Introduction to Robust Statistics

Klaus Nordhausen

Department of Mathematics and Statistics
University of Turku

Autumn 2015

Goal of robust estimation

Recall, the goal of robust estimation is:

- To find an estimate that is “good” at the parametric target model.
- The estimate should remain “good” in a neighborhood of the specified target.

We need ways to quantify “good”!

Location-scale

Consider again the location-scale model

$$x = \sigma y + \mu$$

where most times we will assume now that $y \sim N(0, 1)$ and the goal is to estimate the location μ and/or the scale σ.

In the following $\hat{\mu}$ and $\hat{s}$ denote arbitrary estimates based on a sample $x_1, \ldots, x_n$.

What are the minimum requirements a statistic should fulfill to qualify as an estimate?

Mean squared error

One way to measure how good an estimate $\hat{\theta}$ approximates a true parameter θ is the mean squared error (MSE).

$$\mathrm{MSE}(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$$

It is well-known that the MSE can be decomposed into

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta})^2,$$

where $\mathrm{bias}(\hat{\theta}) = E(\hat{\theta} - \theta)$.

Hence a good estimate is preferably unbiased and has a small variance.

Variance of sample mean and median

Assume that we are in the location-scale model where y is symmetric with density f.

Then both sample mean and sample median are estimates for the location μ.
Under very general conditions it holds then, that:

- $\bar{x} \sim N(\mu, \sigma^2/n)$
- $\text{med}(\mathbf{x}) \sim N(\mu, \frac{1}{n}\frac{1}{4f(\mu)^2})$.

Hence at the normal model $\text{med}(\mathbf{x}) \sim N(\mu, \frac{\sigma^2}{n}\frac{\pi}{2})$.

This gives an asymptotic relative efficiency (ARE)

$$\text{ARE}(\text{med}, \bar{x}) = \frac{\text{avar}(\bar{x})}{\text{avar}(\text{med})} = \frac{2}{\pi} \approx 0.6366.$$

Variance of sample mean and median II

Expressing the t-distribution in a location-scale family one has however:

- For a sample from a Cauchy distribution $\bar{x} \sim Cauchy(\mu, \sigma^2)$ whereas $\text{med}(\mathbf{x}) \approx N(\mu, \frac{\pi^2\sigma^2}{4n})$
- Hence for Cauchy data: $\text{ARE}(\text{med}, \bar{x}) = \infty$ or $\text{ARE}(\bar{x}, \text{med}) = 0$.
- For the general case with ν degrees of freedom:

$$\text{ARE}(\text{med}, \bar{x}) = \frac{4}{(\nu-2)\pi} \frac{\Gamma((\nu+1)/2)^2}{\Gamma(\nu/2)}$$

ν	≤ 2	3	4	5
$\text{ARE}(\text{med}, \bar{x})$	∞	1.621	1.125	0.960
$\text{ARE}(\bar{x}, \text{med})$	0	0.617	0.888	1.041

Efficiencies in contaminated normal model

Tukey, one of the founders of robust statistics compared some efficiencies in the normal contamination model with $G \sim N(\mu, \sigma^2)$ and $H \sim N(\mu, (3\sigma)^2)$.

- For $\epsilon > 0.10$, $\text{ARE}(\text{med}, \bar{x}) > 1$
- For $\epsilon > 0.01$, $\text{ARE}(\text{mad}, s) > 1$

Linear transformations of location estimates

From a “good” location estimate one should also expect that it does not depend on the measured coordinate system.

An estimator $\hat{\mu}$ is shift (location) equivariant if for $x_i^* = x_i + c$

$$\hat{\mu}(x_1^*, \ldots, x_n^*) = \hat{\mu}(x_1, \ldots, x_n) + c$$

for every constant c.

An estimator $\hat{\mu}$ is scale equivariant if for $x_i^* = cx_i$

$$\hat{\mu}(x_1^*, \ldots, x_n^*) = c\hat{\mu}(x_1, \ldots, x_n)$$

for every constant c.

Linear transformations of location estimates II

An estimator $\hat{\mu}$ is affine equivariant if for $x_i^* = c_1 x_i + c_2$

$$\hat{\mu}(x_1^*, \ldots, x_n^*) = c_1 \hat{\mu}(x_1, \ldots, x_n) + c_2$$

for all constants c_1 and c_2.

Linear transformations of scale estimates

From a “good” scale estimate one should also expect that it does neither depend on the measured coordinate system.

An estimator $\hat{s}$ is shift (location) equivariant if for $x_i^* = x_i + c$

$$\hat{s}(x_1^*, \ldots, x_n^*) = \hat{s}(x_1, \ldots, x_n)$$

for every constant c.

An estimator $\hat{\mu}$ is scale equivariant if for $x_i^* = cx_i$

$$\hat{s}(x_1^*, \ldots, x_n^*) = abs(c)\hat{s}(x_1, \ldots, x_n)$$

for every constant c.

Linear transformations of scale estimates II

An estimator $\hat{s}$ is affine equivariant if for $x_i^* = c_1 x_i + c_2$

$$\hat{s}(x_1^*, \ldots, x_n^*) = abs(c_1)\hat{s}(x_1, \ldots, x_n)$$

for all constants c_1 and c_2.

Functional approach

We have actually not yet specified how we “see” an estimator. In robust statistics it is custom to use a functional approach.

This means every parameter is seen as a functional (function) of the probability distribution underlying the data.

The estimate obtained of the parameter if interest is then the functional applied to the empirical distribution.

Functional approach II

To formalize the concept. Consider a random variable X with probability distribution P_θ and cdf F. And $\mathcal{P}$ is the space of all probability distributions under consideration.
Then in many cases the parameter θ can be seen as the function

$$\theta = T(P) = T(F) = T(X)$$

defined on $\mathcal{P}$.

When then the empirical cdf F_n is used it is expected that $T(F_n)$ tends to $T(F)$ as $n \to \infty$.

Mean and standard deviation as functionals

The mean functional is then

$$\mu(P) = \int x dP = E(X)$$

with estimate

$$\mu(P_n) = \int x dP_n = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and similarly

$$s(P) = \sqrt{\int x^2 dP - E(X)^2}$$

with estimate

$$s(P_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}$$

More on functionals

In general if $T(P) = \int h(x)dP$, where $h(\cdot)$ is an arbitrary P-integrable function, then an empirical counterpart of $T(P)$ is

$$T(P_n) = \frac{1}{n}\sum_{i=1}^{n} h(x_i).$$

More on functionals

On the other side, given a statistical estimate, it is usually possible to find the corresponding statistical functional.

The geometric mean of a sample is defined as

$$T(P_n) = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

which can be reformulated as

$$T(P_n) = \exp \{ \log \{ T(P_n) \} \} = \exp \left\{ \frac{1}{n} \sum_{i=1}^{n} \log(x_i) \right\} = \exp \left\{ \int \log(x) dP_n \right\}$$

and hence the corresponding functional is

$$T(P) = \exp \left\{ \int \log(x) dP \right\}.$$

Other location estimates as functionals

The median can also be easily written as a statistical functional.

$$T(P) = \text{med}(P) = \text{med}(F) = F^{-1/2}(0.5).$$

The α trimmed mean as a functional is

$$T(P) = T(F) = T_{TM}(F) = \frac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} F^{-1}(s) ds.$$

The α winsorized mean as a functional is

$$T(P) = T(F) = T_{WM}(F) = (1-2\alpha) T_{TM}(F) + \alpha F^{-1}(\alpha) + \alpha F^{-1}(1-\alpha).$$

Fisher consistency

Let $x_1, \ldots, x_n$ be a sample form a probability distribution function P. The estimate $\theta_n = T(P_n)$ is said to be Fisher consistent if it satisfies $T(P) = \theta$.

Fisher consistency vs unbiasedness

In robust statistics Fisher consistency plays a major role and is the most natural requirement for an estimate. It is considered even more important than unbiasedness!

For example the functional for the standard deviation does not correspond to the unbiased definition as the divisors differ.

For robustness the key feature of the functional approach is that it is then possible to investigate the behavior of a functional in a neighborhood of the distribution of interest.

Tukey's sensitivity curve

To measure the effect of a single observation (usually an outlier) on a estimate Tukey defined the sensitivity curve SC.

Let T_n be the estimate based on the sample $x_1, \ldots, x_n$ and the T_{n+1} the corresponding estimate after adding as $n+1$th observation the value x^*.

The sensitivity curve is then defined as

$$\mathrm{SC}_{n+1}(x^*) = (n+1)(T_{n+1} - T_n).$$

which is equivalent to saying that

$$T_{n+1} = T_n + \frac{1}{n+1}\,\mathrm{SC}_{n+1}(x^*).$$

Sensitivity of some location estimates

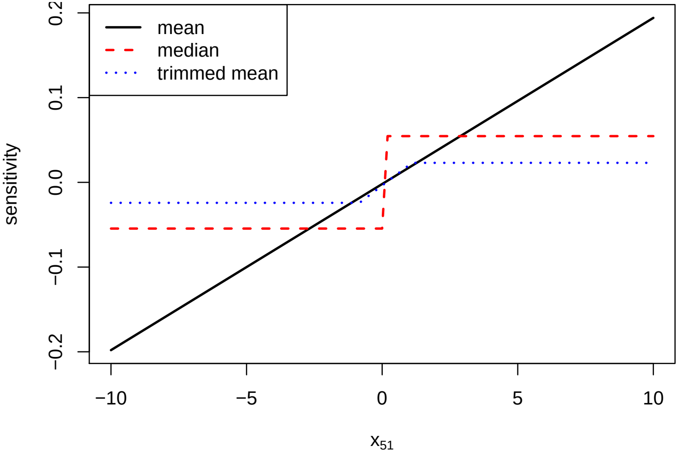
Consider the following simulation study.

Given a sample from $N(0,1)$ of size 50, compute the mean, median and trimmed mean with $\alpha = 0.1$.

Adding an observation x_{51} and letting it range from -10 to 10, we are interested in the difference

$$\hat{\mu}(x_1, \ldots, x_{50}, x_{51}) - \hat{\mu}(x_1, \ldots, x_{50})$$

Sensitivity of some location estimates II

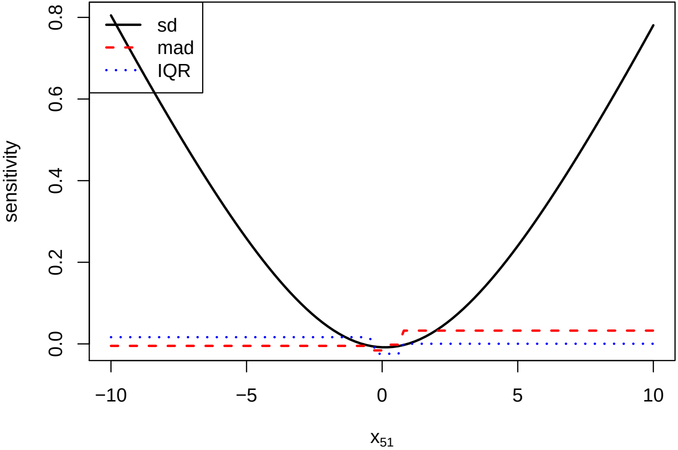


Sensitivity of some scale estimates

Consider then in the exact same setup as previously the three scale estimates standard deviation, median absolute deviation and inter quartile range (IQR). All standardized in such a way to return σ in a normal model.

$$\hat{s}(x_1, \ldots, x_{50}, x_{51}) - \hat{s}(x_1, \ldots, x_{50})$$

Sensitivity of some scale estimates II



Further sensitivity of some location estimates

Consider now the following simulation study.

Given a sample from $N(0,1)$ of size 500, compute the mean, median and trimmed mean with $\alpha = 0.1$.

Adding now however m observations $x_{500+1}, \ldots, x_{500+m}$ each having the value 100, again we are interested in the difference

$$\hat{\mu}(x_1, \ldots, x_{500}, x_{500+1}, \ldots, x_{500+m}) - \hat{\mu}(x_1, \ldots, x_{500})$$

Further sensitivity of some location estimates II

m	mean	median	trimmed mean
0	0.02	-0.04	0.02
1	0.22	-0.04	0.02
2	0.42	-0.04	0.03
3	0.62	-0.03	0.03
4	0.82	-0.03	0.04
5	1.01	-0.03	0.04
10	1.98	-0.02	0.06
50	9.11	0.10	0.21
100	16.69	0.28	8.57
250	33.35	0.68	29.38
500	50.01	51.91	50.19
600	54.56	100.00	55.86

Further sensitivity of some scale estimates

Consider then again the exact same setup as previously and the three scale estimates standard deviation, median absolute deviation and inter quartile range (IQR). All standardized in such a way to return σ in a normal model.

$$\hat{s}(x_1, \ldots, x_{500}, x_{501}, \ldots, x_{500+m}) - \hat{s}(x_1, \ldots, x_{500})$$

Further sensitivity of some scale estimates II

m	SD	MAD	IQR
0	1.01	0.94	0.96
1	4.58	0.94	0.96
2	6.38	0.94	0.97
3	7.77	0.94	0.97
4	8.94	0.95	0.98
5	9.96	0.95	0.98
10	13.91	0.95	0.99
50	28.78	1.09	1.11
100	37.30	1.26	1.33
250	47.17	2.01	74.37
500	50.02	71.31	74.16
600	49.81	0.00	74.05

Influence function

The influence function (IF) of an estimator is the asymptotic version of the sensitivity curve.

The influence function is the most famous concept of robustness in statistics.

It describes the effect of infinitesimal contamination of the model, standardized by the mass of the contamination.

The most desired property of the influence function is boundedness. Estimators with bounded influence function are called B-robust.

Preliminaries for influence function

Let $\hat{T}$ be the functional $T(P_n) = T(F_n)$ which estimates $T(F)$.

Consider the special contamination model

$$F_\epsilon = (1-\epsilon)F + \epsilon\delta_{x^*},$$

where δ_{x^*} is the point mass distribution at x^*.

It is assumed that the functional is continuous in the sense that

$$T(F_\epsilon) \rightarrow T(F) \quad \text{as } \epsilon \rightarrow 0.$$

This is known as qualitative robustness.

The goal is then to compare the functional values $T(F)$ vs $T(F_\epsilon)$.

Definition of influence function

The influence function for a functional T is defined as

$$IF(x^*, T, F) = \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = \frac{\partial}{\partial \epsilon} T(F_\epsilon)|_{\epsilon=0}$$

Note that since here distances between functionals are considered, the derivative “used” here is the so-called Gâteaux derivative.
The limit is usually taken from the right side.

Influence function of the mean

So for the mean, we have

$$T(F) = E_F(X)$$

and

$$\begin{aligned} T(F_\epsilon) &= E_{F_\epsilon}(X) \\ &= (1-\epsilon)E_F(X) + \epsilon E(\delta_{x^*}) \\ &= (1-\epsilon)T(F) + \epsilon x^* \end{aligned}$$

hence

$$IF(x^*, T, F) = \lim_{\epsilon \to 0} \frac{(1-\epsilon)T(F) + \epsilon x^* - T(F)}{\epsilon} = x^* - T(F)$$

Influence function of the median

So for the median, we have

$$T(F) = F^{-1}(0.5).$$

and the IF can be shown to be

$$IF(x^*, T, F) = [2f(T(F))]^{-1} \operatorname{sign}(x^* - T(F))$$

Influence function of the trimmed mean

The α trimmed mean as a functional is

$$T_{TM}(F) = \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(s)ds.$$

The corresponding influence function is

$$IF(x^*, T_{TM}, F) = \begin{cases} \frac{1}{1-2\alpha}(F^{-1}(\alpha) - T_{WM}(F)) & \text{for } x^* < F^{-1}(\alpha) \\ \frac{1}{1-2\alpha}(x^* - T_{WM}(F)) & \text{for } F^{-1}(\alpha) \leq x^* \leq F^{-1}(1-\alpha) \\ \frac{1}{1-2\alpha}(F^{-1}(1-\alpha) - T_{WM}(F)) & \text{for } x^* > F^{-1}(1-\alpha) \end{cases}$$

Influence function of the winsorized mean

Recall, the α winsorized mean as a functional is

$$T_{WM}(F) = (1-2\alpha)T_{TM}(F) + \alpha F^{-1}(\alpha) + \alpha F^{-1}(1-\alpha).$$

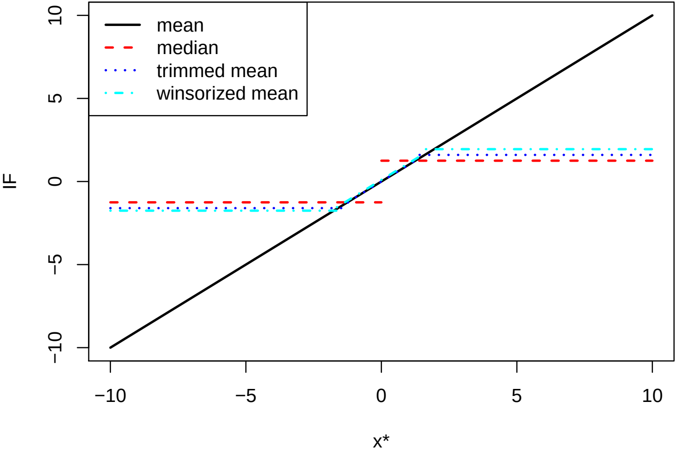
Define first

$$C(F) = T_{WM}(F) - \frac{\alpha^2}{f(F^{-1}(f(F^{-1}(\alpha))))} - \frac{\alpha^2}{f(F^{-1}(1-\alpha))}$$

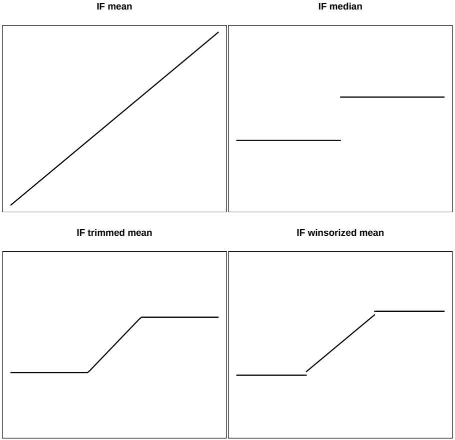
Then the influence function is

$$IF(x^*, T_{WM}, F) = \begin{cases} F^{-1}(\alpha) - \frac{\alpha}{f(F^{-1}(\alpha))} - C(F) & \text{for } x^* < F^{-1}(\alpha) \\ x^* - C(F) & \text{for } F^{-1}(\alpha) < x^* < F^{-1}(1-\alpha) \\ F^{-1}(1-\alpha) - \frac{\alpha}{f(F^{-1}(1-\alpha))} - C(F) & \text{for } x^* > F^{-1}(1-\alpha) \end{cases}$$

IF if some location estimates ($\alpha = 0.1$)



IF if some location estimates II ($\alpha = 0.1$)



Global and local sensitivity

Besides the main interest if whether or not the IF is bounded, there are two popular quantitative characteristics of robustness for a given function based on the influence function.

- Global sensitivity: is the maximum absolute value of the influence function in x^*

$$\gamma^* = \sup_{x^*} |IF(x^*, T)|$$

- Local sensitivity: is the effect of replacing x^* by y^*

$$\lambda^* = \sup_{x^*, y^*, x^* \neq y^*} \left| \frac{IF(x_1^*, T) - IF(x_2^*, T)}{x_1^* - x_2^*} \right|$$

Global and local sensitivity for the mean

The influence function for the mean is:

$$IF(x^*, T, F) = x^* - T(F)$$

Hence

- Global sensitivity:

$$\gamma^* = \infty$$

- Local sensitivity

$$\lambda^* = 1$$

Therefore the mean is in general not robust, but it is not sensitive to local changes.