion négligeable. Si l'on prenait, par exemple, une fourchette de plus ou moins 5%, on considérerait comme négligeable une différence de proportion qui serait supérieure à -.05 et inférieure à .05. Sur la courbe, cela correspondrait aux valeurs situées entre les deux verticales:

```
12.591    +                       *****
          !                    ***********
          !                   *************
          !                 *****************
          !                ******************* 
          !     !         *********************   !
          !     !       ************************* !
6.2957    +     !      ***************************!
          !     !     ****************************!
          !     !   ******************************!
          !     !  *******************************!
          !     !*********************************!
          !     !*********************************!
          !     !*********************************!
 .0000    ++----!----+---------+---------+--------!+---------+
              -.05          -.00687             .05
                 -.03632              .02257              .08146
```

On voit facilement que la plus grande partie de la probabilité est comprise entre ces deux limites. Pour en savoir la valeur exacte, il suffit de demander, dans le paragraphe INF, ce qui est à droite de la limite -.05 et de le diminuer de ce qui est à droite de la limite .05. Pour obtenir les probabilités correspondantes aux deux limites, on fait la demande suivante:

```
- I . B . F -
ENTREZ VOTRE DEMANDE (FIN& POUR FINIR):
COMMANDES LUES:
INF;LIM=-.05,.05;

 ...   .90000 = PROB (DELTA >  -.0327187) *
 ...   .95000 = PROB (DELTA >  -.0442148) *
 ...   .99000 = PROB (DELTA >  -.0657611) *
 ... PROB (DELTA > -.0500000) =  .96620 *
 ... PROB (DELTA >  .05000000) =  .09170 *
```

Pour obtenir la valeur cherchée, il suffit de faire la différence des deux dernières lignes: .96620 - .09170 = .87, ce qui nous donne la probabilité pour que la différence de proportion soit dans la fourchette. On a donc environ 9 chances sur 10 d'avoir raison en affirmant, au vu de l'observation, qu'il y a un écart négligeable selon l'opinion politique en ce qui concerne la crainte pour les retraites.

Cette conclusion est beaucoup plus intéressante que la première inférence sur les mêmes effectifs qui concluait que l'on ne pouvait pas conclure (ce qui est le propre du résultat non significatif). Ici la conclusion est possible, c'est que l'écart est négligeable.

On remarquera cependant que la fourchette choisie de plus ou moins 5% est certainement trop large (souvenons-nous que dans l'exemple précédent un écart de 3% a été jugé significatif et que nous n'avons de ce fait aucune envie de le considérer comme négligeable). La valeur de 5% a été retenue pour les besoins de l'expérimentation du programme. Dans la pratique, on prendra souvent une fourchette beaucoup plus faible.

## Pour mieux comprendre

L'idée sous-jacente de toute inférence statistique en général est que si l'on tire au hasard des échantillons d'une population, un grand nombre de ces échantillons, du fait du simple jeu de la combinatoire, auront une composition proche de celle de la population. Par exemple, sur une population de 100 personnes, dont 80 ont les yeux foncés et 20 les yeux bleus, il est bien sûr possible d'en tirer 10 qui aient tous les yeux bleus. Mais dans toutes les manières possibles de tirer 10 personnes de 100 (c'est à dire, l'univers de tous les échantillons possibles), la proportion d'échantillons qui n'auront que des yeux bleus est assez restreinte. Si l'on fait l'histogramme du nombre d'échantillons qui ont une fréquence donnée, on trouve quelque chose qui est assez proche d'une courbe de densité. Autour de la vraie valeur, la densité d'échantillons est forte; aux extrêmes, elle est faible.

Dans le cas de l'inférence bayésienne, la situation est différente. A partir d'une observation, on détermine la distribution de probabilité du paramètre que l'on veut observer, ici une fréquence (ou une différence de fréquence). Les courbes de densité données par le programme visualisent cet aspect des chose. Elle nous indiquent, pour toute la gamme des fréquences, quelle probabilité a le paramètre que l'on étudie.

Si l'on reprend la courbe du premier exemple, l'échelle des fréquences va de .43 à gauche à .97 à droite. Pour ces deux extrêmes, la densité de probabilité est très faible, ce qui signifie qu'il y a une faible probabilité pour que la fréquence des yeux foncés soit inférieure à 43% (en fait, elle est de 1% nous dit le programme puisque .99 est la probabilité pour que la fréquence soit supérieure à .4395, ce qui était résumé par l'énoncé suivant: .99 = PROB (PHI > .4395). De même à droite, il n'y a que 1% de probabilité pour que la proportion des yeux foncés soit supérieure à 97%. On constate sur la courbe que la densité de probabilité est très forte au milieu, aux alentours de l'observation.

Sur une telle courbe, le produit d'une densité par une étendue de fréquence nous donne directement une valeur approchée de la probabilité pour que la fréquence soit comprise entre les deux bornes. C'est pour cette raison que sur le graphique l'aire comprise sous la courbe est noircie; elle correspond à la probabilité recherchée. On déduit de cette valeur l'état d'incertitude sur la vraie valeur de la fréquence dans la population de référence. Si la probabilité est suffisamment forte pour que le paramètre soit supérieur à une certaine valeur, on aura confiance dans le résultat.

Par exemple, examinons la partie gauche de la courbe et cherchons la valeur de la probabilité pour que la fréquence soit inférieure à la valeur .65187 (qui correspond à une densité de 1.68).

```
1.68 !                        *
     !                     *  +
     !                  *     +
     !               *        +
     !            *           +
     !        *               +
     !*                       +
0    ++---------+---------+
  .43951               .65187
           .54569
```

Sur cet extrait de la première courbe de densité, on approxime la courbe par une droite. La surface de ce triangle rectangle est égale à la moitié du produit des côtés de l'angle droit, soit en hauteur 1.68 et en largeur .65187, moins une valeur légèrement supérieur à .439, soit approximativement .45. La surface du triangle est égale à (.65 - .45)x1.68/2= .17, ce qui veut dire que la probabilité que la fréquence soit inférieure à .65 est à peu près égale à 17%. On en déduit,à l'inverse, que la probabilité que la fréquence soit SUPERIEURE à .65 est le complément à 100 de .17, soit .83. Nous pouvons retrouver ce résultat par le calcul direct en faisant dans l'étape INF LIM=.65. Le programme nous donne la valeur de la probabilité correspondante égale à .83. On n'a donc raisonnablement que 17 chances sur 100 d'avoir tort et 83 d'avoir raison en affirmant que la vraie valeur est supérieur à .65.

En général, on prend comme seuil de confiance 9 chances sur 10 ou 95 sur 100, ou encore 99 sur 100. Mais il ne faut pas oublier que vouloir un seuil de confiance très élevé, c'est s'interdire des conclusions qu'il serait pourtant raisonnable d'accepter. Dans l'action de tous les jours, nous faisons des choix raisonnables dès que nous nous sortons du choix au hasard (50 contre 50). Avoir 2 chances sur 3 d'avoir raison est souvent considéré dans l'action pratique comme raisonnable; c'est d'ailleurs les chances que Napoléon exigeait pour engager une bataille d'après Joseph Caillaux (règle qu'il appliqua pour son propre compte en refusant d'engager les hostilités avec l'Allemagne à Agadir en 1911). On comprend que si on était plus exigeant pour engager une bataille, on ne livrerait que des combats gagnés d'avance, ce qui n'est pas toujours possible. Pour que les adversaires acceptent de s'affronter, il faut qu'ils estiment avoir des chances raisonnables de gagner et ce ne peuvent être des probabilités trop fortes.

## Références bibliographiques

Jean-Marc Bernard, "Méthodes d'inférence bayésienne sur des fréquences", *Informatique et Sciences Humaines*, n. 68-69, mars - juin 1986, pp. 89-133, numéro spécial sur la statistique bayésienne.

Philippe Cibois, "Analyse secondaire d'une enquête sur la natalité", *Informatique et Sciences Humaines*, n. 70-71, septembre - décembre 1986, pp. 15-30, numéro spécial sur la méthode Tri-deux.

Pour l'apprentissage de l'utilisation d'une courbe de densité, on pourra consulter: Marie-Paule et Bruno Lecoutre, *Enseignement programmé sur l'utilisation d'une table de la distribution normale*, Paris, SEDES-CDU, 1979, 45 pp.

====================================