# Contents

# Preface

[The book introduces robust statistics in Stata from an applied perspective. We review existing commands and present a variety of new tools, give advice on how to choose among the different estimators and illustrate how they can be applied in practice. After a general introduction the book first discusses robust estimation of univariate location and scale and, along the way, briefly introduces the basic concepts of robust statistics. The book then moves on to simple and multiple robust regression and models for qualitative dependent variables, each time reviewing (briefly) the theory, presenting the algorithms, commands, and implementation details, and providing applied examples. Furthermore, we discuss multivariate identification of outliers and present robust versions of factor models] ...

# Notation and typography

### Stata code, datasets, programs, and references to manuals

In this book we assume that you are somewhat familiar with Stata, that you know how to input data and to use previously created datasets, create new variables, run regressions, and the like. Generally, we use the `typewriter font` to refer to Stata commands, syntax, and variables. A "dot" prompt followed by a command indicates that you can type verbatim what is displayed after the dot (in context) to replicate the results in the book.

The data we use in this book are freely available for you to download, using a net-aware Stata, from the Stata Press website, `http://www.stata-press.com`. In fact, when we introduce new datasets, we merely load them into Stata the same way that you would. For example,

```
. use http://www.stata-press.com/data/!!!/football.dta, clear
```

In addition, the Stata packages presented in this book may be obtained by typing

```
. ssc install robstat
  (output omitted)
. ssc install robreg
  (output omitted)
. ssc install robmv
  (output omitted)
```

Also say what other packages need to be installed (if any), e.g. `moremata`, I think.

Throughout the book, we often refer to the Stata manuals using [R], [P], etc. For example, [R] **regress** refers to the *Stata Reference Manual* entry for `regress`, and [P] **matrix** refers to the entry for `matrix` in the *Stata Programming Manual*.

### Mathematical and statistical symbols

We also assume that you have basic knowledge of mathematics and statistics, although we tried to keep the exposition as simple and non-technical as possible. Below is a list of some mathematical and statistical symbols that we will frequently use in the book.

| | |
|---|---|
| $X, Y, Z, \ldots$ | random variables |
| $x_i, y_i, z_i, \ldots$ | realizations (observations) of random variables |

| | |
|---|---|
| $n$ | number of observations |
| $x_{(i)}$ | $i$th order statistic of $x_1, \ldots, x_n$ ($i$th observation in the list of observations sorted in ascending order) |
| $F(x)$ | cumulative distribution function of a random variable; ... |
| $f(x)$ | density ... |
| $F'(x)$ | first derivative of function $F(x)$, that is $F'(x) = dF(x)/dx = f(x)$; we use $'$ for both the first derivative of a function and the transposition of a vector or matrix |
| $\mathcal{N}(\mu, \sigma)$ | normal distribution with mean $\mu$ and standard deviation $\sigma$ |
| $\mathcal{N}(0, 1)$ | standard normal distribution |
| $\lvert x \rvert$ | absolute value of $x$ |
| $\lVert \mathbf{x} \rVert$ | Euclidean norm of vector $\mathbf{x} = (x_1, \ldots, x_p)^t$, that is, $\lVert \mathbf{x} \rVert = \sqrt{x_1^2 + \cdots + x_p^2}$ |
| $\lceil x \rceil$ | smallest integer greater or equal to $x$ |
| $\lfloor x \rfloor$ | largest integer smaller or equal to $x$ |
| $\mathbf{x}^t$, $\mathbf{X}^t$ | transposition of a vector or a matrix |
| i.i.d. | independent and identically distributed |
| $\lim_{x \to y} g(x)$ | limiting value of function $g(x)$ as $x$ approaches $y$ |
| $\sup_x g(x)$ | supremum (least upper bound) of function $g(x)$ with respect to argument $x$ |
| $\text{sign}(x)$ | the sign of $x$; to be precise, $\text{sign}(x) = -1$ if $x < 0$, $\text{sign}(x) = +1$ if $x > 0$, $\text{sign}(x) = 0$ if $x = 0$ |
| $X \sim F$ | random variable $X$ is distributed as $F$ |
| $X \approx F$ | random variable $X$ is approximately distributed as $F$ |
| ... | ... |

# Part I

# Introduction

# Chapter 1

# Introduction

## 1.1 Motivation

Linear least-squares (LS) regression is, without doubt, the workhorse of data analysis in social sciences, economics and related fields. The reasons for the popularity of LS regression are obvious. The procedure convinces by its formal and practical simplicity. LS regression is easy to implement from a technical point of view and its results, the estimated regression coefficients, are easy to interpret. Furthermore, LS regression is easy to teach because its math is relatively simple and is didactically convenient because LS solutions for small datasets can easily be computed manually for purpose of exercise and understanding. From a statistical point of view, LS regression is favorable because it can be shown that under the assumption of homoscedastic (i.e., equal-variance) and normally distributed errors the LS estimator is the best (i.e., most efficient) unbiased estimator (BUE) for the coefficients of a linear regression model. That is, among all possible unbiased estimators, the LS estimator has the smallest sampling variance under these conditions.[1] Also under relaxed assumption, such as non-normal or heteroscedastic (i.e., non-equal-variance) errors, the LS estimator is consistent and has, in many cases, good efficiency properties.[2] For example, in case of homoscedastic non-normal errors, the LS estimator is the beast linear unbiased estimator (BLUE), that is, has the smallest sampling variance among all "linear" unbiased estimators.[3]

The outstanding usefulness of LS regression should not be challenged here. It is im-

[1] Noting the equivalence between the LS estimator and the arithmetic mean, the BUE property of the LS estimator is not much of a surprise given the fact that Carl Friedrich Gauß derived the normal distribution as a justification for the arithmetic mean. That is, the normal distribution is *defined* as the distribution under which the LS procedure leads to the best unbiased estimator for the expected value (for historical background see Huber, 1972).

[2] Although in the later case, the ordinary LS estimate of the sampling variance is biased and needs to be adjusted by applying heteroscedasticity-robust variance estimation; see White, 1980.

[3] The term "linear" does not refer to the fact that the coefficients of a linear regression model are to be estimated. An estimator is said to be *linear* if it is a linear function of the observations $Y_1, \ldots, Y_n$ of the dependent variable of the regression model. More precisely, $\widehat{\boldsymbol{\beta}}$ is a linear estimator of the regression parameters vector $\boldsymbol{\beta} \in \mathbb{R}^p$ if there exists a matrix $\mathbf{A} \in \mathbb{R}^{p \times n}$ such that $\widehat{\boldsymbol{\beta}} = \mathbf{AY}$ with $\mathbf{Y} = (Y_1, \ldots, Y_n)^t$.

portant, however, to realize that LS regression may not always be the best—or at least not the only—choice for analyzing a given dataset. The restrictiveness of the conditions under which the LS estimator is deemed best—homoscedasticity and normality of errors—implies that situations are possible in which alternative estimators can be valuable. For example, as mentioned above, if the errors are homoscedastic but non-normal, the LS estimator may be the best linear unbiased estimator, but this also means that there can be non-linear estimators that, depending on the nature of the deviation from normality, substantially outperform the LS estimator in terms of efficiency.[4] In particular, in case of distributions with heavy tails, that is, if extreme values are more frequent than in a normal distribution (an example being the $t$-distribution with few degrees of freedom), the efficiency of the LS estimator can quickly become poor. Furthermore, the LS estimator may yield misleading results if the data are "contaminated" by erroneous observations or, more generally, by a secondary data-generating process.

### Efficiency under alternative error distributions

Assume, for now, that the data are not contaminated and, more or less, follow a uniform data-generating process that can be described by a linear regression model. Why, under such a condition, can a low efficiency of the LS estimator be a problem? Although the LS estimator is unbiased, more efficient estimators would be preferable because the precision of an estimator has a direct effect on the value of the results. For example, the power of a significance test and, therefore, the potential of the test to find an existing relation, decisively depends on the efficiency of the employed estimator.

In the context of error distributions with heavy tails the efficiency argument can also be motivated as follows. Although the LS estimator is unbiased on average, there is a good chance for a single sample—and in practice often only one sample is available—to contain extreme values that bias the regression results in one or the other direction. Robust regression methods that are less sensitive to such outliers will typically provide more valid results in such situations, being closer to the true value of the parameter to be estimated.

Figure 1.1 shows two examples of data sets that have been generated according to model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

with $\beta_0 = \beta_1 = 0$ (that is, the "true" regression function is a horizontal line at $y = 0$) and $\epsilon$ following a $t$ (Student) distribution with two degrees of freedom, that is $\epsilon \sim t_2$. Included as lines are the estimated regression fits using LS estimation, as well as two robust estimators (an M estimator and an MM estimator). As is evident, the LS solution is affected by the outliers and suggests a positive relation between $X$ and $Y$ in the two examples, whereas the two robust estimator are relatively stable. Robust methods, so to say, contain a safeguard against extreme data constellations that can occur at random due to sampling or a stochastic data-generating process. As a diagnostic by-product, robust methods inform about whether given data are characterized by an anomalous

[4] The limitation to linear estimators is not much less restrictive than the limitation to normal errors.

constellation or not, because only in the former case the results from LS and the robust methods will substantially differ.

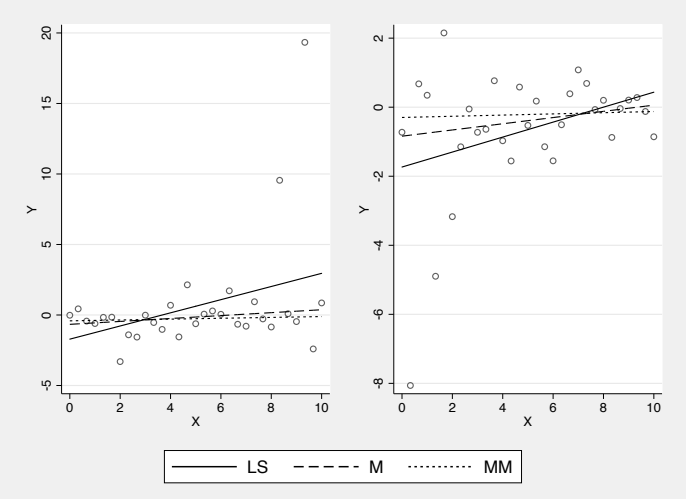


Figure 1.1: Example scatter plots with outliers and different regression fits

**Bias due to data contamination**

Now assume that the data are "contaminated", that is, that the majority of data points follows a well-defined model, but that there are also some observations that come from a different distribution. For example, while collecting the data, coding errors could have occurred for some of the observations. In a study by Jasso (1985) on the relation between marital duration and coital frequency there were four observations with a value of 88 for the monthly coital frequency. Although such values would not be impossible (as argued by Jasso, 1985), the observations were highly suspicious as no other values of comparable magnitude existed in the data. As argued by Kahn and Udry (1986), the four observations probably were miscoded missing values, whose designated value was 99. The problem with such miscoded observations is that they can have strong effects on the results provided by a LS regression. That is, regression results and the substantive conclusions drawn from them may differ depending on whether the miscoded observations are kept in the data or not. It seems important to use methods for data analysis that are able to identify such problems because, in the words of Anscombe (1973, 18), "[w]e are usually happier about asserting a regression relation if the relation is still apparent after a few observations (any ones) have been deleted—that is, we are happier if the regression relation seems to permeate all the observations and does not derive largely from one or two."

Conceptually, contamination can be understood as a situation in which the observed data are the result of a mixture of two or more data-generating processes. In the case

of coding errors there may be a main process of substantive interest (e.g., the relation between marital duration and coital frequency), as well as a secondary process (data miscoding by interviewers) that leads to observations that follow a different distribution and have a different interpretation. LS regression will not be able to distinguish the two processes and its results will be valid for neither one of the processes. If, however, the data are dominated by one of the processes (that is, if one of the processes is responsible for the bulk of the data) and the two processes do lead to distinguishable data structures, statistical procedures to identify the main process are possible. This is where robust regression comes in. One of the goals of robust regression techniques is to provide estimates that are resistant against partial contamination of the data. Robust methods are supposed to correctly identify the primary relation in the data even if, for example, parts of the data are glaringly erroneous.

An illustrative example comes from astronomy. Figure 1.2 shows the Hertzsprung-Russell diagram of star cluster CYG OB1 (see Rousseeuw and Leroy 1987, 27). Displayed is the logarithm of the light intensity of the stars against the logarithm of their effective surface temperature (using a reversed axis). Furthermore, the graph shows as lines the results of three different regression estimators, the LS estimator (solid line), a low breakdown point M estimator (dashed line), and a high breakdown point MM estimator (dotted line). The results from the LS estimator and the low breakdown point M estimator are almost identical. They are strongly influenced by the group of four stars in the upper right corner of the diagram. In contrast, the high breakdown point MM estimator completely ignores the four outliers and adequately captures the trend in the main part of the data. Hence, at least one of the two employed robust estimators successfully identified the main process (due to the estimator's high breakdown point; see below).

Again, from a diagnostic perspective, the interesting cases are the ones in which LS regression and robust estimators lead to differing results. Substantial differences between robust regression and the LS estimator indicate that the data cannot be fully descried by a uniform model and that a part of the observations stands in stark contrast to the main trend in the data. With the help of the residuals from robust regression, the atypical observations can be identified and, for example, be subjected to a separate analysis. In this way, robust regression can contribute to a better understanding of the data and, potentially, give way to new insights and new hypotheses. In fact, according to Kruskal (1960, 1), the atypical observations may prove to be the most interesting part of the data: "An apparently wild (or otherwise anomalous) observation is a signal that says: 'Here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study.'" The four outliers in figure 1.2, by the way, are not errors. The explanation is that there are two different types of stars: main-sequence stars and giants. That is, conceptually, the observation stem from two different populations.

### Goals and use of robust regression

To summarize, we can state that robust regression estimators (1) should achieve good efficiency also in case of non-normal errors and (2) should be resistant against contam-

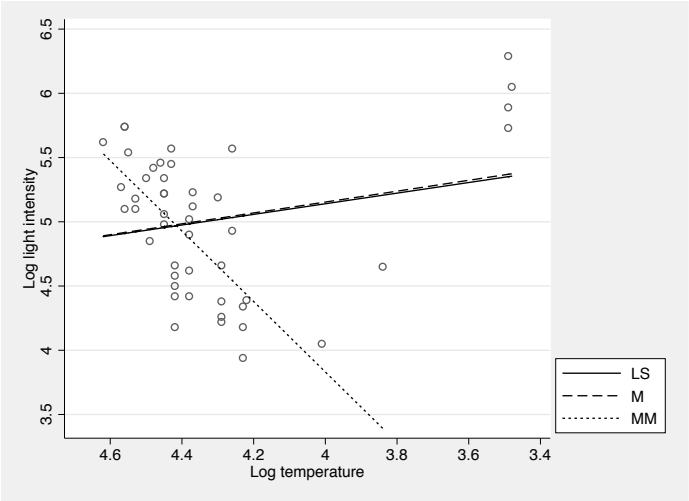


Figure 1.2: Hertzsprung-Russell diagram of the star cluster CYG OB1 including different regression fits (source: Rousseeuw and Leroy, 1987, 27)

ination of the data by outliers. The maximum proportion of contamination a robust estimator is able to absorb is called the *breakdown point*.

Both aspects can be formalized with the help of the viewpoint coined by Huber (1964) that observed data follow a mixture distribution

$$F_\varepsilon = (1-\varepsilon)F_{\boldsymbol{\theta}} + \varepsilon G$$

where $F_{\boldsymbol{\theta}}$ is the distribution of interest according to the supposed model, $G$ is an arbitrary alternative distribution, and $\varepsilon \in [0,1]$ determines the mixing proportion. For example, in line with the assumptions of classic linear regression, $F_{\boldsymbol{\theta}}$ could be a distribution according to the linear model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where $X$ has a given distribution and $\epsilon$ is an independent and identically normally distributed error term. The distribution of the observed data, however, is contaminated by observations from an unspecified alternative distribution $G$ and does not fully follow this model. The goal of robust regression now is to deliver reasonable results for $F_{\boldsymbol{\theta}}$ even if the model is somewhat misspecified, that is, if $\varepsilon > 0$. In the words of Heritier et al. (2009, 7), robust methods are "a set of statistical tools for correct estimation and inference about $F_{\boldsymbol{\theta}}$ when the data-generating process is $F_\varepsilon$, not only when $\varepsilon = 0$, as with classical methods, but also for relatively small $\varepsilon$ and *any* $G$. As a by-product, data not fitting $F_{\boldsymbol{\theta}}$ exactly can be easily identified, and the model can possibly be changed and refitted". In addition, to be of diagnostic value, robust estimators should be serious competitors of classic methods in case of $\varepsilon = 0$. In particular, robust estimators should

achieve good "gaussian efficiency", that is, they should achieve a high relative efficiency compared to LS estimation in the ideal case of normally distributed errors.[5]

Yet, robust regression should be seen as a complement and not so much as a substitute to LS regression. In our view, the main use of robust regression lies in its diagnostic potential. Classic regression techniques may lead to meaningful results in many situations, but a comparison to robust results is always advisable. Before drawing far-reaching conclusions based on classic methods one should evaluate whether the conclusions are "robust", that is, whether methods that rely on less restrictive assumptions and are less affected by outliers and atypical data constellations come to the same conclusions.

If classic procedures and robust regression lead to substantially diverging or even contradicting results, the robust results can provide an immediate contribution to a better understanding of the data. As a by-product of robust estimation, observations that do not fit the supposed model can easily be identified, offering clues about possible misspecification, the nature of outliers, and alternative data-generating processes. Compared to classic regression diagnostics for the identification of influential observations (see Belsley et al., 1980; Cook and Weisberg, 1982; Chatterjee and Hadi, 1988; Fox, 1991) robust regression methods have the advantage that they can also identify "masked" multiple outliers that would go undetected by classic diagnostics. However, robust techniques are no panacea and cannot, for example, fully replace diagnostic methods that are concerned with the identification of structural misspecification (such as omitted variable bias, wrong functional form, or missing interaction terms).

[Should there also be some text giving a brief historical account of the development of robust statistics and robust regression?]

## 1.2 What is covered in this book?

...

## 1.3 Robust statistics in Stata

- Summary of existing tools
- Brief presentation of our new packages; basic usage and syntax

[5] Note that the estimation of "robust standard errors" is not the primary concern of robust regression. The term "robust standard errors" refers to estimators for the sampling variances of the coefficient estimates that are consistent also if the assumption of identically distributed errors is violated (i.e., if the errors are heteroscedastic; see White, 1980). To prevent false conclusions with respect to confidence intervals and significance tests, it is always a good idea to consider "robust standard errors", be it with classic regression or with robust regression.

# Part II

# Robustness theory and basic robust statistics

# Chapter 2

# Basic concepts in estimation

An estimation problem in statistics may have many potential solutions. To separate useful estimation strategies from approaches that are less feasible, criteria have to be defined by which different estimators can be evaluated and compared. In this chapter we first review a number of basic criteria typically used in classic statistics. We then discuss additional criteria that are important in the context of robust statistics. Our discussion in this chapter is conceptual in nature; it is supposed to establish a theoretical basis for the specific robust estimators that are discussed in the subsequent chapters from an applied perspective.

## 2.1 Classical properties of estimators

The goal of statistical estimation is to obtain a reasonable value for the unknown *paramater* of a statistical model, based on data whose properties are assumed to be consistent with the suggested model. Let $\mathcal{X}^{(n)} = \{X_1, \ldots, X_n\}$ be a set of $n$ random variables $X_1, \ldots, X_n$ that have a joint probability distribution $P_{\boldsymbol{\theta}}^{(n)}$ depending on the unknown parameter $\boldsymbol{\theta}$. Note that, depending on context, $\boldsymbol{\theta}$ may be scalar, or it may be a vector of multiple parameters, that is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t$. For simplicity, we consider here that the observations $X_i$ are univariate. However, the concepts introduced hereafter may be (quite easily) generalized in the case of multivariate observations $\mathbf{X}_i$, that is, in the case $\mathbf{X}_i$ is a vector of random variables ($\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})^t$).

Since $\boldsymbol{\theta}$ is unknown, this actually leads to assume that the joint distribution of the random variables $X_i$, $i = 1, \ldots, n$, belongs to the (parametric) *statistical model* $\mathcal{P}^{(n)} = \{P_{\boldsymbol{\theta}}^{(n)} | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta}$ is the set of possible values of $\boldsymbol{\theta}$. The goal is to estimate $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ based on a realization of $\mathcal{X}^{(n)}$. In general, we will consider the case in which the statistical model $\mathcal{P}^{(n)}$ conforms to simple random sampling (SRS), that is, a situation in which the random variables $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.). In this situation, each $X_i$ independently follows a common probability distribution $P_{\boldsymbol{\theta}}$, which can be characterized by the distribution function $F_{\boldsymbol{\theta}}(\mathbf{x}) = \Pr(X_i \leq x)$ (or simply $F$ when there is no risk of confusion about the parameter we have to estimate).

An *estimator* of $\boldsymbol{\theta}$ can then be defined as follows: An estimator of the parameter $\boldsymbol{\theta}$ is any statistic $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ taking its value in $\boldsymbol{\Theta}$.

The value of $\widehat{\boldsymbol{\theta}}$ provided by a particular realization $x_1, \ldots, x_n$ of the random variables $X_1, \ldots, X_n$ is called an *estimate* of $\boldsymbol{\theta}$. Note that, for simplicity, we will often use the notation $x_i$ ($i \in \{1, \ldots, n\}$) to designate the $i$th random observation as well as a realization of it (i.e., a specific value); the context will always clearly indicate if we have to consider $x_i$ as a random variable or as a particular value.

The definition above contains no indication of the quality of an estimator; any statistic $\widehat{\boldsymbol{\theta}}$ that provides a value in $\boldsymbol{\Theta}$ is a valid estimator. To narrow down the set of estimators to estimators that can be considered useful we need quality criteria. Classic quality criteria are unbiasedness, efficiency, and consistency.

### 2.1.1 Unbiasedness

From a good estimator one may expect that, on average, it gives the "correct" answer. Let us denote by $E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)}))$ the expectation of statistic $\widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ when $\mathcal{X}^{(n)} \sim P_{\boldsymbol{\theta}}^{(n)}$. Think of $E_{\boldsymbol{\theta}}$ as the average value we would obtain for $\widehat{\boldsymbol{\theta}}$ from a large number of repeated realizations of $\mathcal{X}^{(n)}$, given that for each repetition $\mathcal{X}^{(n)}$ follows distribution $P_{\boldsymbol{\theta}}^{(n)}$ (as, for example, in repeated random sampling from the same population).

Unbiasedness can then be defined as follows: The estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ is called unbiased if

$$E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta} \qquad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and all } n.$$

That is, no matter the sample size $n$, estimator $\widehat{\boldsymbol{\theta}}$ will, on average across a large number of repeated samples, provide the correct value of $\boldsymbol{\theta}$ (given that our assumptions about the joint distribution of $\mathcal{X}^{(n)}$ are correct, such as, e.g., independent sampling of observations). The difference

$$B_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$$

is called the *bias* of estimator $\widehat{\boldsymbol{\theta}}$. The absence of bias indicates that the sampling distribution of $\widehat{\boldsymbol{\theta}}$ has a mean that coincides with the value of the parameter of interest.

Zero bias is often difficult to achieve in small samples. Therefore, another useful criterion is *asymptotic unbiasedness*: The estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ is called asymptotically unbiased if

$$\lim_{n \to \infty} E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta} \qquad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

That is, an estimator is asymptotically unbiased if the bias vanishes with increasing sample size. An important question in this context is, of course, how fast the bias vanishes (or how large the sample size has to be for the bias to be negligible).

### 2.1.2 Efficiency

For a specific estimation problem, several (asymptotically) unbiased estimators may exist. To choose the best among them we need further information about the performance of the different estimators. Furthermore, there may also be situations in which a biased

estimator is to be preferred over an unbiased estimator. A key aspect in this regard is the *efficiency* of an estimator. Efficiency has to do with how spread out about $\boldsymbol{\theta}$ the sampling distribution of the estimator is. The smaller the dispersion of estimator $\widehat{\boldsymbol{\theta}}$ around the true value $\boldsymbol{\theta}$ in repeated samples, the more "efficient" (or precise) is the estimator.

### Mean squared error

First consider the case of a *scalar* parameter $\theta$. The precision of estimator $\widehat{\theta}$ can be measured by its *mean squared error* (MSE):

$$\text{MSE}_\theta(\widehat{\theta}) = E_\theta\Big((\widehat{\theta} - \theta)^2\Big).$$

A small mean squared error for $\widehat{\theta}$ means that the sampling distribution of $\widehat{\theta}$ is well concentrated around the exact value of the parameter to estimate and hence that the estimator $\widehat{\theta}$ has a good precision.

It is easy to show that

$$\text{MSE}_\theta(\widehat{\theta}) = \text{Var}_\theta(\widehat{\theta}) + \Big(B_\theta(\widehat{\theta})\Big)^2.$$

That is, the mean squared error of an estimator can be decomposed into its variance and its squared bias. Hence, if $\widehat{\theta}$ is unbiased, $\text{MSE}_\theta(\widehat{\theta})$ is simply equal to $\text{Var}_\theta(\widehat{\theta})$.

### Relative efficiency

An estimator $\widehat{\theta}_A$ of $\theta$ is more precise—we will say *more efficient*—than another estimator $\widehat{\theta}_B$ if

$$\text{MSE}_\theta(\widehat{\theta}_A) \leq \text{MSE}_\theta(\widehat{\theta}_B) \qquad \text{for all } \theta \in \Theta$$

and

$$\text{MSE}_\theta(\widehat{\theta}_A) < \text{MSE}_\theta(\widehat{\theta}_B) \qquad \text{for at least one } \theta \in \Theta.$$

In general, we consider the "large-sample" sampling distributions of asymptotically unbiased estimators. If, for large $n$, the estimators $\widehat{\theta}_A$ and $\widehat{\theta}_B$ are approximately $\mathcal{N}(\theta, \text{Var}(\widehat{\theta}_A))$ and $\mathcal{N}(\theta, \text{Var}(\widehat{\theta}_B))$, respectively, we define the *asymptotic relative efficiency* (ARE) of $\widehat{\theta}_B$ with respect to $\widehat{\theta}_A$ as the ratio

$$\text{ARE}_\theta(\widehat{\theta}_B, \widehat{\theta}_A) = \frac{\text{Var}(\widehat{\theta}_A)}{\text{Var}(\widehat{\theta}_B)}$$

(see Serfling, 1980). If $\widehat{\theta}_B$ is (asymptotically) less efficient than $\widehat{\theta}_A$, then

$$\text{ARE}_\theta(\widehat{\theta}_B, \widehat{\theta}_A) \leq 1$$

for all $\theta \in \Theta$, with strict inequality holding for at least some value of $\theta$.

**Efficiency of the maximum likelihood estimator**

Let us consider the case in which the random variables $X_1, \ldots, X_n$ of the sample $\mathcal{X}^{(n)}$ are i.i.d. with a common distribution function $F_\theta$ and a common density function $f_\theta$ that satisfies some differentiability conditions with respect to $\theta$. Suppose also that the *Fisher information*

$$\mathcal{I}(F_\theta) = E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right)$$

is strictly positive and finite. Then it follows that

(i) for large $n$, the maximum likelihood estimator $\widehat{\theta}_{\text{ML}}$ of $\theta$ is approximately distributed as $\mathcal{N}\big(\theta, (n\mathcal{I}(F_\theta))^{-1}\big)$

(ii) for a wide class of estimators $\widehat{\theta}$ that are approximately distributed as $\mathcal{N}(\theta, V)$, a *lower bound* to $V$ is $(n\mathcal{I}(F_\theta))^{-1}$

(see Lehmann and Casella, 1988). In this situation,

$$\text{ARE}_\theta(\widehat{\theta}, \widehat{\theta}_{\text{ML}}) = \frac{(n\mathcal{I}(F_\theta))^{-1}}{V} \leq 1 \tag{2.1}$$

for all $\theta \in \Theta$, making $\widehat{\theta}_{\text{ML}}$ the most (asymptotically) efficient among the given class of estimators $\widehat{\theta}$. Note, however, as will be discussed later, that (2.1) does not necessarily make $\widehat{\theta}_{\text{ML}}$ the estimator of choice, when certain other considerations are taken into account.

**Notation in the multidimensional case**

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t$ is a vector of parameters we define the *mean squared error matrix* of the estimator $\widehat{\boldsymbol{\theta}}$ as follows:

$$\text{MSE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}\Big((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^t\Big).$$

If $\widehat{\boldsymbol{\theta}}$ is unbiased, the mean squared error matrix simply coincides with the covariance matrix of the estimator.

If, for large $n$, the $p$-variate estimators $\widehat{\boldsymbol{\theta}}_A$ and $\widehat{\boldsymbol{\theta}}_B$ are approximately normally distributed with mean $\boldsymbol{\theta}$ and nonsingular covariance matrices $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$, respectively, it is usual to define the *asymptotic relative efficiency* (ARE) of $\widehat{\boldsymbol{\theta}}_B$ with respect to $\widehat{\boldsymbol{\theta}}_A$ as the ratio of the *generalized variances* (determinants of the covariance matrices), raised to the power $1/p$, that is

$$\text{ARE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_B, \widehat{\boldsymbol{\theta}}_A) = \left(\frac{\det(\boldsymbol{\Sigma}_A)}{\det(\boldsymbol{\Sigma}_B)}\right)^{1/p}.$$

If $\text{ARE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_B, \widehat{\boldsymbol{\theta}}_A) \leq 1$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, with strict inequality holding for at least some value $\boldsymbol{\theta}$, estimator $\widehat{\boldsymbol{\theta}}_B$ is (asymptotically) less efficient than estimator $\widehat{\boldsymbol{\theta}}_A$.

Here again the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\text{ML}}$ of $\boldsymbol{\theta}$ appears as the most (asymptotically) efficient estimator among a wide class of (asymptotically) unbiased estimators $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Moreover, for large $n$,

$$\widehat{\boldsymbol{\theta}}_{\text{ML}} \approx \mathcal{N}\big(\boldsymbol{\theta}, (n\mathbf{I}(F_{\boldsymbol{\theta}}))^{-1}\big)$$

where "$\approx$" stands for "is approximately distributed as" and $\mathbf{I}(F_{\boldsymbol{\theta}})$ is the $p \times p$ Fisher information matrix with its elements defined as

$$\mathcal{I}_{ij}(F_{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}\bigg(\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(X) \times \frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{\theta}}(X)\bigg).$$

### 2.1.3 Consistency

High efficiency and (asymptotic) unbiasedness are desired properties for an estimator. But other properties are still required if we want the estimator to provide valid statistical inference for $\boldsymbol{\theta}$.

The sampling distribution of $\widehat{\boldsymbol{\theta}}$ generally depends on the size $n$ of the sample, e.g. via its mean, its variance and (or) other characteristics. How does the distribution of $\widehat{\boldsymbol{\theta}}$ evolve when $n$ increases? This question will first lead us to the properties of *(asymptotic) consistency —convergence in probability—* and *Fisher consistency*. We will then consider the concept of *convergence in distribution*.

#### (Asymptotic) consistency: Convergence in probability

An estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ is *consistent* or *asymptotic consistent* if, for $n \to \infty$, $\widehat{\boldsymbol{\theta}}$ *converges in probability* to $\boldsymbol{\theta}$. That is, for any $\epsilon > 0$ and for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$:

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}}^{(n)}\Big(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \epsilon\Big) = 1. \tag{2.2}$$

This type of convergence means that, when the sample size grows, the probability that the estimator $\widehat{\boldsymbol{\theta}}$ takes a value arbitrarily close to the exact value of the parameter and, consequently, provides a "good" estimate of the parameter $\boldsymbol{\theta}$, grows to 1.

In other terms, a consistent estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is an estimator that becomes more and more precise when the number of observations increases. It also means that the sampling distribution of $\widehat{\boldsymbol{\theta}}$ becomes more and more concentrated around the true value of the parameter being estimated. This property of consistency is of course a highly desirable property for an estimator.

#### Convergence in quadratic mean

Another type of convergence is closely related to the notion of consistency: the *convergence in quadratic mean*. For simplicity, consider the case of a scalar parameter $\theta$: $\widehat{\theta} = \widehat{\theta}(\mathcal{X}^{(n)})$ converges in quadratic mean to $\theta$ if

$$\text{MSE}_\theta(\widehat{\theta}) = E_\theta\Big((\widehat{\theta} - \theta)^2\Big) \overset{n\to\infty}{\longrightarrow} 0.$$

Since $\text{MSE}_\theta(\widehat{\theta}) = \text{Var}_\theta(\widehat{\theta}) + \Big(B_\theta(\widehat{\theta})\Big)^2$, it is clear that, if $\widehat{\theta}$ is asymptotically unbiased and has a variance that tends to zero when $n$ tends to infinity, then $\widehat{\theta}$ converges to $\theta$ in quadratic mean. Moreover, it is quite easy to prove that the convergence in quadratic mean implies the convergence in probability. This result is intuitive since $\text{MSE}_\theta(\widehat{\theta})$ is nothing but a measure of the dispersion of the sampling distribution of $\widehat{\theta}$ around the exact value of the parameter to estimate.

In practice, it is often easier to establish the consistency of $\widehat{\theta}$ by verifying that $\widehat{\theta}$ converges in quadratic mean to $\theta$ instead of considering directly the equality (2.2).

**Fisher consistency**

In statistics, *Fisher consistency* is another desirable property of an estimator asserting that if the estimator were calculated using the entire population rather than a sample, the true value of the estimated parameter would be obtained (Fisher, 1992).

Suppose we have a statistical sample $\mathcal{X}^{(n)} = \{X_1, \ldots, X_n\}$ where each $X_i$ has a distribution characterized by the cumulative distribution $F_{\boldsymbol{\theta}}$ which depends on an unknown parameter $\boldsymbol{\theta}$. If the estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ can be represented as a *functional* of the empirical distribution function[1] $F^{(n)}$, that is,

$$\widehat{\boldsymbol{\theta}} = \mathbf{T}(F^{(n)}),$$

$\widehat{\boldsymbol{\theta}}$ is said to be *Fisher consistent* if

$$\mathbf{T}(F_{\boldsymbol{\theta}}) = \boldsymbol{\theta}.$$

If the strong law of large numbers can be applied, the empirical distribution function $F^{(n)}$ converges pointwise to $F_{\boldsymbol{\theta}}$ when $n$ tends to infinity, allowing us to express Fisher consistency as the following convergence property: The estimator $\widehat{\boldsymbol{\theta}} = \mathbf{T}(F^{(n)})$ is Fisher consistent if

$$\mathbf{T}\Big(\lim_{n\to\infty} F^{(n)}\Big) = \boldsymbol{\theta}.$$

Fisher consistency and (asymptotic) consistency are distinct concepts, although both aim to define a desirable property of an estimator. While many estimators are consistent in both senses, neither definition encompasses the other: There exist estimators that are (asymptotically) consistent but not Fisher consistent and, inversely, there are estimators that are Fisher consistent but not (asymptotically) consistent.

### 2.1.4 Convergence in distribution

A last type of convergence is important when we have to provide a confidence interval for (each component of) $\boldsymbol{\theta}$ or to define a statistical test to solve a testing problem for the unknown parameter $\boldsymbol{\theta}$: The *convergence in distribution*. Since this property concerns

[1] The empirical distribution function $F^{(n)}$ is the distribution function associated with the discrete probability distribution allocating a probability mass of $1/n$ at each observation $X_i$, $i = 1, \ldots, n$, of the sample $\mathcal{X}^{(n)}$.

usually a function of an estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ than the estimator $\widehat{\boldsymbol{\theta}}$ itself, we formulate it in a general way for a statistic $\mathbf{U}^{(n)} = \mathbf{U}(\mathcal{X}^{(n)})$.

The statistic $\mathbf{U}^{(n)}$ *converges in distribution* to the probability distribution $\mathcal{L}$ if, for $n \to \infty$, the distribution function $F_{\mathbf{U}^{(n)}}$ of $\mathbf{U}^{(n)}$ converges to the distribution function $F_{\mathcal{L}}$ associated with $\mathcal{L}$ in any point of continuity of $F_{\mathcal{L}}$. That is, convergence in distribution is given if

$$F_{\mathbf{U}^{(n)}}(\mathbf{u}) \overset{n\to\infty}{\longrightarrow} F_{\mathcal{L}}(\mathbf{u})$$

for any continuity point $\mathbf{u}$ of the distribution function $F_{\mathcal{L}}$.

As a shorthand, we will write

$$\mathbf{U}^{(n)} \overset{d}{\to} \mathcal{L}$$

to denote convergence in distribution for $n \to \infty$. In practice, convergence in distribution means that, for large $n$, we may consider that $\mathbf{U}^{(n)}$ is approximately distributed as $\mathcal{L}$, that is,

$$\mathbf{U}^{(n)} \approx \mathcal{L}.$$

In practice, it is useful to consider an estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such that a certain function of $\widehat{\boldsymbol{\theta}}$ converges in distribution to a well known probability distribution $\mathcal{L}$ as, for example, a normal distribution, a Student distribution, a chi-square or a Fisher distribution. Indeed it allows to develop inference procedures for $\boldsymbol{\theta}$, based on $\widehat{\boldsymbol{\theta}}$, that are valid when the sample size $n$ is not too small.

### 2.1.5 Other aspects

Depending on context, a number of other criteria can be important. For example, from a practical perspective, *computational complexity* can be a relevant criterion to choose between different estimators. In general, estimators that require operations in the order of $n^2$ (that is, if the number of required computational operations grows quadratically with the sample size) lead to prohibitive computational costs in large samples. In many cases it is possible to design alternative estimators (or improved computational algorithms for a given estimator) that only require operations in the order of $\ln n$ and are thus much more efficient (with respect to computer time) in large samples. [Maybe expand a bit on this.]

Furthermore, consistent estimators may differ in their *rate of convergence*, that is, in how fast the mean squared error diminishes with growing sample size.

If $\text{MSE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})$ is of order $n^{-\nu}$ (with $\nu > 0$ such that the mean squared error of $\widehat{\boldsymbol{\theta}}$ tends to zero when $n$ tends to infinity, ensuring the convergence of $\widehat{\boldsymbol{\theta}}$ in quadratic mean and, hence, in probability), the rate of convergence of $\widehat{\boldsymbol{\theta}}$ is said to be equal to $n^{\nu/2}$. This rate of convergence is actually the factor by which we have to multiply $\widehat{\boldsymbol{\theta}}$ to obtain a mean squared error that does not depend anymore on the sample size $n$. Most of estimators in parametric models have a mean squared error of order $n^{-1}$ and, consequently, enjoy a rate of convergence equal to $\sqrt{n}$. It is the case, for example, for the least squares estimator of the regression coefficients vector in the classical linear regression model. The squared root of $n$ is certainly considered as the usual rate of convergence in the parametric context. In the nonparametric context, we may encounter estimators whose

rate of convergence is smaller than $\sqrt{n}$. This is the case, for instance, for a kernel estimator of a density function $f$: The kernel estimator of $f(x)$ has a mean squared error of order $n^{-4/5}$ and hence a rate of convergence equal to $n^{2/5}$.

Naturally, an estimator with a faster rate of convergence is usually to be preferred over an estimator with a slower rate of convergence.

Finally, an estimator should be *equivariant* to transformations of data. That is, a transformation of the data should affect the estimator $\widehat{\boldsymbol{\theta}}$ in the same functional way as it affects the true parameter $\boldsymbol{\theta}$. For example, let $\theta_A$ be the expected value of variable $X_A$ and $\theta_B$ be the expected value of variable $X_B$. If $X_B$ can be expressed as a linear combination of $X_A$, that is, $X_B = a + bX_A$, then $\theta_B = a + b\theta_A$. In this case, also $\widehat{\theta}_B = a + b\widehat{\theta}_A$ should hold. In other words, whether you express your data in Dollars or in Euros, whether you express your data in degrees Fahrenheit or degrees Celsius should only affect the scaling of your estimator, but should not affect your results otherwise.

## 2.2 Measures of robustness

Intuitively, the classical approach to statistics is about defining estimators that have desirable properties under a specified model. The goal of robust methods, however, is to develop estimators that perform well also in the "neighborhood" of such a model. This leads to the proposition of so called "robust" estimators, that are, for instance, not affected too strongly or too quickly by the presence of outliers. Although outliers are only one of the main concerns of robust methods, we will make our first steps into robustness theory by presenting some basic concepts for measuring the degree to which estimators are affected by atypical observations.

**Example**

Consider the following observations of the grades achieved by $n = 25$ students in fifth year of primary school (on a scale of 0 to 10):

| | | | | |
|---|---|---|---|---|
| 6.00 | 6.50 | 7.00 | 7.00 | 7.00 |
| 7.00 | 7.00 | 7.50 | 7.50 | 8.00 |
| 8.00 | 8.00 | 8.50 | 8.50 | 8.50 |
| 8.50 | 9.00 | 9.00 | 9.50 | 9.50 |
| 9.50 | 9.50 | 9.50 | 9.50 | 10.00 |

If we calculate on these data two *measures of location*, the mean and the median, as well as two *measures of scale*, the standard deviation and the interquartile range, the results are as follows:

```
. drop _all
. matrix x = (6.00, 6.50, 7.00, 7.00, 7.00,    ///
>             7.00, 7.00, 7.50, 7.50, 8.00,    ///
>             8.00, 8.00, 8.50, 8.50, 8.50,    ///
>             8.50, 9.00, 9.00, 9.50, 9.50,    ///
>             9.50, 9.50, 9.50, 9.50, 10.00)'
. quietly svmat x
```

```
. tabstat x, statistics(mean median sd iqr)
    variable |      mean       p50        sd       iqr
-------------+--------------------------------------------
          x1 |      8.22       8.5  1.137248       2.5
-------------------------------------------------------
```

Now, imagine the dot separating the decimals in the last observation is mistakenly removed, so that the last observation is coded as 1000. In this case the results are the following:

```
. replace x1 = 1000 in l
(1 real change made)
. tabstat x1, statistics(mean median sd iqr)
    variable |      mean       p50        sd       iqr
-------------+--------------------------------------------
          x1 |     47.82       8.5  198.3737       2.5
-------------------------------------------------------
```

As is evident, the mean and the standard deviation strongly increased due to the introduction of the erroneous observation, whereas the median and the interquartile range remained unchanged. The example illustrates the fact that one single outlier may "break" the mean and the standard deviation, but does not affect the median or the interquartile range. Hence, these two later statistics can be considered as being more robust to erroneous data than the two first ones.

◁

How can the degree of robustness of different statistics be quantified? How can we compare the robustness of different estimators from various viewpoints? These are questions we will address in the rest of this section.

In robust estimation theory it is common to consider *parameters* as *functionals*. More precisely, the functional by which a parameter[2] $T$ is defined is a rule that maps every distribution function $F$ into a real number, that is, $T = T(F)$.[3] Often, a natural *estimate* $T^{(n)}$ of the parameter $T(F)$ based on sample $\mathcal{X}^{(n)} = \{x_1, \ldots, x_n\}$—where $x_1, \ldots, x_n$ are realizations of $n$ independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ of distribution $F$—may be defined as the value of the functional at the empirical distribution $F^{(n)}$.[4] That is, $T^{(n)} = T(F^{(n)})$. For example, if

$$T(F) = \int_{-\infty}^{\infty} x \, dF(x) = \mu$$

[2]For the sake of simplicity, we will only consider here *scalar* parameters.

[3]$F$ is the cumulative distribution function of a random variable $X$. Evaluated at position $x$, the function returns the probability that the random variable will take on a value lower than or equal to $x$, that is, $F(x) = \Pr(X \leq x)$. In order to avoid unnecessary technical difficulties we will generally assume in this chapter that the distribution $F$ is continuous with density $f$. The density is the first derivative of $F$; it is nonnegative and integrates to one.

[4]The empirical distribution function $F^{(n)}$ is the distribution function associated with the discrete probability distribution allocating a probability mass of $1/n$ at each observation $x_i$, $i = 1, \ldots, n$, of the sample $\mathcal{X}^{(n)}$.

is the expected value of the distribution $F$, then

$$T^{(n)} = T(F^{(n)}) = \int_{-\infty}^{\infty} x\, dF^{(n)}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \mu^{(n)}$$

is the arithmetic mean of a sample $\mathcal{X}^{(n)}$ from $F$. Likewise, if $T(F) = F^{-1}(0.5) = Q_{0.5}$ is the median of the distribution $F$, then $T^{(n)} = T(F^{(n)}) = (F^{(n)})^{-1}(0.5) = Q_{0.5}^{(n)}$ is the empirical median of a sample $\mathcal{X}^{(n)}$ from $F$.

The robustness of a statistic (or estimator) $T^{(n)}$ may be analyzed in a very intuitive way by studying how a contamination of the sample $\mathcal{X}^{(n)}$ affects $T^{(n)}$. This empirical approach leads to the notions of the *sensitivity curve* and the *finite-sample breakdown point* of $T^{(n)}$. But it is also of great interest to consider the limiting case where $n$ tends to infinity. As the sample size $n$ grows, the empirical distribution function $F^{(n)}$ approaches the underlying population distribution function $F$, and the empirical measures of robustness of the statistic $T^{(n)}$ move in a natural way to the concepts of the *influence function* and the *asymptotic breakdown point* of the functional $T$.

### 2.2.1 The sensitivity curve and the influence function

#### The sensitivity curve

The *sensitivity curve* (SC) is an empirical tool to quantify the robustness of a statistic in a given sample. Consider a data set $\mathcal{X}^{(n)} = \{x_1, \ldots, x_n\}$ and the statistic $T^{(n)} = T^{(n)}(x_1, \ldots, x_n) = T(F^{(n)})$. To study the impact of a potential outlier on this statistic, we may analyze the change in the value of the statistic once we add an extra data point $x$, where $x$ is varied between $-\infty$ an $+\infty$. Hence, the *(standardized) sensitivity curve* of statistic $T^{(n)}$ for the sample $\mathcal{X}^{(n)}$ is defined as

$$\text{SC}(x; T^{(n)}, \mathcal{X}^{(n)}) = \frac{T^{(n+1)}(x_1, \ldots, x_n, x) - T^{(n)}(x_1, \ldots, x_n)}{\frac{1}{n+1}} \,.$$

That is, for each value of $x$ we compare the statistic in the "contaminated" sample to its value in the original sample, and rescale the difference by dividing the difference by $1/(n+1)$, the proportion of contamination.

**Example**

Consider a data set $\mathcal{X}^{(n)}$ of $n = 20$ (rounded) random numbers from a $\mathcal{N}(0, 1)$ (standard normal) distribution:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| −0.49 | 0.14 | 1.54 | 0.63 | −0.87 | −0.86 | 1.65 | −0.55 | 0.91 | −0.03 |
| −0.61 | 0.22 | −1.61 | 0.15 | 0.36 | 1.96 | 1.04 | 0.24 | −0.45 | 0.98 |

Figure 2.1 shows the standardized sensitivity curves of the mean and the median; Figure 2.2 displays the standardized sensitivity curves of the standard deviation and the interquartile range. As is evident, the mean and the standard deviation have unbounded sensitivity curves (the curves go off to minus or plus infinity as the outlier moves away

from the center of the uncontaminated data), whereas the sensitivity curves of the median and the interquartile range are bounded. The "classic" location and scale measures may be completely perturbed by the presence of one single outlying observation—this illustrates the non-robust character of these two statistics—, while the impact of the additional outlying observation on the quantile-based location and scale measures remains very limited.

◁

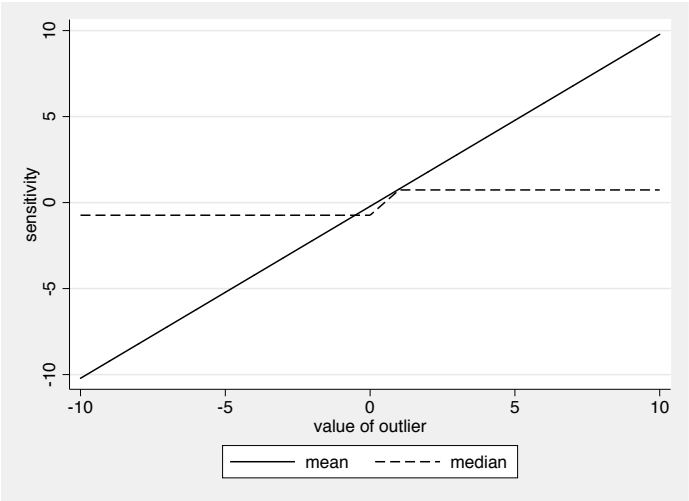


Figure 2.1: Standardized sensitivity curves of the mean and the median for a sample of $n = 20$ random $\mathcal{N}(0, 1)$ numbers

### The influence function

An intuitive way to introduce the *influence function* (IF) of functional $T$ at some distribution $F$ is to think of the influence function as an asymptotic version of the sensitivity curve of statistic $T^{(n)} = T(F^{(n)})$ when the sample size $n$ grows, so that the empirical distribution function $F^{(n)}$ tends to the underlying population distribution function $F$ (cf. Hampel, 1974). More precisely, the influence function is defined as

$$\begin{aligned} \text{IF}(x; T, F) &= \lim_{n\to\infty} \frac{T\Big(\Big(1 - \frac{1}{n+1}\Big)F + \frac{1}{n+1}\Delta_x\Big) - T(F)}{\frac{1}{n+1}} \\ &= \lim_{\varepsilon\to 0} \frac{T((1-\varepsilon)F + \varepsilon\Delta_x) - T(F)}{\varepsilon} \end{aligned}$$

where $\Delta_x$ is a probability distribution with all its mass at point $x$. That is, the influence function measures the effect on $T$ of a perturbation of $F$ obtained by adding a small probability mass at point $x$. The expression of $\text{IF}(x; T, F)$ can be found for most

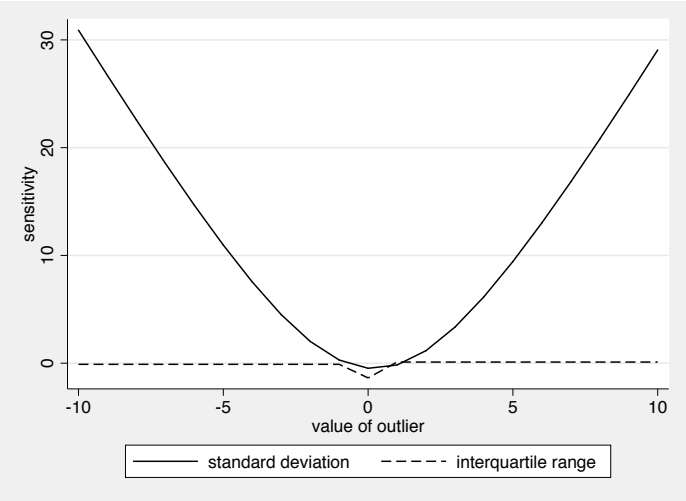


Figure 2.2: Standardized sensitivity curves of the standard deviation and the interquartile range for a sample of $n = 20$ random $\mathcal{N}(0, 1)$ numbers

functionals $T$. In chapter 3 we will provide the influence functions of various measures of location, scale, skewness and tails heaviness.

### The gross-error sensitivity

Since $\text{IF}(x; T, F)$ quantifies the influence on $T$ of an infinitesimal contamination of the distribution $F$ at point $x$, it is a *local* measure of robustness. It may be completed by a more global measure, the *gross-error sensitivity* of $T$ at distribution $F$, defined as

$$\gamma^*(T, F) = \sup_x |\text{IF}(x; T, F)|.$$

$\gamma^*(T, F)$ evaluates to the biggest influence an outlier can have on the functional $T$. With respect to robustness it is desirable to use an estimator that is associated with a functional $T$ for which $\gamma^*(T, F)$ is finite (that is, for which the influence function is bounded).

### The local-shift sensitivity

The local-shift sensitivity is another tool related to the influence function; it aims at measuring the effect of "wiggling" an observation, that is, of a small perturbation as opposed to gross error. This is useful to asses the effects of rounding, grouping, or other local inaccuracies.

Jumps in the influence function (IF) indicate that a small fluctuation of the value of $x$ can cause an abrupt change in the estimate. Hence, from the perspective of robustness, we prefer a continuous IF with an appropriately bounded derivative (wherever the

derivative exists). To appreciate this kind of characteristic of the *IF*, we may determine the *local-shift sensitivity*:

$$\lambda^*(T,F) = \sup_{x \neq y} \frac{|\text{IF}(y;T,F) - \text{IF}(x;T,F)|}{|y-x|} \ .$$

**The asymptotic variance of an estimator**

The influence function may also be used as a heuristic tool to determine the asymptotic variance of the estimators. Indeed, under some regularity conditions for the functional $T$, we have, under $F$,

$$\sqrt{n}\Big(T(F^{(n)}) - T(F)\Big) \xrightarrow{d} \mathcal{N}(0, \text{ASV}(T,F))$$

where

$$\text{ASV}(T,F) = \int_{-\infty}^{\infty} \text{IF}(x;T,F)^2 \, dF(x) \tag{2.3}$$

(cf. Hampel et al., 1986, p. 85 and 226). Consequently, under $F$, the interval

$$\left[T(F^{(n)}) - z_{1-\alpha/2}\sqrt{\frac{\text{ASV}(T,F)}{n}} \ , \ T(F^{(n)}) + z_{1-\alpha/2}\sqrt{\frac{\text{ASV}(T,F)}{n}}\right]$$

where $z_{1-\alpha/2}$ is the quantile of order $(1-\alpha/2)$ of the $\mathcal{N}(0,1)$ distribution, provides an asymptotic confidence interval for the parameter $T(F)$, at a confidence level of $(1-\alpha)$. If the distribution $F$, and hence the asymptotic variance $\text{ASV}(T,F)$, are not known, it is still possible to obtain a confidence interval for $T(F)$ by an appropriate resampling method.

1: `robstat` estimates variances based on the "empirical" IF; this works well. Should maybe be explained here...

## 2.2.2 The breakdown point

The sensitivity curve shows how an estimator reacts to the introduction of one single outlier. Some estimators cannot resist even against a single outlier. As we have seen, this is the case for the mean and the standard deviation. Other estimators, such as the median and the interquartile range, are robust against this type of contamination because their sensitivity curve (SC) is bounded. Possibly, however, the number of outliers in a sample is so large that even estimators with a bounded SC can no longer resist their effect. Hence, to evaluate different estimators, it is important to know what the amount of contamination is an estimator can tolerate. The *breakdown point* is a measure for such *resistance* of an estimator. It quantifies, roughly, the smallest amount of contamination in the sample that may cause the estimator to take on arbitrary values. Its definition is as follows.

**The finite-sample breakdown point**

The breakdown point $\epsilon^{*(n)}(T^{(n)}; \mathcal{X}^{(n)})$ of the statistic $T^{(n)} = T^{(n)}(x_1, \ldots, x_n) = T(F^{(n)})$ at the sample $\mathcal{X}^{(n)} = \{x_1, \ldots, x_n\}$ refers to the smallest proportion of observations in

$\mathcal{X}^{(n)}$ that need to be replaced to cause the value of the statistic to be arbitrarily large or small, and hence, to make the statistic worthless or meaningless. Note that, typically, $\epsilon^{*(n)}$ is independent of $x_1, \ldots, x_n$.

More formally, for a univariate location estimator $T^{(n)}$, which breaks down if its absolute value becomes arbitrarily large, we may define the (finite-sample) breakdown point as follows (see Hampel and Stahel, 1982; Donoho and Huber, 1983). In a given sample $\mathcal{X}^{(n)} = \{x_1, \ldots, x_n\}$, let us replace $m$ data points $x_{i_1}, \ldots, x_{i_m}$ by arbitrary values $y_1, \ldots, y_m$; let us call the new data set $\mathcal{Z}^{(n)} = \{z_1, \ldots, z_n\}$. Then the *(finite-sample gross-error) breakdown point* of the estimator is

$$\epsilon^{*(n)}(T^{(n)}; \mathcal{X}^{(n)}) = \min\left\{ \frac{m}{n} \middle| \max_{i_i,\ldots,i_m} \sup_{y_1,\ldots,y_m} \left| T^{(n)}(z_1, \ldots, z_n) \right| = \infty \right\}.$$

Following the same idea, we will say that a scale estimator breaks down if it takes on a value that is arbitrarily large (scale explosion) or close to zero (scale implosion). Furthermore, a skewness or kurtosis estimator, which is bounded by $[-1, 1]$, breaks down if the absolute value of the estimate attains the value of 1.

**Example**

If the $i$th observation among $x_1, \ldots, x_n$ goes to infinity, the mean $\mu^{(n)}$ and the standard deviation $\sigma^{(n)}$ go to infinity as well. This means that the finite-sample breakdown point of these two statistics is $1/n$. In contrast, the finite-sample breakdown point of the median $Q_{0.5}^{(n)}$ is $\frac{(n/2)}{n}$ if $n$ is even and $\frac{(n+1)/2}{n}$ if $n$ is odd. That is, half the data or a bit more must be replaced to make the median take on arbitrary values. The finite-sample breakdown point of the interquartile range $\text{IQR}^{(n)}$ is equal to $\frac{\lfloor n/4 \rfloor + 1}{n}$, where $\lfloor n/4 \rfloor$ denotes the integer part of $n/4$. That is, a bit more than one fourth of the data needs to be replaced to make the IQR break down.

◁

**The asymptotic breakdown point**

The *asymptotic breakdown point* $\epsilon^*(T, F)$ of the functional $T$ under the distribution $F$ is defined as

$$\epsilon^*(T, F) = \lim_{n \to \infty} \epsilon^{*(n)}(T^{(n)}; \mathcal{X}^{(n)})$$

with the $x_i$'s sampled from $F$ (cf. Hampel, 1971).

[I would delete subsections "2.2.3 Gaussian efficiency" and "2.2.4 Aspects of interpretation". These subsections do not refer to robustness in the presence of outliers; there are related with classical properties of estimators, presented previously. I have modified the subsection "Summary" in order to take into account some ideas that play an important role in the choice of a robust estimator.]

### 2.2.3 Summary

How do we choose a good robust estimator? We are clearly interested in estimators with

1. a *bounded* (low gross-error sensitivity) and *smooth* (low local-shift sensitivity) *influence function*

2. and a *high breakdown point*.

Moreover, we are looking for robust estimators that enjoy good efficiency for a wide variety of distributions. In general, however, compromises between robustness and efficiency must be made to achieve good overall performance, as is shown in the following section.

Finally, we also wish to use asymptotically and Fisher consistent robust estimators, whose rate of convergence is not smaller than the usual rate of convergence—equal to $\sqrt{n}$—in the parametric context.

# Chapter 3

# Basic robust statistics

Many measures of location, scale, skewness and kurtosis or heaviness of the tails have been proposed and studied in the statistical literature. The present chapter is devoted to the comparison of the (asymptotic) Gaussian efficiency and robustness performance of three different classes of estimators: (i) "classic" estimators, based on (centered) moments of the distribution $F_n$; (ii) estimators built from specific quantiles of the distribution; (iii) estimators defined on the basis of pairwise comparisons or combinations of the observations. In addition, we will discuss robust test of normality and robust boxplots.

## 3.1 Robust estimation of location

There is apparent consensus in applied statistics about the fact that the sample mean and the sample median are two complementary location estimators: the mean is very efficient in case of Gaussian (i.e. normally distributed) data but fragile to outliers (and problematic in case of highly asymmetric data) while the median is very robust (and meaningful in case of asymmetry) but rather inefficient. Both are extensively used in practice. Let us briefly recall their respective properties and introduce two other frequently used location estimators.

### 3.1.1 The mean and the $\alpha$-trimmed mean

The *mean* corresponds to the functional $\mu = \mu(F) = \int_{-\infty}^{\infty} x\, dF(x)$; its empirical counterpart is $\mu_n = \mu(F_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$. It is well known that this location estimator is the most efficient estimator for Gaussian data; its asymptotic variance is $\text{ASV}(\mu, F) = \sigma^2$, where $\sigma^2$ denotes the variance of the distribution $F$ (taking $F = \Phi$, the standard normal distribution, we have $\text{ASV}(\mu, \Phi) = 1$). Unfortunately, the mean lacks robustness. Indeed, one single outlying observation can move this estimator towards an arbitrarily large (absolute) value: its asymptotic breakdown point $\epsilon^*(\mu, F)$ is equal to 0. Likewise, its influence function is unbounded, leading to an infinite gross-error sensitivity:

2: Comment on integrals: Why not $\int x f(x)\, dx$? Maybe show somewhere that $\int f(x)\, dx = \int dF(x)$

3: Why "under the standard normal distribution"? Isn't the influence function always the same irrespective of $F$?

IF$(x; \mu, F) = x - \mu$ (see Wilcox, 2005, p. 25). Figure 3.1a shows the influence function of $\mu$ under the standard normal distribution.

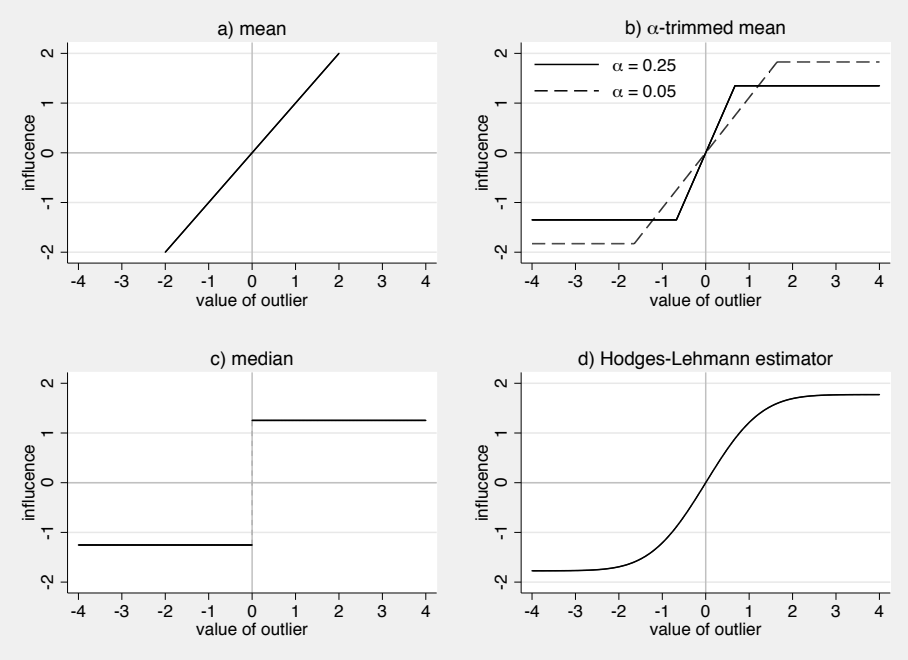


Figure 3.1: Influence functions of $\mu$, $\mu^{0.25}$, $\mu^{0.05}$, $Q_{0.5}$, and HL under the standard normal distribution

A simple and classical way to "robustify" the sample mean consists in discarding a certain proportion $\alpha$ ($0 \leq \alpha < 0.5$) of the smallest and of the biggest observations in the sample. This leads to the *$\alpha$-trimmed mean* defined by

$$\mu_n^\alpha = \frac{1}{n - 2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n - \lfloor \alpha n \rfloor} x_{(i)}$$

where $\lfloor x \rfloor$ denotes the integer part of $x$ and $x_{(i)}$ is the $i$th order statistic (the observation at the $i$th position in the list of sorted observations in ascending order). Note that the sample mean $\mu_n$ is a special case of the $\alpha$-trimmed mean $\mu_n^\alpha$ corresponding to $\alpha = 0$. The functional associated with this location estimator is

$$\mu^\alpha(F) = \frac{1}{1 - 2\alpha} \int_{Q_\alpha}^{Q_{1-\alpha}} x \, dF(x)$$

where $Q_\alpha = F^{-1}(\alpha)$ and $Q_{1-\alpha} = F^{-1}(1 - \alpha)$ are the $\alpha$ and $(1 - \alpha)$ quantiles of distribution $F$.

The influence function of this functional has the advantage to be bounded. If $F$ is symmetric, then

$$\text{IF}(x;\mu^\alpha,F) = \begin{cases} \frac{1}{1-2\alpha}\big(F^{-1}(\alpha)-\mu\big) & \text{if } x < F^{-1}(\alpha) \\ \frac{1}{1-2\alpha}(x-\mu) & \text{if } F^{-1}(\alpha) \le x \le F^{-1}(1-\alpha) \\ \frac{1}{1-2\alpha}\big(F^{-1}(1-\alpha)-\mu\big) & \text{if } x > F^{-1}(1-\alpha) \end{cases}$$

(see, e.g., Staudte and Sheather, 1990). As an example, see Figure 3.1b for the influence functions of $\mu^{0.05}$ and $\mu^{0.25}$ under the standard Gaussian distribution $\Phi$.

Moreover, the asymptotic breakdown point of $\mu^\alpha$ is equal to $100\alpha\%$. Clearly, hence, the proportion $\alpha$ of trimming appears as a parameter allowing to choose the level of robustness of the trimmed mean. Of course, this gain in robustness goes hand in hand with a loss in efficiency. Yet, this loss is not as large as one may fear. Using the fact that

$$\text{ASV}(\mu^\alpha,F) = \int_{-\infty}^{\infty} \text{IF}(x;\mu^\alpha,F)^2\,dF(x)$$

we obtain, for example,

4: How did you compute these numbers? Are there closed form solutions?

$$\text{ASV}(\mu^{0.05},\Phi) \approx 1.0263 \qquad \text{ASV}(\mu^{0.10},\Phi) \approx 1.0604 \qquad \text{ASV}(\mu^{0.25},\Phi) \approx 1.1952$$

for the standard normal distribution. Hence, the asymptotic Gaussian relative efficiency of $\mu^\alpha$ with respect to the mean, defined as $\text{ASV}(\mu,\Phi)/\text{ASV}(\mu^\alpha,\Phi) = 1/\text{ASV}(\mu^\alpha,\Phi)$, reaches 97% for $\alpha = 0.05$, 94% for $\alpha = 0.10$, and, despite the fact that half of the sample is discarded, 84% for $\alpha = 0.25$.

### 3.1.2 The median

The *median*—the quantile of order 0.5—corresponds to the functional $Q_{0.5}(F) = F^{-1}(0.5)$. Its empirical version $Q_{0.5;n}$ is simply the sample median $F_n^{-1}(0.5)$, typically computed as

$$Q_{0.5;n} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

where $x_{(i)}$ again denotes the $i$th order statistic among $x_1,\ldots,x_n$.

The median performs better than the mean from the robustness point of view (see, e.g., Staudte and Sheather, 1990, p. 56 and 59). First of all, its influence function is given by

$$\text{IF}(x;Q_{0.5},F) = \begin{cases} -\frac{1}{2f(F^{-1}(0.5))} & \text{if } x < F^{-1}(0.5) \\ 0 & \text{if } x = F^{-1}(0.5) \\ \frac{1}{2f(F^{-1}(0.5))} & \text{if } x > F^{-1}(0.5) \end{cases}$$

where $f$ is the density function associated with $F$. In particular, as displayed in Figure 3.1c, $\text{IF}(x;Q_{0.5},\Phi) = \text{sign}(x)\sqrt{\pi/2}$ for standard Gaussian data, since $f\big(\Phi^{-1}(0.5)\big) = \sqrt{\pi/2}$. This influence function is bounded, leading to a bounded gross-error sensitivity,

but it has a discontinuity at $x = 0$. Furthermore, the asymptotic breakdown point of the median is equal to 50% and is thus higher than the asymptotic breakdown point of the $\alpha$-trimmed mean, regardless of the value of $\alpha$.

5: Say why the discontinuity is a problem.

Finally, the asymptotic variance of the median is given as

$$\text{ASV}(Q_{0.5}, F) = \frac{1}{4f(F^{-1}(0.5))^2}$$

Hence, relative Gaussian efficiency compared to the mean is equal to

$$\frac{\text{ASV}(\mu, \Phi)}{\text{ASV}(Q_{0.5}, \Phi)} = \frac{1}{\pi/2} = \frac{2}{\pi} \approx 64\%$$

### 3.1.3 The Hodges-Lehmann estimator

Hodges and Lehmann (1963) have introduced an alternative location estimator that has the advantage of a bounded, continuous and smooth influence function, but also a high asymptotic Gaussian relative efficiency with respect to the sample mean. The *Hodges-Lehmann estimator* at the sample $\mathbf{X}_n$ is defined by

6: Can't we type $\text{HL}_n = \underset{i<j}{\text{med}}\left(\frac{x_i+x_j}{2}\right)$? This would seem easier to understand to me. Also, why only $i < j$? Why not all possible combinations (even including $i = j$)?

$$\text{HL}_n = \text{med}\left\{\frac{x_i + x_j}{2}; i < j\right\}$$

It is the empirical version of the functional $\text{HL} = \text{HL}(F)$, which is defined as the median of the distribution of $(X+Y)/2$, where $X$ and $Y$ are i.i.d. random variables of distribution $F$.

For symmetric $F$, the influence function of HL is given as

7: I use $f(x)^2$ instead of $f^2(x)$. I know the latter is often used in stats literature, but to me the former much is clearer. Furthermore, in the numerator it should be $F(x - \text{HL}(F))$, not $F(x - \mu)$, I think (in Staudte/Sheather 1990:121 $\mu$ is used, but they define $\mu$ as the median of $(X + Y)/2$).

$$\text{IF}(x; \text{HL}, F) = \frac{2F(x - \text{HL}(F)) - 1}{2\int_{-\infty}^{\infty} f(y)^2\, dy}$$

Figure 3.1c presents the influence function for standard Gaussian data. It illustrates that outliers have a bounded influence. Also note that the sensitivity of the Hodges-Lehmann estimator depends on the smoothness of $F$. Hence, the local-shift sensitivity is small as long as the data have a smooth distribution function.

8: The original just said that the sensitivity depends on the smoothness of $F$ without saying why this is relevant. I thus added some text. Please check, whether it is ok.

Because HL combines the robust properties of the median with the efficiency properties of averaging, it performs well for a variety of distributions. Based on (2.3) we obtain

$$\text{ASV}(\text{HL}, F) = \frac{1}{12}\left(\frac{1}{\int_{-\infty}^{\infty} f(y)^2\, dy}\right)^2$$

For example, for standard Gaussian date, $\text{ASV}(\text{HL}, \Phi) = \pi/3 = 1.0472$. Hence, the asymptotic efficiency of the Hodges-Lehmann estimator relative to the mean is equal to $3/\pi \approx 95\%$ at the normal distribution. Moreover, the Hodges-Lehmann estimator reaches a relative efficiency with respect to the mean of at least 86% for any symmetric distribution (see, for example, Staudte and Sheather, 1990, p. 120-121). Compared to the median, the Hodges-Lehmann estimator has a higher Gaussian efficiency (95% vs. 64%) but a lower asymptotic breakdown point (29% vs. 50%).

### 3.1.4 M estimate of location

[to be completed]

### 3.1.5 Summary

Table 3.1 summarizes the different properties of the four location estimators. From the perspective of a good balance between high Gaussian efficiency and a high breakdown point, the Hodges-Lehmann estimator appears to perform particularly well.

Table 3.1: Characteristics of the four location estimators

| Estimator | Class | Gaussian efficiency | Asymptotic breakdown point | Bounded influence function |
|---|---|---|---|---|
| mean $\mu_n$ | moment | 100% | 0% | no |
| $\alpha$-trimmed mean $\mu_n^\alpha$ | moment | $\alpha = 0.05$: 97%<br>$\alpha = 0.10$: 94%<br>$\alpha = 0.25$: 84% | $100\alpha\%$ | yes |
| median $Q_{0.5;n}$ | quantile | 64% | 50% | yes |
| Hodges-Lehmann $\text{HL}_n$ | pairwise | 95% | 29% | yes |

## 3.2 Robust estimation of scale

### 3.2.1 The standard deviation

The classic statistic to estimate the scale parameter of a distribution is the *standard deviation* $\sigma_n$, corresponding to the functional

$$\sigma = \sigma(F) = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 \, dF(x)}$$

At sample $\mathbf{X}_n$, $\sigma_n$ is typically computed as

9: I changed this to the usual $1/(n-1)$ variant; if we use the $1/n$ version we need to explain why.

$$\sigma_n = \sigma(F_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_n)^2}$$

with $\mu_n$ as defined above. The standard deviation is the most efficient estimator of the scale parameter $\sigma$ in case of Gaussian data (note that $\text{ASV}(\sigma, \Phi) = 0.5$). However, just like the mean, the standard deviation is very fragile to outliers. Its influence function, given as

10: Show why $\text{ASV}(\sigma, \Phi) = 0.5$

$$\text{IF}(x; \sigma, F) = \frac{1}{2\sigma}\left(x^2 - 2\mu x + \mu^2 - \sigma^2\right)$$

is unbounded and its asymptotic breakdown point is equal to 0% (e.g., Rousseeuw and Croux, 1993, p. 1275). The influence function for standard Gaussian data, given as $\text{IF}(x; \sigma, \Phi) = \frac{1}{2}(x^2 - 1)$, is displayed in Figure 3.2a.

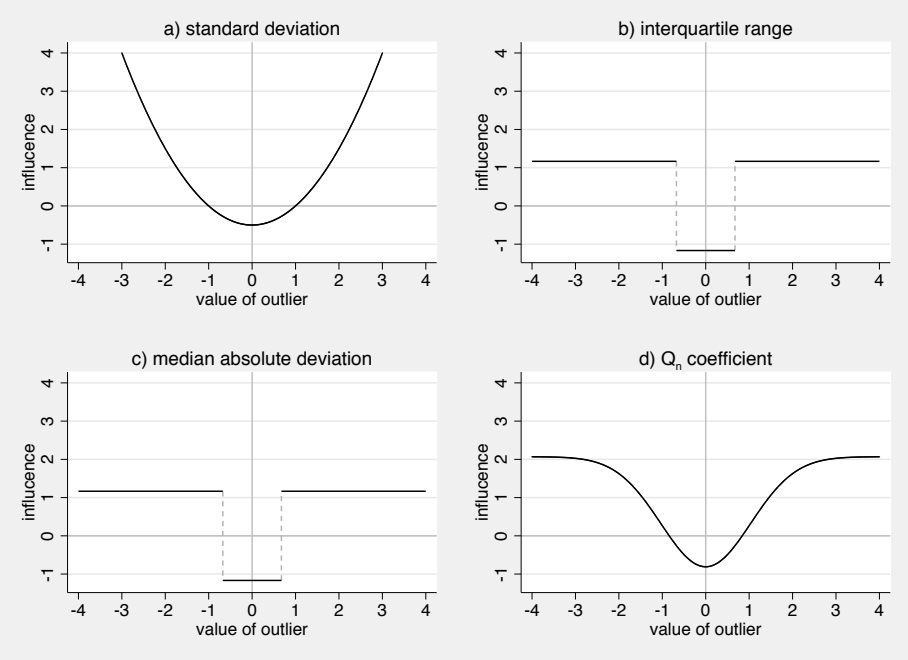


Figure 3.2: Influence functions of $\sigma$, $\text{IQR}_c$, MAD and $Q$ under the standard normal distribution

### 3.2.2 The interquartile range

A common alternative scale measure, defined on the basis of quantiles, is the *interquartile range*

$$\text{IQR} = Q_{0.75} - Q_{0.25}$$

where $Q_{0.25}$ and $Q_{0.75}$ are the first and third quartiles of distribution $F$. Instead of IQR one frequently uses the *corrected interquartile range* $\text{IQR}_c$ defined as

$$\text{IQR}_c = d \cdot \text{IQR}$$

where $d$ is a constant chosen to make the estimator $\text{IQR}_{c;n}$ Fisher-consistent for the scale parameter of the underlying distribution. For example, for a Gaussian distribution, use $d = 1/\big(\Phi^{-1}(0.75) - \Phi^{-1}(0.25)\big) \approx 0.7413$ to make the corrected interquartile range a consistent estimator for the usual scale parameter $\sigma$.

The influence function of the interquartile range is bounded, but discontinuous (see, for example, Wilcox, 2005, p. 35–36). It is given as

$$\text{IF}(x; \text{IQR}, F) = \begin{cases} \frac{1}{f(F^{-1}(0.25))} - C & \text{if } x < F^{-1}(0.25) \\ -C & \text{if } F^{-1}(0.25) \leq x \leq F^{-1}(0.75) \\ \frac{1}{f(F^{-1}(0.75))} - C & \text{if } x > F^{-1}(0.75) \end{cases}$$

where

$$C = \frac{1}{4}\left[\frac{1}{f(F^{-1}(0.25))} + \frac{1}{f(F^{-1}(0.75))}\right]$$

Note that $\text{IF}(x; \text{IQR}_c, F) = d \cdot \text{IF}(x; \text{IQR}, F)$. Figure 3.2b displays the influence function of $\text{IQR}_c$ for standard Gaussian data. Like the two quartiles $Q_{0.25}$ and $Q_{0.75}$, IQR and $\text{IQR}_c$ have an asymptotic breakdown point equal to 25%. This gain in robustness with respect to the standard deviation is accompanied by a high loss in Gaussian efficiency. Specifically, $\text{ASV}(\text{IQR}_c, \Phi) = 1.3605$ so that the asymptotic Gaussian efficiency of the corrected interquartile range with respect to the standard deviation is equal to $\text{ASV}(\sigma, \Phi)/\text{ASV}(\text{IQR}_c, \Phi) = 0.5/1.3605 \approx 37\%$.

11: Give details about $\text{ASV}(\text{IQR}_c, \Phi)$. How is it computed?

### 3.2.3 The median absolute deviation

Another robust alternative is the *median absolute deviation*, whose empirical version is defined as

$$\text{MAD}_n = d \cdot \underset{i}{\text{med}} \left| x_i - \underset{j}{\text{med}}\, x_j \right|$$

That is, the median absolute deviation is equal to the (rescaled) median of the absolute deviations from the median of the variable of interest. For Gaussian distributions, we need to set $d = 2/\big(\Phi^{-1}(0.75) - \Phi^{-1}(0.25)\big) \approx 1.4826$ to make $\text{MAD}_n$ Fisher-consistent for the scale parameter $\sigma$.

The functional corresponding to $\text{MAD}_n$ is

$$\text{MAD} = \text{MAD}(F) = d \cdot G_F^{-1}(0.5)$$

where $G_F$ is the distribution function of $|X - Q_{0.5}(F)| = |X - F^{-1}(0.5)|$ with $X$ as a random variable of distribution $F$. In other words, MAD is equal to $d$ times the median of the distribution associated with $|X - Q_{0.5}(F)|$, the magnitude of the difference between $X$ and its median.

The median absolute deviation may appear more attractive than the interquartile range for certain purposes. It has the same asymptotic Gaussian efficiency as the corrected interquartile range, but performs better in terms of robustness (e.g., Rousseeuw and Croux, 1993, p. 1273–1274): Its asymptotic breakdown point is as high as 50%. The influence function of MAD is given as $d \cdot \text{IF}(x; G_F^{-1}(0.5), F)$ with

$$\text{IF}(x; G_F^{-1}(0.5), F) = \frac{\text{sign}\big(|x - F^{-1}(0.5)| - G_F^{-1}(0.5)\big) - C \cdot \text{sign}\big(x - F^{-1}(0.5)\big)}{2\big[f\big(F^{-1}(0.5) + G_F^{-1}(0.5)\big) + f\big(F^{-1}(0.5) - G_F^{-1}(0.5)\big)\big]}$$

where

$$C = \frac{f\big(F^{-1}(0.5) + G_F^{-1}(0.5)\big) - f\big(F^{-1}(0.5) - G_F^{-1}(0.5)\big)}{f(F^{-1}(0.5))}$$

(see, e.g., Wilcox, 2005). In particular, under the standard Gaussian distribution (with $F = \Phi$ and $f = \phi$), we have

12: Is it only for Gaussian data that the IF for IQR and MAD is the same, or is this true for any symmetric distribution?

$$\text{IF}(x; \text{MAD}, \Phi) = 1.4826 \cdot \frac{\text{sign}\big(|x| - \Phi^{-1}(0.75)\big)}{4\,\phi(\Phi^{-1}(0.75))}$$

13: In Rousseeuw/Croux the IF of MAD is closer to that one of Q; check that...

which is displayed in Figure 3.2c.

Despite its good robustness properties, MAD is primarily useful for *symmetric* distributions. In fact, the MAD corresponds to finding the symmetric interval around the median that contains 50% of the data (50% of the probability mass), which does not appear to be a very sensible approach for asymmetric distributions. The interquartile range does not have this restriction, as the quartiles need not be equally far away from the median.

### 3.2.4 The $Q_n$ coefficient

Finally, a very interesting but relatively unknown scale estimator is the $Q_n$ statistic introduced by Rousseeuw and Croux (1993):

$$Q_n = d \cdot \{|x_i - x_j|; i < j\}_{(k)}$$

where $d$ is a constant factor allowing $Q_n$ to be a Fisher-consistent estimator for the scale parameter of the underlying distribution $F$ and $k = \binom{h}{2} \approx \binom{n}{2}/4$, with $h = \lfloor n/2 \rfloor + 1$ ($h$ is roughly half the number of observations). In other words, if we omit the constant $d$, the statistic $Q_n$ corresponds approximately to the 0.25 quantile of the $\binom{n}{2}$ distances $|x_i - x_j|$, $i < j$. Fisher-consistency for the scale parameter $\sigma$ at Gaussian distributions can be achieved by setting $d = 1/\big(\sqrt{2}\Phi^{-1}(5/8)\big) \approx 2.2191$.

14: In Rousseeuw/Croux the value is 2.2219, which seems to be an error.

The functional counterpart of $Q_n$ is

$$Q = Q(F) = d \cdot H_F^{-1}(0.25)$$

where $H_F$ is the distribution function of $|X - Y|$ with $X$ and $Y$ being two independent random variables of distribution $F$. Note that $Q(F_n)$ is not exactly the same as $Q_n$, where we take an order statistic among $\binom{n}{2}$ elements instead of $n^2$ elements, but asymptotically this makes no difference.

The scale estimator $Q_n$ has globally better properties than the previous scale estimators we have presented. Like the MAD, it has an asymptotic breakdown point equal to 50%. Yet, unlike the MAD, $Q_n$ is not slanted towards symmetric distributions. Moreover, its influence function is not only bounded, but also smooth:

15: I use numerical integration to obtain the value of the denominator (used for the graph and for the computation of the efficiency). Is there also a closed-form solution?

$$\text{IF}(x; Q, F) = d \cdot \frac{0.25 - F(x + d^{-1}) + F(x - d^{-1})}{\int_{-\infty}^{\infty} f(y + d^{-1})\, dF(y)}$$

[I think the above is only valid if $X$ has a scale of 1. Probably the IF should be

$$\text{IF}(x; Q, F) = d \cdot \frac{0.25 - F(x + Q/d) + F(x - Q/d)}{\int_{-\infty}^{\infty} f(y + Q/d)\, dF(y)}$$

Furthermore, wouldn't the denominator have to be different if the distribution of $X$ is asymmetric? Intuitively, it seems odd that only $f(y+Q/d)$ appears in the denominator and not also $f(y-Q/d)$.] Figure 3.2d displays the influence function of $Q_n$ for standard Gaussian data.

Finally, $Q_n$ is asymptotically more efficient than the median absolute deviation under Gaussian distributions. In particular, numerical integration of $\int_{-\infty}^{\infty} \text{IF}(x; Q, \Phi)^2\, d\Phi(x)$ yields $\text{ASV}(Q, \Phi) \approx .6089$, corresponding to an asymptotic Gaussian relative efficiency with respect to the standard deviation of $\text{ASV}(\sigma, \Phi)/\text{ASV}(Q, \Phi) \approx 0.5/.6089 \approx 82\%$, which is surprisingly high.

16: I get 0.6089, not 0.6077.

17: The following output shows my computations. Results should be precise (e.g., they do not change if I increase the number of integration points to 100000). This is just for you. It will be removed later on.

```
. mata
------------------------------------------------- mata (type end to exit) -----
: d = 1/(sqrt(2)*invnormal(5/8))
: d
  2.219144466
: function myf(x) return(normalden(x+sqrt(2)*invnormal(5/8))*normalden(x))
: // note: normalden(37)=2.12e-298, normalden(38)=0
: dd = mm_integrate_sr(&myf(), -38, 38, 1000, 1)
: dd
  .2681315272
: function myf1(x, d, dd)
> {
>     return( (d * (0.25 - normal(x + 1/d) + normal(x - 1/d)) / dd)^2 *
>               normalden(x))
> }
: mm_integrate_sr(&myf1(), -38, 38, 1000, 1, d, dd)
  .6089006937
: end
--------------------------------------------------------------------------------
```

### 3.2.5 M estimate of scale

[to be completed]

### 3.2.6 Summary

To conclude, Table 3.2 summarizes the different properties of the presented estimators of scale. As is evident, the $Q_n$ coefficient has superior properties in terms of robustness and efficiency compared to the other robust estimators.

Table 3.2: Characteristics of the four scale estimators

| Estimator | Class | Gaussian efficiency | Asymptotic breakdown point | Bounded influence function |
|---|---|---|---|---|
| standard deviation $\sigma_n$ | moment | 100% | 0% | no |
| interquartile range $\text{IQR}_n$ | quantile | 37% | 25% | yes |
| median absolute deviation $\text{MAD}_n$ | quantile | 37% | 50% | yes |
| $Q_n$ coefficient | pairwise | 82% | 50% | yes |

## 3.3 Robust estimation of skewness

### 3.3.1 The Fisher coefficient

As far as skewness is concerned, the most classic estimator is the *Fisher coefficient*. Given sample $\mathbf{X}_n$ the Fisher coefficient is defined as

$$\gamma_{1;n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_n}{\sigma_n}\right)^3$$

with $\mu_n$ and $\sigma_n$ as the sample mean and the standard deviation. This estimator is associated with the functional

$$\gamma_1 = \gamma_1(F) = \mu_3(F)/\sigma(F)^3 \qquad \text{where} \quad \mu_3(F) = \int_{-\infty}^{\infty}(x - \mu(F))^3\, dF(x)$$

which is equal to zero at symmetric $F$.

Since the Fisher coefficient relies on the mean and the standard deviation, it is not surprising that its resistance to outliers is poor. More precisely, its asymptotic breakdown point is equal to 0% and its influence function is unbounded (see, for example, Groeneveld, 1991). For a *symmetric* distribution $F$, assuming $\mu(F) = 0$ and $\sigma(F) = 1$ without loss of generality, the influence function of the Fisher coefficient is given as

18: What exactly does this mean? Is the IF always like that irrespective of $\mu$ and $\sigma$ or is it different, but does not change shape? (In this case: What would be the IF for $\mu \neq 0$ and $\sigma \neq 1$?)

$$\text{IF}(x; \gamma_1, F) = x^3 - 3x$$

See Figure 3.3a for a graphical display. The influence function for an *asymmetric* distribution is more complex. However, although no longer being an odd function of $x$, it has a quite similar form to that found for symmetric $F$. Also note, for comparisons with other skewness estimators, that $\text{ASV}(\gamma_1, \Phi) = 6$.

19: How is this found?

### 3.3.2 Yule and Kendall, and Hinkley skewness measures

Alternative estimators of skewness, such as $(\mu_n - \text{mode}_n)/\sigma_n$ and $(\mu_n - Q_{0.5;n})/\sigma_n$ as proposed by Karl Pearson, are just as fragile with respect to outliers as the standard skewness estimator.

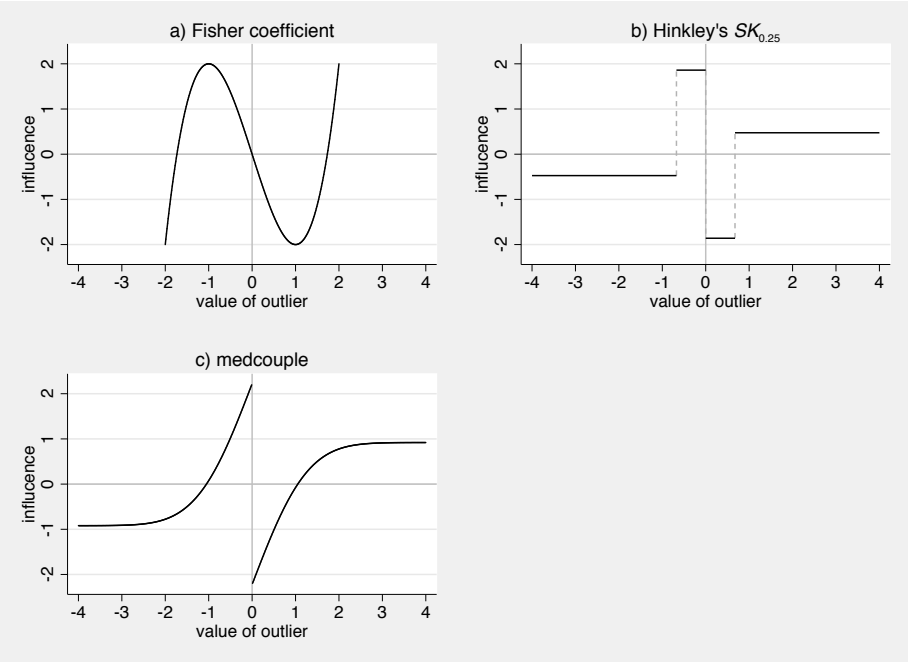


Figure 3.3: Influence functions of $\gamma_1$, $\text{SK}_{0.25}$ and MC under the standard normal distribution

Fortunately, robust alternatives based on quantiles are available. For example, Yule and Kendall (1968) have proposed the skewness measure

$$\text{SK}_{\text{YK}} = \frac{(Q_{0.75} - Q_{0.5}) - (Q_{0.5} - Q_{0.25})}{Q_{0.75} - Q_{0.25}} = \frac{Q_{0.25} + Q_{0.75} - 2Q_{0.5}}{Q_{0.75} - Q_{0.25}}$$

where $Q_{0.25}$, $Q_{0.5}$, and $Q_{0.75}$ are the three quartiles.

Hinkley (1975) generalized this formula to other quantiles:

$$\text{SK}_p = \frac{(Q_{1-p} - Q_{0.5}) - (Q_{0.5} - Q_p)}{Q_{1-p} - Q_p} = \frac{Q_p + Q_{1-p} - 2Q_{0.5}}{Q_{1-p} - Q_p},$$

where $Q_p$ and $Q_{1-p}$ are the quantiles of order $p$ and $1-p$ (with $0 < p < 0.5$). Hinkley's measure $\text{SK}_p$ is equal to zero for symmetric distributions, is positive for right tailed (left skewed) and negative for left tailed (right skewed) distributions. It has a much smaller asymptotic variance under the standard normal distribution than $\gamma_1$. For instance, $\text{ASV}(\text{SK}_{0.25}, \Phi) = 1.8421$. The asymptotic breakdown point of $\text{SK}_p$ is equal to $100p\%$ (in particular, the Yule and Kendall skewness estimator, corresponding to $p = 0.25$, has an asymptotic breakdown point equal to 25%).

20: How is the ASV derived? Plus: what is the expected value of $Q_{0.25}$ for Gaussian data?

For a *symmetric* distribution $F$ with density $f$ and, without loss of generality, $\mu(F) =$

$F^{-1}(0.5) = 0$ and $\sigma(F) = 1$, the influence function of $\text{SK}_p$ is

$$\text{IF}(x; \text{SK}_p, F) = \begin{cases} \frac{-1/f(0)}{F^{-1}(1-p) - F^{-1}(p)} & \text{if } 0 \leq x < F^{-1}(1-p) \\ \frac{1/f\big(F^{-1}(1-p)\big) - 1/f(0)}{F^{-1}(1-p) - F^{-1}(p)} & \text{if } F^{-1}(1-p) \leq x \end{cases}$$

for $x \geq 0$ and $\text{IF}(x; \text{SK}_p, F) = -\text{IF}(-x; \text{SK}_p, F)$ for $x < 0$. The influence function for standard Gaussian data is displayed in Figure 3.3b. In case of an *asymmetric* distribution the influence function is much more complex and is no longer an odd function of $x$ (see Groeneveld, 1991, p. 101).

### 3.3.3 The medcouple

As usual when working with quantiles, the influence function of $\text{SK}_p$ is not smooth. To tackle this problem, Brys et al. (2004) propose to replace the quantiles $Q_p$ and $Q_{1-p}$ in $\text{SK}_p$ by actual data points and introduce a new skewness measure called *medcouple*. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be the $n$ order statistics associated to the sample $\mathbf{X}_n$ and $Q_{0.5;n}$ be the sample median. The medcouple is then defined as

$$\text{MC}_n = \operatorname*{med}_{x_{(i)} \leq Q_{0.5;n} \leq x_{(j)}} h\big(x_{(i)}, x_{(j)}\big)$$

where, for all $x_{(i)} \neq x_{(j)}$, the kernel function $h$ is given as

$$h\big(x_{(i)}, x_{(j)}\big) = \frac{\big(x_{(j)} - Q_{0.5;n}\big) - \big(Q_{0.5;n} - x_{(i)}\big)}{x_{(j)} - x_{(i)}}$$

For the special case $x_{(i)} = x_{(j)} = Q_{0.5;n}$, the kernel is defined as follows: let $m_1 < \cdots < m_k$ denote the indices of the order statistics that are tied to the median $Q_{0.5;n}$ (that is $x_{(m_l)} = Q_{0.5;n}$ for all $l = 1, \ldots, k$). Then,

$$h\big(x_{(m_i)}, x_{(m_j)}\big) = \begin{cases} -1 & \text{if } i + j < k + 1 \\ 0 & \text{if } i + j = k + 1 \\ 1 & \text{if } i + j > k + 1 \end{cases}$$

Due to the denominator it is clear that $h\big(x_{(i)}, x_{(j)}\big)$, and hence $\text{MC}_n$, will always lie between $-1$ and 1 (similar to $\text{SK}_p$).

The functional form of the medcouple is simply defined at any continuous distribution $F$ as

$$\text{MC} = \text{MC}(F) = \operatorname*{med}_{X \leq Q_{0.5} \leq Y} h(X, Y)$$

where $Q_{0.5} = F^{-1}(0.5)$ is the median of $F$ and $X$ and $Y$ are i.i.d. random variables of distribution $F$. The kernel $h$ is the same as above with the finite-sample median $Q_{0.5;n}$ replaced by $Q_{0.5}$. This functional MC is equal to zero in case of a symmetric distribution $F$. It is positive for right tailed (left skewed) and negative left tailed (right skewed) distributions.

The asymptotic breakdown point of the medcouple is equal to 25%, which is the same as for the quartile skewness $\text{SK}_{0.25} = \text{SK}_{\text{YK}}$. The advantage of MC, however, lies in the fact that its influence function resembles a smoothed version of the influence function of $\text{SK}_p$ $(0 < p < 0.5)$. In particular, for standard Gaussian $F$, the influence function is given as

21: Why $\text{SK}_p$? Why not $\text{SK}_{0.25}$?

$$\text{IF}(x; \text{MC}, \Phi) = \pi\left(2\Phi(x) - 1 - \frac{\text{sign}(x)}{\sqrt{2}}\right)$$

(see Figure 3.3c). This leads to an asymptotic variance for Gaussian data of

22: Is 1.25 an exact result? How do you compute that? Plus: what is the expected value of medcouple for Gaussian data?

$$\text{ASV}(\text{MC}, \Phi) = \int_{-\infty}^{\infty} \text{IF}(x; \text{MC}, \Phi)^2 \, d\Phi(x) = 1.25$$

### 3.3.4 Summary

Table 3.3 provides an overview of the properties of the discussed skewness estimators. As can be seen, both proposed robust estimators are more robust and, at the same time, more efficient than the standard skewness coefficient.

23: How to compute efficiency? Need to rescale...

Table 3.3: Characteristics of the three skewness estimators

| Estimator | Class | Gaussian efficiency[a] | Asymptotic breakdown point | Bounded influence function |
|---|---|---|---|---|
| Fisher coefficient $\gamma_{1;n}$ | moment | 100% | 0% | no |
| Hinkley's $\text{SK}_{0.25;n}$ | quantile | ??? | 25% | yes |
| medcouple $\text{MC}_n$ | pairwise | ??? | 25% | yes |

[a] relative to the efficiency of the Fisher coefficient

## 3.4 Robust estimation of the tails heaviness

### 3.4.1 The classical kurtosis coefficient

The classical *kurtosis coefficient* is defined by the functional

$$\gamma_2 = \gamma_2(F) = \mu_4(F)/\sigma(F)^4$$

with

$$\mu_4(F) = \int_{-\infty}^{\infty} (x - \mu(F))^4 \, dF(x)$$

Given a sample $\mathbf{X}_n$, $\gamma_{2;n}$ is computed as

$$\gamma_{2;n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_n}{\sigma_n}\right)^4$$

with $\mu_n$ and $\sigma_n$ as the sample mean and the standard deviation.

The kurtosis coefficient is often considered as a measure of the tail heaviness of a distribution relative to that of the normal distribution. In particular, $\gamma_2$ is equal to three in case of distributions with a tails heaviness similar to the normal distribution, is larger than three for leptokurtic distributions (i.e., distributions with heavier tails than the normal distribution) and is smaller than three for platokurtic distributions (i.e., distributions with lighter tails than the normal distribution). However, since the coefficient also measures the peakedness of a distribution, there is no agreement on what the kurtosis really estimates. Another disadvantage of the kurtosis is that its interpretation, and consequently its use, is restricted to symmetric distributions (due of its intrinsic comparison with the symmetric normal distribution). Moreover, as usual for estimators relying on the mean and the standard deviation, the kurtosis coefficient is very sensitive to outliers in the data. This is reflected in the asymptotic breakdown point being equal to zero. The influence function is unbounded and is given as

$$\text{IF}(x; \gamma_2, F) = (z^2 - \gamma_2)^2 - \gamma_2(\gamma_2 - 1) - 4\gamma_1 z,$$

where $z = (x - \mu(F))/\sigma(F)$, $\gamma_1 = \mu_3(F)/\sigma(F)^3$ and $\gamma_2 = \mu_4(F)/\sigma(F)^4$ (see Ruppert, 1987). See Figure 3.4a for a graphical display of the influence function for standard Gaussian $F$, which is given as $\text{IF}(x; \gamma_2, \Phi) = (x^2 - 3)^2 - 6$. The form of the influence function indicates that contamination at the center has far less influence than that in the extreme tails. This suggests that $\gamma_2$ is primarily a measure of tail behavior, and only to a lesser extent of peakedness. The asymptotic variance of $\gamma_2$ for Gaussian data is given as $\text{ASV}(\gamma_2, \Phi) = 24$.

24: Closed form solution?

### 3.4.2 The quantile and medcouple tail weight measures

[There should also be a section on the quantile tail weight measure defined as

$$\text{QW}_p = \frac{(Q_{1-p/2} - Q_{0.5+p/2}) + (Q_{0.5-p/2} - Q_{p/2})}{Q_{1-p} - Q_p}$$

because this is used in one of the normality tests. See Brys et al. (2008, p. 432).]

To overcome the problems of the kurtosis coefficient, Brys et al. (2006) have proposed two measures of *left* and *right* tail weight for univariate continuous distributions. As discussed below, these measures have the advantage that they can be applied to symmetric as well as asymmetric distributions that do not need to have finite moments. Moreover, their interpretation is unambiguous and they are robust against outlying values.

More precisely, Brys et al. (2006) defined *left* and *right* tail measures as measures of skewness that are applied to the half of the probability mass lying on the left side or on the right side of the median $Q_{0.5}$ of the distribution $F$, respectively. As candidate measures of skewness they use both, $\text{SK}_p$ $(0 < p < 0.5)$ and MC (see above).

Recall that $\text{SK}_p$ is a measure of skewness of the distribution $F$ around $Q_{0.5}$, involving the quantiles $Q_p$ and $Q_{1-p}$ of orders $p$ and $(1-p)$ of $F$. Applying $-\text{SK}_{p/2}$ to the left half of the distribution $F$ (i.e., to $x < Q_{0.5}$) leads to the *Left Quantile Weight*

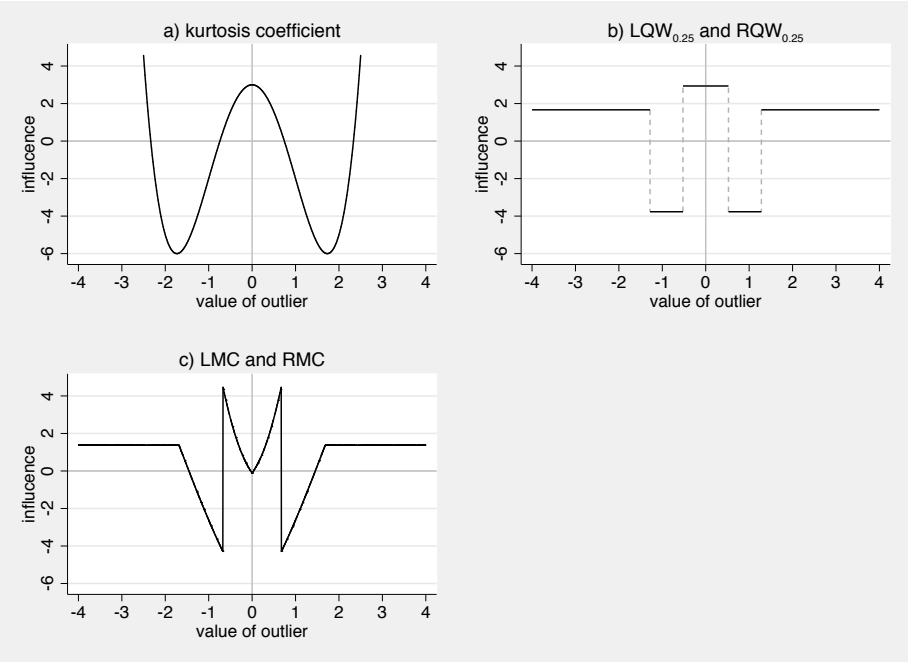


Figure 3.4: Influence functions of $\gamma_2$, $\text{LQW}_{0.25}$ and $\text{RQW}_{0.25}$, LMC and RMC under the standard normal distribution

$\text{LQW}_p = \text{LQW}_p(F)$, which corresponds to (the opposite of) a measure of skewness of the left half of $F$ around the first quartile $Q_{0.25}$ involving the quantiles $Q_{p/2}$ and $Q_{0.5-p/2}$:

$$\text{LQW}_p = -\frac{(Q_{0.5-p/2} - Q_{0.25}) - (Q_{0.25} - Q_{p/2})}{Q_{0.5-p/2} - Q_{p/2}} = -\frac{Q_{p/2} + Q_{0.5-p/2} - 2Q_{0.25}}{Q_{0.5-p/2} - Q_{p/2}}.$$

Similarly, applying $\text{SK}_{p/2}$ to the right half of the distribution $F$ (i.e., to $x > Q_{0.5}$) provides the *Right Quantile Weight* $\text{RQW}_p = \text{RQW}_p(F)$ which corresponds to a measure of skewness of the right half of $F$ around the third quartile $Q_{0.75}$ involving the quantiles $Q_{0.5+p/2}$ and $Q_{1-p/2}$:

$$\text{RQW}_p = \frac{(Q_{1-p/2} - Q_{0.75}) - (Q_{0.75} - Q_{0.5+p/2})}{Q_{1-p/2} - Q_{0.5+p/2}} = \frac{Q_{0.5+p/2} + Q_{1-p/2} - 2Q_{0.75}}{Q_{1-p/2} - Q_{0.5+p/2}}.$$

With $p = 1/4 = 0.25$, we obtain

$$\text{LQW}_{0.25} = -\frac{Q_{0.125} + Q_{0.375} - 2Q_{0.25}}{Q_{0.375} - Q_{0.125}}$$

$$\text{RQW}_{0.25} = \frac{Q_{0.625} + Q_{0.875} - 2Q_{0.75}}{Q_{0.875} - Q_{0.625}}$$

Note that the sample versions $\text{LQW}_{p;n}$ and $\text{RQW}_{p;n}$ are easily found by using the quantiles of $F_n$, the empirical distribution function of $\mathbf{X}_n$.

As with LQW and RQW, we can also apply the MC to each side of the distribution, leading to the *Left Medcouple* (LMC) and to the *Right Medcouple* (RMC), defined as

$$\text{LMC} = \text{LMC}(F) = -\text{MC}(x < Q_{0.5})$$

$$\text{RMC} = \text{RMC}(F) = \text{MC}(x > Q_{0.5})$$

By using $\text{MC}_n$, the finite-sample version of MC, we obtain the finite-sample versions $\text{LMC}_n$ and $\text{RMC}_n$.

Since both the quantile and the medcouple tail weight measures only depend on quantiles, they are given for any distribution, even for distributions without finite moments. Note that LQW and RQW require to fix the parameter $p$ in advance (depending on the degree of robustness one wants to attain), whereas LMC and RMC do not require any additional parameter to be set.

Some general properties of the tail weight measures are as follows. Let $X$ be a random variable with continuous distribution $F_X$. Furthermore, let W stand for any of the defined tail weight measures; let LW stand for left tailed measures and RW for the right tailed measures. Then:

- Like the skewness measures $\text{SK}_p$ and MC, W is location and scale invariant. That is, $\text{W}(F_{aX+b}) = \text{W}(F_X)$.
- $\text{LW}(F_{-X}) = \text{RW}(F_X)$.
- If $F$ is symmetric, then $\text{LW}(F) = \text{RW}(F)$.
- $\text{W} \in [-1, 1]$.

The medcouple tail weight measures can resist up to 12.5% outliers in the data. In a similar way, it can be shown that the left and right quantile tail weight measures $\text{LQW}_p$ and $\text{RQW}_p$ have an asymptotic breakdown point equal to $(100p/2)\%$; in particular, $\text{LQW}_{0.25}$ and $\text{RQW}_{0.25}$ have the same asymptotic breakdown point as LMC and RMC, and $\text{LQW}_{0.125}$ and $\text{RQW}_{0.125}$ have an asymptotic breakdown point of 6.25%.

Moreover, the influence functions of LMC and RMC are smooth versions of the influence functions of $\text{LQW}_{0.25}$ and $\text{RQW}_{0.25}$, as shown in Figure 3.4b and c for standard Gaussian $F$.[1] All these influence functions are bounded. More precisely, if $F$ is a continuous distribution with density $f$ and if we denote, as usual, the quantile of order $p$ of $F$ by $Q_p = F^{-1}(p)$, we have

$$\text{IF}(x; \text{LQW}_p, F) = 2\,\frac{\begin{array}{r}\big(\text{IF}(x; Q_{0.25}, F) - \text{IF}(x; Q_{0.5-p/2}, F)\big)\big(Q_{0.25} - Q_{p/2}\big) \\ -\big(\text{IF}(x; Q_{p/2}, F) - \text{IF}(x; Q_{0.25}, F)\big)\big(Q_{0.5-p/2} - Q_{0.25}\big)\end{array}}{\big(Q_{0.5-p/2} - Q_{p/2}\big)^2}$$

[1] Figure 3.4 shows only the left part (defined on $\mathbb{R}^-$) of the influence functions of LMC and $\text{LQW}_{0.25}$, and the right part (defined on $\mathbb{R}^+$) of the influence functions of RMC and $\text{RQW}_{0.25}$.

and

$$\text{IF}(x; \text{RQW}_p, F) = 2 \frac{\begin{array}{r}\big(\text{IF}(x; Q_{0.5+p/2}, F) - \text{IF}(x; Q_{0.75}, F)\big)\big(Q_{1-p/2} - Q_{0.75}\big) \\ -\big(\text{IF}(x; Q_{0.75}, F) - \text{IF}(x; Q_{1-p/2}, F)\big)\big(Q_{0.75} - Q_{0.5+p/2}\big)\end{array}}{\big(Q_{1-p/2} - Q_{0.5+p/2}\big)^2}$$

with

$$\text{IF}(x; Q_p, F) = \frac{p - \text{I}[x < Q_p]}{f(Q_p)}$$

The expression of the influence functions of LMC and RMC is more complex and can be found in Brys et al. (2006, p. 740–741).

Finally, the asymptotic variances of the left and right tail weight measures under Gaussian distributions are much smaller than the variance of the classical kurtosis coefficient $\gamma_2$. In particular:

$$\text{ASV}(\text{LQW}_{0.25}, \Phi) = \text{ASV}(\text{RQW}_{0.25}, \Phi) = 3.71$$
$$\text{ASV}(\text{LQW}_{0.125}, \Phi) = \text{ASV}(\text{RQW}_{0.125}, \Phi) = 2.23$$

and

$$\text{ASV}(\text{LMC}, \Phi) = \text{ASV}(\text{RMC}, \Phi) = 2.62$$

25: (maybe have a look and then check computation of graph; it seems IF has a discontinuity; need to use shortdash)

26: How are these numbers computed? Furthermore, is it a fair comparison. That is, are L/RQW and L/RMC similar in size to the kurtosis for Gaussian data or do they have to be rescaled? Only if the measures have the same size the variances can be compared.

### 3.4.3 Summary

Table 3.4 summarizes the properties of the presented tails heaviness estimators. [Another sentence needed here!]

27: How to compute efficiency? (need to rescale)

Table 3.4: Characteristics of the three tails heaviness estimators

| Estimator | Class | Gaussian efficiency[a] | Asymptotic breakdown point | Bounded influence function |
|---|---|---|---|---|
| kurtosis coefficient $\gamma_{2;n}$ | moment | 100% | 0% | no |
| $\text{LQW}_{0.25;n}$ and $\text{RQW}_{0.25;n}$ | quantile | ??? | 12.5% | yes |
| $\text{LMC}_n$ and $\text{RMC}_n$ | pairwise | ??? | 12.5% | yes |

[a] relative to the efficiency of the kurtosis coefficient

## 3.5 Variance estimation

[to be completed]

## 3.6 Example

28: The examples will be replaced later by the `robstat` command. Possibly, it would be good to split the example into parts and include the parts at appropriate placed in the sections above.

As an illustrative example we will generate two datasets (of size $n = 1000$), one drawn from a standard normal distribution and one drawn from a chi-square distribution with one degree of freedom. All descriptive statistics presented above will be calculated for both samples. To simplify interpretation we will present the excess kurtosis rather than the kurtosis (in other words, the reported tail heaviness statistics are equal to zero for the normal distribution). We then contaminate the datasets by replacing a random selection of 5% of the observations by value 5 for the normally distributed sample and by $F^{-1}_{\chi^2_1}(\Phi(5))$ for the chi-square distributed sample, $F^{-1}_{\chi^2_1}$ and $\Phi$ are the quantile function of the $\chi^2_1$ distribution and the normal cumulative distribution function, respectively. In this way the degree of outlyingness is comparable between the two setups.

When we compare the classical, quantile-based and pairwise-based estimates obtained for the normally distributed dataset free of outliers (see the upper part of table 3.5), we do not see big differences between the three types of approaches. Indeed, the estimates all point towards a symmetrical distribution centered at zero with non-excessive tails and a dispersion of about one. If one looks at these statistics for the case of the chi-squared distributed data (see the lower part of table 3.5), the location estimate is, as expected, not the same since the mean is more attracted by the tail than the robust competitors. A similar phenomenon occurs for skewness and tails heaviness.

Table 3.5: The estimates of location, scale, skewness and tails heaviness in the original (uncontaminated) datasets

| | | | | Tails heaviness | |
|---|---|---|---|---|---|
| | Location | Scale | Skewness | Left | Right |
| *Normally distributed sample* | | | | | |
| Classical | 0.015 | 0.977 | −0.006 | 0.157 | |
| Quantile-based | 0.031 | 0.946 | −0.088 | 0.025 | 0.0139 |
| Pairwise-based | 0.017 | 0.973 | −0.024 | 0.038 | 0.075 |
| *Chi-square distributed sample* | | | | | |
| Classical | 0.993 | 1.406 | 2.322 | 6.323 | |
| Quantile-based | 0.414 | 0.908 | 2.491 | −0.439 | 0.135 |
| Pairwise-based | 0.652 | 0.496 | 0.539 | −0.491 | 0.190 |

When the dataset is contaminated by a small portion of outliers, the classical statistics change substantially, while the effect on their robust equivalent is only marginal (table 3.6). For example, for the normal case, classical statistics would point towards a right-tailed skewed distribution with relatively large dispersion and big-tail heaviness. The robust counterparts would still point towards the standard normal distribution. A similar phenomenon is observable for the chi-square.

```
set seed 1234
```

Table 3.6: The estimates of location, scale, skewness and tails heaviness in the contaminated datasets

| | | | | Tails heaviness | |
|---|---|---|---|---|---|
| | Location | Scale | Skewness | Left | Right |
| *Normally distributed sample* | | | | | |
| Classical | 0.263 | 1.447 | 1.530 | 3.465 | |
| Quantile-based | 0.086 | 1.012 | 0.016 | 0.023 | 0.055 |
| Pairwise-based | 0.109 | 1.072 | 0.041 | 0.030 | 0.145 |
| *Chi-square distributed sample* | | | | | |
| Classical | 1.011 | 1.416 | 2.311 | 6.276 | |
| Quantile-based | 0.437 | 0.923 | 2.305 | −0.458 | 0.135 |
| Pairwise-based | 0.672 | 0.518 | 0.522 | −0.492 | 0.189 |

```
clear
set obs 1000

drawnorm z
gen x=invchi2(1,uniform())

**** UNCONTAMINATED NORMAL  ******

qui sum z, d
local meanN=r(mean)
local medianN=r(p50)

local sdN=r(sd)
local iqrN=(r(p75)-r(p25))*0.7413
local skewN=r(skewness)
local kurtN=r(kurtosis)-3

centile z, centile(25 50 75)
local qskewN=(r(c_1)+r(c_3)-2*r(c_2))/((r(c_2)-r(c_1))/1.349)

qui hl z
local hlN=e(hl)

qui qn z
local qnN=e(qn)

qui medcouple z, lmc rmc
local mcN=e(mc)
local lmcN=e(lmc)
local rmcN=e(rmc)

qui centile z, centile(12.5 25 37.5 62.5 75 87.5)
local lqwN=-(r(c_1)+r(c_3)-2*r(c_2))/(r(c_3)-r(c_1))
local rqwN=(r(c_4)+r(c_6)-2*r(c_5))/(r(c_6)-r(c_4))

di in r "Location parameters Normal"

di "Mean: " `meanN'
di "Median: " `medianN'
di "Hodges-Lehman: " `hlN'

di in r"Scale parameters Normal"
di "Standard deviation: " `sdN'
di "Iqr: " `iqrN'
```

```
di "Qn: " `qnN'

di in r "Skewness parameters Normal"
di "Skewness: " `skewN'
di "Quantile skewness: " `qskewN'
di "Mecouple: " `mcN'

di in r "Tail heavyness parameters Normal"
di "Kurtosis: " `kurtN'
di "Quantile right heaviness: " `rqwN'
di "Quantile left heaviness: " `lqwN'
di "Right mecouple: " `rmcN'
di "Left mecouple: " `lmcN'

**** UNCONTAMINATED CHI2 *****

qui sum x, d
local meanCHI2=r(mean)
local medianCHI2=r(p50)

local sdCHI2=r(sd)
local iqrCHI2=(r(p75)-r(p25))*0.7413
local skewCHI2=r(skewness)
local kurtCHI2=r(kurtosis)-3

centile x, centile(25 50 75)
local qskewCHI2=(r(c_1)+r(c_3)-2*r(c_2))/((r(c_2)-r(c_1))/1.349)

qui hl x
local hlCHI2=e(hl)

qui qn x
local qnCHI2=e(qn)

qui medcouple x, lmc rmc
local mcCHI2=e(mc)
local lmcCHI2=e(lmc)
local rmcCHI2=e(rmc)

qui centile x, centile(12.5 25 37.5 62.5 75 87.5)
local lqwCHI2=-(r(c_1)+r(c_3)-2*r(c_2))/(r(c_3)-r(c_1))
local rqwCHI2=(r(c_4)+r(c_6)-2*r(c_5))/(r(c_6)-r(c_4))

di in r "Location parameters CHI2"

di "Mean: " `meanCHI2'
di "Median: " `medianCHI2'
di "Hodges-Lehman: " `hlCHI2'

di in r"Scale parameters CHI2"
di "Standard deviation: " `sdCHI2'
di "Iqr: " `iqrCHI2'
di "Qn: " `qnCHI2'

di in r "Skewness parameters CHI2"
di "Skewness: " `skewCHI2'
di "Quantile skewness: " `qskewCHI2'
di "Mecouple: " `mcCHI2'

di in r "Tail heavyness parameters CHI2"
di "Kurtosis: " `kurtCHI2'
di "Quantile right heaviness: " `rqwCHI2'
di "Quantile left heaviness: " `lqwCHI2'
di "Right mecouple: " `rmcCHI2'
di "Left mecouple: " `lmcCHI2'

**** CONTAMINATED NORMAL  ******

replace z=5 in 1/50
```

```
qui sum z, d
local meanNc=r(mean)
local medianNc=r(p50)
local sdNc=r(sd)
local iqrNc=(r(p75)-r(p25))*0.7413
local skewNc=r(skewness)
local kurtNc=r(kurtosis)-3
centile z, centile(25 50 75)
local qskewNc=(r(c_1)+r(c_3)-2*r(c_2))/((r(c_2)-r(c_1))/1.349)
qui hl z
local hlNc=e(hl)
qui qn z
local qnNc=e(qn)
qui medcouple z, lmc rmc
local mcNc=e(mc)
local lmcNc=e(lmc)
local rmcNc=e(rmc)
qui centile z, centile(12.5 25 37.5 62.5 75 87.5)
local lqwNc=-(r(c_1)+r(c_3)-2*r(c_2))/(r(c_3)-r(c_1))
local rqwNc=(r(c_4)+r(c_6)-2*r(c_5))/(r(c_6)-r(c_4))
di in r "Location parameters Normal"
di "Mean: " `meanNc'
di "Median: " `medianNc'
di "Hodges-Lehman: " `hlNc'
di in r"Scale parameters Normal"
di "Standard deviation: " `sdNc'
di "Iqr: " `iqrNc'
di "Qn: " `qnNc'
di in r "Skewness parameters Normal"
di "Skewness: " `skewNc'
di "Quantile skewness: " `qskewNc'
di "Mecouple: " `mcNc'
di in r "Tail heavyness parameters Normal"
di "Kurtosis: " `kurtNc'
di "Quantile right heaviness: " `rqwNc'
di "Quantile left heaviness: " `lqwNc'
di "Right mecouple: " `rmcNc'
di "Left mecouple: " `lmcNc'
**** CONTAMINATED CHI2 *****
replace x=invchi2(1,normal(5)) in 1/50
qui sum x, d
local meanCHI2c=r(mean)
local medianCHI2c=r(p50)
local sdCHI2c=r(sd)
local iqrCHI2c=(r(p75)-r(p25))*0.7413
local skewCHI2c=r(skewness)
local kurtCHI2c=r(kurtosis)-3
centile x, centile(25 50 75)
local qskewCHI2c=(r(c_1)+r(c_3)-2*r(c_2))/((r(c_2)-r(c_1))/1.349)
qui hl x
local hlCHI2c=e(hl)
qui qn x
local qnCHI2c=e(qn)
```

```
qui medcouple x, lmc rmc
local mcCHI2c=e(mc)
local lmcCHI2c=e(lmc)
local rmcCHI2c=e(rmc)
qui centile x, centile(12.5 25 37.5 62.5 75 87.5)
local lqwCHI2c=-(r(c_1)+r(c_3)-2*r(c_2))/(r(c_3)-r(c_1))
local rqwCHI2c=(r(c_4)+r(c_6)-2*r(c_5))/(r(c_6)-r(c_4))
di in r "Location parameters CHI2c"
di "Mean: " `meanCHI2c'
di "Median: " `medianCHI2c'
di "Hodges-Lehman: " `hlCHI2c'
di in r"Scale parameters CHI2c"
di "Standard deviation: " `sdCHI2c'
di "Iqr: " `iqrCHI2c'
di "Qn: " `qnCHI2c'
di in r "Skewness parameters CHI2c"
di "Skewness: " `skewCHI2c'
di "Quantile skewness: " `qskewCHI2c'
di "Mecouple: " `mcCHI2c'
di in r "Tail heavyness parameters CHI2c"
di "Kurtosis: " `kurtCHI2c'
di "Quantile right heaviness: " `rqwCHI2c'
di "Quantile left heaviness: " `lqwCHI2c'
di "Right mecouple: " `rmcCHI2c'
di "Left mecouple: " `lmcCHI2c'
```

## 3.7 Robust tests of normality

[Need to revise this section to include all normality tests computed by `robstat`. Also need to include details on the Wald-style variant of the tests.]

The estimates of location, scale, skewness and tails heaviness can be used to characterize the underlying distribution. In particular, they can be used to test for normality. For example, Jarque and Bera (1980) have proposed a normality test relying on the classical skewness and kurtosis coefficients. More precisely, under the normality assumption ($\gamma_1 = 0$ and $\gamma_2 = 3$), we have

$$\sqrt{n}\begin{bmatrix} \gamma_{1;n} \\ \gamma_{2;n} - 3 \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}\right)$$

which leads to the Jarque-Bera test statistic

$$T = n\left(\frac{\gamma_{1;n}^2}{6} + \frac{(\gamma_{2;n} - 3)^2}{24}\right) \approx \chi_2^2$$

The Jarque-Bera test is a very popular and interesting test for normality. It has been shown that, for a wide range of alternative distributions, it outperforms tests such as the Kolmogorov-Smirnov test, the Cramér-von Mises test and the Durbin test. Unfortunately, despite its good power properties and computational simplicity, the Jarque-Bera test is highly sensitive to outliers because it is constructed from the moment-based skewness and kurtosis measures.

Robust alternatives to the Jarque-Bera test have been proposed and studied in Brys et al. (2008). The authors start from the fact that the Jarque-Bera test can be seen as a special case of the following general testing procedure. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)'$ be a vector of estimators of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ (a vector of characteristic parameters of the underlying distribution) such that, under the null hypothesis of normality,

$$\sqrt{n}\Big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big)' \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$$

Then, the general test consists in rejecting, at level $\alpha$, the null hypothesis of normality if

$$T = n\Big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big)' \boldsymbol{\Omega}^{-1} \Big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big) > \chi^2_{p;1-\alpha}$$

where $\chi^2_{p;1-\alpha}$ is the $(1-\alpha)$-quantile of the chi-square distribution with $p$ degrees of freedom. Brys et al. (2008) then propose to use, in this general testing procedure, the robust skewness estimator $\text{MC}_n$ or the tails heaviness estimators $\text{LMC}_n$ and $\text{RMC}_n$.

Three tests have been studied. The first one is only based on the skewness estimator $\text{MC}_n$ (the medcouple). In this case, $k = 1$, $\widehat{\theta} = \text{MC}_n$ and $\Omega = 1.25$. The second one is based on the left and right tail heaviness estimators $\text{LMC}_n$ (left medcouple) and $\text{RMC}_n$ (right medcouple). In this case, $k = 2$, $\widehat{\boldsymbol{\theta}} = (\text{LMC}_n, \text{RMC}_n)'$, $\boldsymbol{\theta} = (0.199, 0.199)'$ and

$$\boldsymbol{\Omega} = \begin{bmatrix} 2.62 & -0.0123 \\ -0.0123 & 2.62 \end{bmatrix}$$

The third test combines $\text{MC}_n$, $\text{LMC}_n$ and $\text{RMC}_n$. In this case, $k = 3$, $\widehat{\boldsymbol{\theta}} = (\text{MC}_n, \text{LMC}_n, \text{RMC}_n)'$, $\boldsymbol{\theta} = (0, 0.199, 0.199)'$ and

$$\boldsymbol{\Omega} = \begin{bmatrix} 1.25 & 0.323 & -0.323 \\ 0.323 & 2.62 & -0.0123 \\ -0.323 & -0.0123 & 2.62 \end{bmatrix}$$

This last test seems to have the best overall performance.

### Example

29: The example will be revised so that the relevant Stata code is visible.

We will analyze the body weight of 64 different animal species. The dataset we use is available online.[2] These data have been made available by Rice University, University of Houston Clear Lake and Tufts University.

To start the analysis, we first calculate the classic estimators of location, scale, skewness and kurtosis using Stata's `summary` command with the `detail` option. The results are presented in the first line of Table 3.7. In the second line of the table we present the results obtained using the commands for robust estimators (that is, `hl`, `qn`, `medcouple` and `medcouple` with the `lmc` and `rmc` options).

If we would only look at the classic estimators, we would conclude that the average animal weight is very high, but with a huge dispersion. The asymmetry is large and positive and tails are very heavy. When we look at the equivalent robust statistics, we

[2] See `http://onlinestatbook.com/stat\_sim/transformations/body\_weight.html`

Table 3.7: Classic estimates of location, scale, skewness and tails heaviness as well as estimates based on pairwise combinations

| | Location | Dispersion | Skewness | Tails |
|---|---|---|---|---|
| Classic | $\mu_n = 3{,}111{,}355$ | $\sigma_n = 1.3 \times 10^7$ | $\gamma_{1;n} = 5.461$ | $\gamma_{2;n} = 32.77$ |
| Robust | $Q_{0.5;n} = 3{,}500$ $\text{HL}_n = 94{,}307$ | $Q_n = 6{,}667.5$ | $\text{MC}_n = 0.985$ | $\text{LMC}_n = -0.090$ $\text{RMC}_n = 0.915$ |

see that the median weight is much lower than the mean weight. The robust dispersion is also much smaller than that suggested by the standard deviation and right skewness is extreme. As far as the heaviness of the tails is concerned, the right tail is extremely heavy while the left one is similar to the left tail of the normal (even slightly lighter). When looking at the difference between classical and robust estimators, it is evident that outliers are present in the dataset.

A first way to tackle this problem is to transform the data to reduce the excessive importance of very big animals (such as dinosaurs). Given that weights are strictly positive, we consider a logarithmic transformation and redo the above descriptive statistics analysis (see Table 3.8).

Table 3.8: Classic estimates of location, scale, skewness and tails heaviness as well as estimates based on pairwise combinations based on transformed data

| | Location | Dispersion | Skewness | Tails |
|---|---|---|---|---|
| Classic | $\mu_n = 9.313$ | $\sigma_n = 4.135$ | $\gamma_{1;n} = 0.304$ | $\gamma_{2;n} = 2.192$ |
| Robust | $Q_{0.5;n} = 8.161$ $\text{HL}_n = 9.289$ | $Q_n = 4.281$ | $\text{MC}_n = 0.386$ | $\text{LMC}_n = 0.515$ $\text{RMC}_n = 0.241$ |

When we do this transformation, we see that the differences between classic and robust estimators become much smaller. Indeed the mean is only slightly larger than the median, the dispersion estimate is very similar for both methods as well as the skewness estimate that only points towards evidence of very moderate positive skewness. As far as the heaviness of the tails is concerned, the classic estimator is close to 3 which is the value of the kurtosis of the normal distribution and therefore points towards standard tails. Nevertheless when we look at the robust estimate for the latter, there is evidence of a heavy left tail. This last point is very important.

The classic and the robust tests for the normality of the log-transformed body weight variable lead to different findings. The standard Jarque-Bera statistic is 2.726, which is much smaller than the critical value of $\chi^2_{2;0.95} = 5.99$. That is, the standard Jarque-Bera test does not reject the null hypothesis of normality. On the other hand, the robust test statistic involving $\text{MC}_n$, $\text{LMC}_n$ and $\text{RMC}_n$ is equal to 9.266, which is larger than the critical value of $\chi^2_{3;0.95} = 7.815$. That is, the null hypothesis of normality is rejected by the robust test. Even though the logarithmic transformation substantially reduces

the effect of atypical observations, outliers still bias the classic test. In particular, we believe that the heaviness of the left tail is not satisfactorily identified by the classic kurtosis coefficient, and this affects the result of the normality test.

◁

## 3.8 Robust boxplots

As stated by Bruffaerts et al. (2014), among others, the boxplot is without any doubt the most commonly used tool to represent the distribution of the data and identify atypical observations in a univariate dataset. An observation is considered as atypical (or extreme) when it is above the upper whisker or below the lower whisker. An important issue with the standard boxplot is that, as soon as asymmetry or tail heaviness appears, the percentage of values identified as atypical becomes excessive. To cope with this, Hubert and Vandervieren (2008) proposed an *adjusted* boxplot for skewed data. Their idea is to modify the whiskers according to the degree of asymmetry in the data, which can be robustly measured by the medcouple. Alternatively, Bruffaerts et al. (2014) propose to apply a simple rank-preserving transformation on the original data so that the transformed observations can be adjusted by a so-called *Tukey g-and-h distribution*. Using the quantiles of this distribution, it is then relatively easy to recover whiskers of the boxplot related to the original data. Given the result of simulations, the latter seems to be more efficient and we therefore concentrate on that too here.

### 3.8.1 The classic boxplot and the adjusted boxplot

In a univariate setup, an observation is often considered as atypical as soon as its value does not belong to the interval $[Q_{0.25} - 1.5\,\text{IQR}; Q_{0.75} + 1.5\,\text{IQR}]$, where $Q_{0.25}$ and $Q_{0.75}$ are the first and third quartiles, and IQR is the interquartile range. For Gaussian data, approximately 0.7% of the observations will lie outside this interval. Unfortunately, as soon as asymmetry or tail heaviness appears, the percentage of values detected as atypical becomes excessively high. To deal with the above drawbacks of the standard boxplot, Hubert and Vandervieren (2008) suggested to use an alternative boxplot, called the adjusted boxplot, where the interval for the boxplot is

$$\left[Q_{0.25} - 1.5e^{-4\,\text{MC}}\text{IQR},\, Q_{0.75} + 1.5e^{3\,\text{MC}}\text{IQR}\right] \quad \text{if MC} \geq 0$$

and

$$\left[Q_{0.25} - 1.5e^{-3\,\text{MC}}\text{IQR},\, Q_{0.75} + 1.5e^{4\,\text{MC}}\text{IQR}\right] \quad \text{if MC} < 0$$

where MC is the medcouple.

Although this rule works well for most commonly used distributions, it presents some limitations and drawbacks (see Bruffaerts et al., 2014), the most restrictive probably being that it does not deal with excessive tail heaviness. Bruffaerts et al. (2014) deal with most of these limitations. Since this method relies on the Tukey $g$-and-$h$ distribution, we briefly describe this distribution before explaining the methodology of Bruffaerts et al. (2014).

### 3.8.2 The Tukey $g$-and-$h$ distribution

The Tukey $g$-and-$h$ family of distributions covers a large variety of distributions which can substantially differ from normality in both skewness and heaviness of the tails. If $Z$ is a random variable with standard normal distribution, and $g$ and $h$ are two constants ($g \neq 0$, $h \in \mathbb{R}$), then the random variable $Y$ given by

$$Y = \frac{1}{g}(\exp(gZ) - 1)\exp(hZ^2/2)$$

is distributed as a Tukey $g$-and-$h$ distribution, that is $Y \sim T(g, h)$.[3] The constants $g$ and $h$ control the skewness and the tail weight (or elongation) of the distribution, respectively. They can be estimated from the empirical quantiles[4] $Q_{1-p}(\{y_j\})$ and $Q_p(\{y_j\})$ of order $(1-p)$ and $p$ $(0.5 < p < 1)$ of $n$ independent realizations $\{y_1, \ldots, y_n\}$ of $Y$ (see Jiménez and Arunachalam, 2011). Following Jiménez and Arunachalam (2011), $g$ and $h$ can be estimated as follows:

30: Why such complicated notation?

$$\widehat{g} = \frac{1}{z_p}\ln\left(-\frac{Q_p(\{y_j\})}{Q_{1-p}(\{y_j\})}\right) \quad \text{and} \quad \widehat{h} = \frac{2\ln\left(-\widehat{g}\,\frac{Q_p(\{y_j\})Q_{1-p}(\{y_j\})}{Q_p(\{y_j\})+Q_{1-p}(\{y_j\})}\right)}{z_p^2} \tag{3.1}$$

where $z_p$ is the quantile of order $p$ of the standard normal distribution. Note that the chosen order $p$ for the quantiles determines the robustness of the method with respect to outliers. For example, if we set $p = 0.9$, the breakdown point of the estimators of $g$ and $h$ is set to $1-p = 10\%$, as the method provides meaningful results if there are up to 10% of outliers. For estimation purposes, we suggest using $p = 0.9$, except if one believes that the contamination rate is larger than 10%. Working with a lower value for $p$ would increase robustness with respect to outliers, but at the cost of lowering the efficiency. Furthermore, it would not make sense to work with a $p \leq 0.75$, as a contamination of more than 25% would make the first and/or third quartile of the boxplot break down.

### 3.8.3 A generalized boxplot

The method proposed by Bruffaerts et al. (2014) overcomes most of the limitations of the adjusted boxplot. Based on an initial dataset $\{x_1, \ldots, x_n\}$, their procedure is as follows.

1. Reduce the data by a scale factor $s_0$:

$$x_i^* = \frac{x_i}{s_0}, \quad i = 1, \ldots, n$$

with $s_0 = \text{IQR}(\{x_j\})$, where $\text{IQR}(\{x_j\})$ is the interquartile range of the series $\{x_1, \ldots, x_n\}$.

[3] Note that $Y$ is a strictly increasing transformation of $Z$, driven by the values of $g$ and $h$. Hence, for every order $p \in (0, 1)$, $y_p = \frac{1}{g}(\exp(gz_p) - 1)\exp(hz_p^2/2)$, where $y_p$ and $z_p$ are the quantiles of order $p$ of the distributions of $Y$ and $Z$ respectively. This implies in particular that the median $y_{0.5}$ of $Y$ is equal to zero.

[4] The expression $Q_p(\{y_j\})$ denotes the empirical quantile of order $p$ related to the series $\{y_1, \ldots, y_n\}$. The notation $\min(\{x_j\})$ and $\max(\{x_j\})$ is later used to define the minimum and maximum values of the series $\{x_1, \ldots, x_n\}$.

2. Shift the dataset to obtain only strictly positive values: compute

$$r_i = x_i^* - \min(\{x_j^*\}) + \zeta, \quad i = 1, \ldots, n$$

where $\zeta > 0$ is a small quantity. They propose to use $\zeta = 0.1$

3. Standardize the values obtained in step 2 in order to obtain new values belonging to the open interval $(0, 1)$: compute

$$\widetilde{r}_i = \frac{r_i}{\min(\{r_j\}) + \max(\{r_j\})}, \quad i = 1, \ldots, n$$

4. Apply the inverse normal (probit) transformation

$$w_i = \Phi^{-1}(\widetilde{r}_i), \quad i = 1, \ldots, n$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal.

5. Center and reduce the values $w_i$: compute

$$w_i^* = \frac{w_i - Q_{0.5}(\{w_j\})}{\text{IQR}(\{w_j\})/1.3426}, \quad i = 1, \ldots, n$$

where $Q_{0.5}(\{w_j\})$ and $\text{IQR}(\{w_j\})$ are the median and the interquartile range of the series $\{w_1, \ldots, w_n\}$. The constant 1.3426 ensures, in the Gaussian case, the consistency of the scale estimator $\text{IQR}(\{x_j\})$ with the scale parameter $\sigma$ (the standard deviation).

6. Adjust the distribution of the values $w_i^*$, $i = 1, \ldots, n$, by the Tukey $T(\widehat{g}^*, \widehat{h}^*)$ distribution, where $\widehat{g}^*$ and $\widehat{h}^*$ are the estimates of the skewness and tail weight parameters $g$ and $h$ obtained by applying equation (3.1) to the empirical quantiles $Q_{0.1}(\{w_j^*\})$ and $Q_{0.9}(\{w_j^*\})$ of orders 0.1 and 0.9 of the series $\{w_1^*, \ldots, w_n^*\}$.

31: I don't understand. What is exactly done in this step?

7. Determine the quantiles $\xi^*_{\alpha/2}$ and $\xi^*_{1-\alpha/2}$ of orders $\alpha/2$ and $1 - \alpha/2$, $\alpha \in (0, 1)$, of the $T_{\widehat{g}^*, \widehat{h}^*}$ distribution specified in the previous step, where $\alpha$ corresponds to the desired detection rate of atypical values in the absence of contamination with outliers. Let

$$\mathcal{I} = \left\{ i = 1, \ldots, n \middle| w_i^* \notin \left[ \xi^*_{\alpha/2}, \xi^*_{1-\alpha/2} \right] \right\}$$

be the set of indices of the values $w_i^*$ that are detected as atypical in the series $\{w_1^*, \ldots, w_n^*\}$. The values $x_i$ for which $i \in \mathcal{I}$ are considered as atypical observations in the initial dataset.

8. Based on the above steps, it is possible to come up with a *generalized* boxplot which is associated to the original dataset. From the detection bounds $L_-^* = \xi^*_{\alpha/2}$ and $L_+^* = \xi^*_{1-\alpha/2}$ computed in step 7, one can build the respective detection

bounds $B_-^*$ and $B_+^*$ for the related original dataset ($B_-^*$ and $B_+^*$ are the extremities of the lower and upper whiskers of the generalized boxplot):

$$B_\pm^* = \left(\Phi\left(Q_{0.5}(\{w_j\}) + \frac{\text{IQR}(\{w_j\})}{1.3426} L_\pm^*\right) \times \{\min(\{r_j\}) + \max(\{r_j\})\} + \min(\{x_j^*\}) - \zeta\right) s_0$$

**Example**

We illustrate the use of the robust boxplot by an example from Bruffaerts et al., 2014. The data contain daily earnings (in British pounds) of 50 top soccer players.[5] The left panel in Figure 3.5 displays the kernel density estimate of the earnings variable.[6] The right panel displays the three flavors of boxplots. Daily earnings appear to be slightly asymmetrically distributed with a relatively heavy tail. A medcouple measure of 0.12 indicates that the distribution is slightly asymmetric, but not too much. This explains why the upper whisker of the generalized boxplot goes beyond the upper whisker of the two other boxplots.

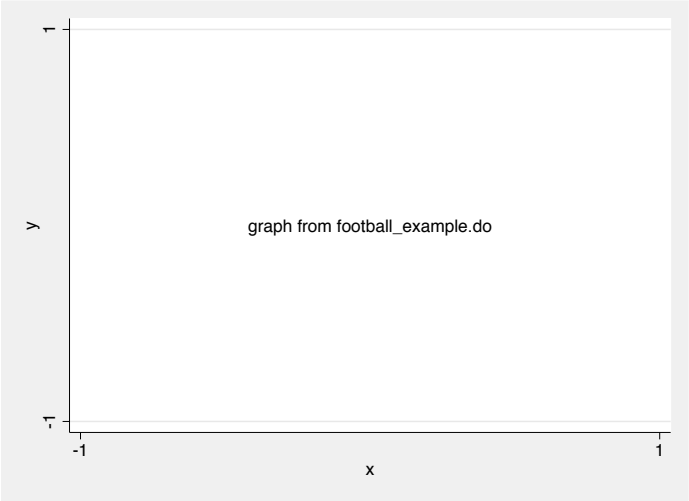


Figure 3.5: Classic, adjusted, and generalized boxplot [will be inserted after revising the boxplot program]

[5] Source: `http://www.paywizard.co.uk/main/pay/vip-celebrity-salary/football-players-salary`

[6] We use an Epanechnikov kernel with Silverman's rule-of-thumb bandwidth.

# References


Anscombe, F. J. 1973. Graphs in Statistical Analysis. *The American Statistician* 27(1): 17–21.

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley series in probability and mathematical statistics, New York: Wiley.

Bruffaerts et al. 2014. Details missing!!! .

Brys, G., M. Hubert, and A. Struyf. 2004. A Robust Measure of Skewness. *Journal of Computational and Graphical Statistics* 13(4): 996–1017.

———. 2006. Robust measures of tail weight. *Computational Statistics & Data Analysis* 50: 733–759.

———. 2008. Goodness-of-fit tests based on a robust measure of skewness. *Computational Statistics* 23: 429–442.

Chatterjee, S., and A. S. Hadi. 1988. *Sensitivity Analysis in Linear Regression*. New York: John Wiley & Sons.

Cook, R. D., and S. Weisberg. 1982. *Residuals and Influence in Regression*. New York: Chapman and Hall.

Donoho, and P. J. Huber. 1983. Details missing!!! .

Fisher, R. A. 1992. Details missing!!! .

Fox, J. 1991. *Regression Diagnostics*. Quantitative applications in the social sciences, Newbury Park, CA: Sage.

Groeneveld. 1991. Details missing!!! .

Hampel, F. R. 1971. Details missing!!! .

———. 1974. Details missing!!! .

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. *Robust Statistics. The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Hampel, F. R., and W. A. Stahel. 1982. Details missing!!! .

Heritier, S., E. Cantoni, S. Copt, and M.-P. Victoria-Feser. 2009. *Robust Statistics in Biostatistics*. West Sussey: Wiley.

Hinkley, D. V. 1975. On power transformations to symmetry. *Biometrika* 62(1): 101–111.

Hodges, J. L., Jr., and E. L. Lehmann. 1963. Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34(2): 598–611.

Huber, P. J. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35(1): 73–101.

———. 1972. The 1972 Wald Lecture. Robust Statistics: A Review. *The Annals of Mathematical Statistics* 43(4): 1041–1067.

Hubert, M., and E. Vandervieren. 2008. An Adjusted Boxplot for Skewed Distributions. *Computational Statistics and Data Analysis* 52: 5186–5201.

Jarque, and Bera. 1980. Details missing!!! .

Jasso, G. 1985. Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences. *American Sociological Review* 50(2): 224–241.

Jiménez, J. A., and V. Arunachalam. 2011. Using Tukey's $g$ and $h$ family of distributions to calculate value-at-risk and conditional value-at-risk. *Journal of Risk* 13(4): 95–116.

Kahn, J. R., and J. R. Udry. 1986. Mariatl Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions. *American Sociological Review* 51(5): 734–737.

Kruskal, W. H. 1960. Some Remarks on Wild Observations. *Technometrics* 2(1): 1–3.

Lehmann, E. L., and G. Casella. 1988. *Theory of Point Estimation*. 2nd ed. Springer.

Rousseeuw, P. J., and C. Croux. 1993. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association* 88(424): 1273–1283.

Rousseeuw, P. J., and A. M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Ruppert. 1987. Details missing!!! .

Serfling, R. 1980. *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons.

Staudte, R. G., and S. J. Sheather. 1990. *Robust Estimation and Testing*. New York: John Wiley & Sons.

White, H. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48(4): 817–838.

Wilcox, R. R. 2005. *Introduction to Robust Estimation and Hypothesis Testing*. 2nd ed. New York: Elsevier Academic Press.

Yule, G. U., and M. G. Kendall. 1968. *An Introduction to the Theory of Statistics*. 14th ed. London: Griffin.