Comparaison de « moyennes » de 2 échantillons aléatoires indépendants

Quel test utiliser ?

Antoine Poncet – Service d' Épidémiologie Clinique

Introduction

- Une question récurrente dans la comparaison de deux « moyennes » est la condition de normalité des variables analysées pour certains tests (test t)
- Plusieurs avis se confrontent:
  - 1) On ne vérifie pas la normalité et on utilise un test paramétrique
    - Quand n est « suffisamment grand », le Théorème Central Limite postule que l'estimateur de la moyenne suit une loi normale et on utilise le test t
    - Limite: qu'est-ce qu'un grand échantillon (n=30?, n=50?)
    - Certains auteurs considèrent que les tests paramétriques sont robustes jusqu'à un certain point à la violation de la normalité
    - Limite: les violations peuvent être de différente nature (aplatissement, asymétrie)
  - 2) On ne vérifie pas la normalité et on utilise un test non paramétrique
    - Les tests non paramétriques s'appliquent quelque soit la forme de la distribution
    - Limite: les tests non paramétriques sont réputés moins puissants que les tests paramétriques si les conditions d'application sont respectées (normalité)
  - 3) On vérifie systématiquement la normalité
    - Utilisation d'un test paramétrique (test t) en cas de non rejet
    - Utilisation de tests non paramétriques ou robustes en cas d'échec
    - Limite: quel test de normalité utiliser, performance de ces tests réputée médiocre

→ L'objectif de ce travail est d'explorer les performances de ces approches

Plan

I. Objectifs

II. Présentation

- des tests
- de la méthode

III. Résultats

IV. Conclusion / Discussion

Objectifs (1/2)

- Explorer les erreurs de type I et II
  - de 5 différents tests :
    - Tests paramétriques : test t et test z
    - Tests non paramétriques : test de MW et test exact
    - Test robuste
  - sous diverses distributions:
    - Normale: distribution « idéale »
    - Uniforme: distribution symétrique
    - Seminormale: distribution asymétrique
    - Bimodale: le second mode correspond à des « outliers »
    - Lognormale: distribution asymétrique très skewed
  - Sous diverses tailles d'échantillon
  - Sous différentes tailles d'effet

Objectifs (2/2)

- Comparer 2 statistiques de test: celle du t et du MW
- Explorer la puissance de différentes stratégies
  - Stratégie 1 : t test d'emblée
  - Stratégie 2 : MW d'emblée
  - Stratégie 3 :
    - Test de Normalité de Kolmogorov-Smirnov (sur chacun des 2 groupes)
    - Si rejet de l'hypothèse de normalité dans au moins l'un des deux groupes alors utilisation de MW
    - Sinon utilisation du t test

Présentation des tests

- 1. test t
- 2. test z
- 3. test de MW
- 4. test exact
- 5. test robuste

Présentation des tests (1/5)
test t pour échantillons indépendants

- Condition d'application (Robert F. WOOLSON):
  - Les observations de chaque échantillon sont issues d'une distribution normale
  - Les variances des 2 échantillons sont égales
- H0: les moyennes des 2 populations normales sont égales
- Statistique de test :

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S^2_{X_1} + S^2_{X_2}}{n}}} \qquad t \sim T(2n-2)$$

Présentation des tests (2/5)
test Z

- Condition d'application
  - Les observations de chaque échantillon sont issues d'une distribution normale
  - Les écarts types des populations source sont supposées connues
- H0: les moyennes des 2 populations normales sont égales
- Statistique de test

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_{X_1}^2 + S_{X_2}^2}{n}}} \qquad z \sim \mathcal{N}(0,1)$$

Présentation des tests (3/5)
test de Mann Whitney

- Pas d'hypothèse sur la forme des distributions
- Test qui se base sur le rang des observations
- H0: La probabilité qu'une observation aléatoire du 1ère groupe soit plus grande qu'une observation aléatoire du 2nd groupe est 0.5 (AUC-ROC=0.5)
- Attention, H0 n'est pas l'égalité des moyennes, ni l'égalité des médianes, mais plutôt « la médiane des différences est égale à 0 »

$$K = \frac{12}{N(N+1)} \sum_{j=1}^{k} \left\{ \frac{R_j^2}{n_j} - 3(N+1) \right\}$$

$$K \sim \chi^2(k-1)$$

Avec:

- R_j: somme totale des rangs
- n_j: taille d'échantillon
- k: nombre de groupes (dans notre situation, k=2)
- N: taille de l'échantillon total (dans notre situation, N=2n)
- Possibilité d'ajuster la statistique de test en cas d'ex aequo

1947
On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other
Author(s): H. B. Mann and D. R. Whitney

Présentation des tests (4/5)
test exact

- Test basé sur les permutations
  - Aucune hypothèse sur les distributions si $n<50$
  - Approximation normale usuelle (test Z) si $n \geq 50$

- H0: les distributions sont identiques dans les 2 groupes

Présentation des tests (5/5)
test robuste

- Objectif : proposer une alternative qui soit moins impactée par les outliers ou autres violations des hypothèses du modèle (normalité)
- H0: les moyennes des 2 populations sont égales
- Principe: plus un individu est loin de la moyenne de son groupe, plus son poids sera faible (il existe plusieurs fonctions permettant de définir les poids)
- On utilise dans R la fonction lmrob (paramètres par défaut)
- J.W. Tukey (1979):

  "… just which robust/resistant methods you use is not important – what is important is that you use some. It is perfectly proper to use both classical and robust/resistant methods routinely, and only worry when they differ enough to matter. But when they differ, you should think hard."

Méthode

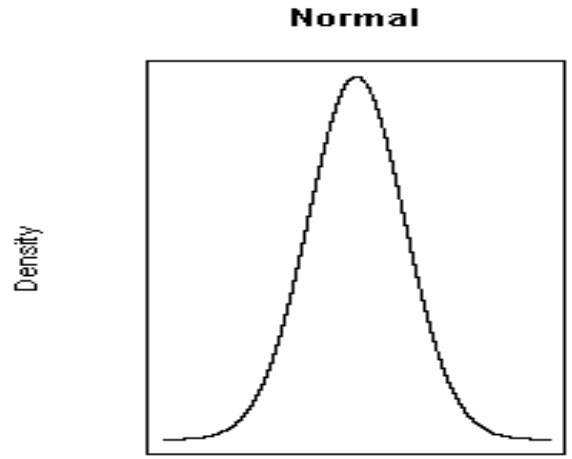
Méthode (1/2)

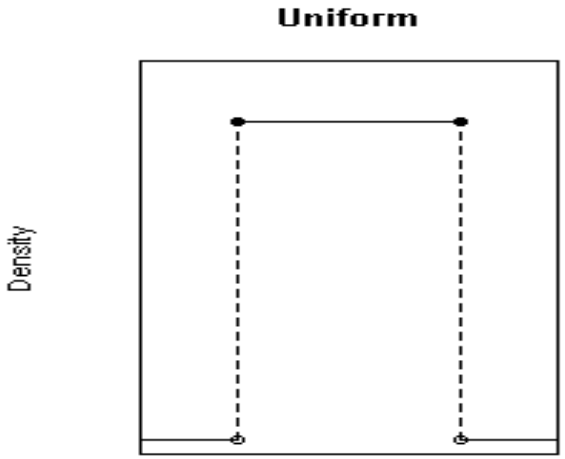
- On utilise des simulations pour répondre aux objectifs. On réplique 1000 échantillons pour chaque condition d'expérience. Les 5 tests sont alors appliqués sur chaque échantillon
- Description des simulations:
  - Génération de la population source selon une certaine distribution (N=10.000.000)
  - Par simplicité, on centre réduit (m=0, sd=1) les distributions générées
  - Sélection aléatoire de l'échantillon (bras 1) d'une certaine taille
  - Sélection aléatoire de l'échantillon (bras 2) de même taille auquel on rajoute une certaine différence

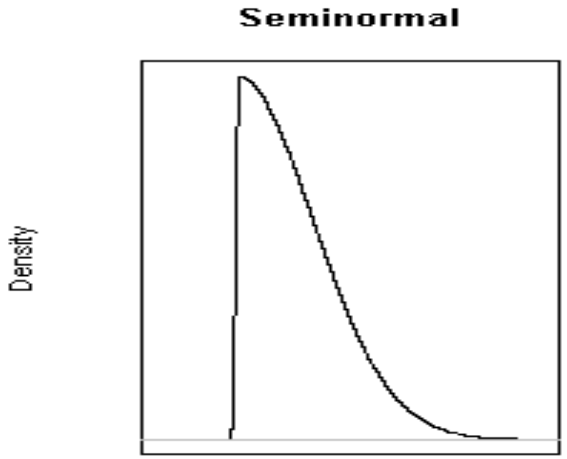
Méthode (2/2)

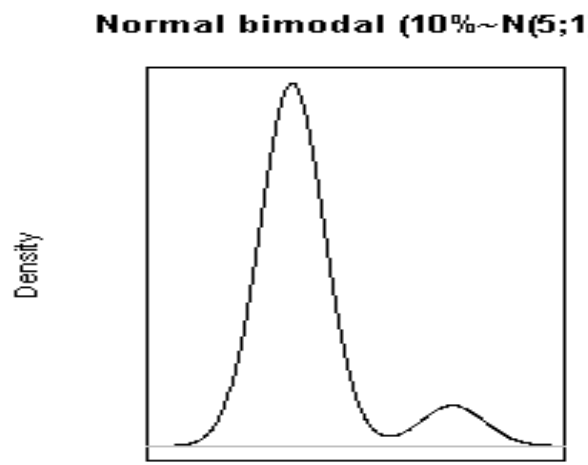
- Condition de simulations
  - Distribution : Normale, Uniforme, Seminormale, Bimodale et Lognormale
  - Taille de l'effet en Ecart Type (ET) : 0, 0.2, 0.5, 0.8 (nulle, faible, modérée, forte selon Cohen)
  - Taille d'échantillon par groupe : 10, 20, 30, 50, 100, 200, 500, 1000

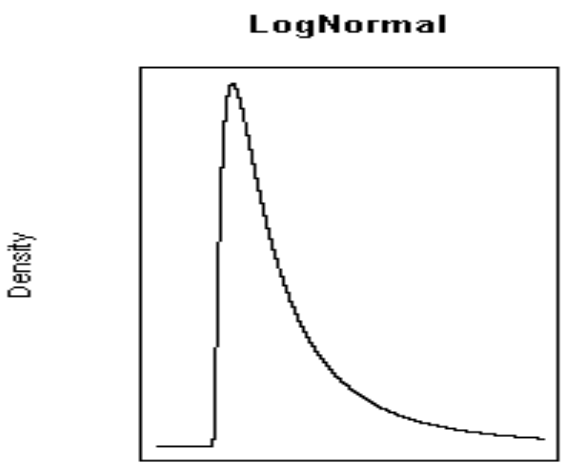
Formes des distributions étudiées





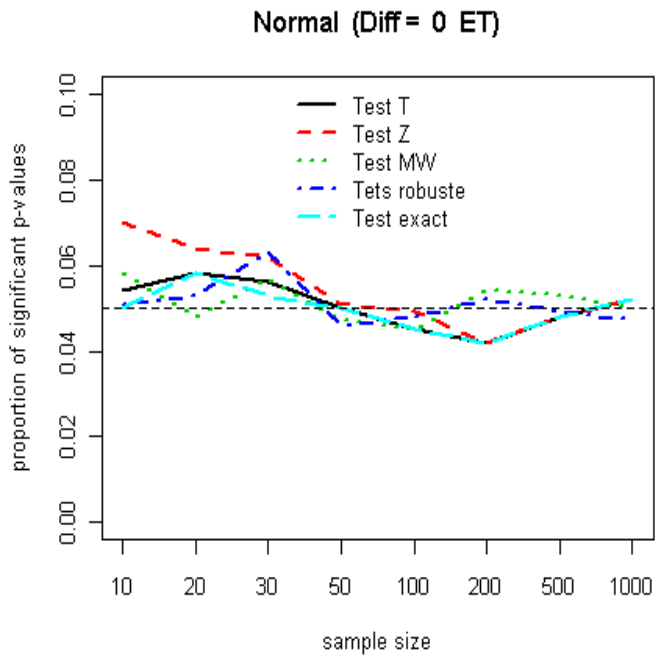


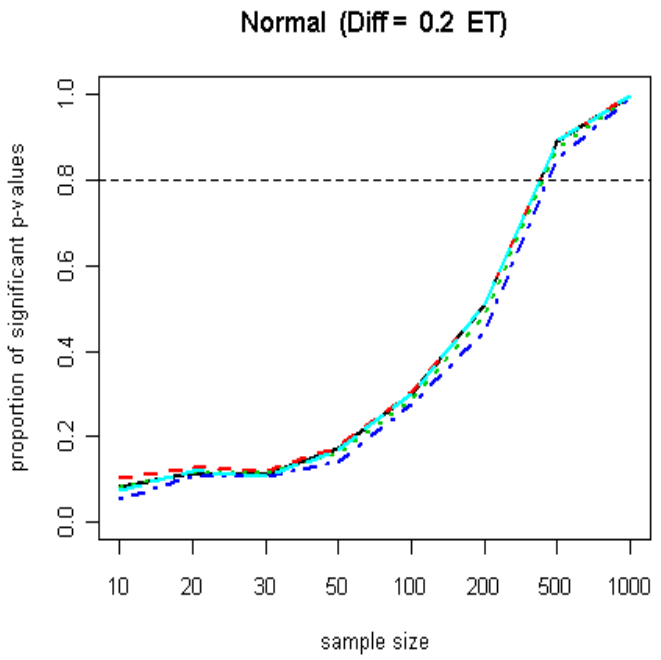


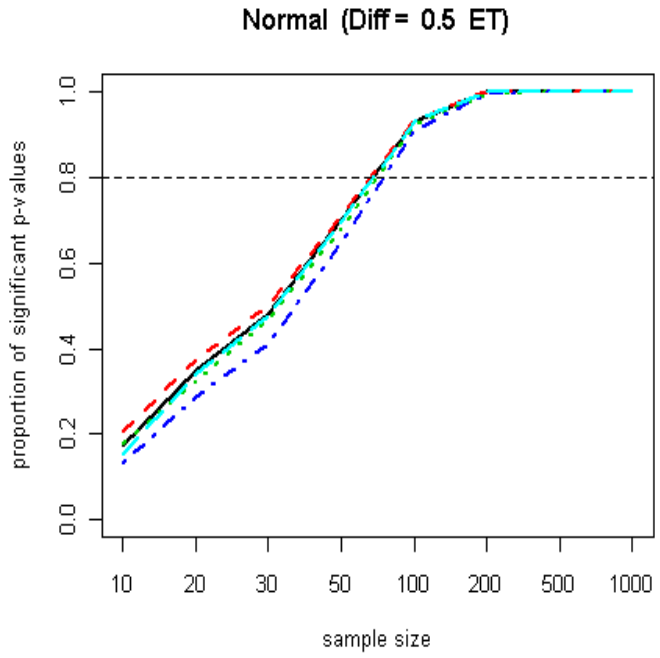


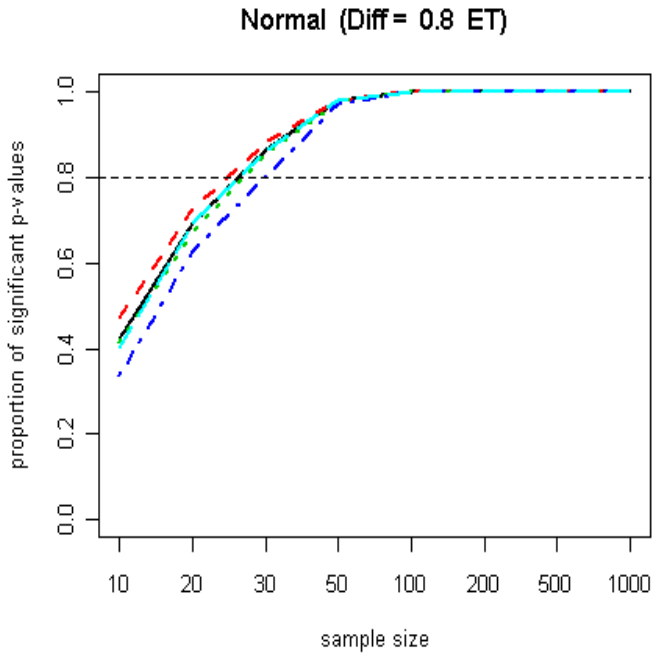
Résultats

- Erreurs de type I et II des 5 tests, distribution par distribution



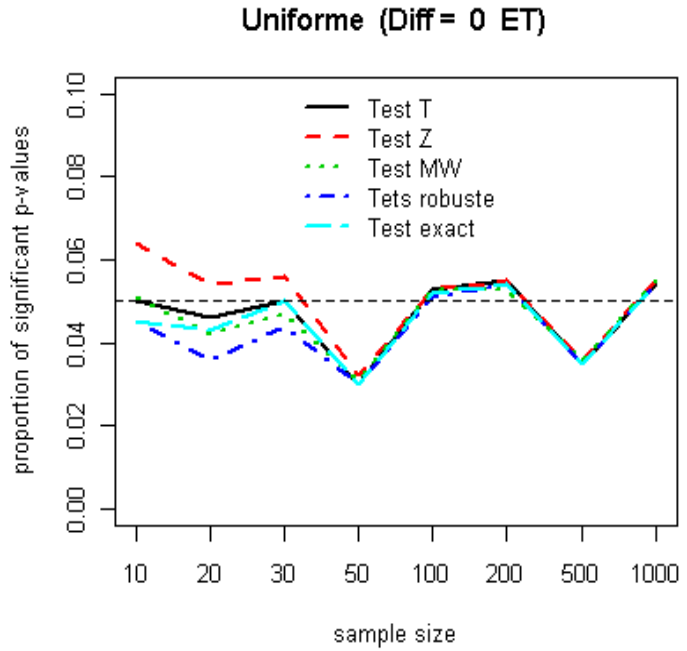


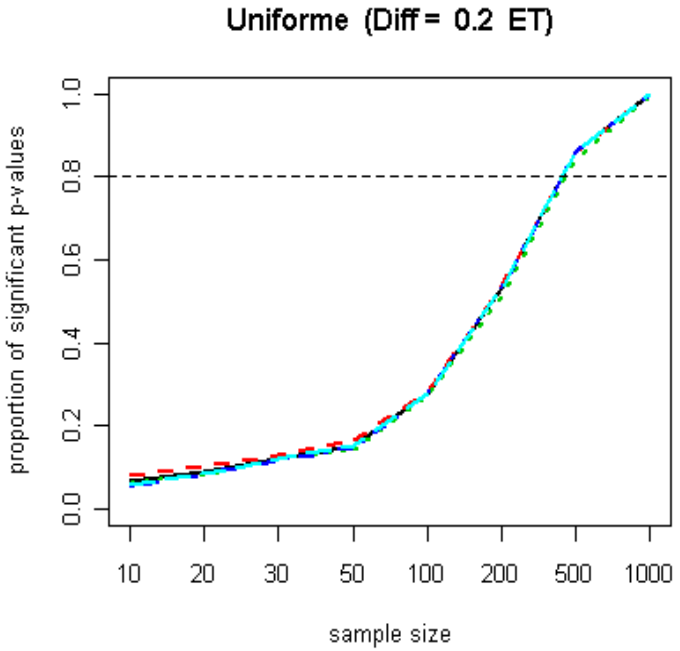


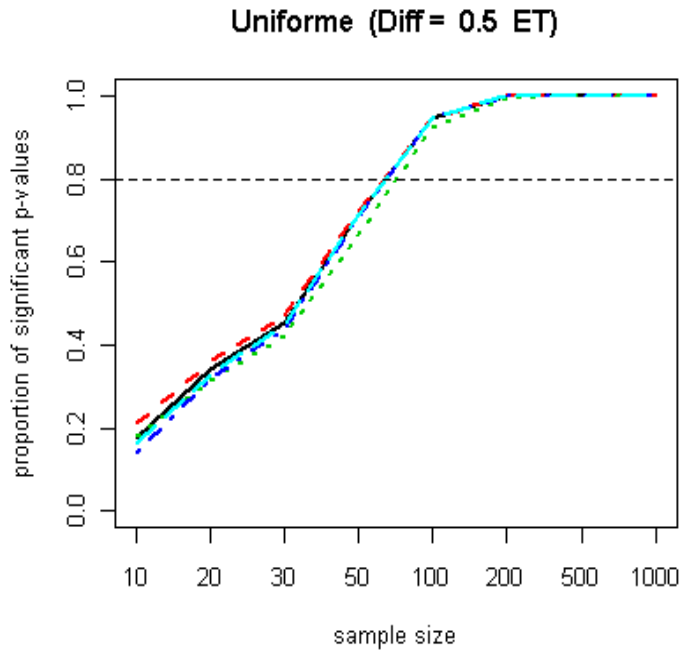


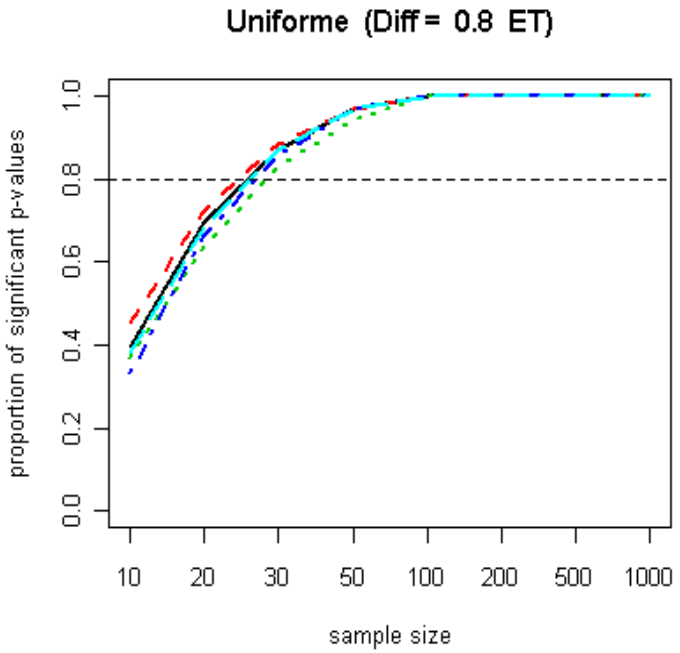
Distribution normale

- Erreur de type 1 sensiblement équivalente
- Différence de puissance faible entre les tests
- Conclusion: les tests sont quasi équivalents pour des distributions normales







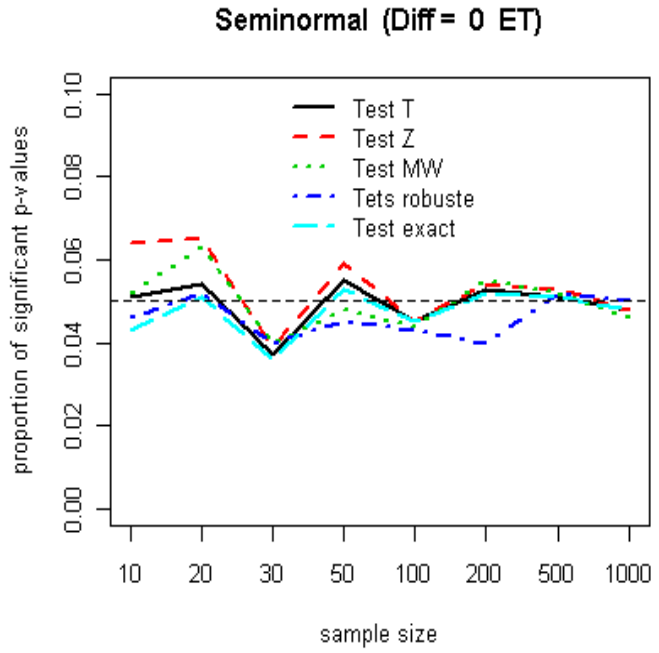


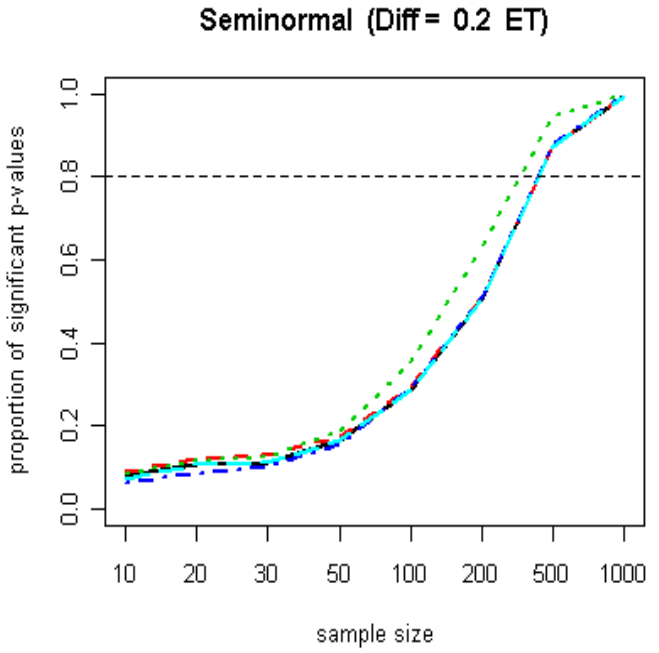
Distribution uniforme

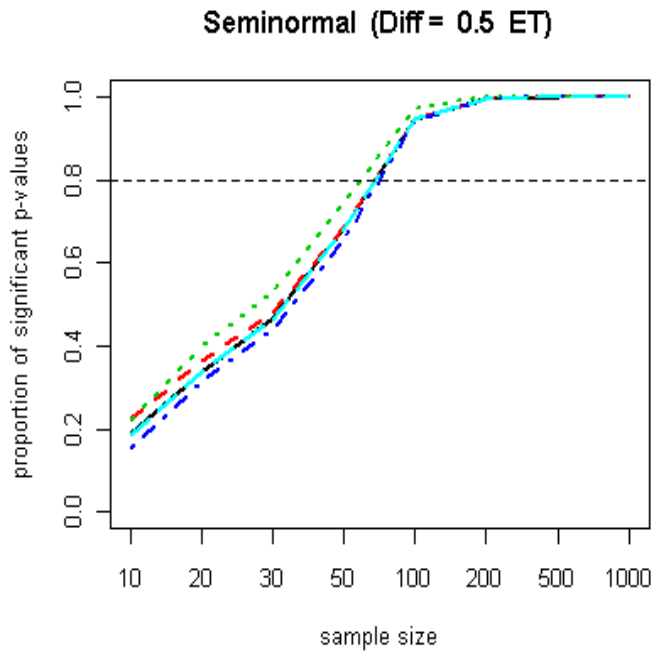
Alpha ≈

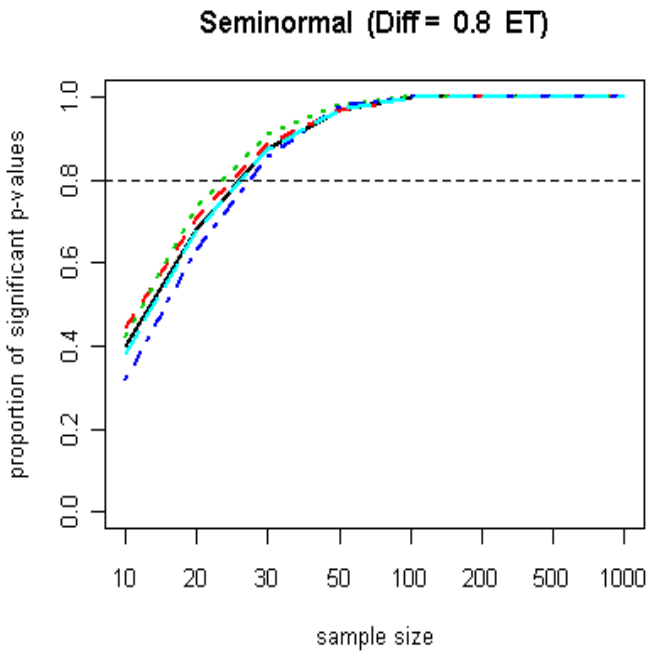
Puissance ≈

les tests sont quasi équivalents pour des distributions uniformes





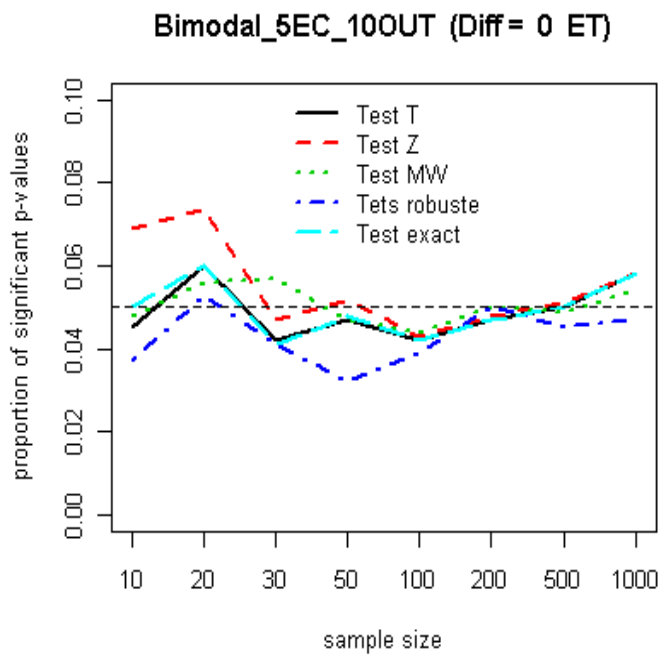


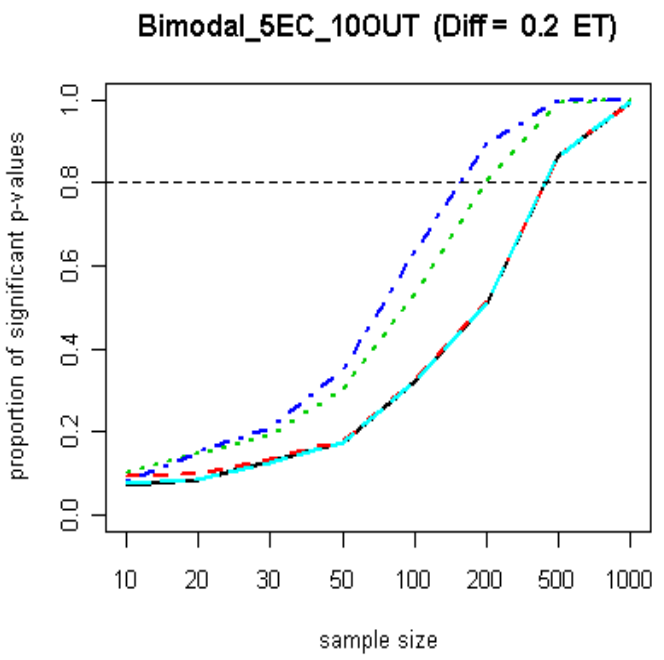


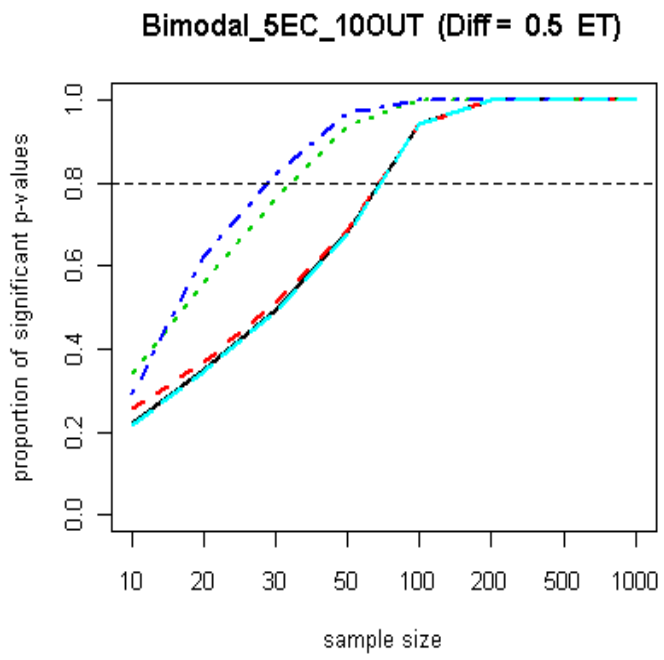
Distribution seminormale

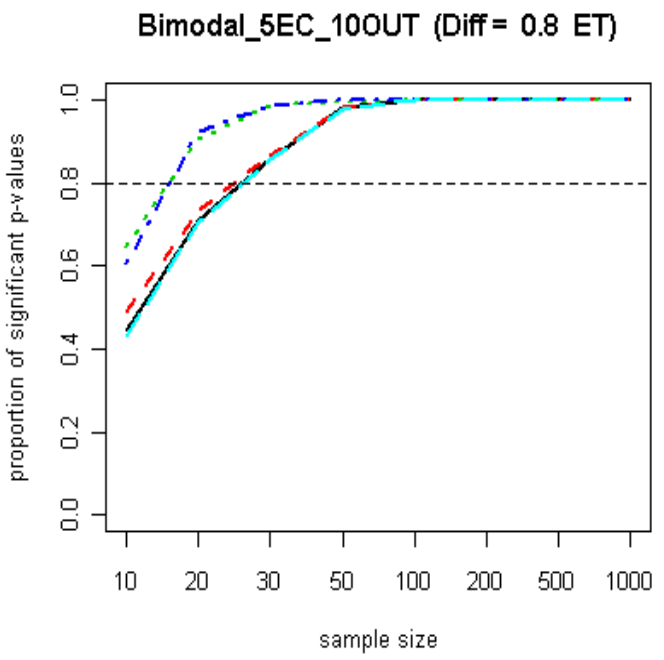
Erreur de type 1 sensiblement équivalente

Puissance sensiblement équivalente sauf pour le test de Mann-Whitney qui est légèrement plus puissant





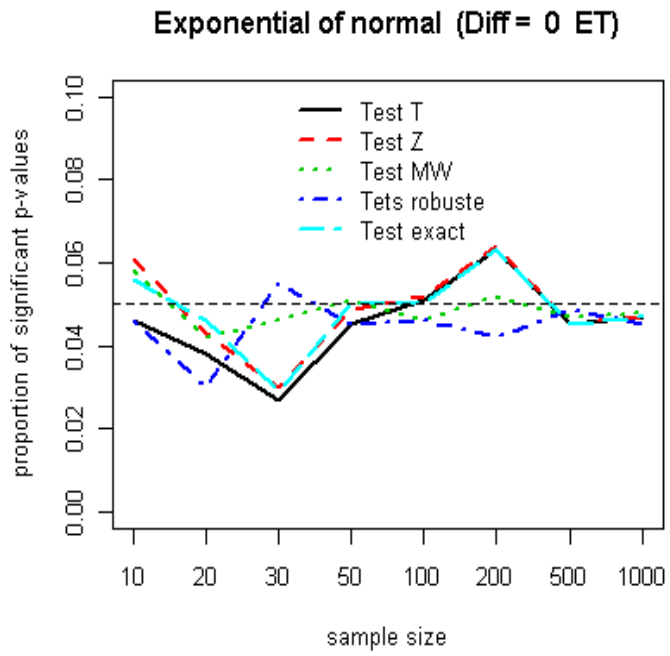


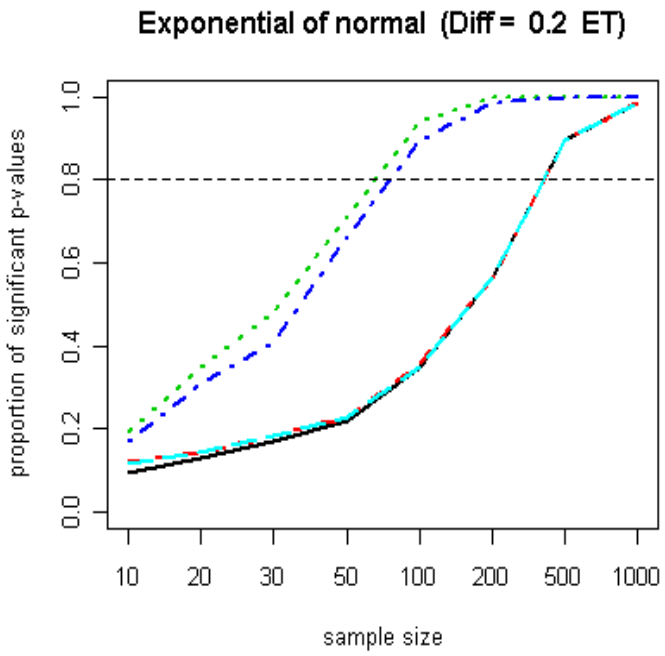


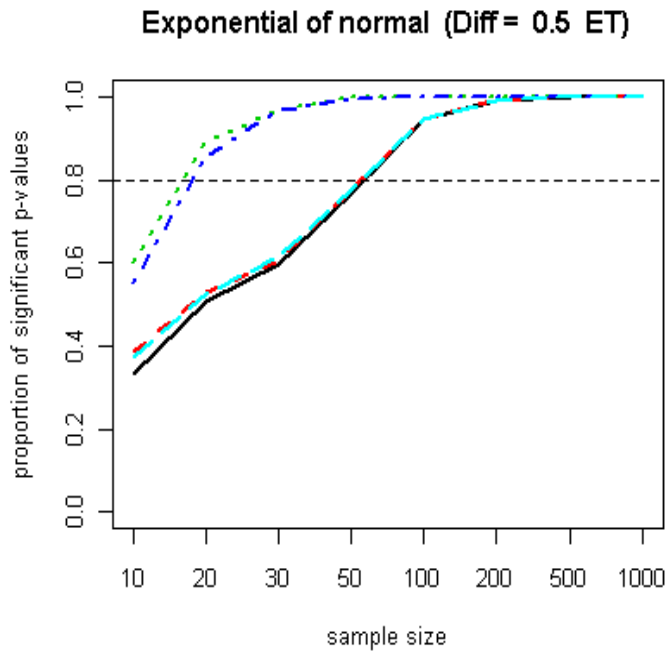
Distribution normale bimodale

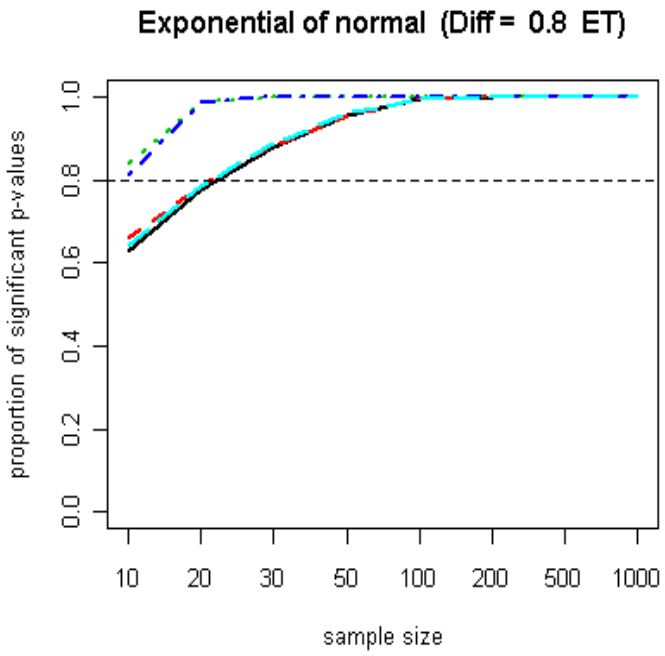
Erreur de type 1 sensiblement équivalente

Test de Mann-Whitney et test robuste sont plus puissants









Distribution lognormale

- Erreur de type 1 sensiblement équivalente
- Puissance plus forte pour le test robuste et de Mann-Whitney
- A l'inverse, le test Z est un peu + conservateur

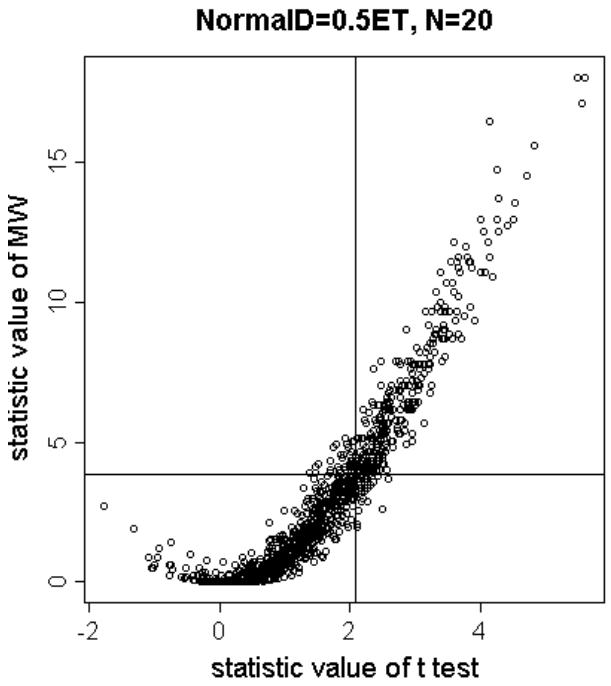
résumé

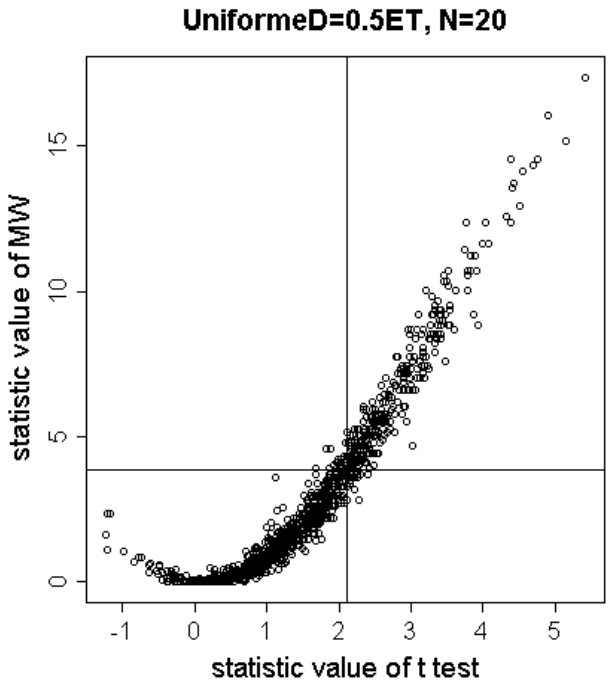
- Sous des distributions normales: (très) faible perte de puissance du test robuste et de MW par rapport aux tests t, Z et exact
- Toute violation de la normalité n'a pas les mêmes impacts sur les erreurs de type I et de type II :
  - Distributions uniformes : résultats identiques que sous distributions normales
  - Distributions seminormale : résultats proches mais avec un faible avantage en faveur du MW
  - Distributions bimodales (outliers) : MW et robuste net gain de puissance avec avantage au robuste entre les 2
  - Distributions lognormales : net gain de puissance pour MW et robuste avec avantage au MW
- L'asymétrie et les outliers jouent sur la puissance des tests et le test de MW se comporte bien dans tous les cas de figure

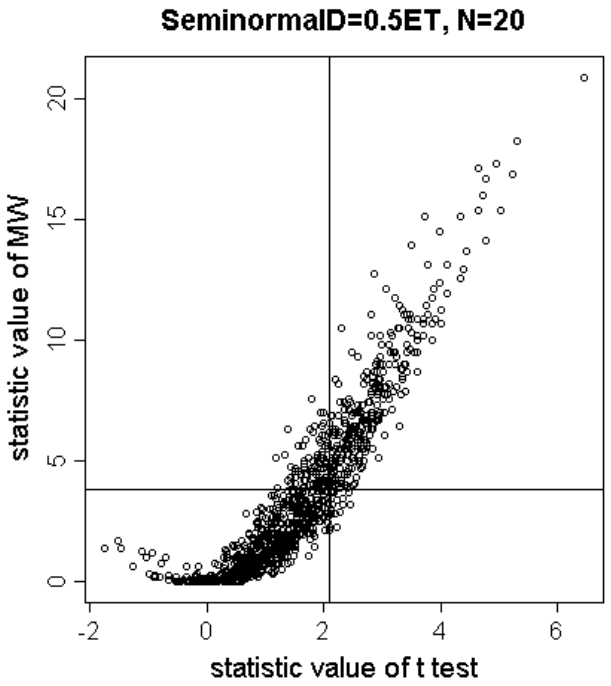
Comparaison des statistiques de test du t test et du test de MW

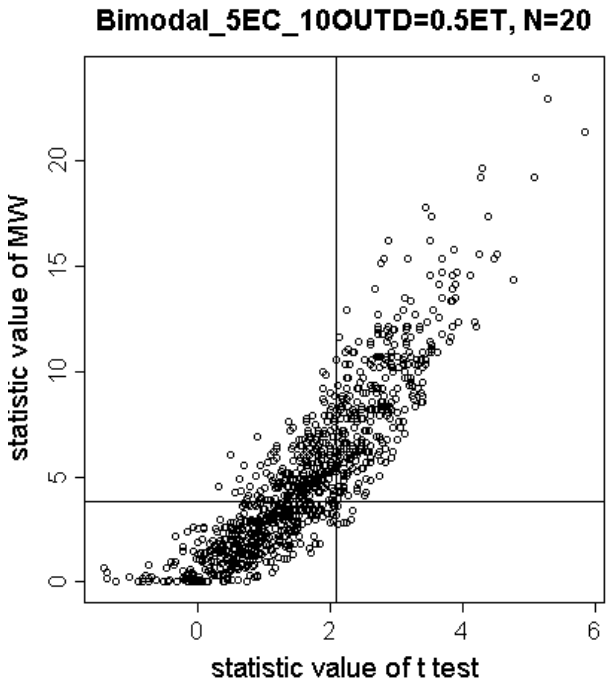
- On cherche à explorer plus en détail les différences observées entre le test t et MW. On se place sous le cas n=20 et diff=0.5 ET

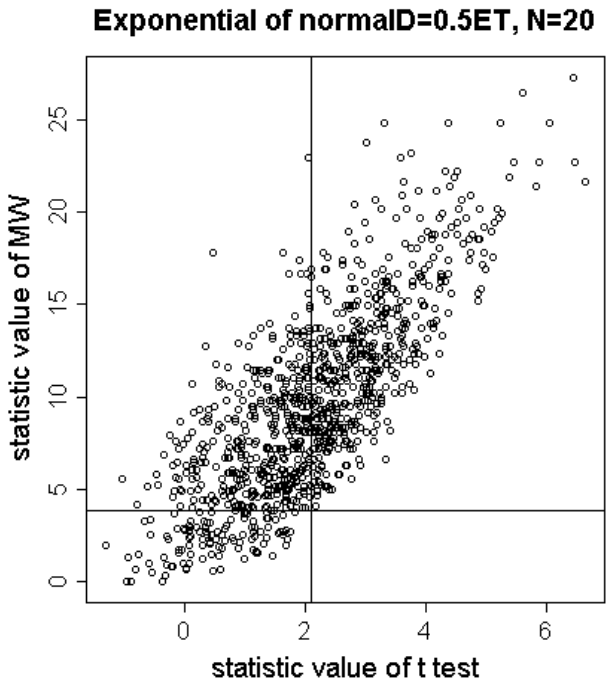
- Pour chaque simulation, on représente conjointement la statistique de test du test t et du test de MW











- Ligne horizontale

  h≈3.841 (valeur critique MW)

- Ligne verticale

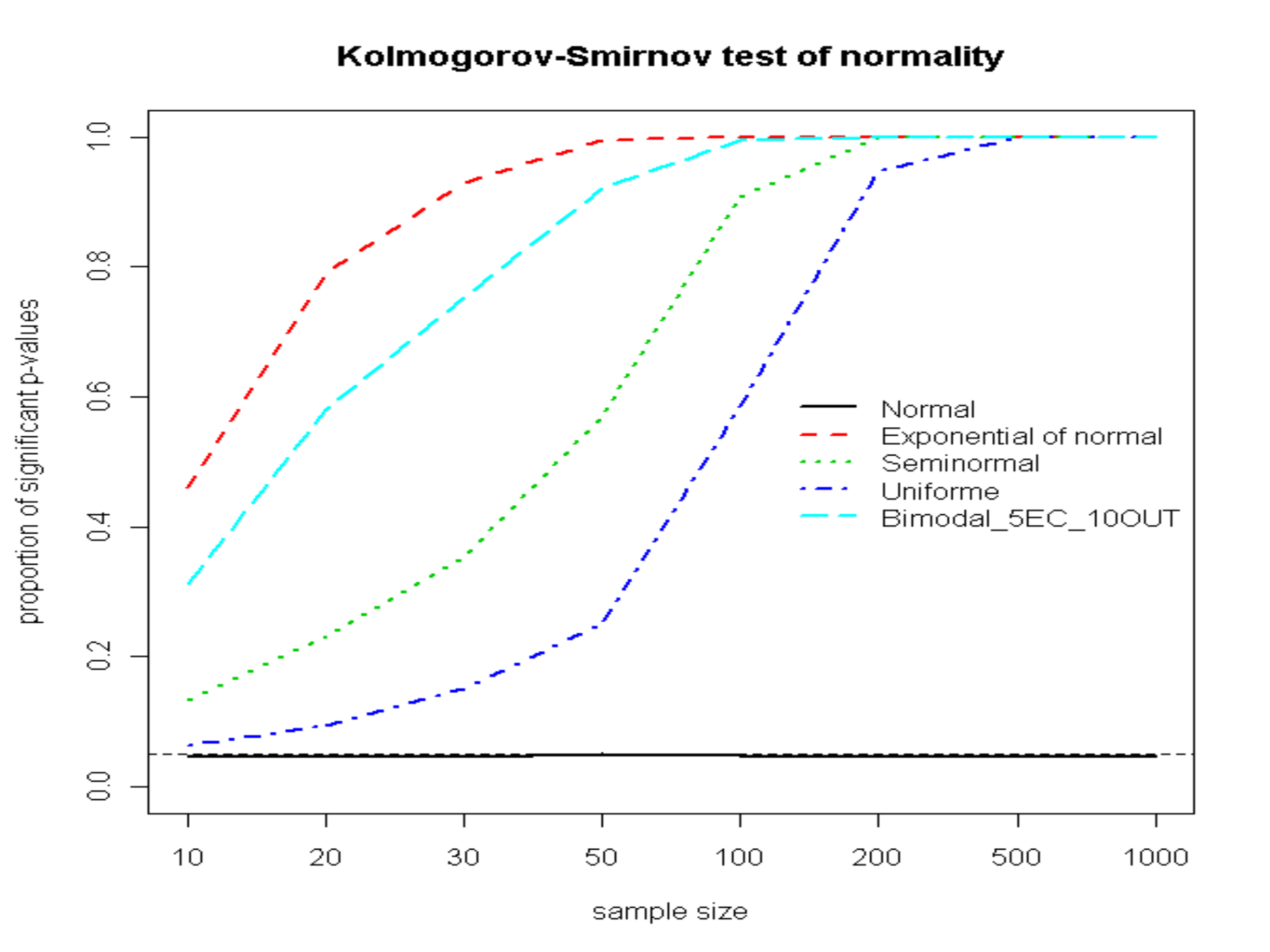
  v ≈ 2.1 (valeur critique t)

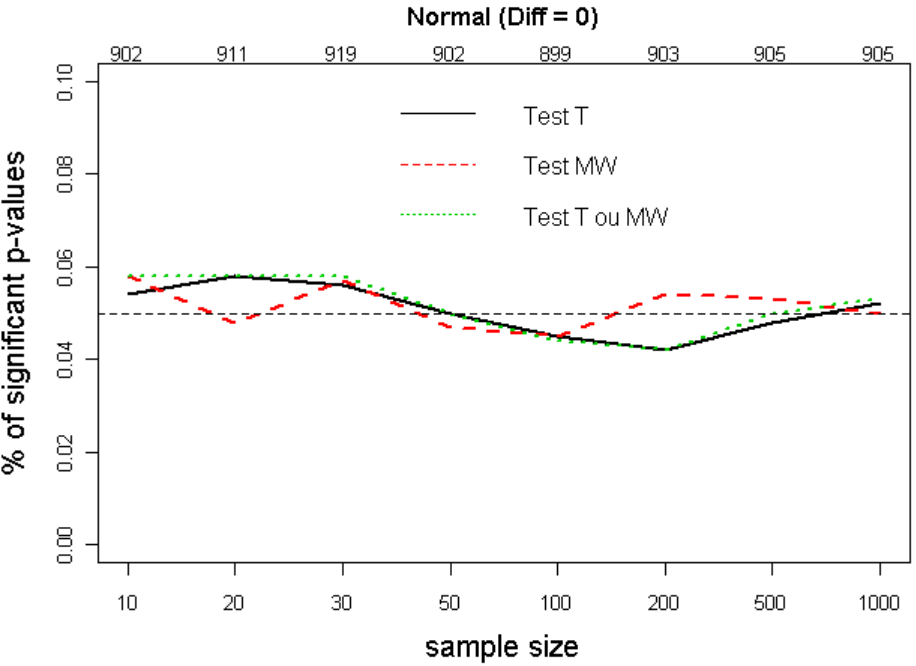
résumé

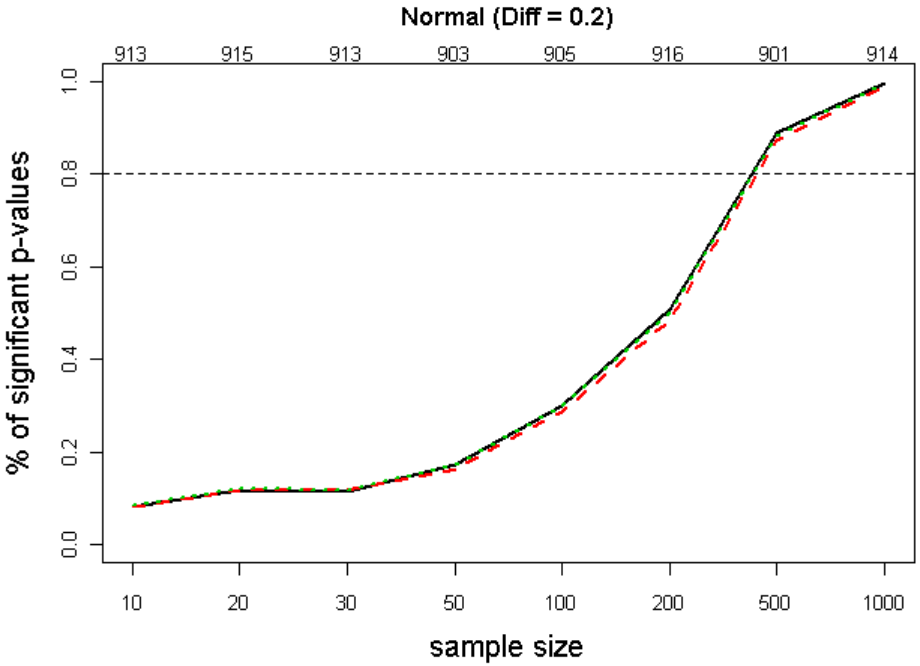
- Sous des conditions de normalité, la concordance entre les deux statistiques est bonne et il y a autant de décisions discordantes « T+ & MW- » que « T- & MW+ »
- Plus on s'éloigne des conditions de normalité en terme d'asymétrie et d'outliers plus on trouve un gain systématique en faveur du MW:
  - Le nombre de cas MW+ et T- augmente
  - Le nombre de cas MW- et T+ diminue

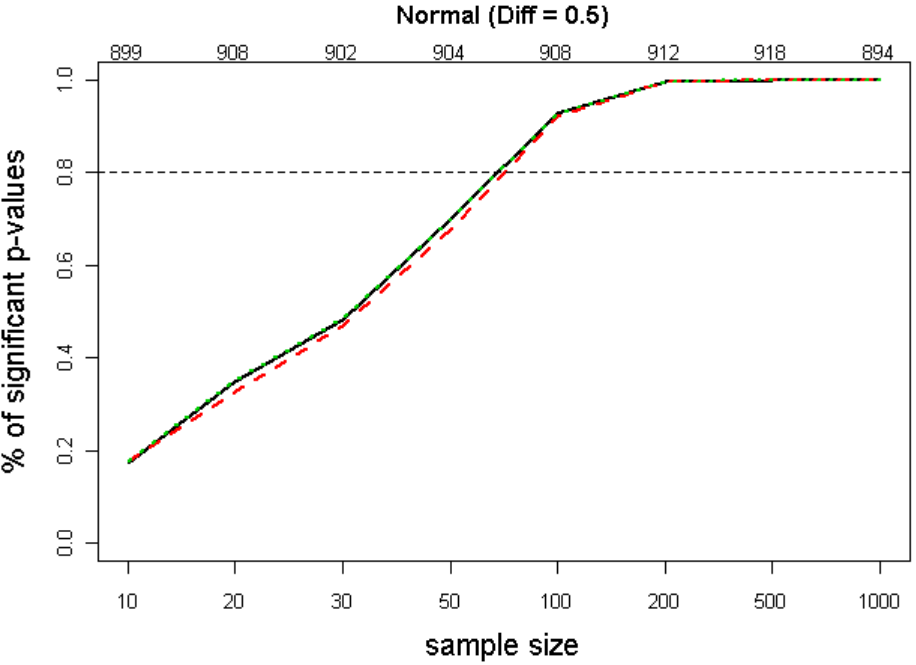
Analyse de la puissance de 3 stratégies

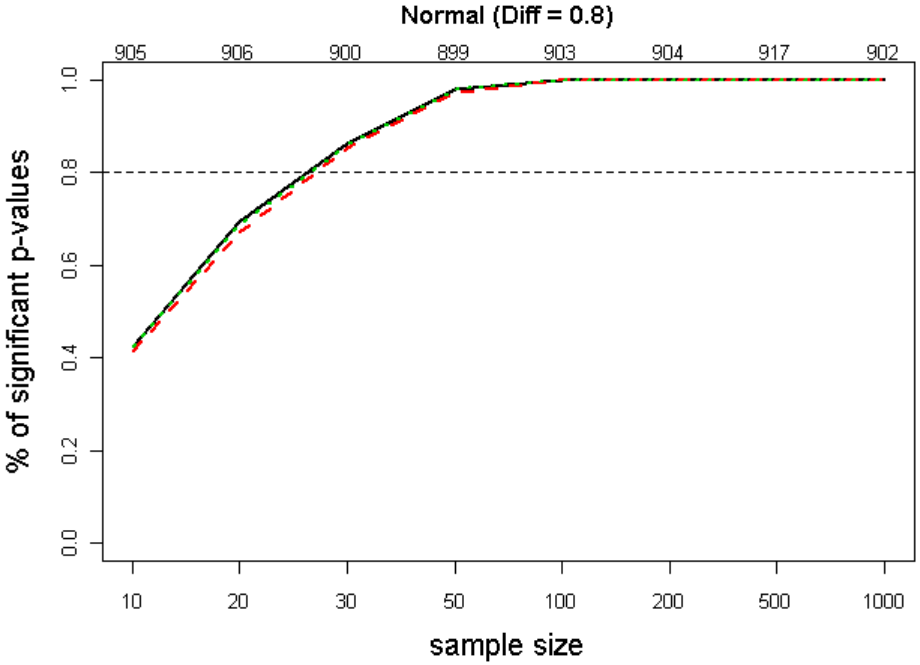
- Stratégie 1: on applique toujours le test t
- Stratégie 2: on applique toujours le test de MW
- Stratégie 3: on applique le test t si le test de normalité (Kolmogorov Smirnov) ne rejette pas H0, le test de MW sinon

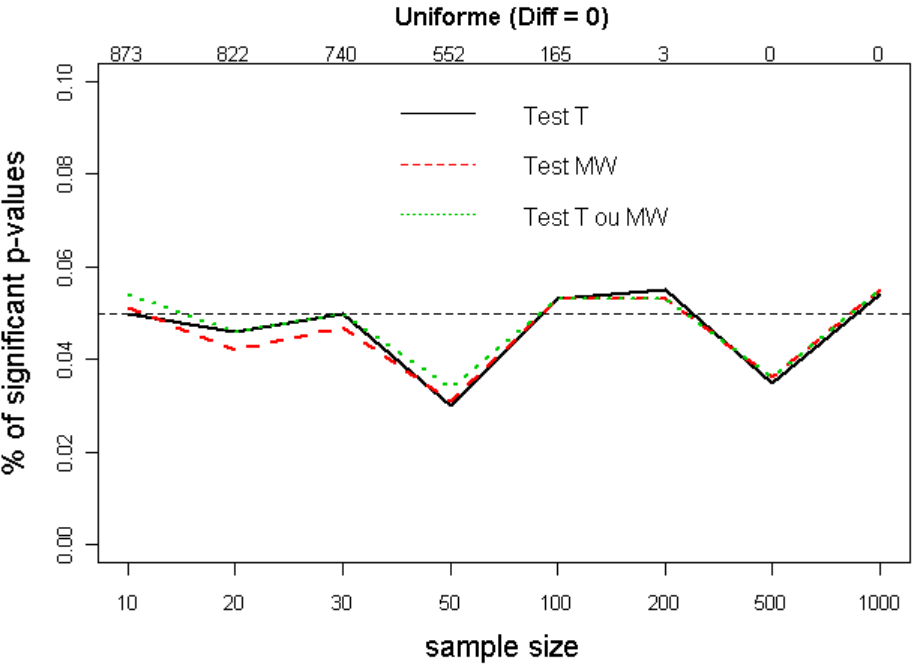

Kolmogorov-Smirnov test of normality
proportion of significant p-values
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
50
100
200
500
1000
sample size
Normal
Exponential of normal
Seminormal
Uniforme
Bimodal_5EC_10OUT

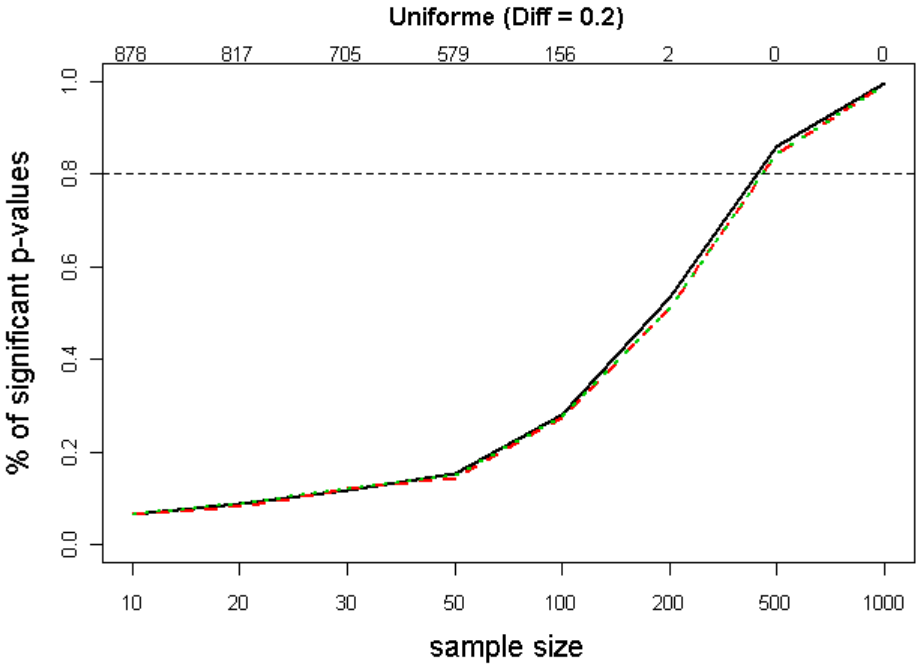

Normal (Diff = 0)
902 911 919 902 899 903 905 905
Test T
Test MW
Test T ou MW
% of significant p-values
0.00 0.02 0.04 0.06 0.08 0.10
10 20 30 50 100 200 500 1000
sample size

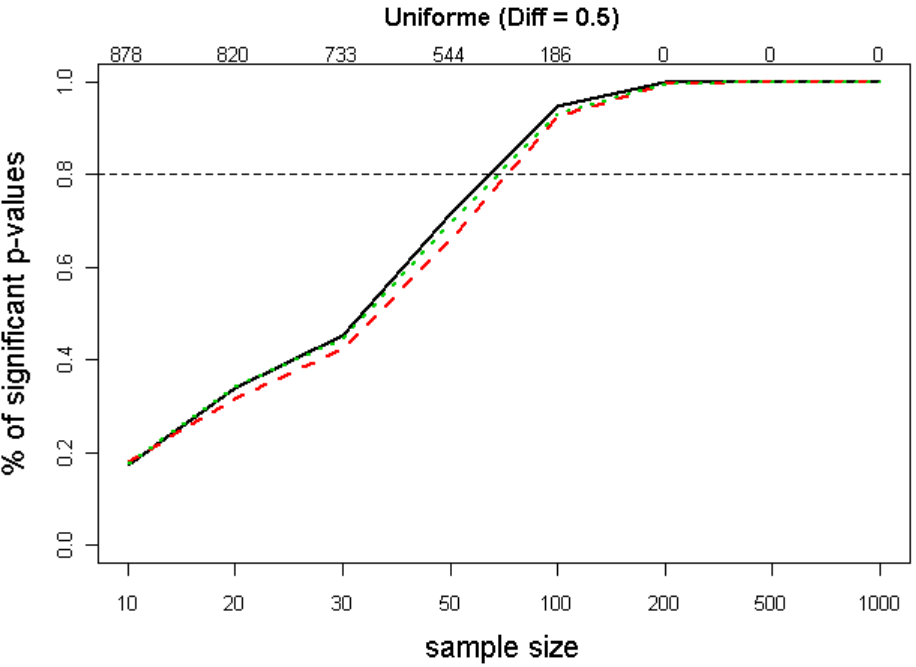

Normal (Diff = 0.2)
913 915 913 903 905 916 901 914
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

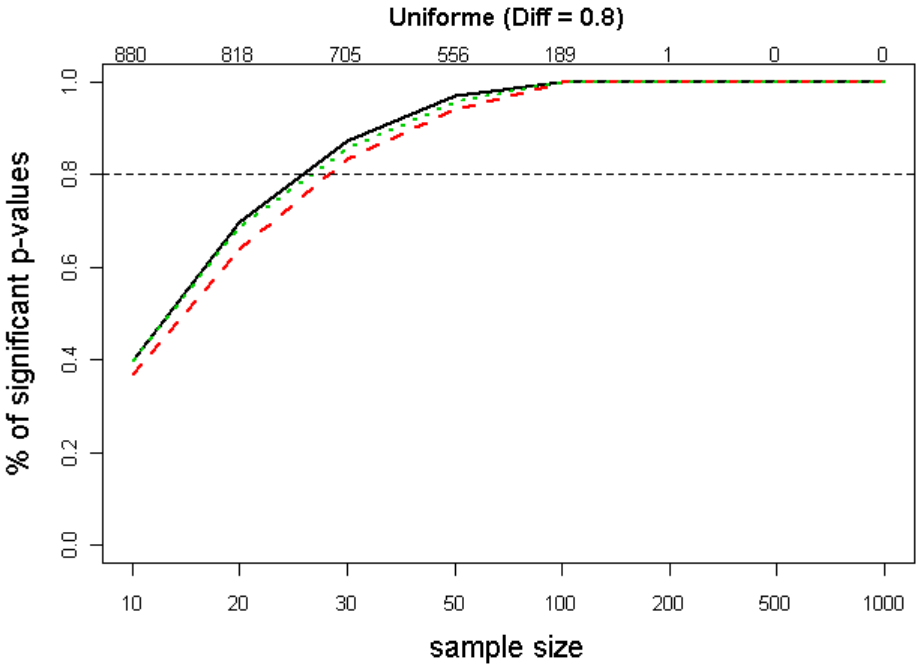

Normal (Diff = 0.5)
899 908 902 904 908 912 918 894
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

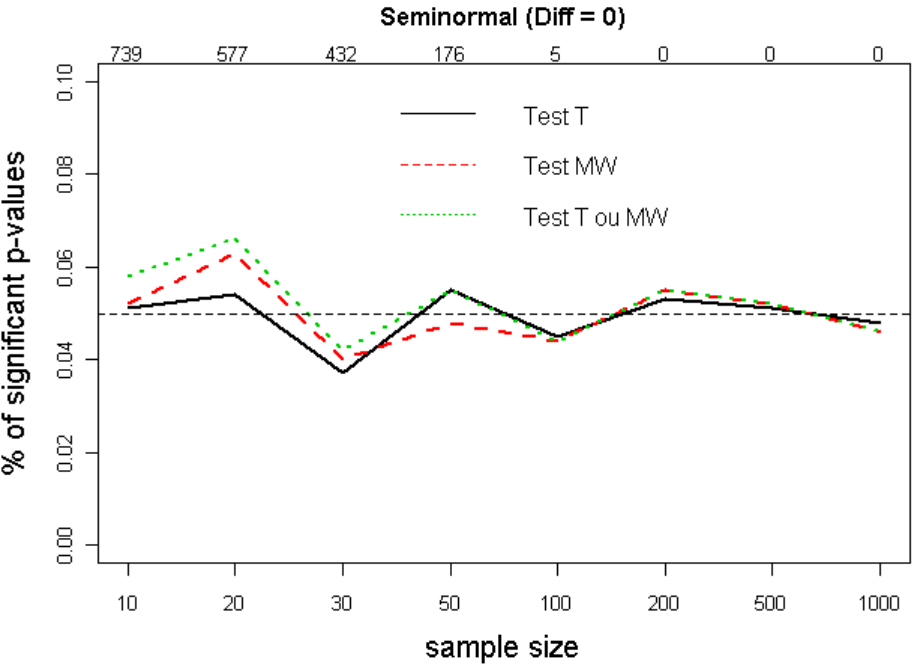

Normal (Diff = 0.8)
905 906 900 899 903 904 917 902
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

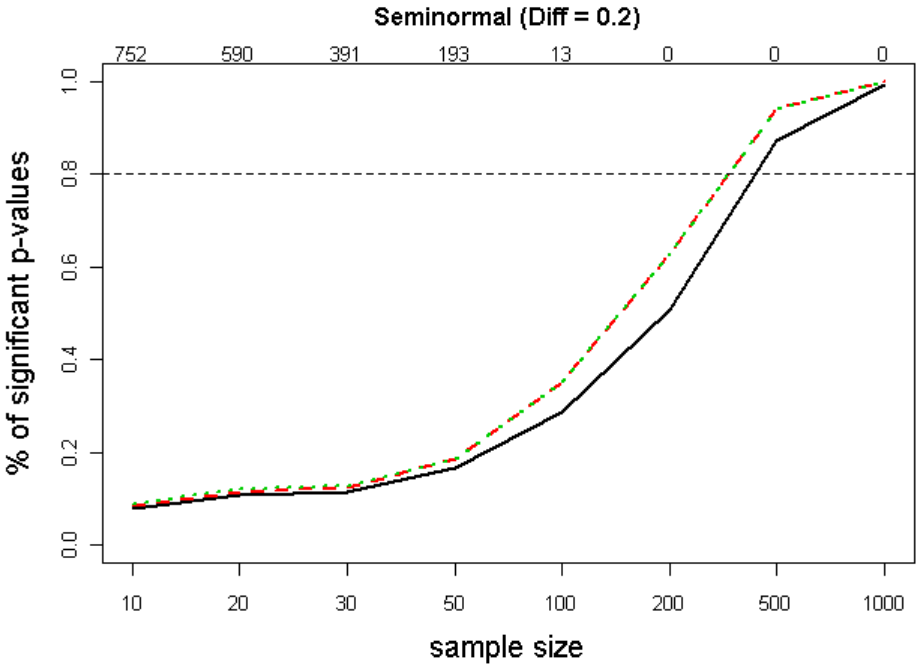

Uniforme (Diff = 0)
873 822 740 552 165 3 0 0
Test T
Test MW
Test T ou MW
% of significant p-values
0.00 0.02 0.04 0.06 0.08 0.10
10 20 30 50 100 200 500 1000
sample size

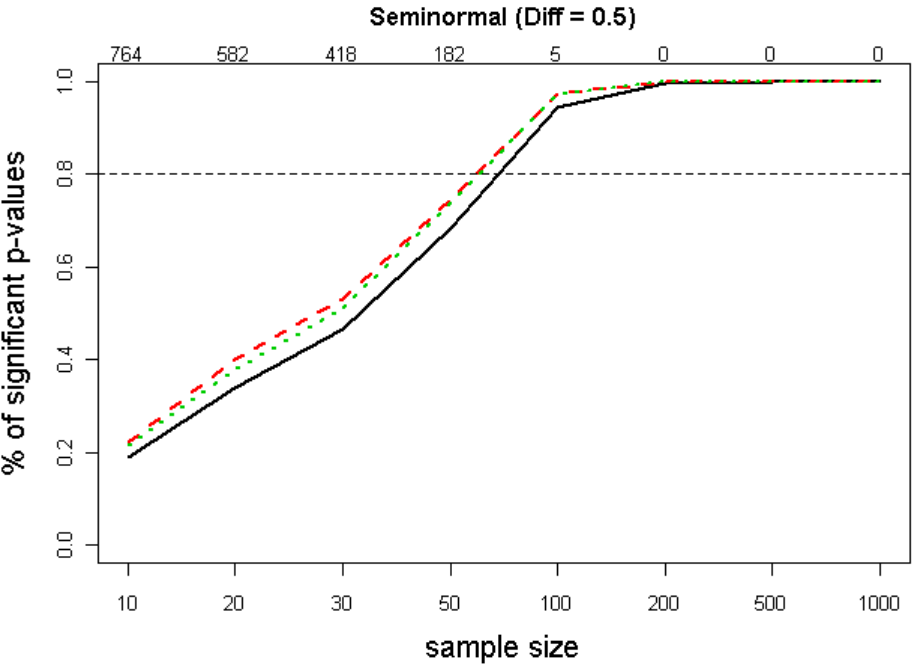

Uniforme (Diff = 0.2)
878 817 705 579 156 2 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

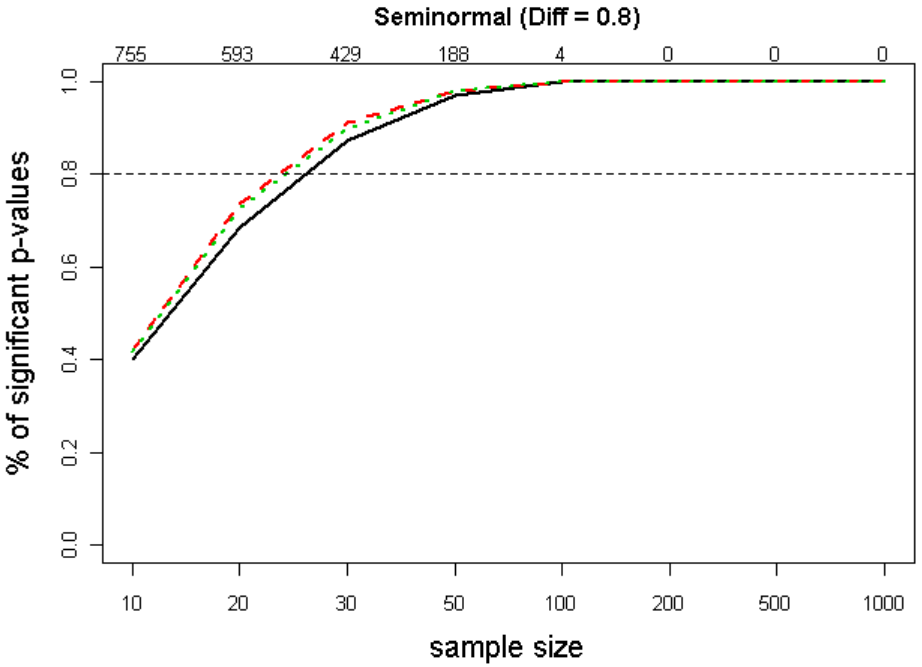

Uniforme (Diff = 0.5)
878 820 733 544 186 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

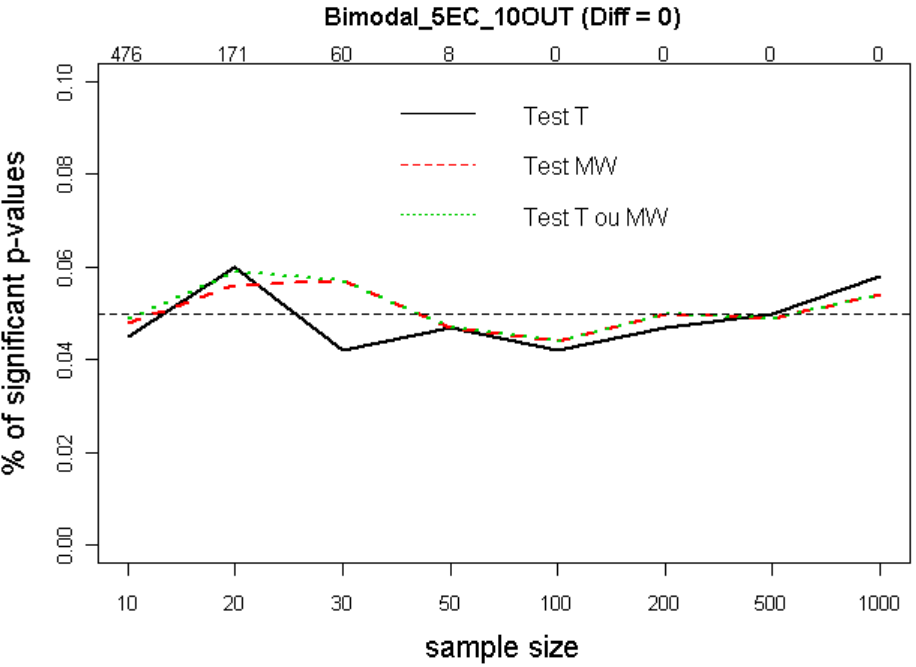

Uniforme (Diff = 0.8)
880 818 705 556 189 1 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

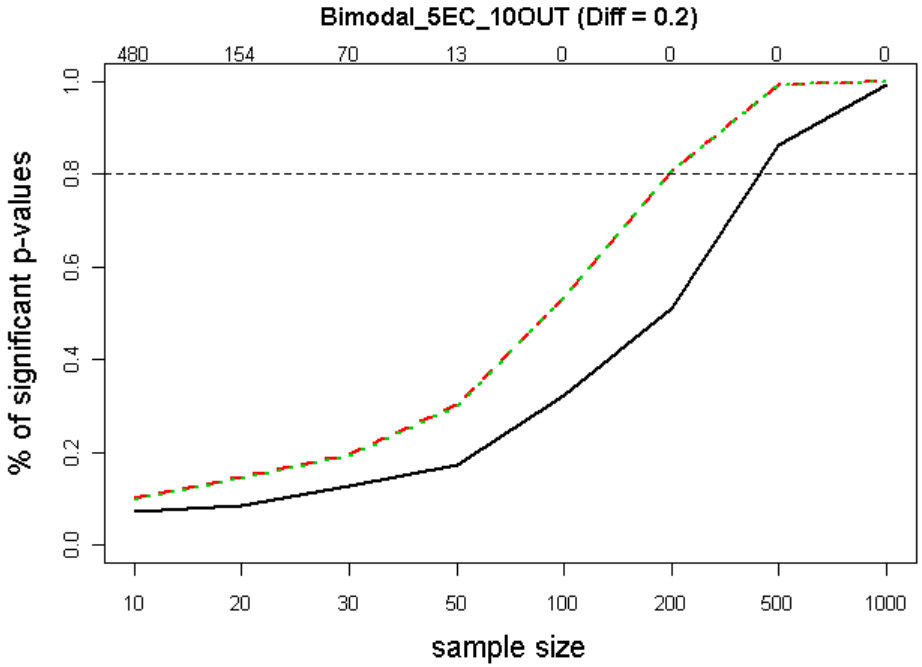

Seminormal (Diff = 0)
739 577 432 176 5 0 0 0
Test T
Test MW
Test T ou MW
% of significant p-values
0.00 0.02 0.04 0.06 0.08 0.10
10 20 30 50 100 200 500 1000
sample size

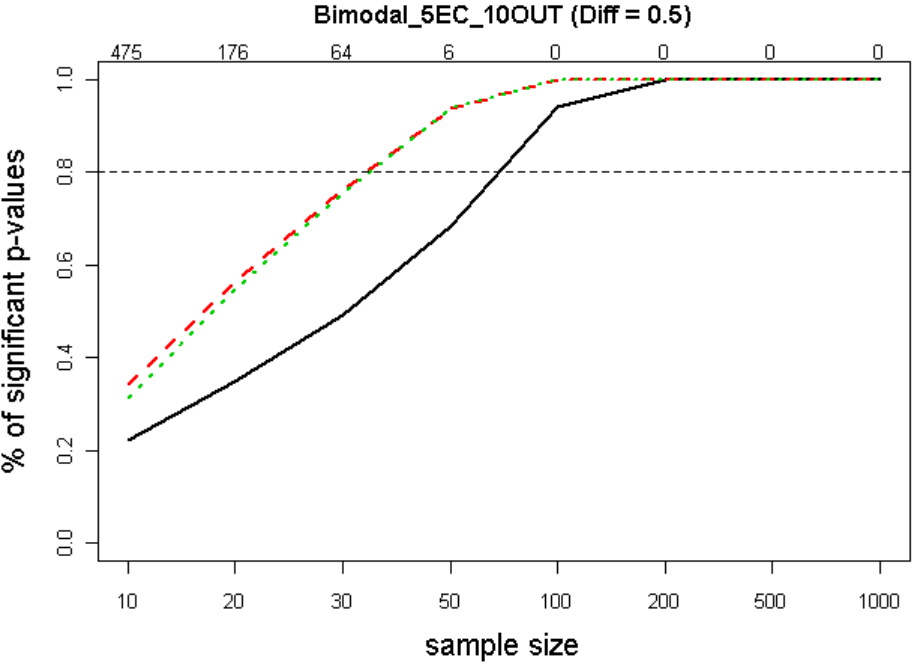

Seminormal (Diff = 0.2)
752 590 391 193 13 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

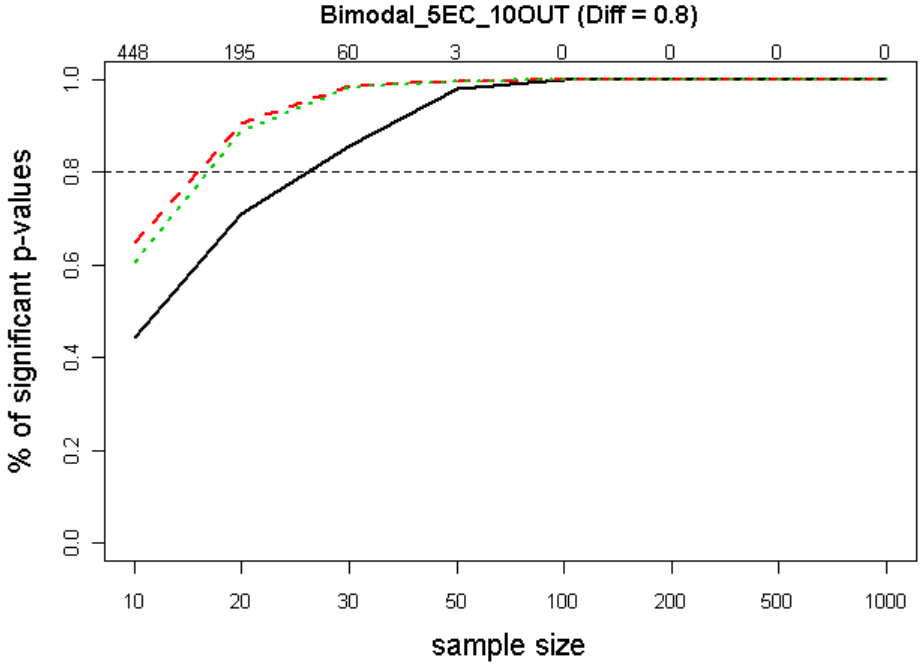

Seminormal (Diff = 0.5)
764 582 418 182 5 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

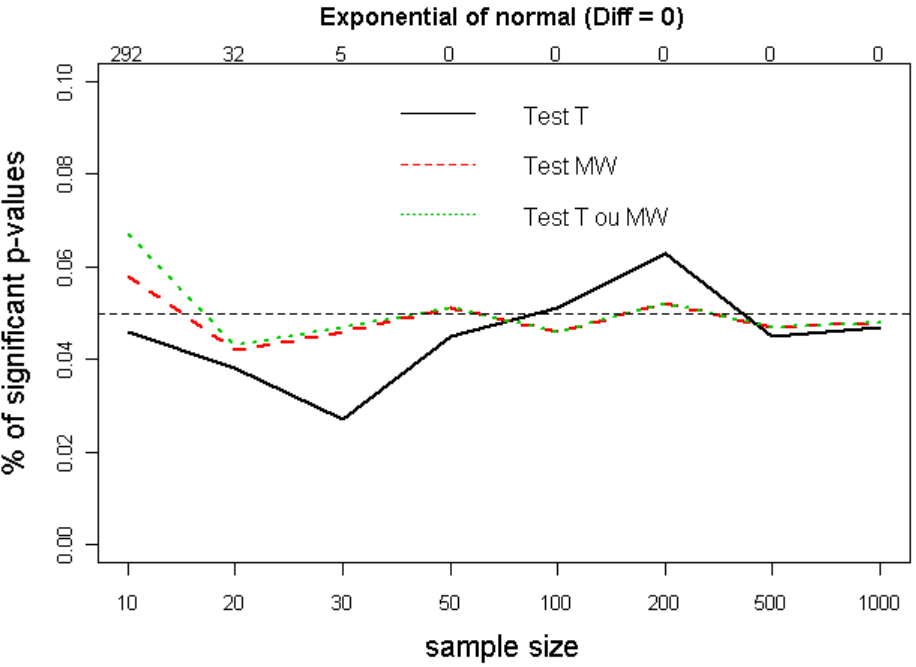

Seminormal (Diff = 0.8)
755 593 429 188 4 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

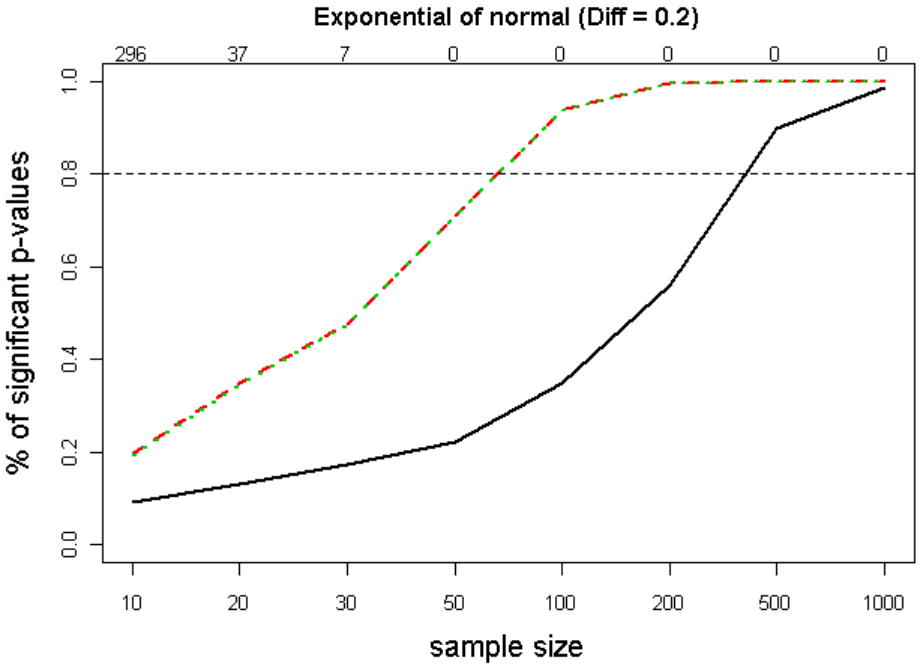

Bimodal_5EC_10OUT (Diff = 0)
476 171 60 8 0 0 0 0
Test T
Test MW
Test T ou MW
% of significant p-values
0.00 0.02 0.04 0.06 0.08 0.10
10 20 30 50 100 200 500 1000
sample size

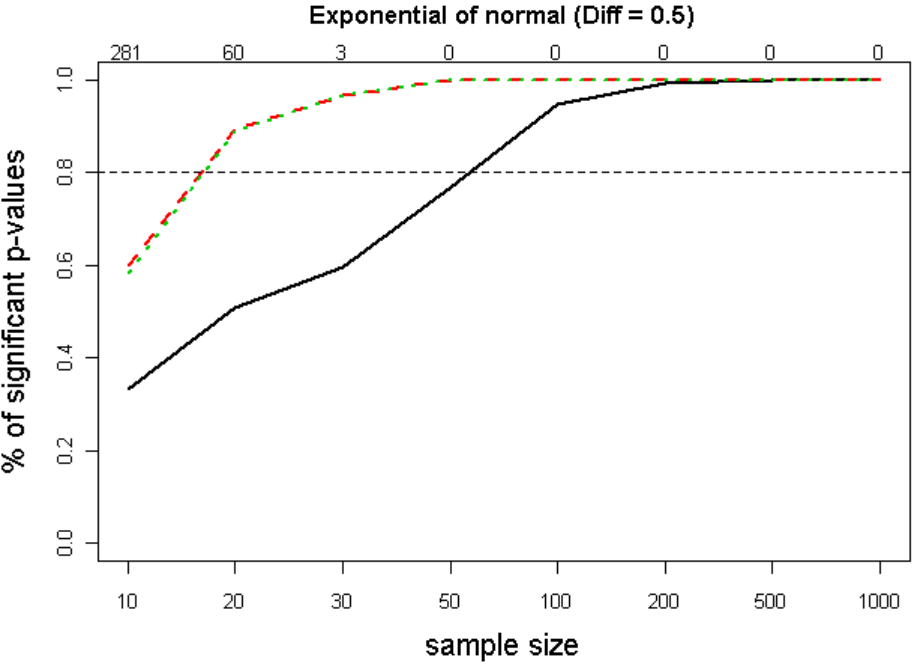

Bimodal_5EC_10OUT (Diff = 0.2)
480 154 70 13 0 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

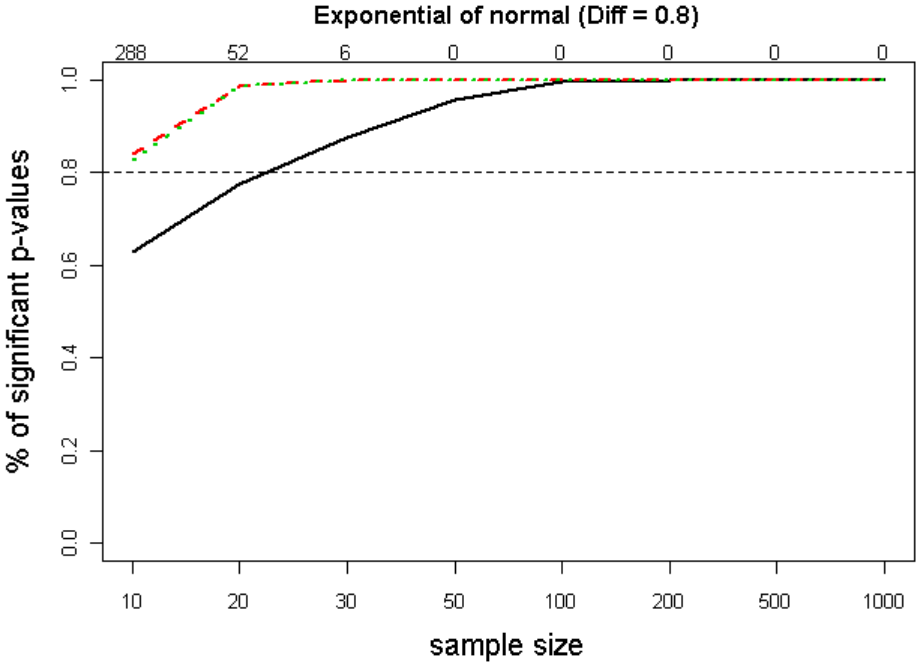

Bimodal_5EC_10OUT (Diff = 0.5)
475 176 64 6 0 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size


Bimodal_5EC_10OUT (Diff = 0.8)
448 195 60 3 0 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size


Exponential of normal (Diff = 0)
292 32 5 0 0 0 0 0
Test T
Test MW
Test T ou MW
% of significant p-values
0.00 0.02 0.04 0.06 0.08 0.10
10 20 30 50 100 200 500 1000
sample size


Exponential of normal (Diff = 0.2)
296 37 7 0 0 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size


Exponential of normal (Diff = 0.5)
281 60 3 0 0 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

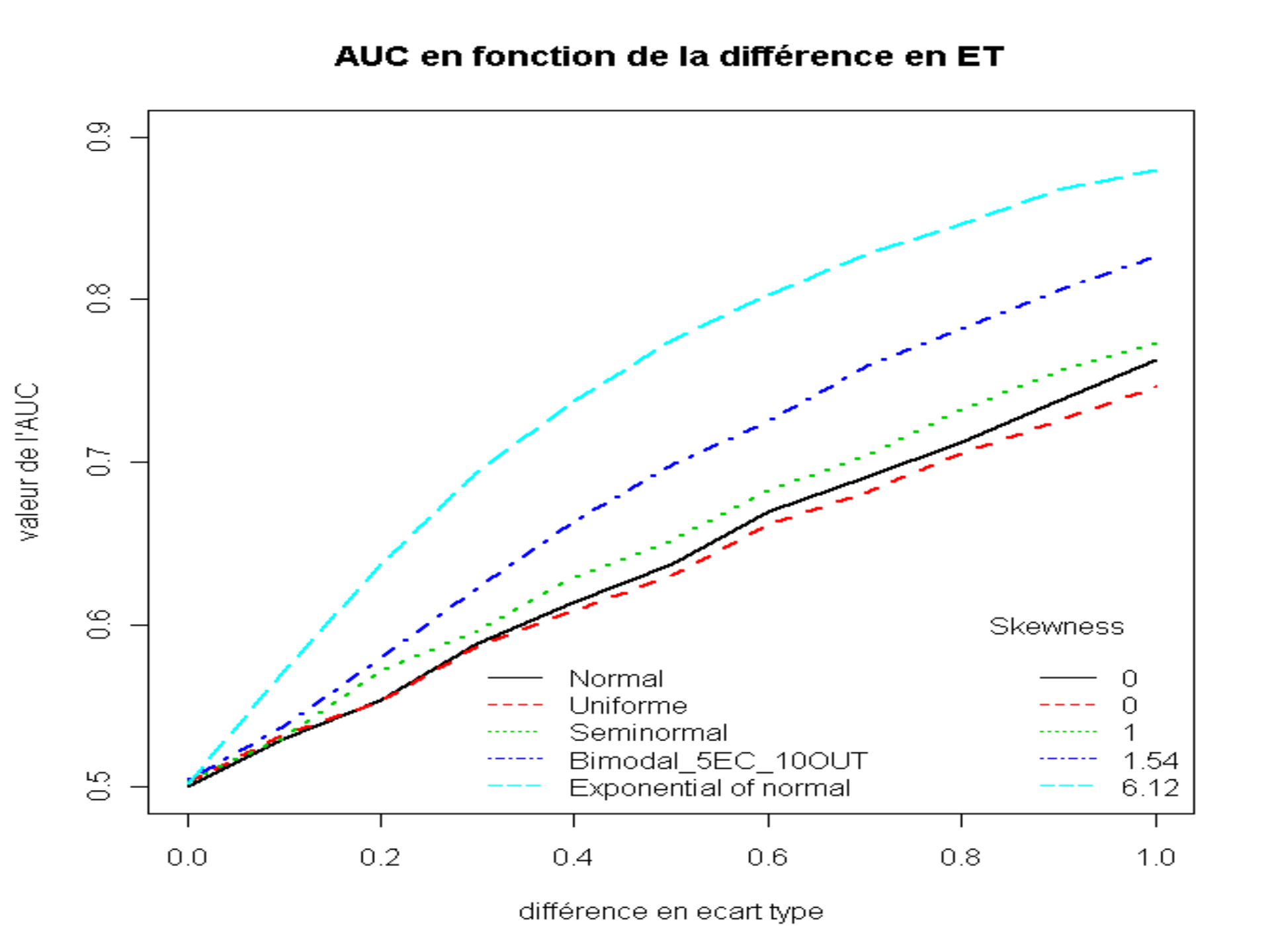

Exponential of normal (Diff = 0.8)
288 52 6 0 0 0 0 0
% of significant p-values
0.0 0.2 0.4 0.6 0.8 1.0
10 20 30 50 100 200 500 1000
sample size

Conclusion des 3 stratégies

- Le test de normalité n’apporte pas de gain notable !!!
- MW forever…

Conclusion

- Sous distribution symétrique (normale ou uniforme): test t et z ont de meilleures performances que les 3 autres tests mais le gain est très faible
- Sous distribution asymétrique ou en présence d'outliers, gain de puissance du MW et du test robuste pouvant être très important par rapport aux tests t, z et exact
- Pas de « N magique » pour le test de Student (le TCL est trop dépendant de la forme de la distribution)
- Test de normalité inutile


AUC en fonction de la différence en ET
valeur de l'AUC
0.5
0.6
0.7
0.8
0.9
0.0
0.2
0.4
0.6
0.8
1.0
différence en ecart type
Normal
Uniforme
Seminormal
Bimodal_5EC_10OUT
Exponential of normal
Skewness
0
0
1
1.54
6.12

Discussion

- Ces résultats confirment certaines publications sur le sujet (sous condition de normalité il faut 0.955 fois moins de sujets avec le test t que le test de MW pour obtenir la même puissance)
- Mais d'autres publications sont en faveur du test t
- Débat ouvert depuis 60 ans !
- On sait que des variances inégales impactent sur les performances du test t et de MW en terme d'erreur de type I. Il s'agit d'une limite de notre étude: on n'a pas fait varier ce paramètre

A guide to choosing an appropriate test

Variances	Distributions	Sample sizes	*t* test	WMW test	Welch's *U* test
Equal	Normal	Equal	*	+	+
		Unequal	*	+	+
	Heavy tailed	Equal	+	*	+
		Unequal	+	*	+
	Skewed	Equal	−	*	−
		Unequal	−	*	−
Unequal	Normal	Equal	+	−	*
		Unequal	−	−	*
	Heavy tailed	Equal	+	−	+
		Unequal	−	−	+
	Skewed[a]	Equal	−	−	−
		Unequal	−	−	−

Symbols: * = method of choice, best properties, + = acceptable, − = not acceptable.

[a] Transformations are recommended.

Points communs avec notre étude:

- Différentes distributions étudiées

Points différents

- Le ratio des sd varie
- Le ratio des effectifs varie
- La taille de l'effet invariante = 0

Limite de cette étude :

- Le critère de jugement est l'erreur de type I (seulement)

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 598-600

1538 – 9472/05/$95.00

Misconceptions Leading to Choosing the t Test Over the Wilcoxon Mann-Whitney Test for Shift in Location Parameter

Shlomo S. Sawilowsky
Wayne State University

There exist many misconceptions in choosing the t over the Wilcoxon Rank-Sum test when testing for shift. Examples are given in the following three groups: (1) false statement, (2) true premise, but false conclusion, and (3) true statement irrelevant in choosing between the t test and the Wilcoxon Rank Sum test.

1. False statement	2. True premise but false conclusion	3.True statement but irrelevant
The Wilcoxon is only for use when the data are originally in the form of ranks	The Wilcoxon's underlying assumptions are weaker (true), therefore the hypothesis being tested is less interesting (false)	The t is a classical test
The Wilcoxon is only for use in the presence of outliers	In terms of central tendency, the Wilcoxon pertains to the median (true), which is less interesting than the median (false)	Results based on the t have been accumulating for almost a century, permitting direct comparison of results over time
The Wilcoxon should only be used for small samples		The hypotheses being tested for the t and Wilcoxon aren't exactly the same
If a modern procedure should be used, it should be a permutation test, not the W		Even its inventor called the Wilcoxon test a « quick and dirty » or « crude » procedure

Présentation des tests (4/5)
test exact

- Test basé sur les permutations
  - Aucune hypothèse sur les distributions si n<50
  - Approximation normale usuelle (test Z) si n≥50

- H0: les distributions sont identiques dans les 2 groupes