STT-3500/STT-6230, Méthodes non paramétriques/ Méthodes non paramétriques avancées

Pierre Duchesne

December 25, 2017

Coordonnées

- Professeur: Pierre Duchesne
- Courriel: duchesne@dms.umontreal.ca
- Téléphone: 514-343-7267
- Bureau: 4251 du Pavillon André-Aisenstadt
- Site web: http://www.dms.umontreal.ca/∼duchesne

Plan de cours

1. Méthodes non paramétriques versus nonparamétriques
2. Statistiques d'ordre et méthodes pour un échantillon
3. Méthodes pour deux échantillons
4. Méthodes pour données appariées
5. Comparaisons de plusieurs groupes
6. Tests d'ajustement
7. Tests sur les chaînes

Barême

Le barême proposé pour STT-3500 est le suivant:

- Examen intra: 30%;
- Examen final: 40%;
- Devoirs: 30%;

Le barême proposé pour STT-6230:

- Examen intra: 30%;
- Examen final: 40%;
- Devoirs: 30%;

Barême

Ouvrages de référence (recommandés):

- Dickinson Gibbons, Jean et Chakraborti, S. (2011). *Nonparametric Statistical Inference*, 5ième édition. CRC Press/Chapman & Hall: Boca Raton, Florida.
- Higgins, James J. (2004). *Introduction to Modern Nonparametric Statistics*, Brooks/Cole: Belmont, California.
- Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, Holden Day: Oakland, California.
- Serfling, Robert, J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons: New York.

Méthodes non paramétriques versus nonparamétriques

Souvent, lors d'analyse de données, on doit vérifier l'hypothèse de normalité.

- La normalité est un modèle pour décrire la distribution de la population.
- Si la normalité est raisonnable, on peut par exemple formuler une vraisemblance normale et utiliser des méthodes reposants sur la normale.
- Lorsque c'est normal, les lois de plusieurs tests sont exactes (en régression, en planification d'expériences).

Que faire si la normalité n'est pas plausible?

La méthodologie des transformations permet souvent de se rapprocher de la normalité.

- En économie, souvent la transformation log est appliquée avant de faire les analyses.
- Famille de transformations puissance.
- Plus systématiquement, famille de transformation de Box-Cox:

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ \log(Y) & \lambda = 0. \end{cases}$$

Méthodes non paramétriques

- Ces méthodes nécessitent des hypothèses minimales sur la forme de la distribution.
- Un exemple d'hypothèse est de présumer que la loi est continue.
- Un autre exemple est que la distribution dépend d'un paramètre de position et d'échelle, mais la distribution proprement dite n'est pas mentionnée.

Famille de distributions avec paramètre de position et d'échelle

Une famille de distributions particulièrement utile est de la forme:

$$f(x|\boldsymbol{\theta}) = \frac{1}{b} h\left(\frac{x-a}{b}\right),$$

où $\boldsymbol{\theta} = (a, b)^\top$. La loi normale est dans cette classe, avec $a = \mu$, $b = \sigma$, et $h(z)$ est la densité de la normale centrée réduite. Les paramètres a et b sont dits paramètres de position et d'échelle, respectivement.

Méthodes statistiques et non paramétriques

En statistique, de manière générale, il y a deux paradigmes:

- Statistique dite classique: $X_1, \ldots, X_n$ est un échantillon aléatoire de taille n, donc composé de variables aléatoires indépendantes et identiquement distribuées (iid), de densité $f(x|\theta)$. Le vecteur des paramètres est θ, qui est fixe mais inconnu.
- Statistique non paramétrique: $X_1, \ldots, X_n$ est un échantillon aléatoire, composé de n variables aléatoires iid dont la densité est inconnue mais on est intéressé à une caractéristique θ de celle-ci. Le θ peut être un scalaire ou un vecteur (on pourrait retrouver la médiane, un paramètre d'échelle, etc.).

Dans les deux cas, on veut faire de l'inférence sur θ.

Validité sous des conditions plus générales

- Souvent, les méthodes non paramétriques sont valides pour une plus grande classe de distributions (par exemple, pas seulement pour la loi normale). Pour illustrer la sensibilité des méthodes classiques, considérons l'exemple suivant.
- Exemple: Soit $X_1, \ldots, X_n$ un échantillon iid de densité $\mathcal{N}(\theta, \sigma^2)$. On désire confronter les hypothèses suivantes:

$$\begin{cases} H_0 : \theta = 0, \\ H_1 : \theta \neq 0. \end{cases}$$

On suppose que σ^2 est inconnu.

Exemple (suite)

- Afin de tester l'hypothèse précédente, le test du rapport de vraisemblance donne le test-t, avec la région critique de niveau α qui est donnée par:

$$\mathcal{C} = \{(x_1, \ldots, x_n) \mid n^{1/2}\frac{|\bar{x}|}{s} \geq t_{n-1;1-\alpha/2}\},$$

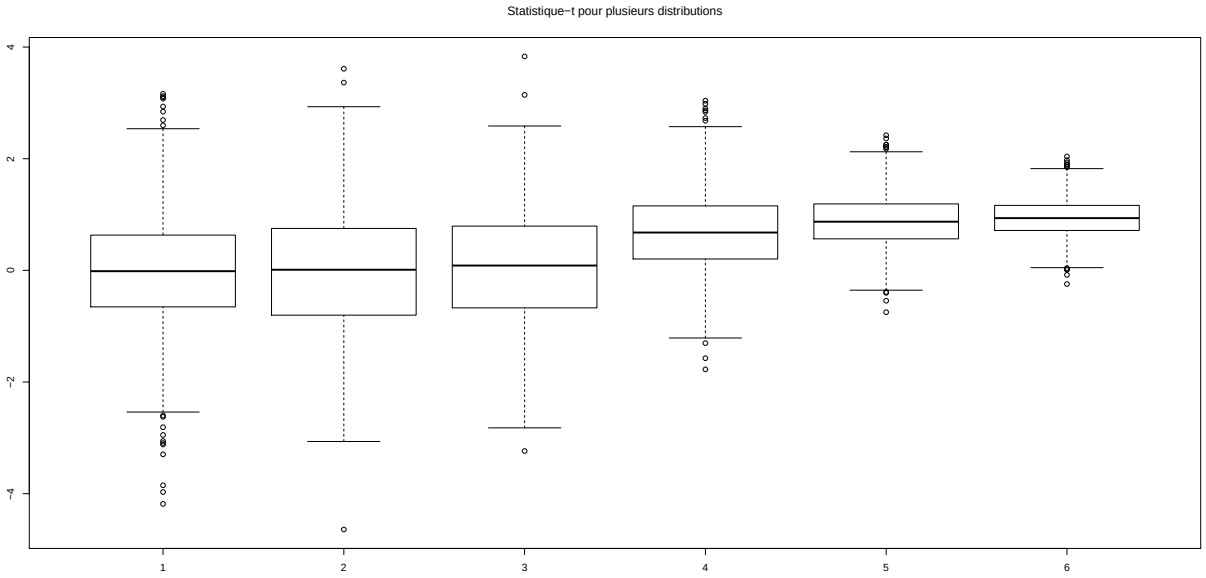
où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ et $t_{n-1;1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ d'une t_{n-1}.

Exemple (suite)

- Afin de vérifier la sensibilité du test-t à l'hypothèse de normalité, on considère une approche par simulation Monte Carlo.
- On considère $n = 20$, et en sélectionne 1000 échantillons de taille n de façon indépendante. On fait le test au niveau 5%.
- On considère les lois $\mathcal{N}(0, 1)$, t_2, t_3. On considère aussi des situations où la 20ième donnée est systématiquement mal enregistrée et vaut 4,8 et 12.
- On calcule le nombre moyen des rejets de l'hypothèse nulle.

Exemple (suite)

distribution	taux de rejet
$\mathcal{N}(0,1)$	0.050
t_2	0.045
t_3	0.046
$x_{20} \equiv 4$	0.039
$x_{20} \equiv 8$	0.009
$x_{20} \equiv 12$	0.000



Loi binomiale

On rappelle que la loi binomiale, notée $X \sim bin(n, \theta)$, avec $\theta \in [0, 1]$, est une variable aléatoire discrète de support $\{0, 1, \ldots, n\}$:

$$p_X(x; \theta) = P(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}.$$

On rappelle que $E(X) = n\theta$ et $\text{var}(X) = n\theta(1 - \theta)$.
Nous allons considérer des méthodes non-paramétriques reposant sur les propriétés de la loi binomiale afin d'étudier la médiane.