

REVUE DE STATISTIQUE APPLIQUÉE

D. BOSQ

Tests du X^2 généralisés. Comparaison avec le test du X^2 classique

Revue de statistique appliquée, tome 37, n° 1 (1989), p. 43-52

http://www.numdam.org/item?id=RSA_1989__37_1_43_0

© Société française de statistique, 1989, 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/>

TESTS DU X^2 GÉNÉRALISÉS COMPARAISON AVEC LE TEST DU X^2 CLASSIQUE

D. BOSQ

Université Pierre et Marie Curie

RÉSUMÉ

Le test du X^2 classique fait partie d'une classe de tests d'adéquation basés sur des systèmes orthogonaux. Des considérations théoriques montrent que l'utilisation de systèmes orthogonaux lisses permet de construire des tests plus efficaces que le test usuel. Nous présentons des simulations qui confirment cette idée.

SUMMARY

The classical X^2 -test belongs to a family of goodness of fit tests based upon orthogonal systems, some theoretical results show that use of smooth orthogonal systems leads to more efficient tests than the PEARSON X^2 -test. We give some simulations which corroborate this fact.

Mots clés : Tests d'adéquation, Test du X^2 , Méthode des projections.

I. Introduction

Une manière naturelle de construire un test d'adéquation consiste à utiliser un estimateur de la loi dont on teste l'ajustement. Le calcul de la distance de la loi présumée à son estimée permettra d'apprécier la validité de l'hypothèse de départ.

Ainsi le célèbre test du X^2 est fondé sur un estimateur élémentaire de la densité : l'histogramme des fréquences. Or il est bien connu que les estimateurs lisses de la densité approchent la loi des observations beaucoup mieux que l'histogramme. C'est le cas notamment des estimateurs à noyau et des estimateurs par projection (cf. BOSQ-LECOUTRE(1987)). Aussi peut-on espérer que les tests associés à ces estimateurs sont plus performants que le test du X^2 usuel.

On trouve des résultats théoriques sur les tests basés sur les estimateurs par projection ("tests hilbertiens") dans BOSQ (1980, 1983), GADIAGA (1982), BLACHER (1985).

Le présent travail est consacré à la comparaison d'un test hilbertien et du test du X^2 par simulation. Nous avons choisi le test basé sur les polynômes de LEGENDRE en raison de sa simplicité de mise en œuvre. Ce test n'est pas nouveau puisqu'il avait déjà été envisagé par NEYMAN en 1937 !

Dans la suite, après quelques rappels sur les tests hilbertiens nous donnons des indications sur le choix des cellules dans le test du X^2 puis sur le choix du nombre de polynômes dans le "test de LEGENDRE". Nous présentons ensuite une comparaison

empirique des deux tests. Les simulations ont été effectuées par IZRAELEWICZ - LAFITTE - LAVALT - ROUBERT (1988).

La supériorité du test de LEGENDRE étant nette, nous espérons que cet article contribuera à développer son utilisation par les praticiens de la statistique.

II. Tests hilbertiens

Soient X_1, \dots, X_n des variables aléatoires réelles observées, supposées indépendantes et de même loi et à valeurs dans un intervalle D de longueur finie ou infinie. Soit m une probabilité sur D . On se propose de tester l'hypothèse $H : L(X_i) = m$ (i.e. la loi de X_i est m).

Pour cela on se donne un entier k et des fonctions numériques bornées f_0, f_1, \dots, f_k définies sur D et vérifiant

$$\int_D f_j(x) f_{j'}(x) dm(x) = \delta_{jj'}, \quad j, j' \in \{0, 1, \dots, k\}$$

où $\delta_{jj'}$ est le symbole de KRONECKER. On suppose que l'espace vectoriel E_k engendré par les f_j contient les constantes.

Le test repose alors sur l'estimation des paramètres

$$a_j(p) = \int_D f_j(x) dp(x); \quad j = 0, 1, \dots, k$$

où p désigne la loi de X_i . Les estimateurs

$$\hat{a}_{jn} = \frac{1}{n} \sum_{i=1}^n f_j(X_i); \quad j = 0, 1, \dots, k$$

sont symétrique sans biais. D'où la statistique de test

$$T_n = n \sum_{j=0}^k (\hat{a}_{jn} - a_j(m))^2$$

et un test de région critique $T_n > c_n$ où c_n est choisi d'une manière convenable.

Comme la densité de m par rapport à elle-même est constante et égale à 1, les a_j vérifient la relation remarquable.

$$\sum_{j=0}^k a_j(m) f_j(x) = 1, \quad x \in D.$$

D'ailleurs on peut toujours remplacer le système (f_j) par un système orthonormal de la forme $1, g_1, \dots, g_k$ engendrant le même espace que les f_j . On a alors $a_0(m) = 1$ et $a_j(m) = 0; j = 1, \dots, k$.

Le test du X^2 correspond au choix d'un système de la forme $[m(B_j)]^{-1/2} 1_{B_j}$, $j = 0, 1, \dots, K$ où les B_j forment une $(k+1)$ - partition de D telle que $m(B_j) > 0$ pour tout j .

Pour effectuer le test de LEGENDRE on commence par amener les observations sur $(-1, +1)$ en posant

$$Y_i = 2F(X_i) - 1; \quad i = 1, \dots, n$$

où F désigne la fonction de répartition de m supposée continue. Sous H les Y_i suivent la loi uniforme U sur $(-1, +1)$.

Le système (f_j) est alors obtenu en orthonormalisant par rapport à U les polynômes $1, x, x^2, \dots$, on obtient pour tout j

$$f_j(x) = (2j + 1)^{1/2} P_j(x), \quad x \in (-1, +1)$$

où

$$P_j(x) = \frac{1}{2^j j!} \frac{d^j}{dx^j} (x^2 - 1)^j, \quad x \in (-1, +1).$$

Voici les premières fonctions f_j :

$$\begin{aligned} f_0(x) &= 1 \\ f_1(x) &= \sqrt{3} x \\ f_2(x) &= \frac{1}{2} \sqrt{5} (3x^2 - 1) \\ f_3(x) &= \frac{1}{2} \sqrt{7} (5x^3 - 3x) \\ f_4(x) &= \frac{3}{8} (35x^4 - 30x^2 - 3) \\ f_5(x) &= \frac{1}{8} \sqrt{11} (63x^5 - 70x^3 + 15x) \\ f_6(x) &= \frac{1}{16} \sqrt{13} (231x^6 - 315x^4 + 105x^2 - 5). \end{aligned}$$

Pour des résultats théoriques sur les tests hilbertiens nous renvoyons à BOSQ (1980, 1983). Nous citons les théorèmes suivants :

Théorème 1

Sous H , T_n converge en loi vers un X^2 à k degrés de liberté. De plus la distance uniforme entre la fonction de répartition de T_n et celle du X^2 est un $O(n^{-1/2})$.

Théorème 2

Si une suite (p_n) d'alternatives est telle que p_n admette une densité h_n par rapport à m et si

- a) Pour tout n , $h_n \in E_k$
- b) $n \int_D (1 - h_n^2(x)) dm(x) \rightarrow \ell^2 > 0$.

Alors T_n converge en loi vers un X^2 décentré à k degrés de liberté et de coefficient de décentrage ℓ^2 . De plus la distance uniforme entre les fonctions de répartition est un $O(n^{-1/2})$.

Ces propositions s'appliquent au test du X^2 et au test de LEGENDRE.

III. Le choix des cellules dans le test du X^2

Des études récentes ont modifié les critères de choix des cellules dans le test du X^2 . Nous donnons maintenant quelques indications sur les règles empiriques énoncées par BEST - RAYNER (1981, 1982, 1985), OOSTERHOOF (1985), KALLENBERG - OOSTERHOOF - SCHIEVER (1985).

Si le choix de cellules équiprobables sous H est considéré comme acquis, c'est le nombre de cellules à choisir qui a été critiqué par les auteurs précités. Ils se sont rendus compte que le nombre $k' = k + 1$ de cellules assurant une bonne puissance du test était beaucoup plus petit que ne l'indiquaient les règles usuelles de MANN - WALD (1942) et KENDALL (1973) qui conseillaient de choisir k' de l'ordre de $n^{2/5}$. Leurs résultats font l'effet d'un "pavé dans la mare" (cf. BEST - RAYNER (1981) : "Are two classes enough for the X^2 goodness of fit test?").

Ainsi les alternatives étant contaminantes et contigües, le niveau à 5 % et n égal à 100, BEST - RAYNER proposent une règle du type suivant : $2 \leq k' \leq 6$ si la queue de l'alternative est plus faible que celle de m, k' voisin de 10 si les queues sont analogues, $15 \leq k' \leq 20$ si la queue de l'alternative est nettement plus lourde. Dans des conditions du même genre la formule de MANN - WALD donne $k' = 23$. La différence est donc sensible.

Les simulations dont nous rendons compte ici confirment la nécessité de choisir k' petit pour avoir un bon test.

IV. Le choix du nombre de polynomes dans le test de Legendre

Le choix de ce nombre est lié au choix de la dimension dans la construction de l'estimateur de la densité par projection. KRONMAL - TARTER (1968) et HALL - DIGGLE (1986) ont proposé des solutions spécifiques mais les simulations qui suivent utilisent une méthode différente.

Une première simulation était destinée à déterminer un ordre de grandeur raisonnable pour le nombre de polynômes de LEGENDRE entrant dans la statistique T_n .

Pour des raisons de commodités le problème a été transposé sur $D = (0, 1)$ en considérant les $F(X_i)$ au lieu des Y_i définis au paragraphe II, ce qui permet de prendre en compte comme alternatives un certain nombre de lois Béta. Dans chaque cas il a été tiré 100 échantillons et la puissance empirique a été calculée en pourcentages de rejets. Voici les résultats obtenus pour un niveau à 5 % et des échantillons de taille 50. A la lecture de ce tableau on constate encore que des valeurs de k' relativement faibles permettent d'obtenir une bonne puissance.

Paramètres de B (r,s)		Nombre de polynômes retenus							
r	s	1	2	3	4	5	6	7	8
0,5	0,5	13	88	87	91	92	93	97	100
0,5	0,8	79	91	92	95	95	95	99	98
0,8	0,8	6	19	22	26	24	25	52	43
0,5	1	98	98	100	98	98	98	100	100
0,8	1	26	23	24	22	21	19	58	81
1	1	4	5	6	8	8	7	30	31
0,9	0,9	9	7	10	10	11	11	38	37
1	0,8	29	29	28	27	28	29	75	73
11	11	3	5	6	4	3	2	29	32

V. Comparaison empirique des tests de Legendre et du X^2

La comparaison des deux tests s'effectue dans les mêmes conditions que précédemment : calcul des puissances empiriques au niveau 5 % et avec des échantillons de taille 50 ou 100.

Compte tenu des résultats préliminaires le nombre de cellules varie de 2 à 10 pour le test du X^2 et le nombre de polynômes de LEGENDRE de 2 à 6. Nous donnons en appendice des courbes de puissance pour différentes hypothèses alternatives H_a . Le nombre de cellules (resp. de polynômes) est en abscisse et les puissances en pourcentage sont en ordonnée. Les courbes de puissance du test de LEGENDRE sont en traits pleins, celles du test du X^2 sont en pointillés.

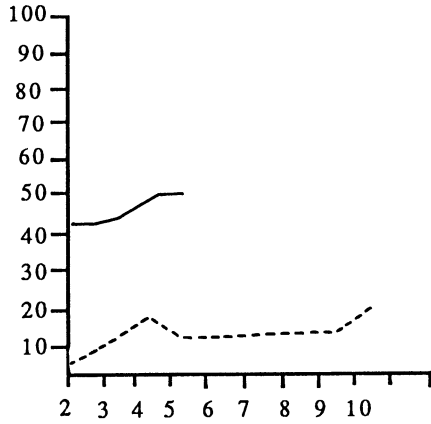
Il apparaît clairement que le test de LEGENDRE est performant et que sa mise en œuvre est simple. Le choix du nombre de polynômes n'est pas un problème majeur car une valeur telle que $k' = 5$ assure une bonne puissance dans des cas très variés.

La comparaison des deux tests montre la supériorité du test de LEGENDRE lorsque la variance est plus grande sous l'alternative que sous H. Dans le cas contraire on peut considérer que les deux tests sont au pire équivalents.

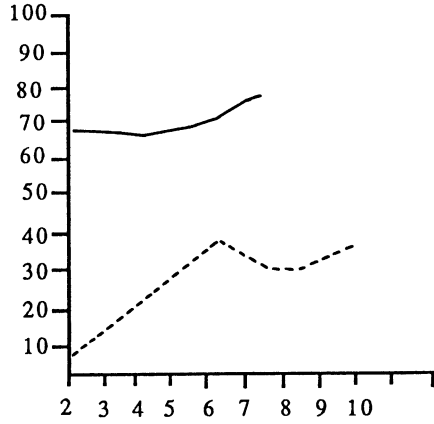
Bibliographie

- D.J. BEST, J.C.W. RAYNER (1981). – Are two classes enough for the X^2 goodness of fit test? *Statistica Neerlandica*, 35, 157-163.
- D.J. BEST, J.C.W. RAYNER (1982). – The choice of class probabilities and number classes for simple X^2 goodness of fit test. *Sankhya*.
- D.J. BEST, J.C.W. RAYNER (1985). – Uniformity testing when alternatives have low order. *Sankhya*.
- R. BLACHER (1985). – A new chi-squared independence test. Rapport de recherche N° 512 IMAG (Grenoble).

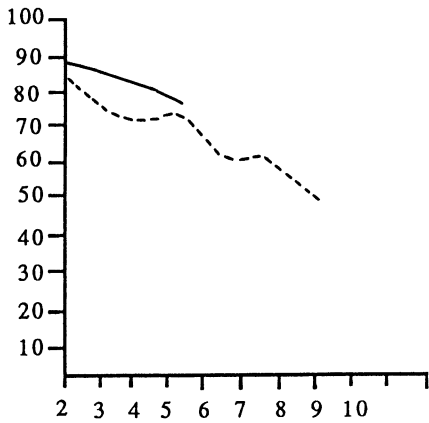
- D. BOSQ (1980). – Sur une classe de tests qui contient le test du X^2 . Publ. ISUP fasc. 1-2, p. 1-16.
- D. BOSQ (1983). – Lois limites et efficacité asymptotique des tests Hilbertiens. *Stat. et An. des données* 8, n° 1, p. 1-40.
- D. BOSQ, J.P. LECOUTRE (1987). – Théorie de l'estimation fonctionnelle 342 p. *Economica* (Paris).
- D. GADIAGA (1982). – Test Hilbertien. Le cas d'un processus stationnaire. Thèse troisième cycle. Université de Lille 1.
- W.C.M. KALLENBERG, J. OOSTERHOOF, B.F. SCHIEVER (1985). – The number of classes in chi-squared goodness of fit test. *JASA*, 80, n° 392, p. 959-968.
- M.G. KENDALL (1973). – *The advanced theory of statistics*. Vol. 2, Charles Griffin (Londres).
- R. KRONMAL, M. TARTER (1968). – The estimation of probability densities and cumulatives by Fourier series methods. *JASA*, 63, p. 925-952.
- P. HALL, J.D. DIGGLE. – The selection of terms in an orthogonal series density estimation. *JASA*, 81, 393, p. 230-234.
- E. IZRAELEWICZ, I. LAFITTE, Z. LAVAULT, B. ROUBERT. – Le test du X^2 et le test de LEGENDRE. Projet ISUP (1988).
- H.B. MANN, A. WALD (1942). – On the choice of the number of class intervals in the application of the chi-square test. *Ann. Math Stat.*, 13, p. 306-317.
- J. NEYMAN (1937). – Smooth test for goodness of fit. *Skand. Aktuar*, 20, p. 149-128.
- J. OOSTERHOOF (1985). – The choice of cells in chi-square tests. *Statistica Neerlandica*, 2, p. 115-128.



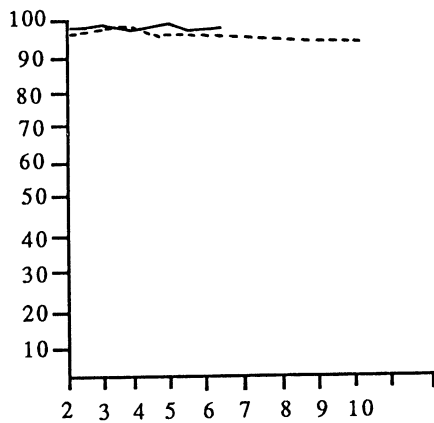
H : N (0,1)
 H_a : N (0,25/16)
 n = 50



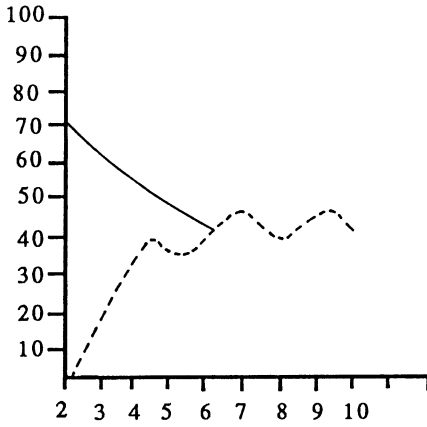
H : N (0,1)
 H_a : N (0,25/16)
 n = 100



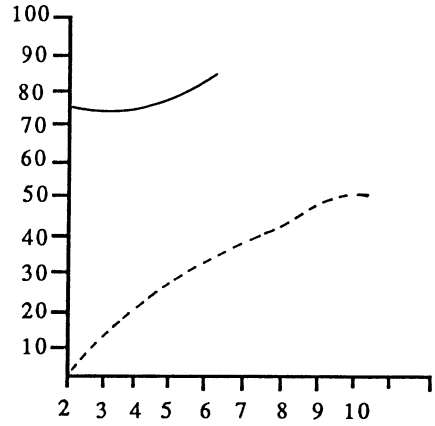
H : N (0,1)
 H_a : N (0,5,1)
 n = 50



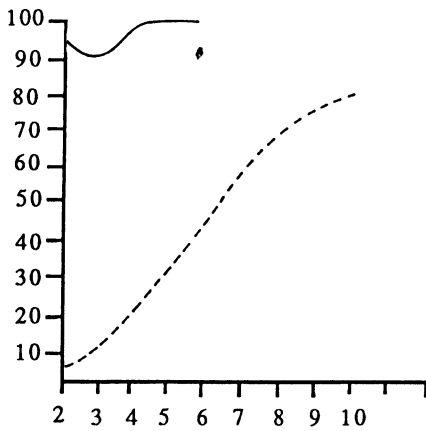
H : N (0,1)
 H_a : N (0,5,1)
 n = 100



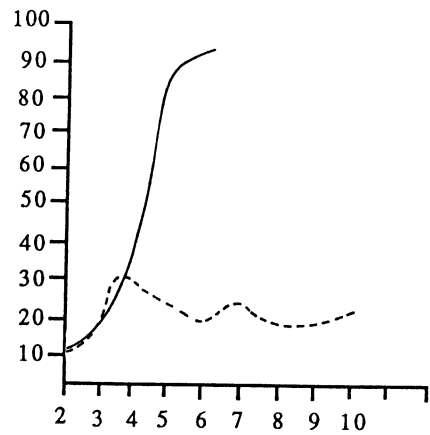
H : N (0,1)
 H_a : N (0,0,5)
 n = 50



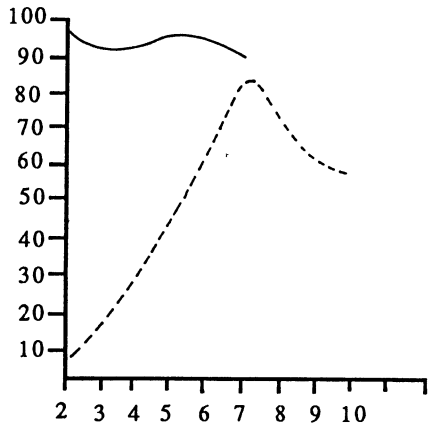
H : N (0,1)
 H_a : N (0,2)
 n = 50



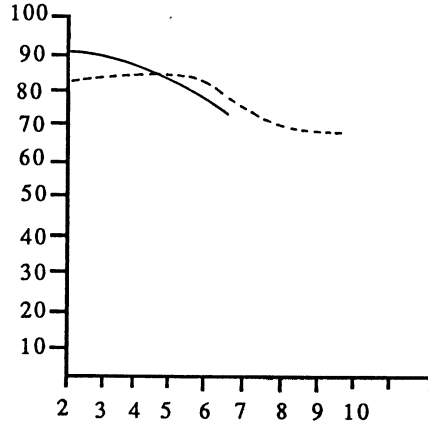
H : N (0,1)
 H_a : STUDENT (1)
 n = 50



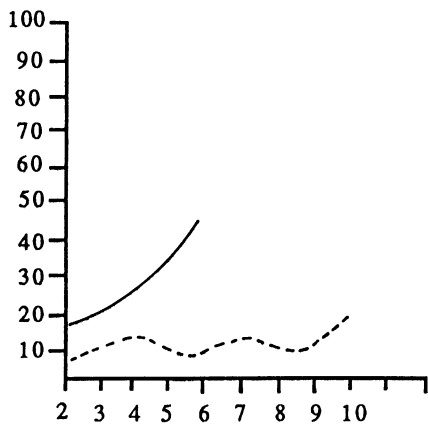
H : N (0,1)
 H_a : CAUCHY (0,0,5)



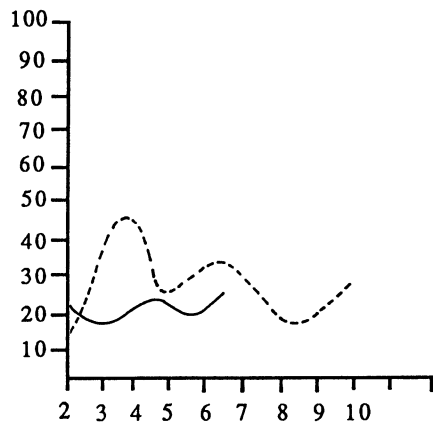
H : CAUCHY (0,1)
 H_a : STUDENT (25)
 n = 50



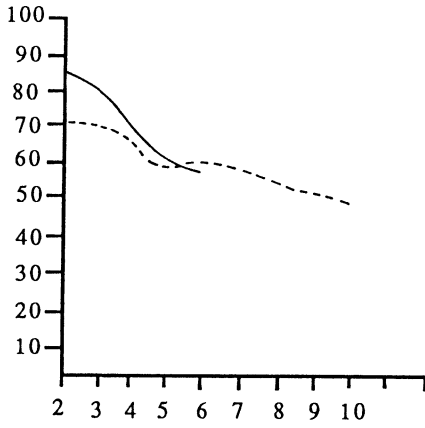
H : U (0,1)
 H_a : BETA (3/2,1)



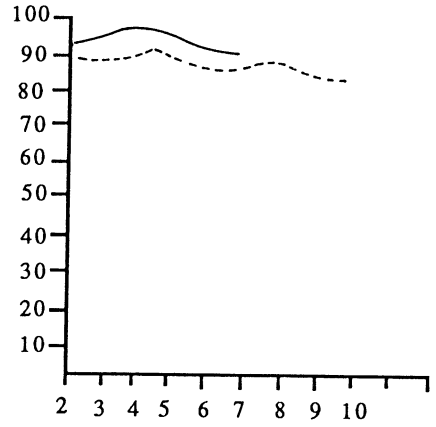
H : N (0,1)
 H_a : 0,9 N (0,1) + 0,1 N (0,25)
 n = 50



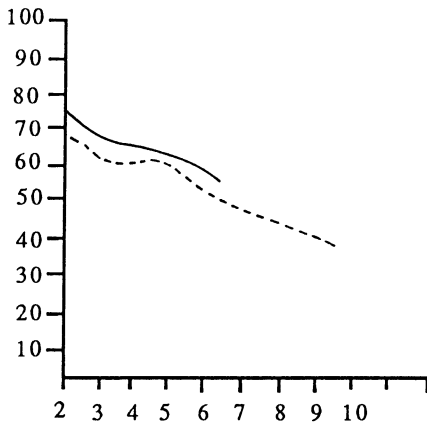
H : N (0,1)
 H_a : 0,8 N (0,1) + 0,2 N (0,0,04)



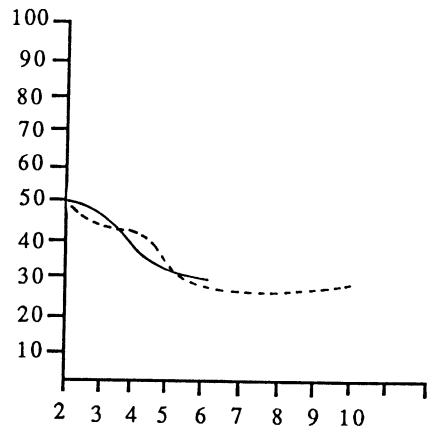
H : U (0,1)
 H_a : BETA (1,1,4)
 n = 100



H : exp (1)
 H_a : exp (0,5)
 n = 50



H : exp (0,5)
 H_a : Γ (0,8,0,6)
 n = 50



H : exp (1)
 H_a : Γ (0,9,1,2)
 n = 50