Détection et traitement des valeurs extrêmes

COURS 12

Aujourd’hui, nous parlons prinicpalement des valeurs extrêmes

1) Définitions
   Qu’entendons-nous exactement par valeur extreme, exactement?

2) Motivation
   Les valeurs extrêmes ont-ells une importance?

3) Détection
   Comment détecter les valeurs extrêmes?

4) Traitement
   Comment prendre en compte les valeurs extremes?

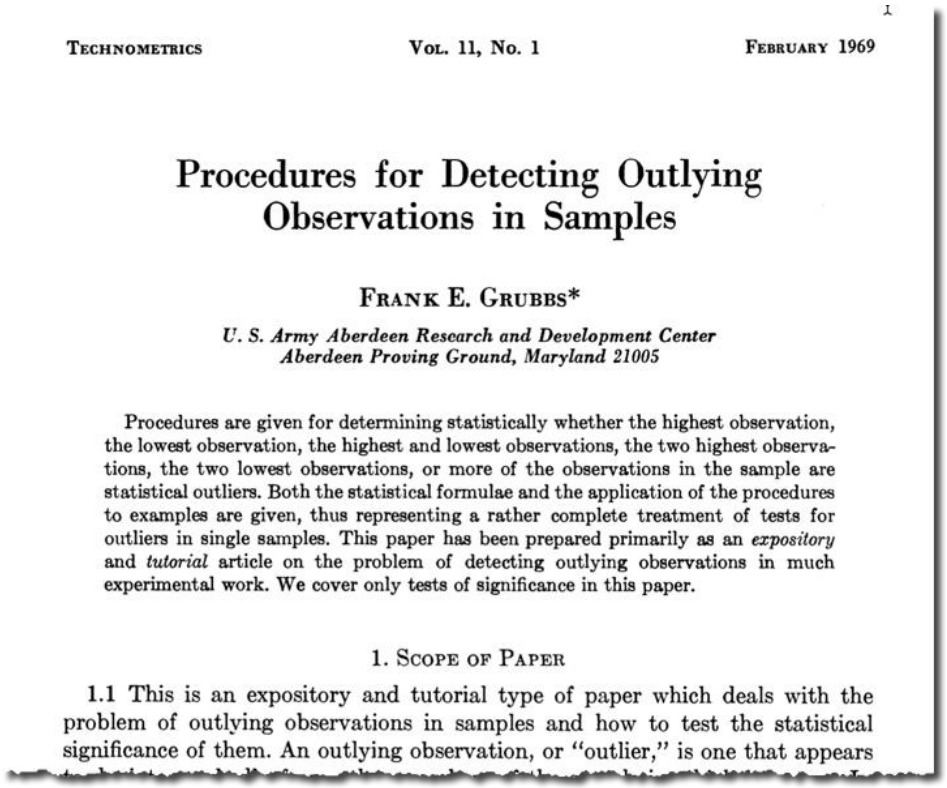
Préambule

- Les suggestions partagées dans ce cours ne se substituent pas aux protocoles mis en oeuvre par les instituts de statistique pour assurer la qualité des données
- Elles sont censées offrir d'offrir des garanties supplémentaires une fois les vérifications "de routine" effectuées
- Utiles tant pour les analystes que pour les producteurs de données

Définitions

Qu’est-ce qu’une valeur extrême?

Une valeur extrême (outlier) est une observation “qui semble dévier notoirement des autres éléments de l’échantillon auquel elle appartient” (Grubbs, 1969)



TECHNOMETRICS VOL. 11, NO. 1 FEBRUARY 1969

Procedures for Detecting Outlying Observations in Samples

FRANK E. GRUBBS*

U. S. Army Aberdeen Research and Development Center
Aberdeen Proving Ground, Maryland 21005

Procedures are given for determining statistically whether the highest observation, the lowest observation, the highest and lowest observations, the two highest observations, the two lowest observations, or more of the observations in the sample are statistical outliers. Both the statistical formulae and the application of the procedures to examples are given, thus representing a rather complete treatment of tests for outliers in single samples. This paper has been prepared primarily as an *expository* and *tutorial* article on the problem of detecting outlying observations in much experimental work. We cover only tests of significance in this paper.

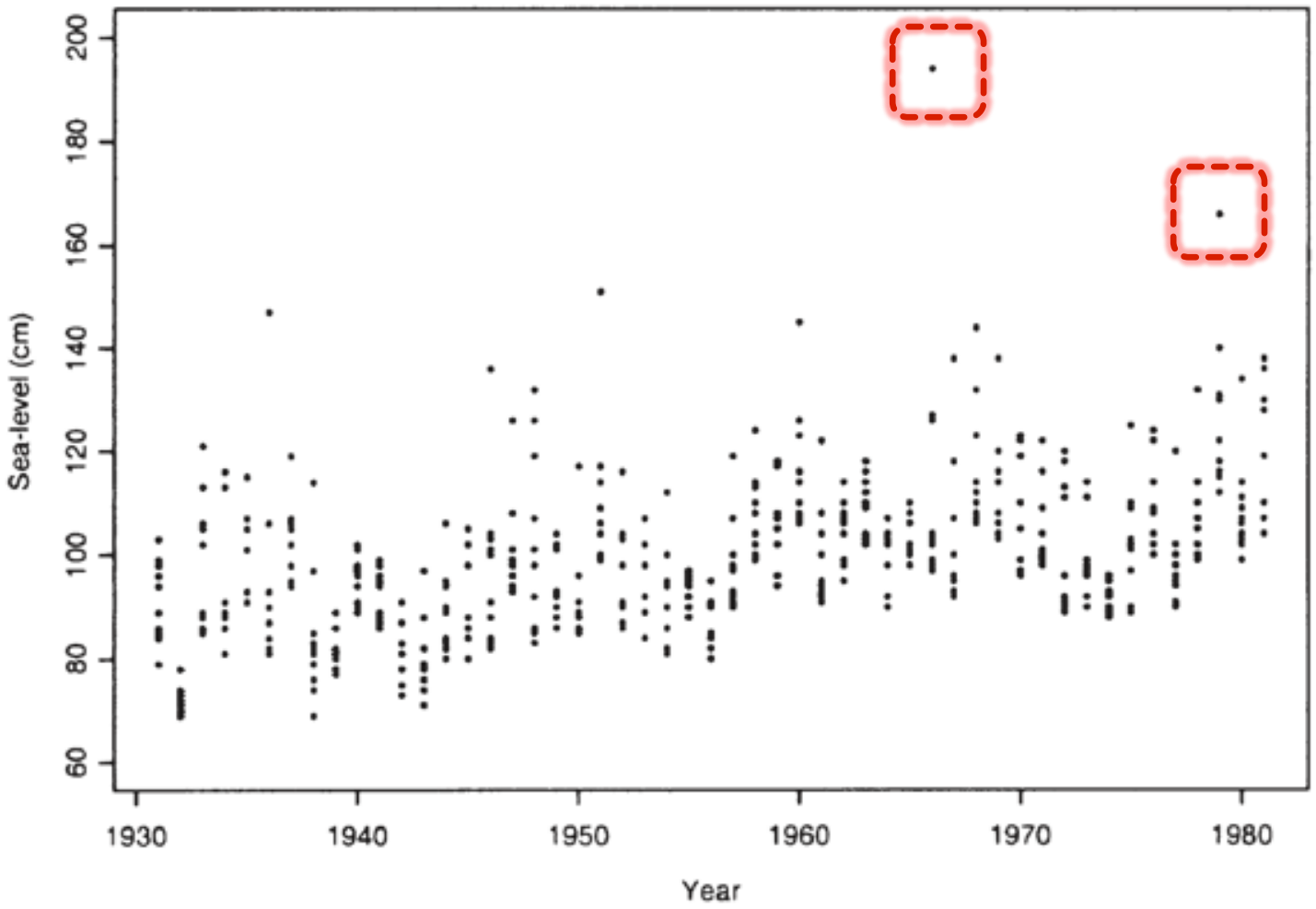
1. SCOPE OF PAPER

1.1 This is an expository and tutorial type of paper which deals with the problem of outlying observations in samples and how to test the statistical significance of them. An outlying observation, or “outlier,” is one that appears

Qu’est-ce qu’une valeur extrême?

- Remarque: On se concentre sur les valeurs extrêmes univariées, celles trouvées quand on regarde la distribution des valeurs dans une seule dimension (Ex: revenu).
- Nous utilisons Venise comme illustration

Plus haut niveau de la mer à Venise



Quelles sont les causes des valeurs extrêmes?

- Erreurs humaines, ex: Erreurs de saisie
- Erreurs d'instrument, ex: Erreurs de mesure
- Erreurs de traitement des données, ex: Manipulation des données
- Erreurs d'échantillonnage, ex: Extraction des données de mauvaises sources
- Une valeur extreme n'est pas une erreur, juste une 'innovation' dans les données

Un dilemme

- Les valeurs extrêmes peuvent être d'authentiques valeurs
- L'arbitrage entre la perte de précision si l'on exclut de ''bonnes'' observations, et le biais des estimateurs si l'on garde les ''mauvaises'' observations
- Le challenge est double:
  1. pour déterminer si une valeur extrême est bonne (authentique) ou mauvaise (erreur)
  2. pour évaluer son impact sur les statistiques d'intérêt

Les valeurs extrêmes ont-elles une importance?

Théorie d'abord

- Trois papiers:

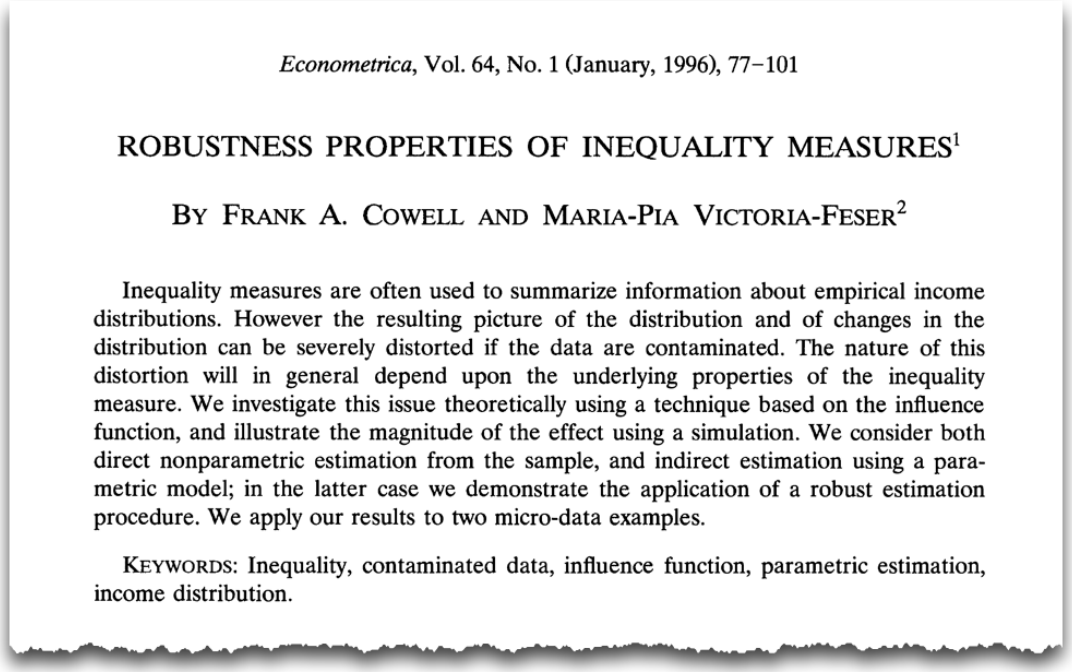
  I. 1996a
     Frank Cowell et Maria-Pia Victoria-Feser

  II. 2007
     Frank Cowell et Emmanuel Flachaire (*)

  III. 1996b
     Frank Cowell et Maria-Pia Victoria-Feser

Valeurs extrêmes et mesures d'inégalité – I

Cowell et Victoria-Feser (1996a)



Econometrica, Vol. 64, No. 1 (January, 1996), 77–101

ROBUSTNESS PROPERTIES OF INEQUALITY MEASURES[1]

BY FRANK A. COWELL AND MARIA-PIA VICTORIA-FESER[2]

Inequality measures are often used to summarize information about empirical income distributions. However the resulting picture of the distribution and of changes in the distribution can be severely distorted if the data are contaminated. The nature of this distortion will in general depend upon the underlying properties of the inequality measure. We investigate this issue theoretically using a technique based on the influence function, and illustrate the magnitude of the effect using a simulation. We consider both direct nonparametric estimation from the sample, and indirect estimation using a parametric model; in the latter case we demonstrate the application of a robust estimation procedure. We apply our results to two micro-data examples.

KEYWORDS: Inequality, contaminated data, influence function, parametric estimation, income distribution.

- C'est un bel article
- Explique pourquoi les valeurs extrêmes (bruits) sont de sérieuses menaces pour la plupart des mesures d'inégalité.
- "si la moyenne doit être estimée à partir de l'échantillon, alors toutes les mesures indépendantes de l'échelle ou de la traduction et décomposables ont une fonction d'influence non limitée" (p. 89)
- Une fonction d'influence non limitée est une catastrophe.

Pourquoi la fonction d'influence non limitée est une catastrophe?

- La fonction IF est une mesure du biais de l'estimateur dû à la presence des valeurs extrêmes.
- Une fonction d'influence non limitée signifie que le biais peut être infiniment grand.
- Si le biais des estimateurs de l'inégalité peut être infiniment grand, alors les valeurs extrêmes sont une priorité pour les producteurs et les utilisateurs de données.

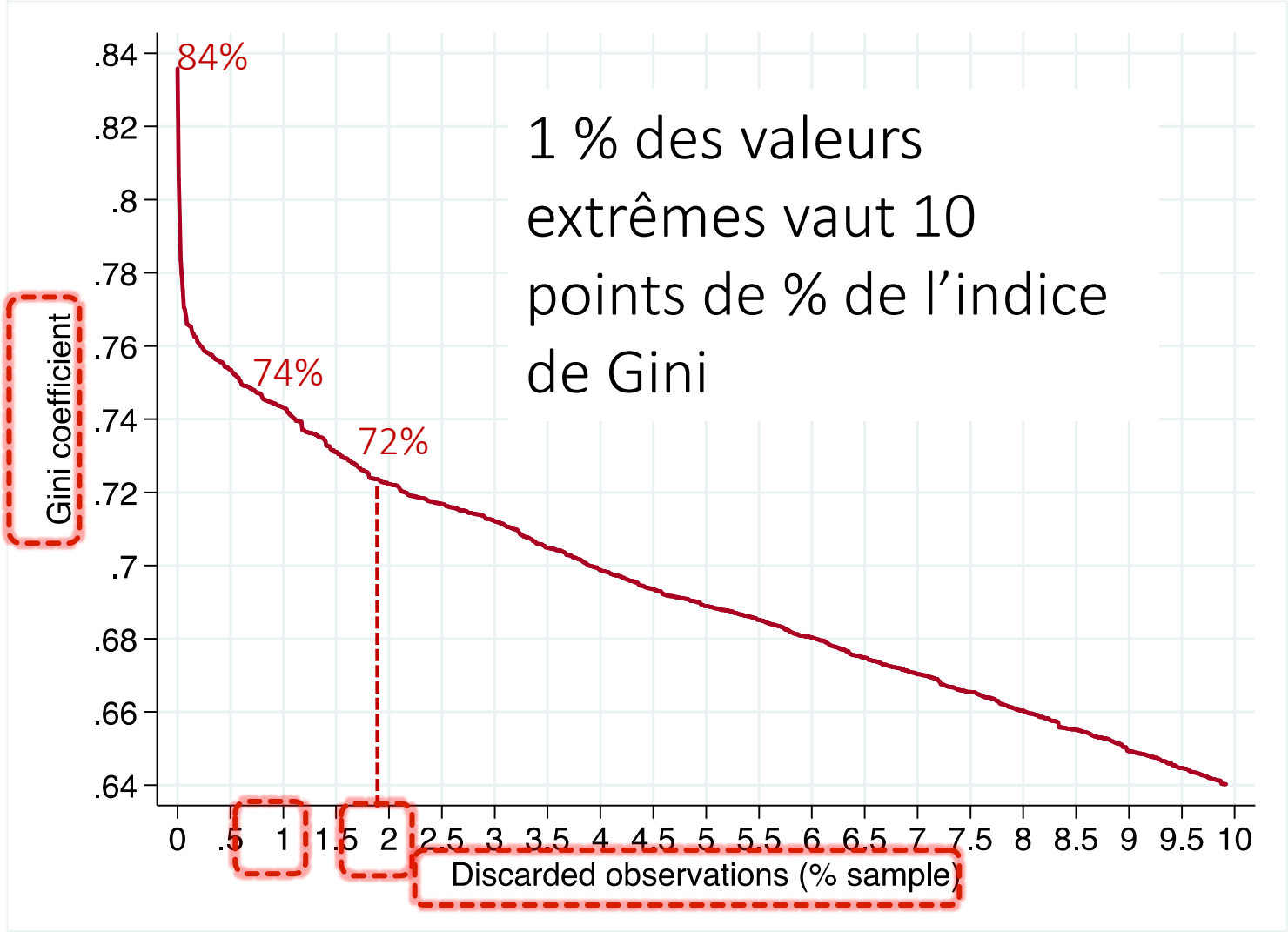
En pratique

Hlasny et Verme (2018: 191)

- De nombreux chercheurs éliminent régulièrement les observations aberrantes ou problématiques ou appliquent un codage de haut niveau sans tenir compte des implications pour la mesure de l'inégalité
- Un exemple pour illustrer

Sensibilité de l'indice de Gini index aux valeurs extrêmes

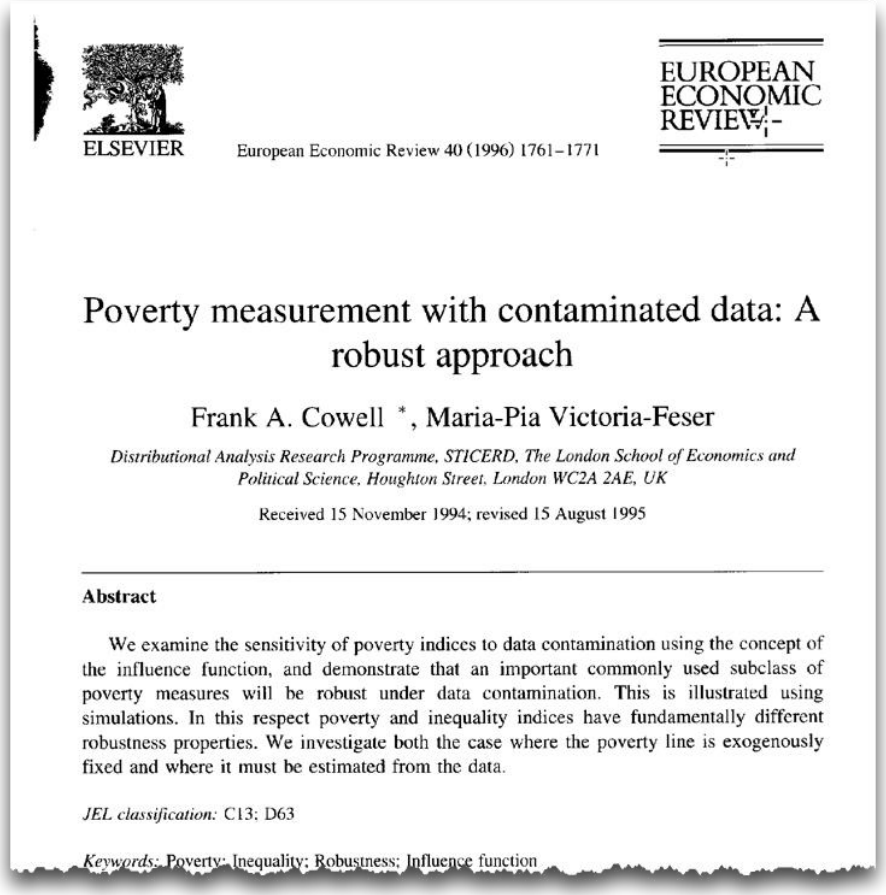
Elimination itérative



C4D2 TRAINING

Valeurs extrêmes et mesures de la pauvreté

Cowell et Victoria-Feser (1996b)



ELSEVIER

European Economic Review 40 (1996) 1761–1771

EUROPEAN ECONOMIC REVIEW

Poverty measurement with contaminated data: A robust approach

Frank A. Cowell *, Maria-Pia Victoria-Feser

Distributional Analysis Research Programme, STICERD, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK

Received 15 November 1994; revised 15 August 1995

Abstract

We examine the sensitivity of poverty indices to data contamination using the concept of the influence function, and demonstrate that an important commonly used subclass of poverty measures will be robust under data contamination. This is illustrated using simulations. In this respect poverty and inequality indices have fundamentally different robustness properties. We investigate both the case where the poverty line is exogenously fixed and where it must be estimated from the data.

JEL classification: C13; D63

Keywords: Poverty; Inequality; Robustness; Influence function

- Explique pourquoi les valeurs extrêmes sont rarement de sérieuses menaces pour la plupart des mesures de la pauvreté.
- Les mesures de la pauvreté ne sont pas sensibles aux valeurs (réelles ou contaminées) des revenus des riches

Récapitulatif

- La réponse à la question de savoir si les valeurs extrêmes ont un intérêt particulier dépend de la statistique d'intérêt
- Inégalité: Aussi bien la théorie (fonction d'influence non limitée) que la pratique (troncature progressive) suggèrent toutes les deux qu'elles comptent (énormément). Ne pas tenir compte de cette question met en danger les comparaisons d'inégalités.
- Pauvreté: pas tellement

Comment détecter les valeurs extrêmes?

Inspection visuelle

(Deaton et Tarozzi 2005)

- “Nos procédures sont en partie graphiques et en partie automatiques. Pour chaque élément, nous traçons des histogrammes et des graphiques à sens unique des logarithmes des valeurs unitaires, en utilisant chacun d'entre eux pour détecter la présence de valeurs aberrantes brutes en vue d'investigations ultérieures. [...] [La méthode automatique] ne supprime pas la nécessité de l'inspection graphique”

Inspection visuelle

Malawi IHS3, Dépenses relatives aux tubercules de manioc

Malawi - Integrated Household Panel Survey 2010-2013 (Short-Term Panel, 204 EAs)

Reference ID	MWI_2010-2013_IHPS_v01_M
Year	2010 - 2013
Country	Malawi
Producer(s)	National Statistical Office - Government of Malawi
Sponsor(s)	Government of Malawi - GovMWI - Funded the study The World Bank - WB - Funded the study Millennium Challenge Corporation - MCC - Funded the study Irish Aid - IA - Funded the study German Development Corporation - GTZ - Funded the stud
Collection(s)	Living Standards Measurement Study (LSMS) Central
Metadata	Documentation in PDF

Created on	Apr 21, 2015
Last modified	Dec 13, 2017

Related Materials | Study Description | Data Dictionar | Get Microdata

Inspection visuelle

Malawi IHS3, Dépenses relatives aux tubercules de manioc

MODULE G: FOOD CONSUMPTION OVER PAST ONE WEEK

DATA ENTRY LINE NUMBER	Over the past one week (7 days), did you or others in your household consume any [. . .]? INCLUDE FOOD BOTH EATEN COMMUNALLY IN THE HOUSEHOLD AND THAT EATEN SEPARATELY BY INDIVIDUAL HOUSEHOLD MEMBERS.	G01 YES. . 1 NO. . . 2>> NEXT ITEM	G02 ITEM CODE	G05 How much did you spend? MK
19	**Roots, Tubers, and Plantains**			
20	Cassava tubers		201	

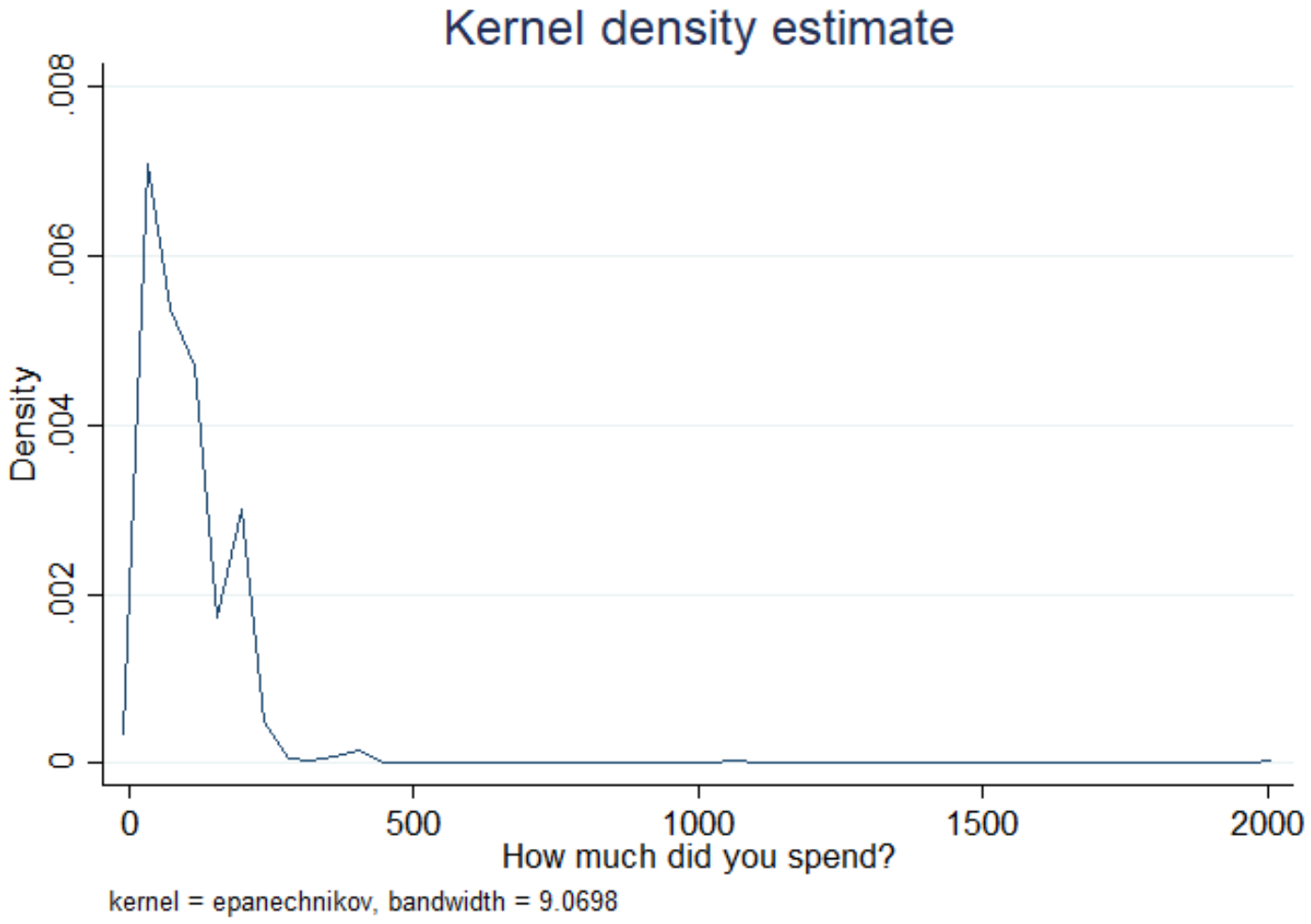
Inspection visuelle

Exemple 1: Regard sur les statistiques descriptives

```
. sum hh_g05 if hh_g02==201,d

                    How much did you spend?
-------------------------------------------------------------
      Percentiles      Smallest
 1%            5              0
 5%           20              0
10%           20              0       Obs                 673
25%           50              0       Sum of Wgt.         673

50%           75                      Mean           94.95097
                        Largest       Std. Dev.      106.2379
75%          100            400
90%          200            400       Variance        11286.5
95%          220           1050       Skewness        10.0151
99%          350           2000       Kurtosis       164.7054
```

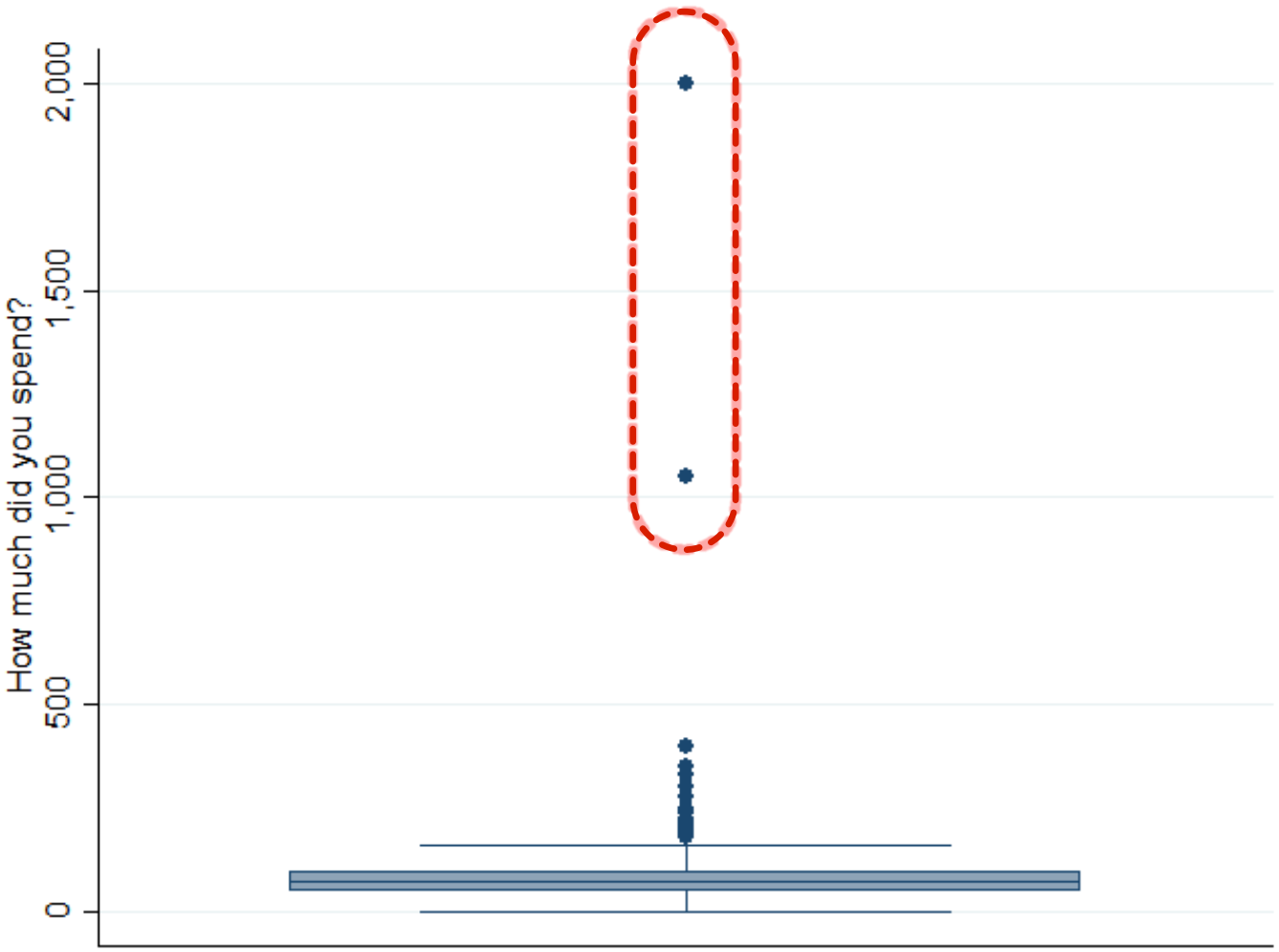
Inspection visuelle

Exemple 2: Graphique de la distribution des données



Inspection visuelle

Exemple 3: Utilisation d'outils graphiques diagnostiques, ex: la boite-à-moustache



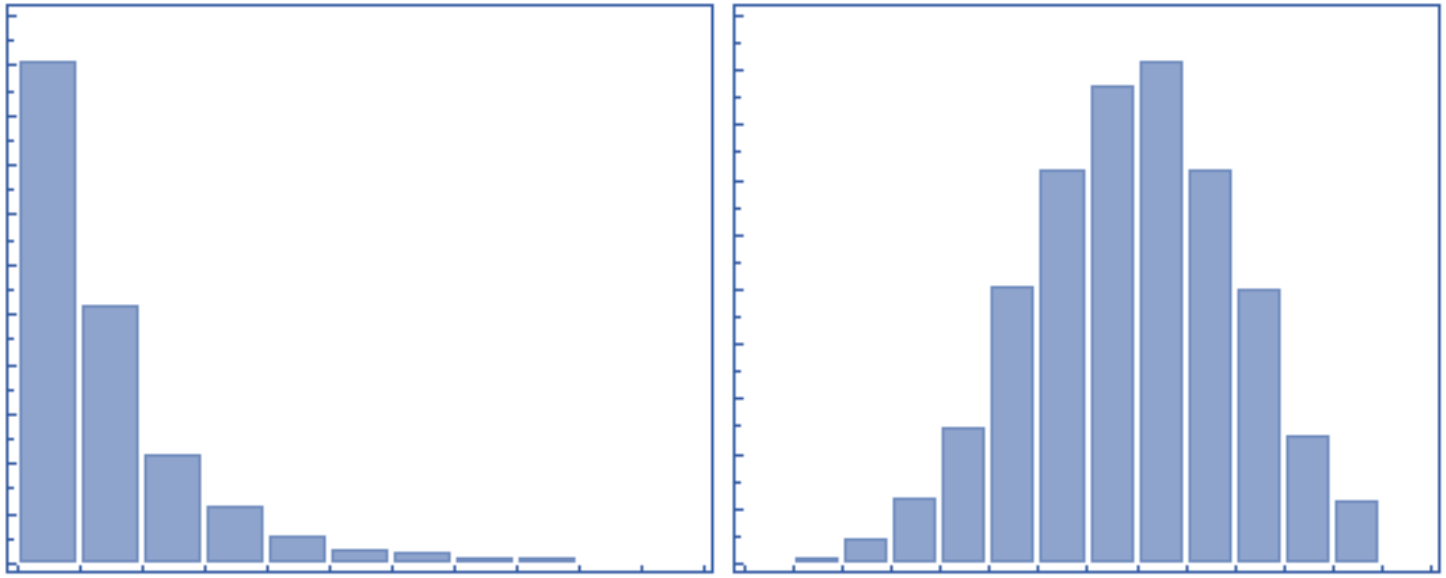
Méthodes statistiques

- La littérature est riche sur les méthodes d'identification des valeurs extrêmes; En pratique, la plupart des méthodes utilisées dans les travaux empiriques dépendent de la distribution sous-jacente des données.

- L'idée est simple:
  - transformer la variable pour induire la normalité
  - fixer des seuils pour identifier les valeurs extrêmes

Transformer la variable pour induire la normalité

- La transformation la plus simple consiste à prendre le logarithme de la variable d'intérêt
- Le logarithme "comprime" davantage les grandes valeurs, de sorte que les distributions aplaties deviennent plus symétriques et plus proches d'une distribution normale.



Définir un seuil

- Il faut définir un seuil pour décider quelle observation peut être considérée comme 'trop extrême' (outlier ou non?)

- Règles des seuils communs : une observation est considérée comme une valeur extrême si elle est éloignée de plus de 2,5, 3, 3,5 écarts types de la moyenne de la distribution

- En formule: x est une valeur extrême si

$$x > \bar{x} + z_\alpha \, s$$

  où z_α est égal à 2,5.

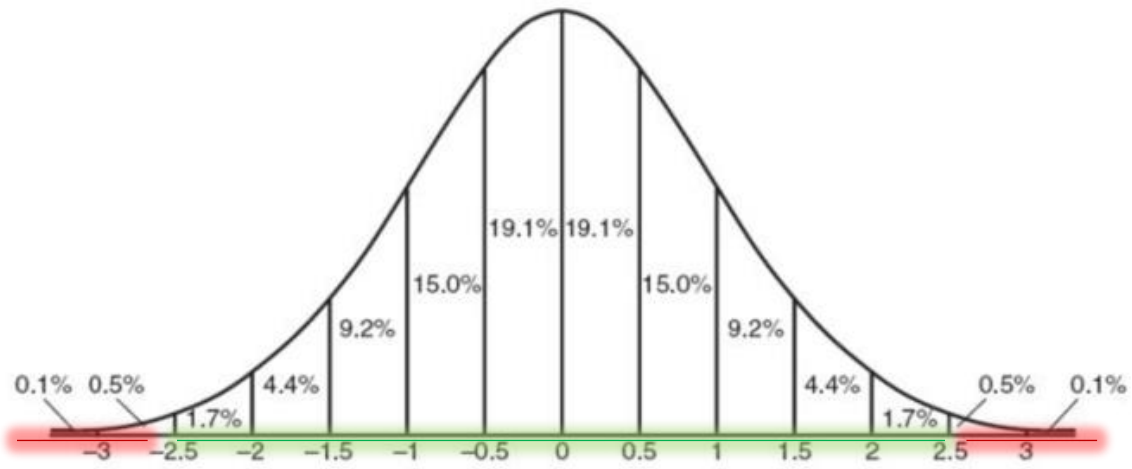
- On peut réécrire le même critère comme suit

$$\frac{x - \bar{x}}{s} > z_\alpha$$

  où le terme de gauche est appelé z-score (une variable avec une moyenne = 0 et une variance = 1)

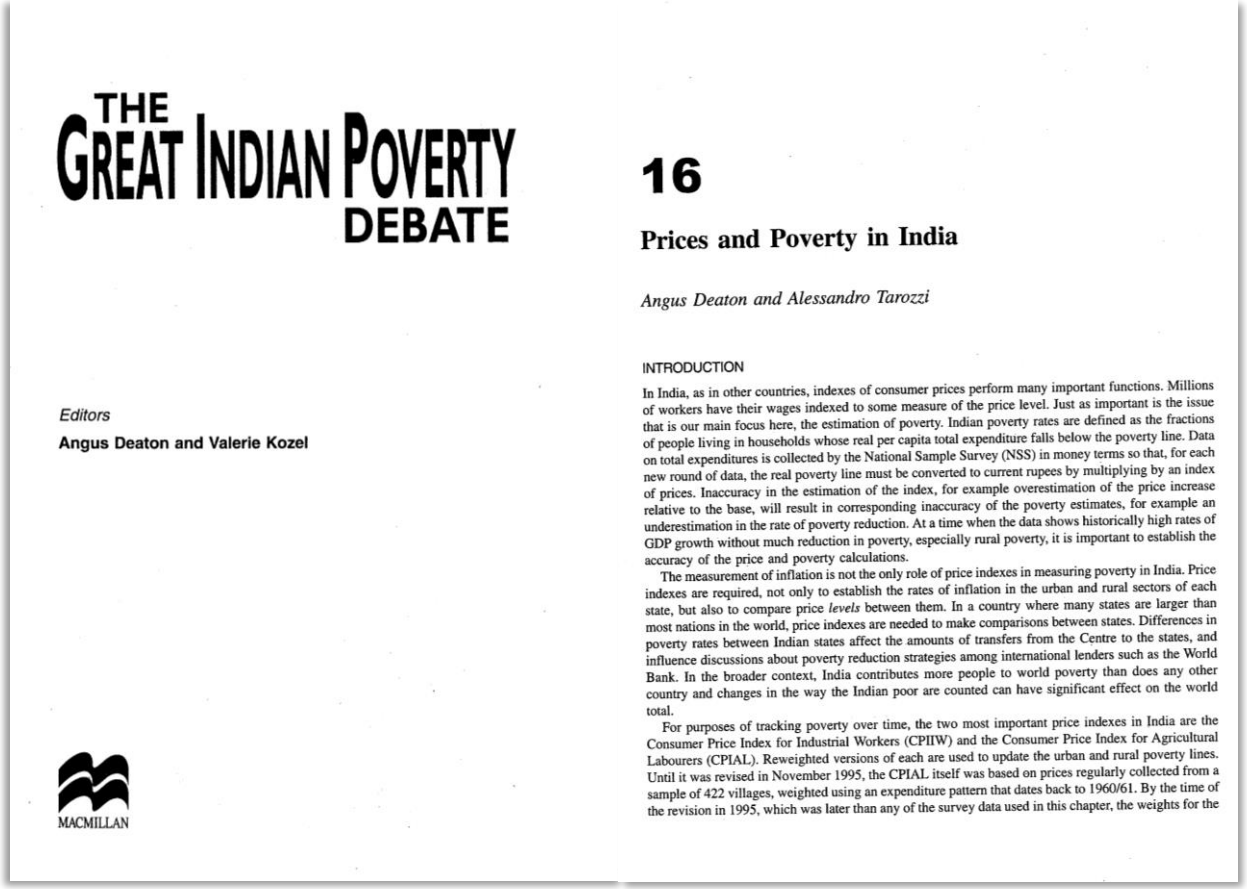
Pourquoi 2,5, 3, 3,5, ou une autre valeur?

- Sous l'hypothèse de normalité :



$z_\alpha = 2{,}5$ implique les valeurs extrêmes qui sont dans la region où $\alpha = 0{,}5$ pourcent des observations se situent normalement.

Deaton et Tarozzi (2005)



THE GREAT INDIAN POVERTY DEBATE

Editors
Angus Deaton and Valerie Kozel

MACMILLAN

16

Prices and Poverty in India

Angus Deaton and Alessandro Tarozzi

INTRODUCTION

In India, as in other countries, indexes of consumer prices perform many important functions. Millions of workers have their wages indexed to some measure of the price level. Just as important is the issue that is our main focus here, the estimation of poverty. Indian poverty rates are defined as the fractions of people living in households whose real per capita total expenditure falls below the poverty line. Data on total expenditures is collected by the National Sample Survey (NSS) in money terms so that, for each new round of data, the real poverty line must be converted to current rupees by multiplying by an index of prices. Inaccuracy in the estimation of the index, for example overestimation of the price increase relative to the base, will result in corresponding inaccuracy of the poverty estimates, for example an underestimation in the rate of poverty reduction. At a time when the data shows historically high rates of GDP growth without much reduction in poverty, especially rural poverty, it is important to establish the accuracy of the price and poverty calculations.

The measurement of inflation is not the only role of price indexes in measuring poverty in India. Price indexes are required, not only to establish the rates of inflation in the urban and rural sectors of each state, but also to compare price *levels* between them. In a country where many states are larger than most nations in the world, price indexes are needed to make comparisons between states. Differences in poverty rates between Indian states affect the amounts of transfers from the Centre to the states, and influence discussions about poverty reduction strategies among international lenders such as the World Bank. In the broader context, India contributes more people to world poverty than does any other country and changes in the way the Indian poor are counted can have significant effect on the world total.

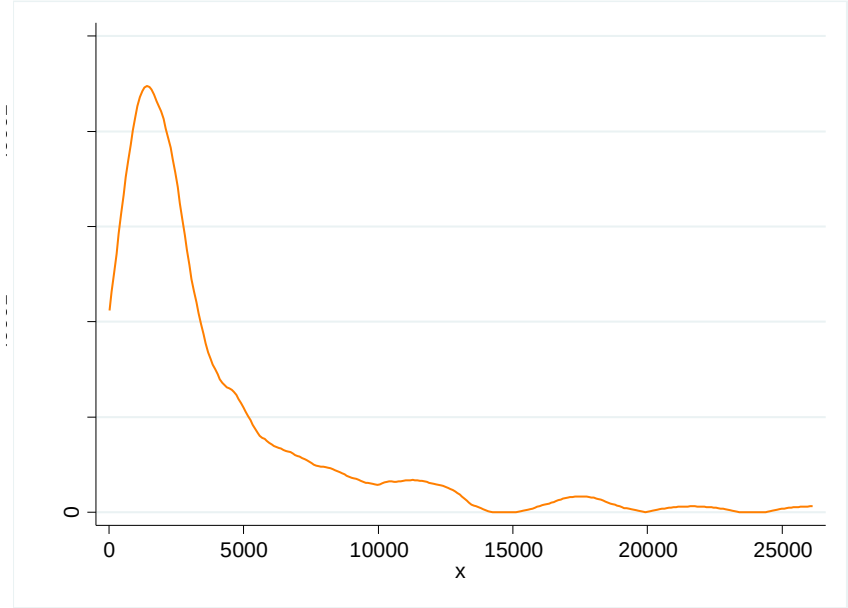
For purposes of tracking poverty over time, the two most important price indexes in India are the Consumer Price Index for Industrial Workers (CPIIW) and the Consumer Price Index for Agricultural Labourers (CPIAL). Reweighted versions of each are used to update the urban and rural poverty lines. Until it was revised in November 1995, the CPIAL itself was based on prices regularly collected from a sample of 422 villages, weighted using an expenditure pattern that dates back to 1960/61. By the time of the revision in 1995, which was later than any of the survey data used in this chapter, the weights for the

Dans le cas de l'Inde, D&T (2005) ont relevé des prix extrêmes dont les logarithmes dépassaient la moyenne des logarithmes de plus de 2,5 écarts types :

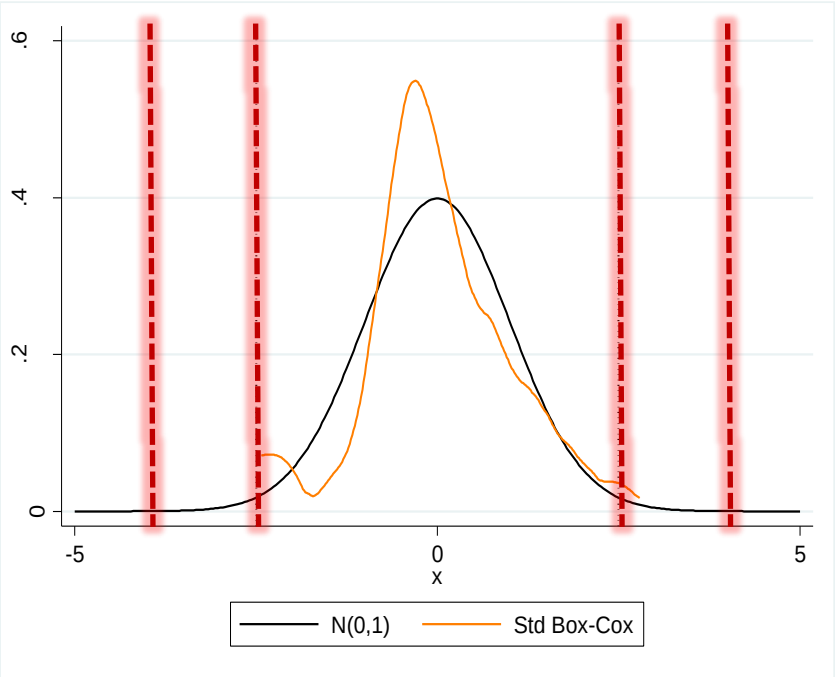
$$\frac{ln(x) - E[ln(x)]}{sd[ln(x)]} > 2.5$$

Transformation et seuils

Données brutes non transformées



Données transformées



Deux questions

1) Quelle est la qualité d'une telle approche?

2) Que faire après avoir repéré les valeurs extrêmes?

1) Quelle est la qualité d'une telle approche?

- La log-transformation est très simple – ex: Comment gérer les valeurs négatives?
- Non recommandé lorsque la distribution des logs ne peut être considérée comme normale
- Pourquoi devrions-nous fixer le seuil en utilisant la moyenne et l'écart-type, qui sont sensibles aux valeurs extrêmes, si c'est exactement ce qui nous préoccupe ?

$$\frac{ln(x) - E[ln(x)]}{sd[ln(x)]} > 2{,}5$$

- On peut mieux faire

Une stratégie populaire

Robustesse

- Bien qu'il n'y ait pas d'accord sur la meilleure méthode, une solution commune consiste à utiliser de solides mesures d'échelle et de localisation pour fixer le seuil de signalement des valeurs aberrantes

- l'idée est de remplacer la moyenne de l'échantillon $\bar{x}$ par un estimateur robuste (par exemple la médiane), et l'écart-type s par un estimateur robuste. Une option populaire est l'Ecart Absolu Médian (EAM).

Robustesse du z-score

L'Ecart Absolu Médian (EAM)

$$z_h = \frac{x_h - \bar{x}}{s}$$

$$z_h = \left| \frac{x_h - med[x_h]}{MAD} \right|$$

$$MAD = b \times med|x - med[x]|$$

b = 1,4826

<u>si</u> la distribution est Gaussienne

Nous pouvons mieux faire

Rousseeuw et Croux (1993, JASA)

Alternatives to the Median Absolute Deviation

Peter J. Rousseeuw and Christophe Croux*

In robust estimation one frequently needs an initial or auxiliary estimate of scale. For this one usually takes the median absolute deviation $\mathrm{MAD}_n = 1.4826\ \mathrm{med}_i \{ | x_i - \mathrm{med}_j x_j | \}$, because it has a simple explicit formula, needs little computation time, and is very robust as witnessed by its bounded influence function and its 50% breakdown point. But there is still room for improvement in two areas: the fact that MAD_n is aimed at symmetric distributions and its low (37%) Gaussian efficiency. In this article we set out to construct explicit and 50% breakdown scale estimators that are more efficient. We consider the estimator $S_n = 1.1926\ \mathrm{med}_i \{ \mathrm{med}_j | x_i - x_j | \}$ and the estimator Q_n given by the .25 quantile of the distances $\{ | x_i - x_j | ; i < j \}$. Note that S_n and Q_n do not need any location estimate. Both S_n and Q_n can be computed using $O(n \log n)$ time and $O(n)$ storage. The Gaussian efficiency of S_n is 58%, whereas Q_n attains 82%. We study S_n and Q_n by means of their influence functions, their bias curves (for implosion as well as explosion), and their finite-sample performance. Their behavior is also compared at non-Gaussian models, including the negative exponential model where S_n has a lower gross-error sensitivity than the MAD.

KEY WORDS: Bias curve; Breakdown point; Influence function; Robustness; Scale estimation.

Rousseeuw et Croux (1993)

- Rousseeuw et Croux (1993) proposent de remplacer l'EAM avec un estimateur différent :

$$S = c \times med\{\mathrm{med}_j |x_j - x_i|\}$$

- Pour chaque j, on calcule la médiane de $|x_j - x_i|$ $(i = 1, \ldots, n)$. On obtient ainsi les chiffres de n, dont la médiane donne notre estimation finale.

$$z_h = \left| \frac{x_h - med[x_h]}{S} \right|$$

c = 1,1926 du modèle gaussien.

Récapitulatif

- On peut mieux faire que “prendre le log et travailler”:

  1. Utiliser d’autres transformations que le logarithme pour normaliser la variable d’intérêt donne souvent de meilleurs résultats

  2. Tester la robustesse du z-score est une meilleure pratique.

- Belotti et al. (2020): `outdetect.ado` aide à mettre en œuvre ces éléments

Prendre le log et travailler vs. z-scores robustes

Pays	Année	Valeurs extrêmes (%)					
Seuil = 3		log-transformation			z-scores robustes		
		Global	Gauche	Droite	Global	Gauche	Droite
		(1)	(2)	(3)	(4)	(5)	(6)
Malawi	2017	0,75	0,14	0,61	0,30	0,22	0,08
Nigeria	2012	1,35	0,11	1,24	0,72	0,32	0,40
India	2012	1,39	0,03	1,36	0,62	0,13	0,49
Pakistan	2014	1,58	0,02	1,56	0,39	0,21	0,18
Guatemala	2014	1,14	0,06	1,08	0,61	0,15	0,46
Peru	2015	0,36	0,09	0,27	0,28	0,16	0,12
Armenia	2013	0,91	0,08	0,83	0,68	0,17	0,51
Georgia	2015	0,75	0,25	0,50	0,73	0,32	0,41

Impact des valeurs extrêmes sur l'indice de Gini

Pays	Année	**Indice de Gini**		
Seuil = 3				
		Brut	Découpage (log)	Découpage (Meil.)
		(7)	(8)	(9)
Malawi	2017	40.6	34.8	36.6
Nigeria	2012	43.7	36.7	38.2
India	2012	39.5	36.2	37.6
Pakistan	2014	32.9	30.0	32.3
Guatemala	2014	37.2	34.7	35.9
Peru	2015	36.8	36.0	36.3
Armenia	2013	28.9	26.7	26.9
Georgia	2015	37.1	35.4	35.6

Comment gérer les valeurs extrêmes?

(en une diapositive)

Traitement des valeurs extrêmes

Trois principales méthodes sont utilisées pour gérer les valeurs extrêmes, hormis leur suppression des données:

1) Réduire la pondération des valeurs extrêmes (pondération de césure)

2) Changer les valeurs des valeurs extrêmes (Winsorisation, Césure, imputation – par exemple via la régression quantile)

3) Utiliser les techniques d'estimation robustes (M-estimation).

- Documentation, transparence et reproductibilité

Leçons apprises

- Les valeurs extrêmes peuvent être de véritables observations… Etre attentif aux données et documenter chaque étape du traitement des données
- Si la mesure de l'inégalité est l'objectif, les valeurs extrêmes sont le pire ennemi (Fonction d'influence non limitée)
- Détection des valeurs extrêmes:
  - Aller au-delà de la stratégie de "prendre le log et travailler". Elle marche bien seulement si l'on peut décrire les données avec une distribution gaussienne. Cependant, les distributions son en general étalées.
  - Dans le cas échéant, utiliser la "meilleure normalisation" disponible, et tester la robustesse du z-score.
- Traitement des valeurs extrêmes: Cela dépend. La regression quantile est un bon candidat.

Bibliographie

Lectures obligatoires

Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data. 3rd edition. J. Wiley & Sons (Chapter 1 & 2)

Lectures recommandées

Alvarez, E., Garcıa-Fernández, R. M., Blanco-Encomienda, F. J., & Munoz, J. F. (2014). The effect of outliers on the economic and social survey on income and living conditions. World Acad. Sci., Eng. Technol., Int. J. Soc., Behav., Educ., Econ., Bus. Ind. Eng, 8, 3276-3280.

Belotti, F., & Vecchi, G. (2019). Take the Log and Run: Outliers and Welfare Measurement, mimeo.

Cowell, F. A., & Flachaire, E. (2007). Income distribution and inequality measurement: The problem of extreme values. Journal of Econometrics, 141(2), 1044-1072.

Cowell, F., & Victoria-Feser, M. (1996). Robustness Properties of Inequality Measures. Econometrica, 64(1), 77-101

Cowell, F. A., & Victoria-Feser, M. P. (1996). Poverty measurement with contaminated data: A robust approach. European Economic Review, 40(9), 1761-1771.

Deaton, A., & Tarozzi, A. (2005). “Prices and Poverty in India.” The Great Indian Poverty Debate. New Delhi : MacMillan.

Dupriez, O. (2007). Building a household consumption database for the calculation of poverty PPPs. Technical note. Available at: http://go. worldbank. org/4YG7I5RGT0.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11(1), 1-21.

Hlasny, V., & Verme, P. (2018). Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data. Econometrics, 6(2), 30.

Mancini, G., & Vecchi, G. (2019). On the Construction of a Welfare Indicator for Inequality and Poverty Analysis, mimeo.

OECD (2013). OECD Guidelines for Micro Statistics on Household Wealth

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. Journal of the American Statistical association, 88(424), 1273-1283.

Merci pour votre attention

Exercices

Exercice 1 – Interroger la littérature

World Academy of Science, Engineering and Technology
International Journal of Economics and Management Engineering
Vol:8, No:10, 2014

The Effect of Outliers on the Economic and Social Survey on Income and Living Conditions

Encarnación Álvarez, Rosa M. García-Fernández, Francisco J. Blanco-Encomienda, Juan F. Muñoz

Abstract—The European Union Survey on Income and Living Conditions (EU-SILC) is a popular survey which provides information on income, poverty, social exclusion and living conditions of households and individuals in the European Union. The EU-SILC contains variables which may contain outliers. The presence of outliers can have an impact on the measures and indicators used by the EU-SILC. In this paper, we used data sets from various countries to analyze the presence of outliers. In addition, we obtain some indicators after removing these outliers, and a comparison between both situations can be observed. Finally, some conclusions are obtained.

Keywords—Headcount index, poverty line, risk of poverty, skewness coefficient.

as poor if his/her income is less than the official poverty line. Poverty line and risk of poverty are common concepts used by poverty studies. Relevant references that define and describe such concepts related to poverty are [2], [3], [8], [12], [13], [16], [19], [22] and [23].

It is quite common that surveys used by poverty studies contain various variables, and some of them can have a strong relationship with respect to the variable of interest. For example, this situation can be observed by the EU-SILC. The variable of interest in this survey is the equivalised net income, since this variable is used for the problem of estimating the poverty line and the poverty risk. The EU-SILC

Résumer les principals conclusions du papier: Les valeurs extrêmes ont-ells une importance? Pourquoi ou Pourquoi pas?

Exercice 2 – Fais-le toi-même….

Anglais

1) Générer une distribution de richesse log-normale

2) Estimer l'indice de Gini

3) Contaminer la distribution avec quelques valeurs extrêmes

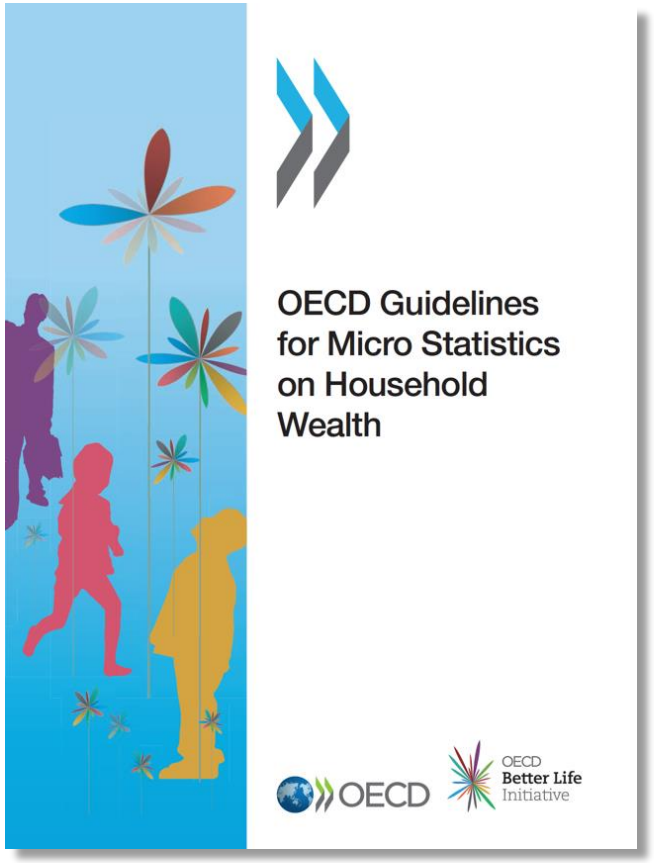
4) Ré-estimer l'indice de Gini

Stata/R/SPSS/Excel/…

```
clear

set obs 5000
set seed 198607
gen n  = rnormal(0,1)
gen ln = exp(n)

* simulate order of magnitude mistake:
* take 100 obs around the median
* of the distribution and multiply
* them by 100

sort ln

gen cont100 = 1
replace cont100 = 100 in 2480/2520
gen ln_cont100 = ln*cont100
```

Exercice 3 – Mesures d'inégalité

- Commenter le tableau 7.3 du rapport de l'OECD (2013) p.172 (Voir slide suivante).
- Que pouvez-vous dire de la sensibilité des des estimateurs au traitement des valeurs extremes?



Exercice 3 – Mesures d'inégalité

OECD (2013)

Table 7.3. **Effect of the treatment of outliers on summary measures of wealth inequality in the United States, 2007**

	Raw	Shave top and bottom 1%	Shave top 1% and bottom 0.5%
Mean	556 846	378 215	559 361
Median	120 780	120 780	123 800
Gini	0.82	0.74	0.81
$\frac{1}{2}CV^2$	18.1	2.4	14.6
P90/P10	30 000	3 369	3 061
P75/P25	26.3	24.5	24.3
P90/P50	7.6	7.0	7.4
n	4 418	3 698	4 359

Source: 2007 Survey of Consumer Finances.