Statistical Inference:
Maximum Likelihood Estimation,
Robust Estimation,
Goodness-of-fit Testing

Brian D. Ripley

Robust estimation

MLEs work well if the assumed model is accurate. However, even the sample mean and sample variance are badly affected by outliers.

Definition:

> *Outliers* are sample values that cause surprise in relation to the majority of the sample. This is not a pejorative term; outliers may be correct, but they should always be checked for transcription errors.

In the past the strategy was for the data analyst to look for outliers but this is an expensive process, and outliers can cause harm before they are really obvious. Robust and resistant statistical methods aim to do as good a job as possible in the presence of departures from the assumed models, in particular long-tailed distributions which give rise to outliers.

Digression: boxplots

Remember the central part of a boxplot goes from the lower hinge to the upper hinge, so has width the IQR. For a normal distribution $IQR \approx 1.35\sigma$. So the whiskers run out to about twice the IQR from the centre (half inside the box, $1.5\times$ outside), that is to $\pm 2.7\sigma$). And for a normal about 0.7% of the population lie outside those limits.

Thus the boxplot regards outliers as points in the outer 1% of a normal population.

As an estimator of spread, $IQR/1.35$ is much less susceptible to outliers than s.

Breakdown points

The sample mean and the sample variance can be made arbitrarily large by changing just one sample value. They have breakdown point zero. The median changes only slightly until half the sample is changed, and so has breakdown point 50%.

In between are trimmed and Winsorized means. An α-trimmed mean has the outer $\alpha n/2$ observations removed from each end and then the mean is taken. An α-Winsorized mean does not remove the end points but shrinks them to the appropriate extreme. They have breakdown point $\alpha/2$.

M-estimators

The problem with MLEs for the normal distribution is that they work well for distributions with shorter tails than the normal but badly for those with longer tails. So perhaps we should use MLEs based on longer-tailed distributions. This is the idea of M-estimators.

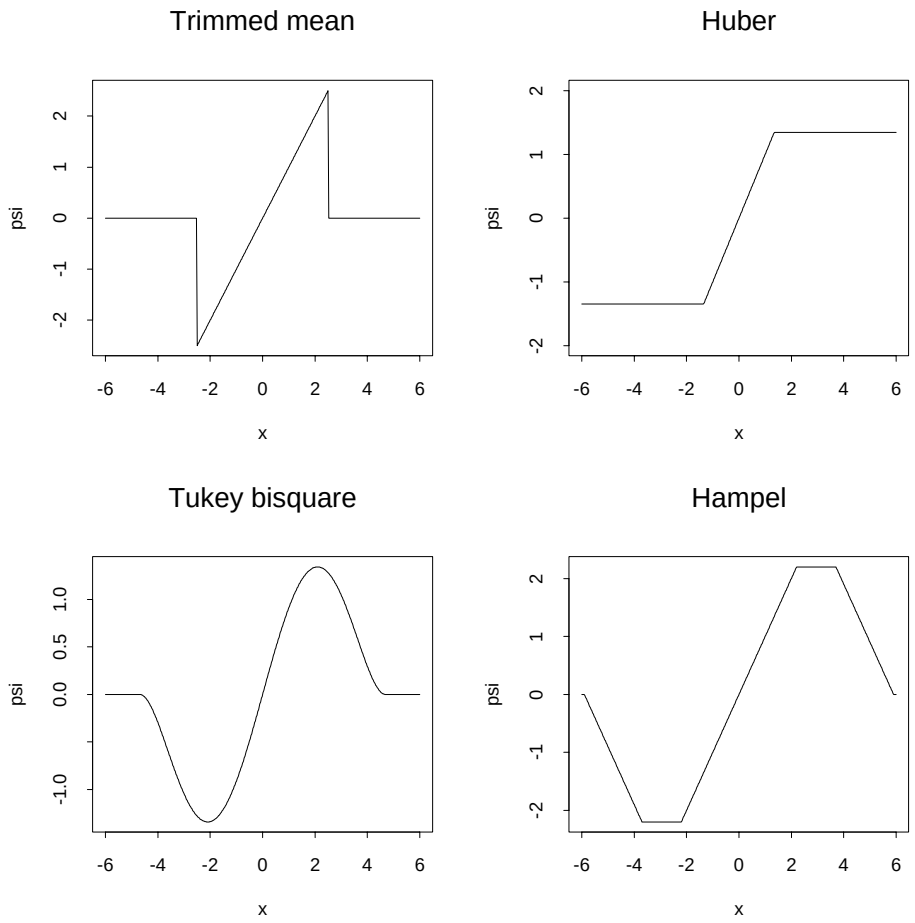
The MLE in a location family would solve

$$\min_{\mu} \sum_i -\log p(y_i - \mu) = \min_{\mu} \sum_i \rho(y_i - \mu)$$

and this makes sense for functions ρ not corresponding to pdfs. Let $\psi = \rho'$ if this exists. Then we will have $\sum_i \psi(y_i - \hat{\mu}) = 0$ or $\sum_i w_i (y_i - \hat{\mu}) = 0$ where $w_i = \psi(y_i - \hat{\mu})/(y_i - \hat{\mu})$. This suggests an iterative method of solution, updating the weights at each iteration.

The mean corresponds to $\rho(x) = x^2/2\sigma^2$, and the median to $\rho(x) = |x|$.

Examples of M-estimators



The ψ-functions for four common M-estimators.

There is a scaling problem here. We can apply the estimator to rescaled results, that is,

$$\min_{\mu} \sum_{i} \rho\left(\frac{y_i - \mu}{cs}\right)$$

for a scale factor s, for example $IQR/1.35$ or the MAD estimator

$$MAD = \underset{i}{\text{median}}\left\{|Y_i - \underset{j}{\text{median}}\left(Y_j\right)|\right\}$$

Then the constant c is chosen as a compromise: it should be large to work well at the Normal, and be small to have a high breakdown point.

Here is an example from analytical chemistry, 24 determinations of copper ($\mu g\, g^{-1}$) in wholemeal flour; these data are part of a larger study that suggests $\mu = 3.68$.

```
> sort(chem)
 [1]  2.20  2.20  2.40  2.40  2.50  2.70  2.80  2.90  3.03
[10]  3.03  3.10  3.37  3.40  3.40  3.40  3.50  3.60  3.70
[19]  3.70  3.70  3.70  3.77  5.28 28.95
```

The sample is clearly highly asymmetric with one value that appears to be out by a factor of 10. It was checked and reported as correct by the laboratory. With such a distribution the various estimators are estimating different aspects of the distribution and so are not comparable. Only for symmetric distributions do all the location estimators estimate the same quantity, and although the true distribution is unknown here, it is unlikely to be symmetric.

```
> mean(chem)
[1] 4.2804
> median(chem)
[1] 3.385
> location.m(chem)
[1] 3.1452
> mad(chem)
[1] 0.52632
> scale.tau(chem)
> unlist(huber(chem))
     mu       s
 3.2067 0.52632
> unlist(hubers(chem))
     mu       s
 3.2055 0.67365
> fitdistr(chem, "t", list(m = 3, s = 0.5), df = 5)
      m       s
 3.1854 0.64217
```