



Tests non paramétriques

Olivier Decourt

Consultant et formateur indépendant

<http://www.faq-sas.com>



Plan



- Rappels : les tests paramétriques
- Les tests sur les rangs
- Les autres tests non-paramétriques



1- Rappels : les tests paramétriques

Principe général d'un test (1/5)

- Ingrédients nécessaires

- Une hypothèse à vérifier (ou à rejeter) : choisissez-la la plus contraignante possible



exemple : « ceci est une pipe »

- Une statistique de test associée à l'hypothèse
- Une distribution est connue pour cette statistique de test *si l'hypothèse est vérifiée*



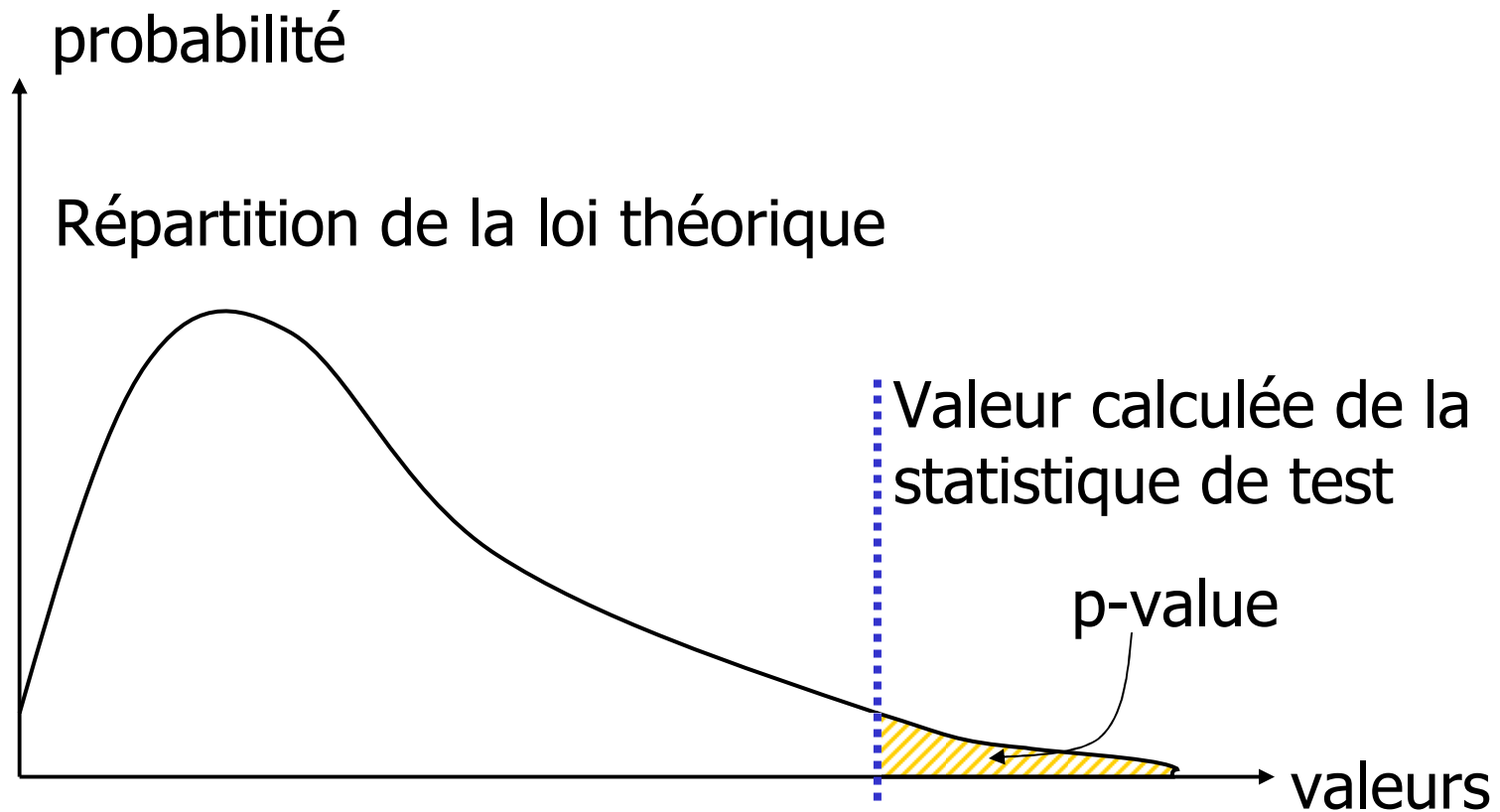
Principe général d'un test (2/5)

- **Raisonnement**
 - Supposons que l'hypothèse est vraie
→ la statistique de test suit une loi connue

 - Calculons la valeur de cette statistique de test

 - Comparons-la avec les valeurs théoriques de la loi

Principe général d'un test (3/5)



Principe général d'un test (4/5)

- P-value et conclusion
 - P-value = 1-probabilité de la valeur obtenue pour la statistique de test
 - C'est la probabilité d'obtenir une valeur de la statistique de test *encore moins favorable à l'hypothèse*
 - \approx probabilité de plausibilité de l'hypothèse (← c'est une approximation)



Principe général d'un test (5/5)

- P-value et conclusion
 - **P-value faible** → hypothèse **rejetée**
 - **P-value forte** → hypothèse **non rejetée**
 - **Attention !**

Une p-value forte n'implique pas une hypothèse acceptée. On sait juste qu'on n'a pas réussi à prouver que l'hypothèse est fausse.





Quelques tests usuels

- Comparaison de moyennes à deux groupes : test de **Student**
- Comparaison de moyennes à plusieurs groupes : test de **Fisher**
- Comparaison de variances : test de **Levene**
- Significativité d'un coefficient de régression : test de **Wald**



Pourquoi « paramétriques » ?

- **Hypothèses** sur la **loi** des données (X) testées
- Généralement, X suit une loi **normale**
 - $E(X)$ suit une loi normale
 - $E(X)/\sigma(X)$ suit une loi de Student
 - $V(X)$ suit une loi du χ^2
 - un rapport de variances suit une loi de Fisher-Snedecor



Limites des tests paramétriques

- Normalité des données
 - Pas toujours assurée (coûts, IMC, ...)
 - Garantie par le théorème central-limite...
*au-delà d'un certain **volume** de données !*
- Volume de données
 - Trop faible → test peu fiable
 - Trop important → test peu fiable !



Le vocabulaire des tests (1/2)

- Erreur de types I et II

		Réalité	
		Hypothèse vérifiée	Hypothèse rejetée
Décision	Acceptation	Acceptation correcte de l'hypothèse.	Erreur de type II (risque bêta)
	Rejet	Erreur de type I (risque alpha)	Rejet correct de l'hypothèse.



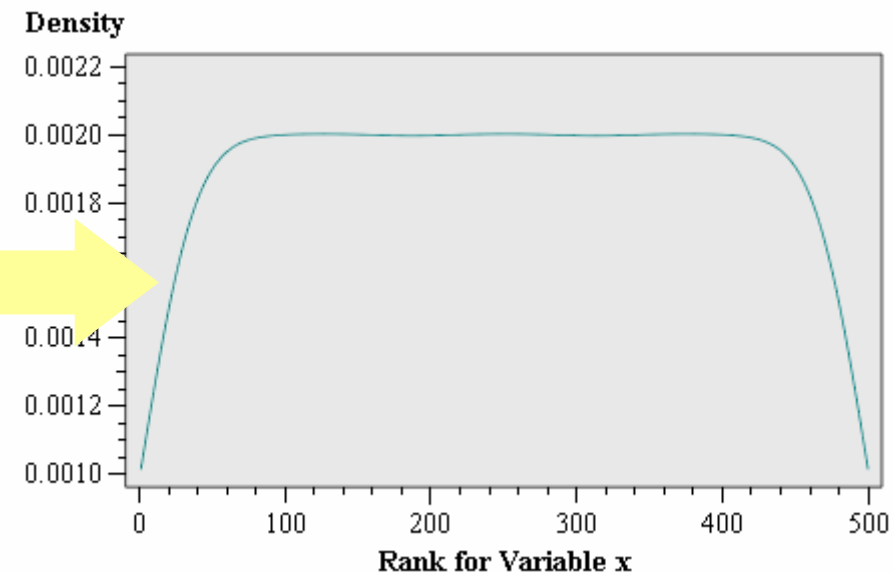
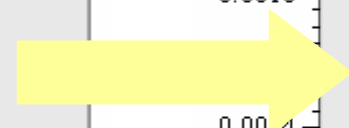
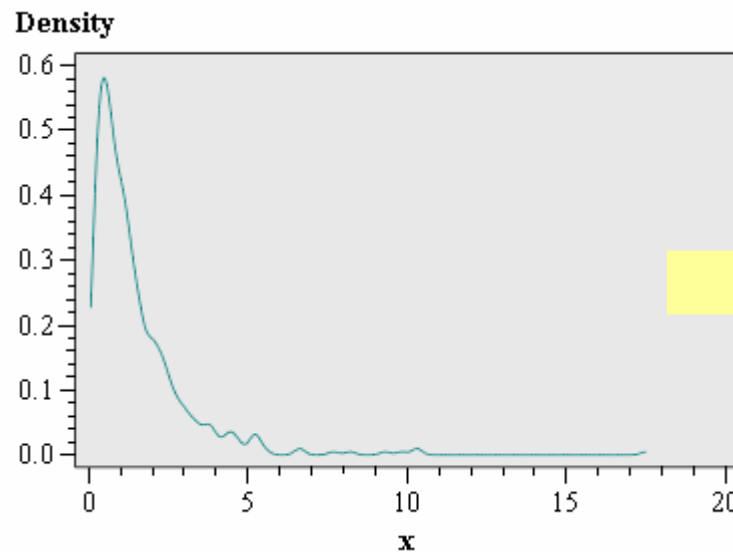
Le vocabulaire des tests (2/2)

- **Puissance** du test : $1 - \text{risque } \beta$
- **Non-rejet** de l'hypothèse
 - Trop de données nuit (augmentation du risque β)
 - Test trop puissant → danger (ex : tests de normalité)
- **Rejet** de l'hypothèse
 - Trop peu de données nuit (augmentation du risque α)
 - Sur un grand volume de données, hypothèse rejetée avec un fort risque β



2- Les tests sur les rangs

Du bon usage des rangs (1/3)



- Moyenne des rangs = médiane
- → tests d'égalités de médianes



Du bon usage des rangs (2/3)

```
PROC RANK DATA = tableSAS1
              OUT = tableSAS2 ;
  VAR variable(s)lue(s) ;
  RANKS variable(s)Créée(s) ;

RUN ;
```



Du bon usage des rangs (3/3)

- Intérêt des tests sur les rangs :
 - Distributions asymétriques
 - Nombre important d'individus atypiques (outliers)



- Attention, test puissant !
 - Difficulté de valider l'hypothèse sur de gros volumes de données



La procédure NPAR1WAY

Principales options
EDF
MEDIAN
WILCOXON

```
PROC NPAR1WAY
```

```
DATA = tableSAS
```

```
options ;
```

```
CLASS variableDeGroupe ;
```

```
VAR variable(s)Numérique(s) ;
```

```
RUN ;
```



Tests sur les rangs (exemples)

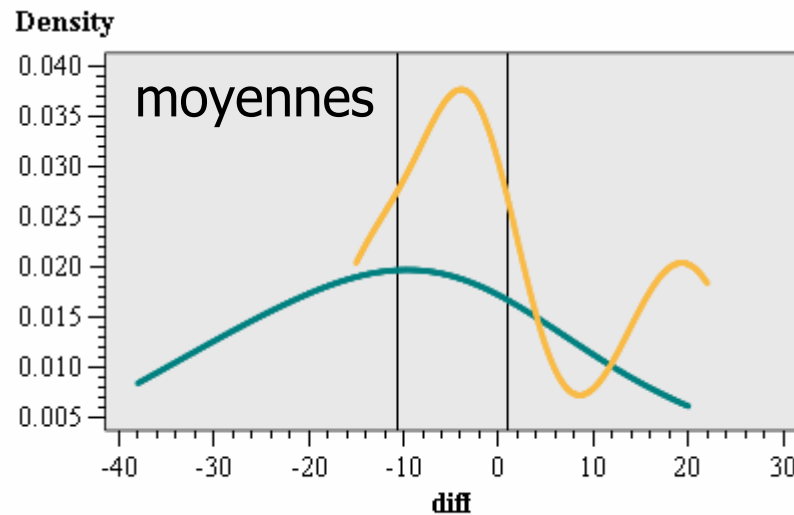
Cas étudié	Calcul paramétrique	Calcul non paramétrique équivalent
Égalité de moyenne à une valeur	Test de Student à 1 échantillon	Test de Mann et Whitney
Comparaison des moyennes de deux groupes	Test de Student à 2 échantillons	Test de Wilcoxon
Comparaison des moyennes de k groupes	Test de Fisher	Test de Kruskal-Wallis
Corrélation de deux variables	Corrélation (r) de Pearson	Corrélation de Spearman



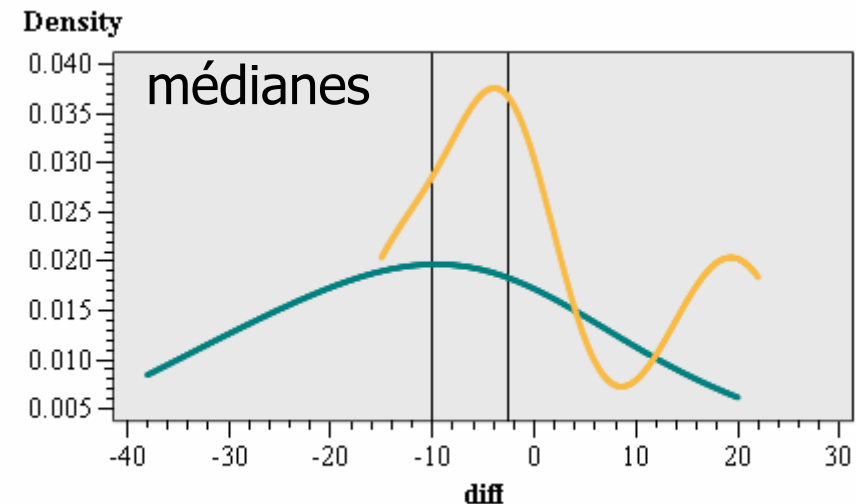
Un exemple concret (1/3)

- Un sujet porteur : le **sexe**
- 2 groupes étudiés (de 10 jeunes) : un **témoin** et un groupe ayant suivi un **stage de sensibilisation** au SIDA
- Étude : **nombre de rapports non protégés** (différence avant/après le stage)

Un exemple concret (2/3)



Groupe ; 0 = contrôle, 1 = stage — 0 — 1



Groupe ; 0 = contrôle, 1 = stage — 0 — 1

Groupe	Moyenne	Médiane
Contrôle	-10,7	-10
Stage	1	-2,5



Un exemple concret (3/3)

- Test de **Student** à 2 groupes :
 - Variances égales (p-value : 0,53)
 - **Moyennes différentes** (de justesse : **0,10 !**)

- Test de **Wilcoxon** :
 - **Médianes proches** (**0,20** ou 0,19 selon le type de test choisi)



3- Les autres tests non-paramétriques



Test d'adéquation à une loi

- Divers tests forcément non-paramétriques

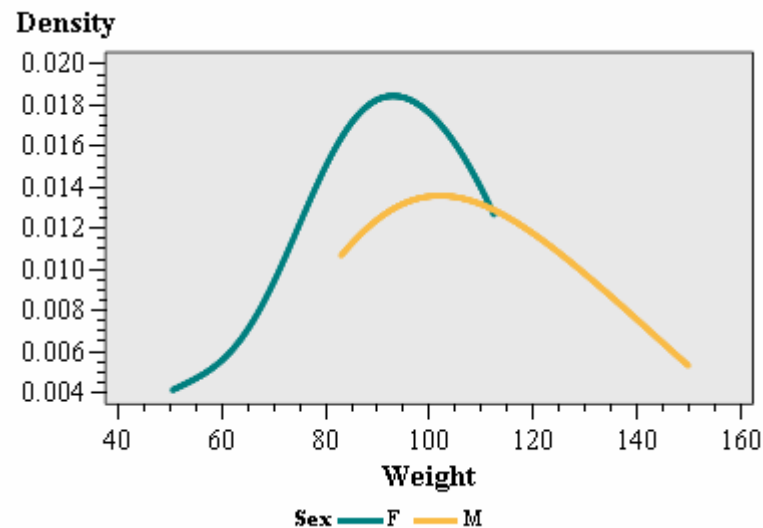
- Shapiro-Wilks
- Kolmogorov-Smirnov
- Anderson-Darling
- Cramer-Von Mises

```
PROC UNIVARIATE  
  DATA = tableSAS  
  NORMAL ;  
  VAR variable ;  
RUN ;
```

- Test de K/S : $D = \sup |F_{\text{théorie}} - F_{\text{obs}}|$

Comparaison de deux distributions

- Test de K/S également disponible pour comparer deux distributions
- PROC NPAR1WAY, option EDF



Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0,18307 9	D	0,36666 7
KSa	0,79802 4	Pr>KSa	0,5474



Association de 2 variables (1/2)

- Test du χ^2 : non-paramétrique...
- ... mais fragile (effectifs théoriques < 5).

- Test exact de Fisher : calcul de toutes les tables ayant les mêmes marges, comparaison de la liaison ligne/colonne dans tous les cas \rightarrow probabilité de la liaison observée

Association de 2 variables (2/2)

Frequency Expected Percent	Table of Sex by Height			
	Sex	Height		Total
		60-	60+	
F	4	5	9	
	3,3158 21,05	5,6842 26,32	47,37	
M	3	7	10	
	3,6842 15,79	6,3158 36,84	52,63	
Total	7	12	19	
	36,84	63,16	100	

WARNING: 50% of the cells have expected counts less than 5, Chi-Square may not be a valid test

Statistic	DF	Value	Prob
Chi-Square	1	0,425	0,515

Fisher's Exact Test	
Table Probability (P)	0,3
Two-sided Pr <= P	0,65



Tests exacts (1/3)

- Principe des tests exacts :
 - tous les cas de figures possibles (autres jeux de données respectant les mêmes contraintes, par exemple les marges)
 - calcul direct de probabilités

- Dans de nombreux cas, calcul coûteux
 - → approximation de Monte-Carlo



Tests exacts (2/3)

- Tests exacts disponibles :
 - PROC FREQ = Fisher, χ^2 de Pearson, Cochran-Armitage, ...
 - PROC NPAR1WAY = Wilcoxon, ...
 - PROC LOGISTIC = Wald des coefficients

... via l'instruction EXACT

Tests exacts (3/3)

Tranche d'âge	Maladie		Total
	0	1	
20-29	9	1	10
30-34	13	2	15
35-39	9	3	12
40-44	10	5	15
45-49	7	6	13
50-54	3	5	8
55-59	4	13	17
60 +	2	8	10
Total	57	43	100

Fisher's Exact Test	
Table Probability (P)	1,87E-10

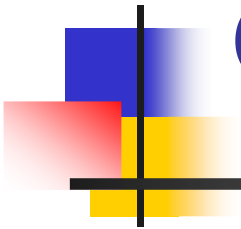
Monte Carlo Estimate for the Exact Test	
Pr <= P	2,00E-04
99% Lower Conf Limit	0
99% Upper Conf Limit	5,64E-04

Cochran-Armitage Trend Test	
Statistic (Z)	-5,1067
One-sided Pr < Z	<,0001
Two-sided Pr > Z	<,0001

Monte Carlo Estimates for the Exact Test	
Two-sided Pr >= Z	
Estimate	0
99% Lower Conf Limit	0
99% Upper Conf Limit	4,60E-04



Conclusion





Conclusion (1/2)

- Limites des tests paramétriques :
 - **Trop peu** de données
 - Distributions **asymétriques**
 - **Moyennes peu fiables**
- Remède : le test non-paramétrique
 - Attention à sa **puissance** !
 - Test exact = bon **complément** d'un test classique

Conclusion (2/2)

- Dans tous les cas :

TROP de données
→ test inadapté

