Contribution

To estimation of a central moment

F. Hroch

On board

The problem:

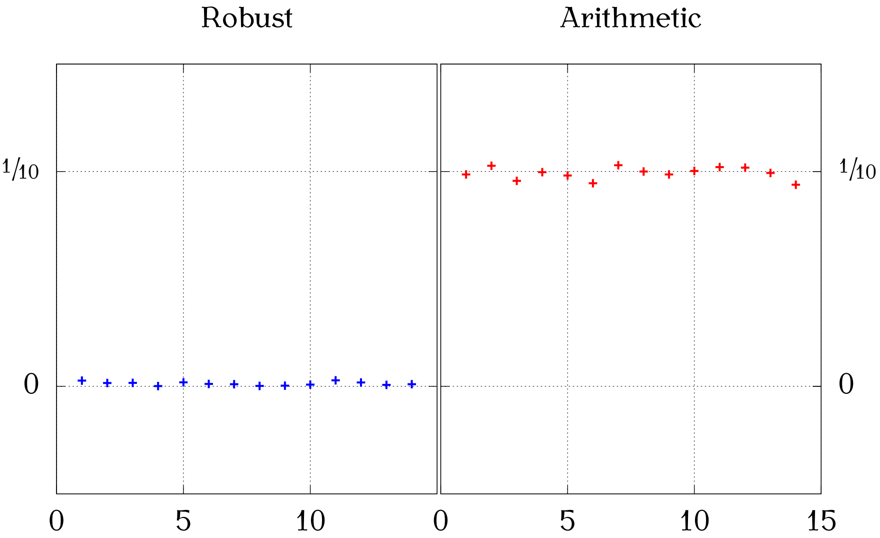
- Real world data – contamined data,
- contamined data – due outliers or another dataset,
- outliers – estimations fails,
- a fail – sinking boat,
- no boat – no live.

The solution:

- Cruel world data – contamined data,
- contamined data – robust estimations,
- robust estimations – unsinkable boat,
- sunny live – true love.

Fascinated by robust algorithms

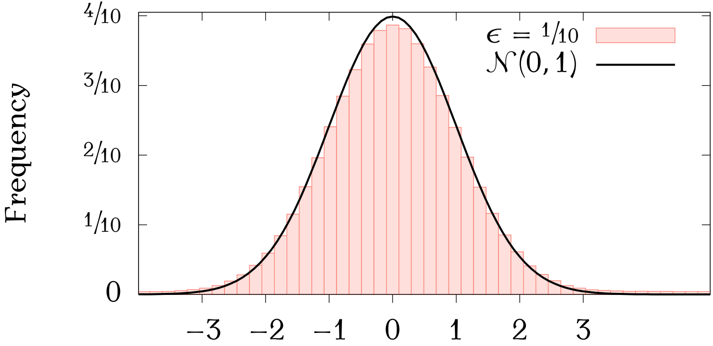
Reconstructing the past



Gross error model

$x_n \in \{(1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(1,10)\}$

ϵ	$\bar{x}$	σ	$\sigma_{\bar{x}}$
0	−0.001	1.0	0.004
1/100	0.008	1.4	0.005
1/10*	0.1	3.3	0.013



*protagonist

Data set (a sample)

$$\{x_1, x_2, \ldots, x_N\}.$$

A probability density of $\mathcal{N}(0,1)$

$$f(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\,.$$

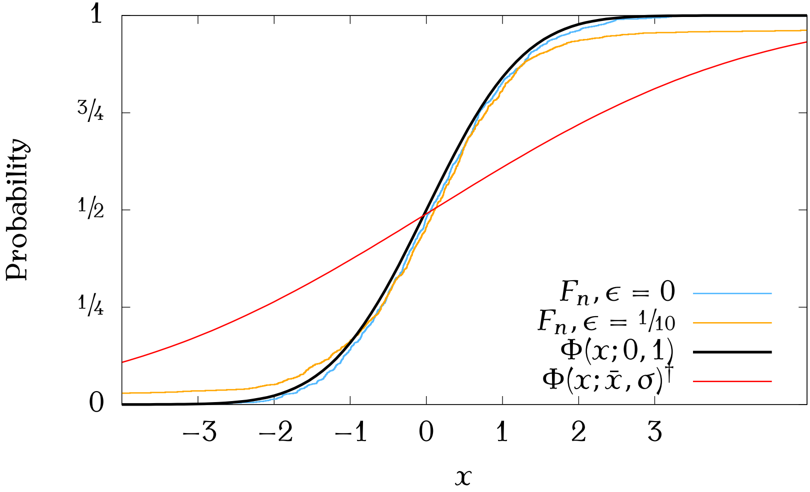
A distribution function (probability)

$$F(x) = \int_{-\infty}^{x} f(u)\ \mathrm{d}u \overset{\mathcal{N}(0,1)}{=} \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right] = \Phi(x).$$

An empirical distribution function

$$F_n = \frac{1}{N}\sum_{i=1}^{n} 1\{x_i < n/N\}, \quad n = 1, \ldots, N.$$

Distribution functions



† $\bar{x} = -0.08, \sigma = 3.3, N = 1000$

Hampel's theorem[‡]

As a tool for robust method recognition

Let the observation x_i be independent, with common distribution F, and let $T_N = T_N(x_1, \dots x_N)$ be a sequence of estimates or test statistics with values in $\mathbb{R}^k$. This sequence is called robust at $F = F_0$ if the sequence of maps of distributions

$$F \to \mathcal{L}_F(T_N)$$

is equicontinous at F_0, that is, if for every $\varepsilon > 0$, there is a $\delta > 0$ and an N_0 such that, for all F and all $N \geq N_0$,

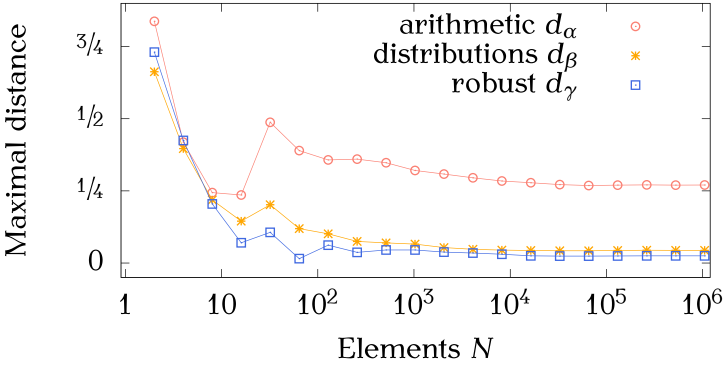
$$d_*(F_0, F) \leq \delta \implies d_*(\mathcal{L}_{F_0}(T_N), \mathcal{L}_F(T_N)) \leq \varepsilon.$$

[‡]Huber & Ronchetti: Robust Statistics (2009)

Hampel's theorem in action, $\epsilon = 1/10$

Analysis of $d_*(F_0, F) \leq \delta \implies d_*(\mathcal{L}_{F_0}(T_N), \mathcal{L}_F(T_N)) \leq \varepsilon$

$$d_\alpha = \max |\Phi(x_n; 0, 1) - \Phi(x_n; \bar{x}, \sigma)|,$$
$$d_\beta = \max |\Phi(x_n; 0, 1) - F_n|,$$
$$d_\gamma = \max |\Phi(x_n; 0, 1) - \Phi(x_n; \tilde{x}, \tilde{\sigma})|.$$



Design of robust statistics

According to Hampel's theorem, or an equivalent condition

R-estimates or Rank estimates replaces data itself by its rank: median, quartile or Wilcoxon test.

L-estimates or Linear combinations of selected statistics.

M-estimates or Maximum likelihood estimates which keeps a spirit of classical estimates: physical and technical applications, multidimensional problems.

- The central point is a robust function $\psi(x)$.
- Replaces least squares by some robust function.
- Reproduces least-squares near minimum.
- A design of robust functions is arbitrary with certain properties.

$$f(x) = \frac{1}{\Gamma}\,\mathrm{e}^{-\varrho(x)}, \qquad \left[\Leftrightarrow \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\right],$$

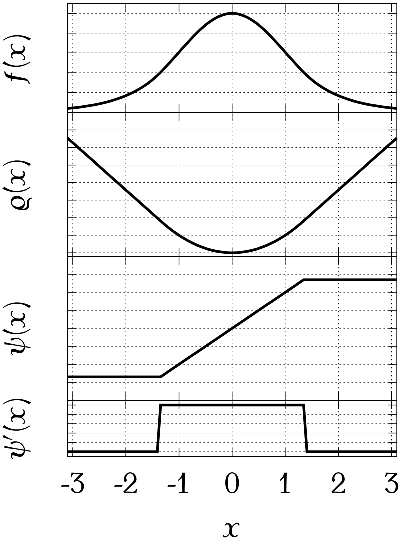
$$\varrho(x) = \int \psi(x)\;\mathrm{d}x, \qquad \left[\Leftrightarrow \frac{x^2}{2}\right],$$

$$\psi(x) = -\,(\ln f)' = -\frac{f'}{f}, \qquad [\Leftrightarrow x].$$

Huber's minimax

$$\psi(x) = \begin{cases} -a, & x < -a, \\ x, & |x| \le a, \\ a, & x > a \end{cases}$$

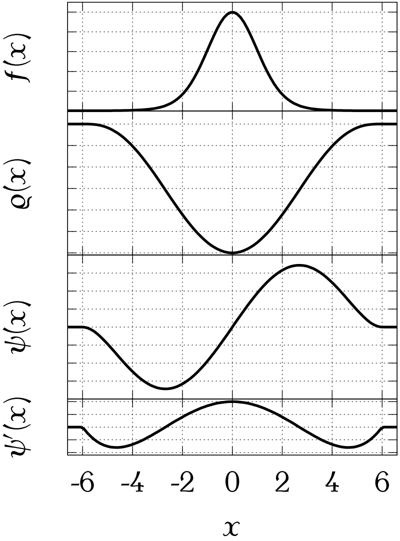
- An equivalent definition is $\psi(x) = \max[-a, \min(a, x)]$,
- an optimal choice $a = 1.345$,
- least-squares near minimum, the absolute value otherwise.
- It is suitable for a theory,
- and sensitive to outliers.



Tukey's biweight

$$\psi(x) = \begin{cases} x[1-(x/a)^2]^2, & |x| \le a, \\ 0, & |x| > a \end{cases}$$

- The 5-order polynomial,
- least-squares near minimum,
- one vanish at infinity,
- an optimal choice $a = 6$.
- It is suitable for real data,
- but a descending function.



A product of independent probabilities

$$P(A \wedge B \wedge \ldots) = P(A)\, P(B) \ldots$$

Lets *suppose the density probability* $f(x_n; \bar{x})$ of an *every point* of data set: there is a such point for

$$\Delta P = \prod_{n=1}^{N} f(x_n; \bar{x})\, \Delta x$$

gets the maximum. If the interval Δx is arbitrary, its is equivalent to find of maximum of the likelihood function[§]

$$L(x_n; \bar{x}) = \prod_{n=1}^{N} f(x_n; \bar{x}).$$

[§]Brandt: Data Analysis: Statistical and Computational Methods for Scientists and Engineers (2014)

Robust mean

By maximum likelihood

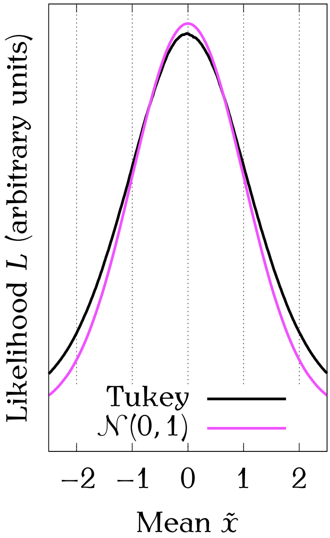
The likelihood

$$L(x_n;\bar{x}) = \prod_{n=1}^{N} f(x_n;\bar{x}),$$

$$L(x_n;\tilde{x}) = \prod_{n=1}^{N} \frac{1}{\Gamma} \exp\left[-\varrho\left(\frac{x_n - \tilde{x}}{s}\right)\right],$$

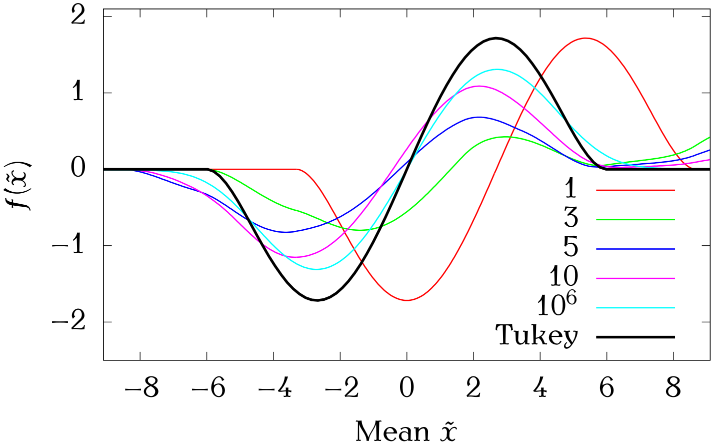
$$\frac{\mathrm{d}\ln L}{\mathrm{d}\tilde{x}} = \frac{1}{s}\sum_{n=1}^{N} \psi\left(\frac{x_n - \tilde{x}}{s}\right) = 0.$$

- ψ is some robust function,
- A solution is given by the non-linear equation against to $\tilde{x}$.
- $s = 1$ (important!).



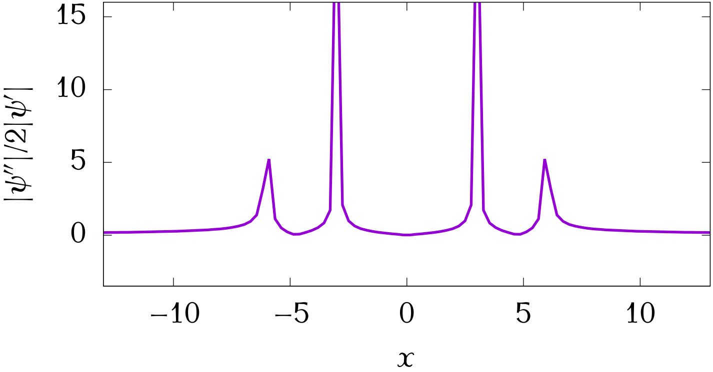
Tukey in action

$$f(\tilde{x}) = \frac{1}{s} \sum_{n=1}^{N} \psi\left(\frac{x_n - \tilde{x}}{s}\right)$$



An approximation error[¶] of Newton's method:

$$\epsilon^{(i+1)} = \frac{|\psi''(x^{(i)})|}{2|\psi'(x^{(i)})|}(\epsilon^{(i)})^2$$



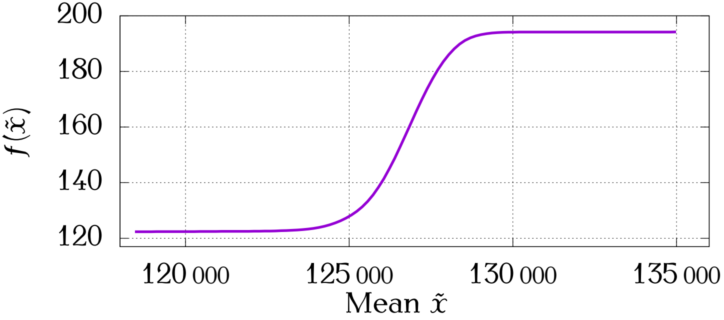
¶ Ralston & Rabinowitz: A First Course in Numerical Analysis (2012)

Bias of Huber's minimax

Another strange protagonist

$$f(\tilde{x}) = \frac{1}{s}\sum_{n=1}^{N} \psi\left(\frac{x_n - \tilde{x}}{s}\right) = \sum_{|(x_n - \tilde{x})/s| \le a} \frac{x_n - \tilde{x}}{s} + a(N_+ - N_-)$$

$$N_+ \overset{?}{\approx} N_-, \quad \text{(a-)symetry}$$



Join estimation of location and scale

Scale does matter; seriously.

$$L(x_n; \tilde{x}, s) = \prod_{n=1}^{N} \frac{1}{\Gamma s} \exp\left[-\varrho\left(\frac{x_n - \tilde{x}}{s}\right)\right].$$

$$\frac{1}{s}\sum_{n=1}^{N} \psi\left(\frac{x_n - \tilde{x}}{s}\right) = 0, \qquad \text{together} \qquad \max_s \left[-\sum_{n=1}^{N} \varrho_n - N \ln \Gamma s\right],$$

where

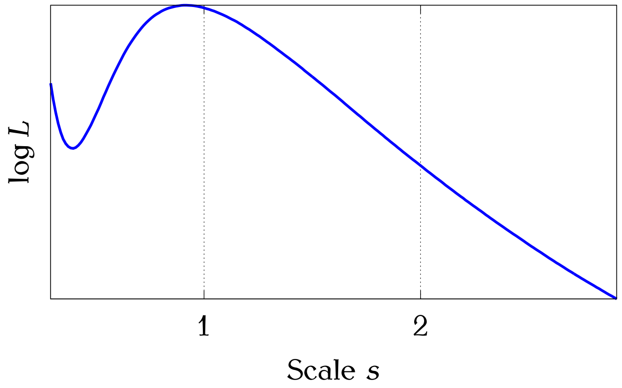
$$r_n = x_n - \tilde{x},$$

$$\varrho_n = \varrho\left(\frac{r_n}{s}\right).$$

Maximum of scale

Our protagonist on the stage again

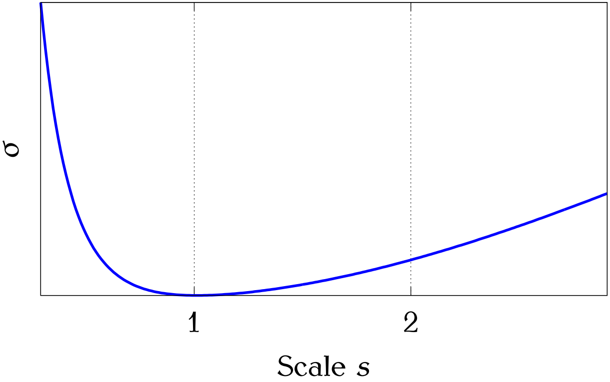
$$\ln L(s) = -\sum_{n=1}^{N} \varrho\left(\frac{x_n - \tilde{x}}{s}\right) - N \ln \Gamma s$$



The dispersion

The actor never disappear

$$\tilde{\sigma}^2 = s^2 \frac{N^2}{N-1} \frac{\sum_{n=1}^{N} \psi^2(r_n/s)}{[\sum_{n=1}^{N} \psi'(r_n/s)]^2}$$



The algorithm

Part I. – An initial estimate

i) Estimate of the location by median $\tilde{x}^{(0)}$

$$\tilde{x}^{(0)} = \text{median}\{x_1, x_2, \ldots, x_N\}.$$

ii) Estimate of s by median of absolute deviations (MAD)

$$s^{(0)} = \frac{\text{median}\{|x_n - \tilde{x}^{(0)}|, n = 1, \ldots, N\}}{\Phi^{-1}(3/4)}.$$

iii) Solve the equation (the initial estimate $\tilde{x}^{(0)} \to \tilde{x}^{(1)}$)

$$\sum_{n=1}^{N} \psi\left(\frac{x_n - \tilde{x}^{(1)}}{s^{(0)}}\right) = 0,$$

for $\tilde{x}^{(1)}$, by a method without derivation.

The algorithm

Part II. – Increasing accuracy

iv) Solve for scale $s^{(1)}$ by finding of maximum of likelihood (with initial $s^{(0)} \to s^{(1)}$)

$$-\sum_{n=1}^{N} \varrho\left(\frac{x_n - \tilde{x}^{(1)}}{s^{(1)}}\right) - \ln s.$$

v) Increase precision of the mean by Newton iterations

$$\tilde{x}^{(i+1)} = \tilde{x}^{(i)} + s^{(1)} \frac{\sum_{n=1}^{N} \psi[(x_n - \tilde{x}^{(i)})/s^{(1)}]}{\sum_{n=1}^{N} \psi'[(x_n - \tilde{x}^{(i)})/s^{(1)}]}, \; i = 1, \ldots$$

vi) Declare results $s = s^{(1)}, \tilde{x} = \tilde{x}^{(i \gg 1)}$.

vii) Compute the standard error, $r_n = x_n - \tilde{x}$:

$$\tilde{\sigma}_{\tilde{x}}^2 = s^2 \frac{N^2}{N-1} \frac{\sum_{n=1}^{N} \psi^2(r_n/s)}{\left[\sum_{n=1}^{N} \psi'(r_n/s)\right]^2}.$$

viii) Compute the standard deviation

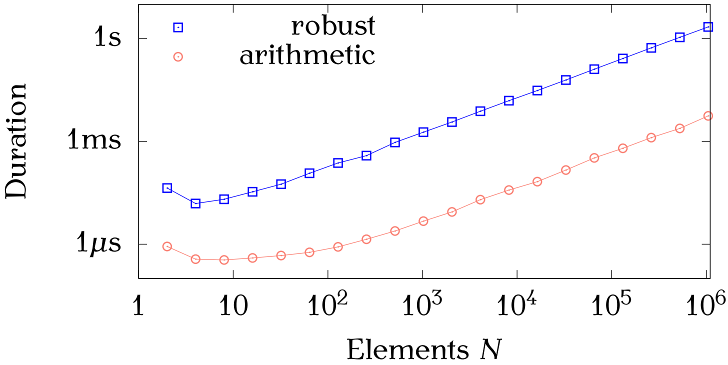
$$\tilde{\sigma} = \sqrt{N}\, \tilde{\sigma}_{\tilde{x}}.$$

dclxvi) A final estimate gives: the standard deviation $\tilde{\sigma}$, parameters of $\mathcal{N}(\tilde{x}, \tilde{\sigma})$, the robust mean and the standard error (no Studentising applied)

$$\tilde{x} \pm \tilde{\sigma}_{\tilde{x}}.$$

Dark side of robust mean

- There is very slow algorithm with rate $1:300$, $\mathcal{O}(n)$
- The algorithm is complicated (advanced numerical methods required, complex logic).
- There is no an explicit form.



Generalizations

Easy:

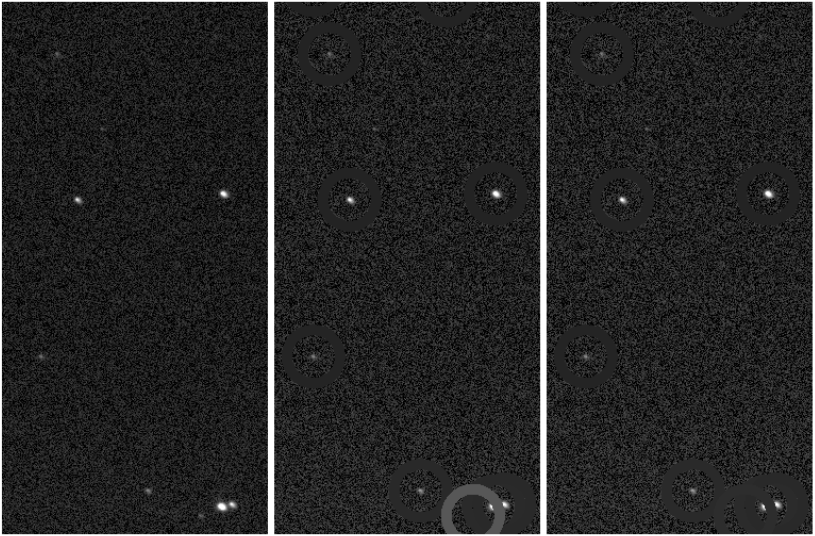
- Weighted Mean
- Multidimensional functions: lines, planes, ...
- Statistical tests (Student).

Hard:

- Non-Gaussian (uniform, Poisson), ...distributions.
- Very limited data sets.
- Data holding some condition(s).

A sky around stars

Revelation of memories



Conclusions

Robustness signifies insensitivity to small deviations from assumptions. – Peter J. Huber

- Robust estimators gives negligible difference between the expected and derived distributions functions.
- Results by maximum likelihood (probability).
- Scale does matter.
- The implementation can be a little bit tricky, whilst usage is common and results are quite reproducible.

The End