Chapter 11

Robust statistical methods

Much of what appears here comes from ideas presented in the book:


There are many definitions in the literature as to what a robust statistical method is. Huber (1981) defines a robust statistical procedure as a method that does not exhibit “sensitivity to small deviations from the assumptions”.

A common motivation comes from introductory statistics. We commonly use the one-sample t-test as a way to test for plausible values of the population mean parameter, $\mu$. The null $t$ distribution (with $n - 1$ degrees of freedom) for the $t$-statistic,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

comes from assuming that sample values, $x_1, \ldots, x_n$, are independent draws from a normal population with mean $\mu$ and finite variance, $\sigma^2$.

In this motivating example, we can think of this normal population as being the assumed distribution. The question we will ask in this set of notes is what happens to the sampling distribution (or some features of the distribution) of an estimator or test-statistic when the actual population distribution is slightly different from what we assumed. A secondary
question is to identify statistics that are “distributionally robust” (Huber 1981). Such statistics should have good performance under the assumed model and under small deviations from this model the performance should only be impaired slightly.

Along the way we have to decide how we are going to measure deviations in the sampling distribution of these statistics when the assumed distribution is not correct. We also have to define what we mean by “under small deviations of the assumed model”. We will begin with the latter.

We do all our computing in this chapter in R. While we solve the exercises we will learn some more complicated ways to use R. We also revisit Monte Carlo experiments and resampling methods.

11.1 The $\epsilon$-contaminated statistical model

Suppose that $F_0(\cdot)$ is the cumulative distribution function (cdf) for the assumed distribution. In the $\epsilon$-contaminated model we assume that we observe independent random draws not just from $F_0$. Some of the time we observe draws from some other distribution, denoted by $H$. For each $x$ value the cdf of the contaminated population is given by

$$F(x) = (1 - \epsilon)F_0(x) + \epsilon H(x).$$

Here $\epsilon$ is the probability of drawing a “bad” observation from a distribution described by the cdf $H$, and $(1 - \epsilon)$ is the probability of drawing an observation from the assumed (“good”) distribution $F_0(x)$.

Suppose that the probability density/mass functions exist for all the cdfs (denoted by $f$, $f_0$ and $h$ respectively). Then we can write the density function as

$$f(x) = (1 - \epsilon)f_0(x) + \epsilon h(x).$$

To simulate from the $\epsilon$-contaminated model we draw a Bernoulli random variable, $Z$, with success probability $\epsilon$. When $Z = 1$ we draw an observation $x$ from a population with cdf $H$, otherwise (when $Z = 0$) we draw $x$ from $F_0$. 
Exercise 1: For some parameter $\mu$ and $\sigma^2 > 0$, suppose that $F_0$ is the cdf for a normal distribution with mean $\mu$ and variance $\sigma^2$ and $H$ is the cdf of a normal distribution with mean $\mu$ and variance $9\sigma^2$. Describe what the contaminated density function looks like in this case.

Exercise 2: Let $\mu_0$ and $\sigma_0^2 > 0$ be the mean and variance of the assumed distribution $F_0$, and let $\mu_H$ and $\sigma_H^2 > 0$ be the mean and variance of the distribution $H$. Assume all the parameters are finite. Remembering that the distributions $F_0$ and $H$ are independent of one another, show that the mean of the $\epsilon$-contaminated distribution $F$ is given by

$$
\mu_F = (1 - \epsilon)\mu_0 + \epsilon\mu_H,
$$

and the variance is

$$
\sigma_F^2 = (1 - \epsilon)\sigma_0^2 + \epsilon\sigma_H^2 + \epsilon(1 - \epsilon)(\mu_H - \mu_0)^2.
$$

Calculate the mean and variance for the special case of Exercise 1.

Exercise 3: Suppose that you have written two R functions to simulate from $F_0$ and $H$, respectively. Using the idea that we can pass functions as arguments to a function, write a function that simulates $n$ independent observations from the contaminated distribution.

Produce graphical summaries of this contaminated distribution in the case of Exercise 1, with $n = 1,000$, $\mu = 0$, $\sigma^2 = 1$, and $\epsilon = 0.05$. Compare this distribution to the case that $\epsilon = 0.10$.

11.2 Comparing estimators

We start with an example involving estimating the center of a distribution. We know that the sample mean is sensitive to outliers and extreme value – it is not a robust measure of the center of a distribution. A more robust measure of the center of the distribution is the median. The key trouble with the median however is that for many assumed distributions (including the normal), the median is not an efficient estimator of the center of the distribution.
We start by carefully defining our notion of efficiency. For some integer $n$, suppose we have a random sample of values $\{x_1, \ldots, x_n\}$ drawn from some population distribution $F$ (which does not have to be the $\epsilon$-contaminated distribution), which depends on some unknown parameter $\theta$.

The **bias** of an estimator, $\hat{\theta}_n$, of $\theta$ is defined to be

$$\text{bias}(\hat{\theta}_n) = E(\hat{\theta}_n - \theta) = E(\hat{\theta}_n) - \theta,$$

where $\theta$ is the true value of the parameter. We say that $\hat{\theta}_n$ is an unbiased estimate of $\theta$ when $\text{bias}(\hat{\theta}_n) = 0$.

We use the mean squared error as a way to compare different estimators, usually picking the estimator that has the smallest mean squared error. The **mean squared error** of the estimator $\hat{\theta}_n$ is

$$\text{MSE}(\hat{\theta}_n) = E\left(\left[\hat{\theta}_n - \theta\right]^2\right).$$

$$= \text{var}(\hat{\theta}_n) + \left[\text{bias}(\hat{\theta}_n)\right]^2,$$

where $\text{var}(\hat{\theta}_n)$ is the variance of the estimator. When $\hat{\theta}_n$ is an unbiased estimate of $\theta$ the mean squared error is equal to the variance. (This is the situation we are most used to.)

Now suppose that we have two estimators of $\theta$, $\hat{\theta}_n$ and $\tilde{\theta}_n$. One definition of the **relative efficiency** (RE) of $\hat{\theta}_n$ relative to $\tilde{\theta}_n$, is

$$\text{RE}(\hat{\theta}_n, \tilde{\theta}_n) = \frac{\text{MSE}(\hat{\theta}_n)}{\text{MSE}(\tilde{\theta}_n)}.$$

Often this ratio is expressed as a percentage. Note that the relative efficiency simplifies to a ratio of variances when both estimators are unbiased for $\theta$.

For many population distributions, $F$, we cannot calculate the relative efficiency at a fixed sample size $n$. Instead it is common to consider the **asymptotic relative efficiency** (ARE), in which we compare ratios of functions of the mean squared errors as $n \to \infty$. This needs careful mathematical argument (see e.g., Chapter 10, and in particular Section 10.1.2 of Casella and Berger, 2002).

At fixed sample sizes all is not lost however. We do have the opportunity to write Monte Carlo experiments that allow us to approximate these quantities, as the next exercise illustrates.
Exercise 4: Suppose we want to write a Monte Carlo experiment to compare the relative performance of the sample mean and sample median as estimates of the population mean for the $\epsilon$-contaminated distribution defined in Exercise 1, when $n = 100$, $\mu = 0$, and $\sigma = 1$. We want to investigate the relationship as we vary $\epsilon$ from 0 to 0.1, in steps of 0.01.

(i) How do you know that both estimators are unbiased estimators of the population mean of the contaminated distribution?

(ii) Write R code to compare the relative efficiency for these two estimators. Produce a graph that summarizes the relative efficiency as a function of $\epsilon$.

(iii) How can you be sure of the relationship between $\epsilon$ and the relative efficiency that you observe in the Monte Carlo experiment is significant? Write any additional R code to answer this part of the exercise.

11.3 M-estimators for estimating the center

More efficient, robust measures of the center, $\mu$, can be obtained by using an M-estimator. Suppose we have a random sample $\{x_i : i = 1, \ldots, n\}$ drawn from a population with pdf $f(x - \mu)$. We have already shown in this class that the maximum likelihood estimate of $\mu$ is given by

$$\max_{\mu} \prod_{i=1}^{n} f(x_i - \mu) = \max_{\mu} \sum_{i=1}^{n} \log f(x_i - \mu)$$

or equivalently

$$\min_{\mu} \sum_{i=1}^{n} - \log f(x_i - \mu).$$

We obtain an M-estimator by replacing $- \log f(\cdot)$ by some other function $\rho(\cdot)$.

If we choose $\rho(x) = x^2$ then the M-estimator is the mean, and for $\rho(x) = |x|$, we obtain the median. For some constant $c$, we obtain the metric-trimming M-estimator with the function

$$\rho(x) = \begin{cases} x^2, & |x| < c, \\ 0 & \text{otherwise}. \end{cases}$$
The metric-winzoring M-estimator due to Huber is obtained with the function

\[
\rho(x) = \begin{cases} 
  x^2, & |x| < c, \\
  c(2|x| - c) & \text{otherwise.}
\end{cases}
\]

As \( c \) goes to zero, we obtain the median, and as \( c \) goes to infinity we obtain the mean. To choose \( c \) in general we need to consider estimates for the spread as well as the center.

### 11.4 Estimating the spread of a distribution

For the spread, we know the variance and standard deviation is not resistant to outliers. A commonly used estimator is the interquartile range, calculated by

\[
\text{IQR} = Q_3 - Q_1.
\]

**Exercise 5:** For the standard normal distribution, show that the IQR is 1.348980 (to 6 decimal places). Hence deduce for a \( N(\mu, \sigma^2) \) distribution that the IQR is \( 1.348980\sigma \).

Using this result we see that an estimate of \( \sigma \) for a normal distribution is

\[
\hat{\sigma} = \frac{\text{IQR}}{1.3490}.
\]

Another estimate of spread is the median absolute deviations (MAD) estimator, defined by

\[
\text{MAD} = \text{median}_{i=1,\ldots,n}\{|x_i - M|\},
\]

where \( M \) is the sample median of the data. We need to multiply the MAD estimate by 1.4826 to get an unbiased estimate of \( \sigma \) for a \( N(\mu, \sigma^2) \) distribution. The asymptotic relative efficiency for the MAD estimator as compared to the sample standard deviation is 37%, but the estimator is very resistant to outliers. We use the `mad` function in R to calculate the MAD estimate of spread. The MAD and IQR estimator are less affected by heavy tails.

**Exercise 6:** Verify that the MAD estimator for a normal distribution is approximately equal to \( \sigma \).
Exercise 7: Think about how you could write a simulation to compare the efficiency of the MAD estimator with the sample standard deviation, for the $\epsilon$-contaminated distribution defined in Exercise 1, with $n = 100$, $\mu = 0$, and $\sigma = 1$. (The true spread of this distribution is given in Exercise 2).

11.5 M-estimators for the center and scale

An M-estimator for the center based on the rescaled data is given by

$$\min_{\mu} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{s}\right).$$

One common choice for $s$ is the MAD estimator. For the metric-trimming and metric-winzoring functions, $\rho(\cdot)$, we then express $c$ in terms of the number of units of $s$, for example, $1.5s$.

The following code calculates the MAD estimator and then metric-winzoring M-estimator of the center for data stored in the vector $x$. In the first line we need to load the MASS library.

```r
library(MASS)
huber(x)
```

Since the MAD estimate is not very efficient we can also use a metric-winzoring M-estimate for the spread: this gives an M-estimator for the center and spread:

```r
hubers(x)
```

For more details on robust statistics (in particular how to calculate standard errors for these estimators) see Huber (1981).

References
