Chapter 2

The Location Model

2.1 Four Essential Statistics

The *location model*

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{2.1}$$

is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. Assume that there is a sample $Y_1, \ldots, Y_n$ of size n where the Y_i are iid from a distribution with median MED(Y), mean $E(Y)$, and variance $V(Y)$ if they exist. Also assume that the Y_i have a cumulative distribution function (cdf) F that is known up to a few parameters. For example, Y_i could be normal, exponential, or double exponential. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$.

By far the most important robust technique for the location model is to make a plot of the data. Dot plots, histograms, box plots, density estimates, and quantile plots (also called empirical cdfs) can be used for this purpose and allow the investigator to see patterns such as shape, spread, skewness, and outliers.

Example 2.1. Buxton (1920) presents various measurements on 88 men from Cyprus. Case 9 was removed since it had missing values. Figure 2.1 shows the dot plot, histogram, density estimate, and box plot for the heights of the men. Although measurements such as height are often well approximated by a normal distribution, cases 62-66 are gross outliers with recorded

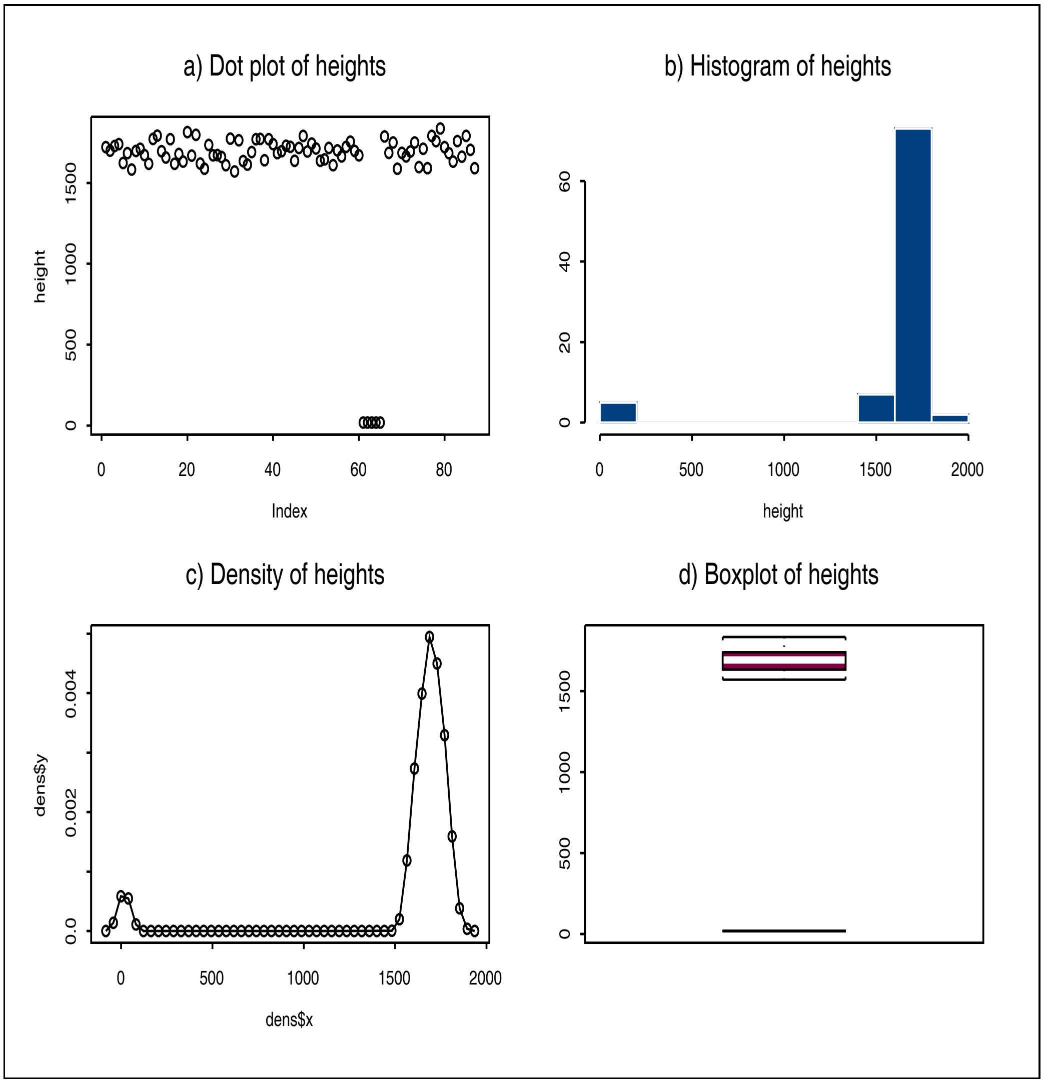

Figure 2.1: Dot plot, histogram, density estimate, and box plot for heights from Buxton (1920).

heights around 0.75 inches! It appears that their heights were recorded under the variable "head length," so these height outliers can be corrected. Note that the presence of outliers is easily detected in all four plots.

Point estimation is one of the oldest problems in statistics and four of the most important statistics for the location model are the sample mean, median, variance, and the median absolute deviation (mad). Let $Y_1, \ldots, Y_n$ be the random sample; ie, assume that $Y_1, ..., Y_n$ are iid.

Definition 2.1. The *sample mean*

$$\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}. \tag{2.2}$$

The sample mean is a measure of location and estimates the population mean (expected value) $\mu = E(Y)$. The sample mean is often described as the "balance point" of the data. The following alternative description is also useful. For any value m consider the data values $Y_i \leq m$, and the values $Y_i > m$. Suppose that there are n rods where rod i has length $|r_i(m)| = |Y_i - m|$ where $r_i(m)$ is the ith residual of m. Since $\sum_{i=1}^n (Y_i - \overline{Y}) = 0$, $\overline{Y}$ is the value of m such that the sum of the lengths of the rods corresponding to $Y_i \leq m$ is equal to the sum of the lengths of the rods corresponding to $Y_i > m$. If the rods have the same diameter, then the weight of a rod is proportional to its length, and the weight of the rods corresponding to the $Y_i \leq \overline{Y}$ is equal to the weight of the rods corresponding to $Y_i > \overline{Y}$. The sample mean is drawn towards an outlier since the absolute residual corresponding to a single outlier is large.

If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. Using this notation, the median

$$\text{MED}_c(n) = Y_{((n+1)/2)} \ \text{ if n is odd,}$$

and

$$\text{MED}_c(n) = (1-c)Y_{(n/2)} + cY_{((n/2)+1)} \ \text{ if n is even}$$

for $c \in [0, 1]$. Note that since a statistic is a function, c needs to be fixed. The *low median* corresponds to $c = 0$, and the *high median* corresponds to $c = 1$. The choice of $c = 0.5$ will yield the sample median. For example, if the data $Y_1 = 1, Y_2 = 4, Y_3 = 2, Y_4 = 5$, and $Y_5 = 3$, then $\overline{Y} = 3$, $Y_{(i)} = i$ for $i = 1, ..., 5$ and $\text{MED}_c(n) = 3$ where the sample size $n = 5$.

Definition 2.2. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \quad \text{if n is odd,} \tag{2.3}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \quad \text{if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used.

Definition 2.3. The *sample variance*

$$S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1}, \tag{2.4}$$

and the *sample standard deviation* $S_n = \sqrt{S_n^2}$.

The sample median need not be unique and is a measure of location while the sample standard deviation is a measure of scale. In terms of the "rod analogy," the median is a value m such that at least half of the rods are to the left of m and at least half of the rods are to the right of m. Hence the number of rods to the left and right of m rather than the lengths of the rods determine the sample median. The sample standard deviation is vulnerable to outliers and is a measure of the average value of the rod lengths $|r_i(\overline{Y})|$. The sample mad, defined below, is a measure of the median value of the rod lengths $|r_i(\text{MED}(n))|$.

Definition 2.4. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|,\ i = 1, \ldots, n). \tag{2.5}$$

Since $\text{MAD}(n)$ is the median of n distances, at least half of the observations are within a distance $\text{MAD}(n)$ of $\text{MED}(n)$ and at least half of the observations are a distance of $\text{MAD}(n)$ or more away from $\text{MED}(n)$.

Example 2.2. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and $\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

Since these estimators are nonparametric estimators of the corresponding population quantities, they are useful for a very wide range of distributions.

Table 2.1: Some commonly used notation.

population	sample
$E(Y)$, μ, θ	$\overline{Y}_n$, $E(n)$ $\hat{\mu}$, $\hat{\theta}$
$\text{MED}(Y), M$	$\text{MED}(n), \hat{M}$
$\text{VAR}(Y), \sigma^2$	$\text{VAR}(n), S^2, \hat{\sigma}^2$
$\text{SD}(Y), \sigma$	$\text{SD}(n), S, \hat{\sigma}$
MAD(Y)	MAD(n)
IQR(Y)	IQR(n)

They are also quite old. Rey (1978, p. 2) quotes Thucydides on a technique used in the winter of 428 B.C. by Greek besiegers. Cities were often surrounded by walls made of layers of bricks, and besiegers made ladders to scale these walls. The length of the ladders was determined by counting the layers of bricks. Many soldiers counted the number of bricks, and the mode of the counts was used to estimate the number of layers. The reasoning was that some of the counters would make mistakes, but the majority were likely to hit the true count. If the majority did hit the true count, then the sample median would equal the mode. In a lecture, Professor Portnoy stated that in 215 A.D., an "eggs bulk" of impurity was allowed in the ritual preparation of food, and two Rabbis desired to know what is an "average sized egg" given a collection of eggs. One said use the middle sized egg while the other said average the largest and smallest eggs of the collection. Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 65) attribute MAD(n) to Gauss in 1816.

2.2 A Note on Notation

Notation is needed in order to distinguish between population quantities, random quantities, and observed quantities. For population quantities, capital letters like $E(Y)$ and MAD(Y) will often be used while the estimators will often be denoted by $\text{MED}(n), \text{MAD}(n)$, $\text{MED}(Y_i, i = 1, ..., n)$, or $\text{MED}(Y_1, \ldots, Y_n)$. The random sample will be denoted by $Y_1, \ldots, Y_n$. Sometimes the observed sample will be fixed and lower case letters will be used. For example, the observed sample may be denoted by $y_1, ..., y_n$ while the estimates may be denoted by $\text{med}(n), \text{mad}(n)$, or $\overline{y}_n$. Table 2.1 summarizes some of this notation.

2.3 The Population Median and MAD

The population median MED(Y) and the population median absolute deviation MAD(Y) are very important quantities of a distribution.

Definition 2.5. The *population median* is any value MED(Y) such that

$$P(Y \leq \text{MED}(Y)) \geq 0.5 \text{ and } P(Y \geq \text{MED}(Y)) \geq 0.5. \tag{2.6}$$

Definition 2.6. The *population median absolute deviation* is

$$\text{MAD}(Y) = \text{MED}(|Y - \text{MED}(Y)|). \tag{2.7}$$

MED(Y) is a measure of location while MAD(Y) is a measure of scale. The median is the middle value of the distribution. Since MAD(Y) is the median distance from MED(Y), at least half of the mass is inside $[\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y)]$ and at least half of the mass of the distribution is outside of the interval $(\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y))$. In other words, MAD($Y$) is any value such that

$$P(Y \in [\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y)]) \geq 0.5,$$

and

$$P(Y \in (\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y))\ ) \leq 0.5.$$

Warning. There is often no simple formula for MAD(Y). For example, if $Y \sim \text{Gamma}(\nu, \lambda)$, then $\text{VAR}(Y) = \nu\lambda^2$, but for each value of ν, there is a different formula for MAD(Y).

MAD(Y) and MED(Y) are often simple to find for location, scale, and location–scale families. Assume that the cdf F of Y has a *probability density function* (pdf) or *probability mass function* (pmf) f. The following definitions are taken from Casella and Berger (2002, p. 116-119) and Lehmann (1983, p. 20).

Definition 2.7. Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = f_Y(w - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is

Table 2.2: MED(Y) and MAD(Y) for some useful random variables.

NAME	Section	MED(Y)	MAD(Y)
Cauchy C(μ, σ)	3.3	μ	σ
double exponential DE(θ, λ)	3.6	θ	0.6931λ
exponential EXP(λ)	3.7	0.6931λ	$\lambda/2.0781$
two parameter exponential EXP(θ, λ)	3.8	$\theta + 0.6931\lambda$	$\lambda/2.0781$
half normal HN(μ, σ)	3.12	$\mu + 0.6745\sigma$	$0.3991\ \sigma$
largest extreme value LEV(θ, σ)	3.13	$\theta + 0.3665\sigma$	0.7670σ
logistic L(μ, σ)	3.14	μ	$1.0986\ \sigma$
normal N(μ, σ^2)	3.19	μ	0.6745σ
Rayleigh R(μ, σ)	3.23	$\mu + 1.1774\sigma$	0.4485σ
smallest extreme value SEV(θ, σ)	3.24	$\theta - 0.3665\sigma$	0.7670σ
t_p	3.25	0	$t_{p,3/4}$
uniform U(θ_1, θ_2)	3.27	$(\theta_1 + \theta_2)/2$	$(\theta_2 - \theta_1)/4$

the *location family* for the random variable $W = \mu + Y$ with *standard pdf* $f_Y(y)$.

Definition 2.8. Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = (1/\sigma) f_Y(w/\sigma)$ indexed by the *scale parameter* $\sigma > 0$, is the *scale family* for the random variable $W = \sigma Y$ with *standard pdf* $f_Y(y)$.

Definition 2.9. Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = (1/\sigma) f_Y((w - \mu)/\sigma)$ indexed by the *location and scale parameters* μ, $-\infty < \mu < \infty$, and $\sigma > 0$, is the *location–scale family* for the random variable $W = \mu + \sigma Y$ with *standard pdf* $f_Y(y)$.

Table 2.2 gives the population mads and medians for some “brand name” distributions. The distributions are location–scale families except for the exponential and t_p distributions. The notation t_p denotes a t distribution with p degrees of freedom while $t_{p,\alpha}$ is the α percentile of the t_p distribution, ie $P(t_p \leq t_{p,\alpha}) = \alpha$. Hence $t_{p,0.5} = 0$ is the population median. The second column of Table 2.2 gives the section of Chapter 3 where the random variable is described further. For example, the exponential (λ) random variable is described in Section 3.7. Table 2.3 presents approximations for the binomial,

Table 2.3: Approximations for MED(Y) and MAD(Y).

Name	Section	MED(Y)	MAD(Y)
binomial BIN(k,ρ)	3.1	$k\rho$	$0.6745\sqrt{k\rho(1-\rho)}$
chi-square χ^2_p	3.5	$p - 2/3$	$0.9536\sqrt{p}$
gamma G(ν, λ)	3.9	$\beta(\nu - 1/3)$	$\lambda\sqrt{\nu}/1.483$

chi-square and gamma distributions.

Finding MED(Y) and MAD(Y) for symmetric distributions and location–scale families is made easier by the following lemma and Table 2.2. Let $F(y_\alpha) = P(Y \leq y_\alpha) = \alpha$ for $0 < \alpha < 1$ where the cdf $F(y) = P(Y \leq y)$. Let $D = \text{MAD}(Y)$, $M = \text{MED}(Y) = y_{0.5}$ and $U = y_{0.75}$.

Lemma 2.1. a) If $W = a + bY$, then $\text{MED}(W) = a + b\text{MED}(Y)$ and $\text{MAD}(W) = |b|\text{MAD}(Y)$.

b) If Y has a pdf that is continuous and positive on its support and symmetric about μ, then $\text{MED}(Y) = \mu$ and $\text{MAD}(Y) = y_{0.75} - \text{MED}(Y)$. Find $M = \text{MED}(Y)$ by solving the equation $F(M) = 0.5$ for M, and find U by solving $F(U) = 0.75$ for U. Then $D = \text{MAD}(Y) = U - M$.

c) Suppose that W is from a location–scale family with standard pdf $f_Y(y)$ that is continuous and positive on its support. Then $W = \mu + \sigma Y$ where $\sigma > 0$. First find M by solving $F_Y(M) = 0.5$. After finding M, find D by solving $F_Y(M + D) - F_Y(M - D) = 0.5$. Then $\text{MED}(W) = \mu + \sigma M$ and $\text{MAD}(W) = \sigma D$.

Proof sketch. a) Assume the probability density function of Y is continuous and positive on its support. Assume $b > 0$. Then

$$1/2 = P[Y \leq \text{MED}(Y)] = P[a + bY \leq a + b\text{MED}(Y)] = P[W \leq \text{MED}(W)].$$

$$1/2 = P[\text{MED}(Y) - \text{MAD}(Y) \leq Y \leq \text{MED}(Y) + \text{MAD}(Y)]$$
$$= P[a + b\text{MED}(Y) - b\text{MAD}(Y) \leq a + bY \leq a + b\text{MED}(Y) + b\text{MAD}(Y)]$$
$$= P[\text{MED}(W) - b\text{MAD}(Y) \leq W \leq \text{MED}(W) + b\text{MAD}(Y)]$$
$$= P[\text{MED}(W) - \text{MAD}(W) \leq W \leq \text{MED}(W) + \text{MAD}(W)].$$

The proofs of b) and c) are similar. QED

Frequently the population median can be found without using a computer, but often the population mad is found numerically. A good way to get a starting value for MAD(Y) is to generate a simulated random sample $Y_1, ..., Y_n$ for $n \approx 10000$ and then compute MAD(n). The following examples are illustrative.

Example 2.3. Suppose the $W \sim N(\mu, \sigma^2)$. Then $W = \mu + \sigma Z$ where $Z \sim N(0, 1)$. The standard normal random variable Z has a pdf that is symmetric about 0. Hence MED(Z) = 0 and MED(W) = μ+σMED(Z) = μ. Let $D = MAD(Z)$ and let $P(Z \leq z) = \Phi(z)$ be the cdf of Z. Now $\Phi(z)$ does not have a closed form but is tabled extensively. Lemma 2.1b) implies that $D = z_{0.75} - 0 = z_{0.75}$ where $P(Z \leq z_{0.75}) = 0.75$. From a standard normal table, $0.67 < D < 0.68$ or $D \approx 0.674$. A more accurate value can be found with the following *R/Splus* command.

```
> qnorm(0.75)
[1] 0.6744898
```

Hence MAD(W) $\approx 0.6745\sigma$.

Example 2.4. If W is exponential (λ), then the cdf of W is $F_W(w) = 1 - \exp(-w/\lambda)$ for $w > 0$ and $F_W(w) = 0$ otherwise. Since $\exp(\log(1/2)) = \exp(-\log(2)) = 0.5$, MED($W$) = $\log(2)\lambda$. Since the exponential distribution is a scale family with scale parameter λ, MAD(W) = $D\lambda$ for some $D > 0$. Hence

$$0.5 = F_W(\log(2)\lambda + D\lambda) - F_W(\log(2)\lambda - D\lambda),$$

or 0.5 =

$$1 - \exp[-(\log(2) + D)] - (1 - \exp[-(\log(2) - D)]) = \exp(-\log(2))[e^D - e^{-D}].$$

Thus $1 = \exp(D) - \exp(-D)$ which may be solved numerically. One way to solve this equation is to write the following *R/Splus* function.

```
tem <- function(D){exp(D) - exp(-D)}
```

Then plug in values D until tem(D) ≈ 1. Below is some output.

```
> mad(rexp(10000),constant=1) #get the sample MAD if n = 10000
[1] 0.4807404
> tem(0.48)
[1] 0.997291
> tem(0.49)
[1] 1.01969
> tem(0.484)
[1] 1.006238
> tem(0.483)
[1] 1.004
> tem(0.481)
[1] 0.9995264
> tem(0.482)
[1] 1.001763
> tem(0.4813)
[1] 1.000197
> tem(0.4811)
[1] 0.99975
> tem(0.4812)
[1] 0.9999736
```

Hence $D \approx 0.4812$ and $\text{MAD}(W) \approx 0.4812\lambda \approx \lambda/2.0781$. If X is a two parameter exponential (θ, λ) random variable, then $X = \theta + W$. Hence $\text{MED}(X) = \theta + \log(2)\lambda$ and $\text{MAD}(X) \approx \lambda/2.0781$. Arnold Willemsen, personal communication, noted that $1 = e^D + e^{-D}$. Multiply both sides by $W = e^D$ so $W = W^2 - 1$ or $0 = W^2 - W - 1$ or $e^D = (1 + \sqrt{5})/2$ so $D = \log[(1 + \sqrt{5})/2] \approx 0.4812$.

Example 2.5. This example shows how to approximate the population median and mad under severe contamination when the "clean" observations are from a symmetric location–scale family. Let Φ be the cdf of the standard normal, and let $\Phi(z_\alpha) = \alpha$. Note that $z_\alpha = \Phi^{-1}(\alpha)$. Suppose $Y \sim (1-\gamma)F_W + \gamma F_C$ where $W \sim N(\mu, \sigma^2)$ and C is a random variable far to the right of μ. Show a)

$$\text{MED}(Y) \approx \mu + \sigma z_{[\frac{1}{2(1-\gamma)}]}$$

and b) if $0.4285 < \gamma < 0.5$,

$$\text{MAD}(Y) \approx \text{MED}(Y) - \mu + \sigma z_{[\frac{1}{2(1-\gamma)}]} \approx 2\sigma z_{[\frac{1}{2(1-\gamma)}]}.$$

Solution. a) Since the pdf of C is far to the right of μ,

$$(1-\gamma)\Phi(\frac{\text{MED}(Y)-\mu}{\sigma}) \approx 0.5,$$

and

$$\Phi(\frac{\text{MED}(Y)-\mu}{\sigma}) \approx \frac{1}{2(1-\gamma)}.$$

b) Since the mass of C is far to the right of μ,

$$(1-\gamma)P[\text{MED}(Y) - \text{MAD}(Y) < W < \text{MED}(Y) + \text{MAD}(Y)] \approx 0.5.$$

Since the contamination is high, $P(W < \text{MED}(Y) + \text{MAD}(Y)) \approx 1$, and

$$0.5 \approx (1-\gamma)P(\text{MED}(Y) - \text{MAD}(Y) < W)$$

$$= (1-\gamma)[1-\Phi(\frac{\text{MED}(Y)-\text{MAD}(Y)-\mu}{\sigma})].$$

Writing $z[\alpha]$ for z_α gives

$$\frac{\text{MED}(Y)-\text{MAD}(Y)-\mu}{\sigma} \approx z\left[\frac{1-2\gamma}{2(1-\gamma)}\right].$$

Thus

$$\text{MAD}(Y) \approx \text{MED}(Y) - \mu - \sigma z\left[\frac{1-2\gamma}{2(1-\gamma)}\right].$$

Since $z[\alpha] = -z[1-\alpha]$,

$$-z\left[\frac{1-2\gamma}{2(1-\gamma)}\right] = z\left[\frac{1}{2(1-\gamma)}\right]$$

and

$$\text{MAD}(Y) \approx \mu + \sigma z\left[\frac{1}{2(1-\gamma)}\right] - \mu + \sigma z\left[\frac{1}{2(1-\gamma)}\right].$$

Application 2.1. *The MAD Method:* In analogy with the method of moments, *robust point estimators* can be obtained by solving MED(n) = MED(Y) and MAD(n) = MAD(Y). In particular, the location and scale parameters of a location–scale family can often be estimated robustly using

Table 2.4: Robust point estimators for some useful random variables.

BIN(k,ρ)	$\hat{\rho} \approx \text{MED}(n)/k$	
C(μ, σ)	$\hat{\mu} = \text{MED}(n)$	$\hat{\sigma} = \text{MAD}(n)$
χ^2_p	$\hat{p} \approx \text{MED}(n) + 2/3$, rounded	
DE(θ, λ)	$\hat{\theta} = \text{MED}(n)$	$\hat{\lambda} = 1.443\text{MAD}(n)$
EXP(λ)	$\hat{\lambda}_1 = 1.443\text{MED}(n)$	$\hat{\lambda}_2 = 2.0781\text{MAD}(n)$
EXP(θ, λ)	$\hat{\theta} = \text{MED}(n) - 1.440\text{MAD}(n)$	$\hat{\lambda} = 2.0781\text{MAD}(n)$
G(ν, λ)	$\hat{\nu} \approx [\text{MED}(n)/1.483\text{MAD}(n)]^2$	$\hat{\lambda} \approx \frac{[1.483\text{MAD}(n)]^2}{\text{MED}(n)}$
HN(μ, σ)	$\hat{\mu} = \text{MED}(n) - 1.6901\text{MAD}(n)$	$\hat{\sigma} = 2.5057\text{MAD}(n)$
LEV(θ, σ)	$\hat{\theta} = \text{MED}(n) - 0.4778\text{MAD}(n)$	$\hat{\sigma} = 1.3037\text{MAD}(n)$
L(μ, σ)	$\hat{\mu} = \text{MED}(n)$	$\hat{\sigma} = 0.9102\text{MAD}(n)$
N(μ, σ^2)	$\hat{\mu} = \text{MED}(n)$	$\hat{\sigma} = 1.483\text{MAD}(n)$
R(μ, σ)	$\hat{\mu} = \text{MED}(n) - 2.6255\text{MAD}(n)$	$\hat{\sigma} = 2.230\text{MAD}(n)$
U(θ_1, θ_2)	$\hat{\theta}_1 = \text{MED}(n) - 2\text{MAD}(n)$	$\hat{\theta}_2 = \text{MED}(n) + 2\text{MAD}(n)$

c_1MED(n) and c_2MAD(n) where c_1 and c_2 are appropriate constants. Table 2.4 shows some of the point estimators and the following example illustrates the procedure. For a location–scale family, asymptotically efficient estimators can be obtained using the cross checking technique. See He and Fung (1999).

Example 2.6. a) For the normal $N(\mu, \sigma^2)$ distribution, MED$(Y) = \mu$ and MAD$(Y) \approx 0.6745\sigma$. Hence $\hat{\mu} = $ MED(n) and $\hat{\sigma} \approx$ MAD$(n)/0.6745 \approx$ 1.483MAD(n).

b) Assume that Y is gamma(ν, λ). Chen and Rubin (1986) showed that MED$(Y) \approx \lambda(\nu - 1/3)$ for $\nu > 1.5$. By the central limit theorem,

$$Y \approx N(\nu\lambda, \nu\lambda^2)$$

for large ν. If X is $N(\mu, \sigma^2)$ then MAD$(X) \approx \sigma/1.483$. Hence MAD$(Y) \approx \lambda\sqrt{\nu}/1.483$. Assuming that ν is large, solve MED$(n) = \lambda\nu$ and MAD$(n) = \lambda\sqrt{\nu}/1.483$ for ν and λ obtaining

$$\hat{\nu} \approx \left(\frac{\text{MED}(n)}{1.483\text{MAD}(n)}\right)^2 \text{ and } \hat{\lambda} \approx \frac{(1.483\text{MAD}(n))^2}{\text{MED}(n)}.$$

c) Suppose that $Y_1, ..., Y_n$ are iid from a largest extreme value distribution, then the cdf of Y is

$$F(y) = \exp[-\exp(-(\frac{y-\theta}{\sigma}))].$$

This family is an asymmetric location-scale family. Since $0.5 = F(\text{MED}(Y))$, $\text{MED}(Y) = \theta - \sigma \log(\log(2)) \approx \theta + 0.36651\sigma$. Let $D = \text{MAD}(Y)$ if $\theta = 0$ and $\sigma = 1$. Then $0.5 = F[\text{MED}(Y) + \text{MAD}(Y)] - F[\text{MED}(Y) - \text{MAD}(Y)]$. Solving $0.5 = \exp[-\exp(-(0.36651 + D))] - \exp[-\exp(-(0.36651 - D))]$ for D numerically yields $D = 0.767049$. Hence $\text{MAD}(Y) = 0.767049\sigma$.

d) Sometimes MED(n) and MAD(n) can also be used to estimate the parameters of two parameter families that are not location–scale families. Suppose that $Y_1, ..., Y_n$ are iid from a Weibull(ϕ, λ) distribution where λ, y, and ϕ are all positive. Then $W = \log(Y)$ has a smallest extreme value SEV($\theta = \log(\lambda^{1/\phi}), \sigma = 1/\phi$) distribution. Let $\hat{\sigma} = \text{MAD}(W_1, ..., W_n)/0.767049$ and let $\hat{\theta} = \text{MED}(W_1, ..., W_n) - \log(\log(2))\hat{\sigma}$. Then $\hat{\phi} = 1/\hat{\sigma}$ and $\hat{\lambda} = \exp(\hat{\theta}/\hat{\sigma})$.

Falk (1997) shows that under regularity conditions, the joint distribution of the sample median and mad is asymptotically normal. See Section 2.9. A special case of this result follows. Let ξ_α be the α percentile of Y. Thus $P(Y \leq \xi_\alpha) = \alpha$. If Y is symmetric and has a positive continuous pdf f, then MED(n) and MAD(n) are asymptotically independent

$$\sqrt{n}\left(\begin{pmatrix} \text{MED}(n) \\ \text{MAD}(n) \end{pmatrix} - \begin{pmatrix} \text{MED}(Y) \\ \text{MAD}(Y) \end{pmatrix} \right) \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_M^2 & 0 \\ 0 & \sigma_D^2 \end{pmatrix} \right)$$

where

$$\sigma_M^2 = \frac{1}{4[f(\text{MED}(Y))]^2},$$

and

$$\sigma_D^2 = \frac{1}{64}\left[\frac{3}{[f(\xi_{3/4})]^2} - \frac{2}{f(\xi_{3/4})f(\xi_{1/4})} + \frac{3}{[f(\xi_{1/4})]^2}\right] = \frac{1}{16[f(\xi_{3/4})]^2}.$$

2.4 Robust Confidence Intervals

In this section, large sample confidence intervals (CIs) for the sample median and 25% trimmed mean are given. The following confidence interval provides considerable resistance to gross outliers while being very simple to compute. The standard error SE(MED(n)) is due to Bloch and Gastwirth (1968), but the degrees of freedom p is motivated by the confidence interval for the trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (eg, $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (eg, $\lceil 7.7 \rceil = 8$).

Application 2.2: inference with the sample median. Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\ \rceil$ and use

$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}).$$

Let $p = U_n - L_n - 1$ (so $p \approx \lceil\ \sqrt{n}\ \ \rceil$). Then a $100(1-\alpha)\%$ confidence interval for the population median is

$$\text{MED}(n) \pm t_{p,1-\alpha/2} SE(\text{MED}(n)). \tag{2.8}$$

Definition 2.10. The symmetrically trimmed mean or the δ *trimmed mean*

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{2.9}$$

where $L_n = \lfloor n\delta \rfloor$ and $U_n = n - L_n$. If $\delta = 0.25$, say, then the δ trimmed mean is called the 25% trimmed mean.

The $(\delta, 1-\gamma)$ *trimmed mean* uses $L_n = \lfloor n\delta \rfloor$ and $U_n = \lfloor n\gamma \rfloor$.

The trimmed mean is estimating a truncated mean μ_T. Assume that Y has a probability density function $f_Y(y)$ that is continuous and positive on its support. Let y_δ be the number satisfying $P(Y \le y_\delta) = \delta$. Then

$$\mu_T = \frac{1}{1-2\delta} \int_{y_\delta}^{y_{1-\delta}} y f_Y(y) dy. \tag{2.10}$$

Notice that the 25% trimmed mean is estimating

$$\mu_T = \int_{y_{0.25}}^{y_{0.75}} 2y f_Y(y) dy.$$

To perform inference, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \le L_n \\ Y_{(i)}, & L_n + 1 \le i \le U_n \\ Y_{(U_n)}, & i \ge U_n + 1. \end{cases}$$

Then the Winsorized variance is the sample variance $S_n^2(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \tag{2.11}$$

The standard error (SE) of T_n is $SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n}$.

Application 2.3: inference with the δ trimmed mean. A large sample 100 $(1 - \alpha)$% confidence interval (CI) for μ_T is

$$T_n \pm t_{p,1-\frac{\alpha}{2}} SE(T_n) \tag{2.12}$$

where $P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ if t_p is from a t distribution with $p = U_n - L_n - 1$ degrees of freedom. This interval is the classical t–interval when $\delta = 0$, but $\delta = 0.25$ gives a robust CI.

Example 2.7. In 1979 an 8th grade student received the following scores for the nonverbal, verbal, reading, English, math, science, social studies, and problem solving sections of a standardized test: 6, 9, 9, 7, 8, 9, 9, 7. Assume that if this student took the exam many times, then these scores would be well approximated by a symmetric distribution with mean μ. Find a 95% CI for μ.

Solution. When computing small examples by hand, the steps are to sort the data from smallest to largest value, find n, L_n, U_n, $Y_{(L_n+1)}$, $Y_{(U_n)}$, p, MED(n) and SE(MED(n)). After finding $t_{p,1-\alpha/2}$, plug the relevant quantities into the formula for the CI. The sorted data are 6, 7, 7, 8, 9, 9, 9, 9. Thus MED$(n) = (8 + 9)/2 = 8.5$. Since $n = 8$, $L_n = \lfloor 4 \rfloor - \lceil \sqrt{2} \rceil = 4 - \lceil 1.414 \rceil = 4 - 2 = 2$ and $U_n = n - L_n = 8 - 2 = 6$. Hence $SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9 - 7) = 1$. The degrees of freedom $p = U_n - L_n - 1 = 6 - 2 - 1 = 3$. The cutoff $t_{3,0.975} = 3.182$. Thus the 95% CI for MED(Y) is

$$\text{MED}(n) \pm t_{3,0.975} SE(\text{MED}(n))$$

$= 8.5 \pm 3.182(1) = (5.318, 11.682)$. The classical t–interval uses $\overline{Y} = (6 + 7 + 7 + 8 + 9 + 9 + 9 + 9)/8$ and $S_n^2 = (1/7)[(\sum_{i=1}^n Y_i^2) - 8(8^2)] = (1/7)[(522 - 8(64)] = 10/7 \approx 1.4286$, and $t_{7,0.975} \approx 2.365$. Hence the 95% CI for μ is $8 \pm 2.365(\sqrt{1.4286/8}) = (7.001, 8.999)$. Notice that the t-cutoff $= 2.365$ for the classical interval is less than the t-cutoff $= 3.182$ for the median interval and that $SE(\overline{Y}) < SE(\text{MED}(n))$. The parameter μ is between 1 and 9 since the test scores are integers between 1 and 9. Hence for this example, the t–interval is considerably superior to the overly long median interval.

Example 2.8. In the last example, what happens if the 6 becomes 66 and a 9 becomes 99?

Solution. Then the ordered data are 7, 7, 8, 9, 9, 9, 66, 99. Hence $\text{MED}(n) = 9$. Since L_n and U_n only depend on the sample size, they take the same values as in the previous example and $SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9 - 8) = 0.5$. Hence the 95% CI for $\text{MED}(Y)$ is $\text{MED}(n) \pm t_{3,0.975} SE(\text{MED}(n)) = 9 \pm 3.182(0.5) = (7.409, 10.591)$. Notice that with discrete data, it is possible to drive $SE(\text{MED}(n))$ to 0 with a few outliers if n is small. The classical confidence interval $\overline{Y} \pm t_{7,0.975} S/\sqrt{n}$ blows up and is equal to $(-2.955, 56.455)$.

Example 2.9. The Buxton (1920) data contains 87 heights of men, but five of the men were recorded to be about 0.75 inches tall! The mean height is $\overline{Y} = 1598.862$ and the classical 95% CI is (1514.206, 1683.518). $\text{MED}(n) = 1693.0$ and the resistant 95% CI based on the median is (1678.517, 1707.483). The 25% trimmed mean $T_n = 1689.689$ with 95% CI (1672.096, 1707.282).

The heights for the five men were recorded under their head lengths, so the outliers can be corrected. Then $\overline{Y} = 1692.356$ and the classical 95% CI is (1678.595, 1706.118). Now $\text{MED}(n) = 1694.0$ and the 95% CI based on the median is (1678.403, 1709.597). The 25% trimmed mean $T_n = 1693.200$ with 95% CI (1676.259, 1710.141). Notice that when the outliers are corrected, the three intervals are very similar although the classical interval length is slightly shorter. Also notice that the outliers roughly shifted the median confidence interval by about 1 mm while the outliers greatly increased the length of the classical t–interval.

Sections 2.5, 2.6 and 2.7 provide additional information on CIs and tests.

2.5 Large Sample CIs and Tests

Large sample theory can be used to construct *confidence intervals* (CIs) and *hypothesis tests*. Suppose that $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and that $W_n \equiv W_n(\boldsymbol{Y})$ is an estimator of some parameter μ_W such that

$$\sqrt{n}(W_n - \mu_W) \xrightarrow{D} N(0, \sigma^2_W)$$

where σ^2_W/n is the asymptotic variance of the estimator W_n. The above notation means that if n is large, then for probability calculations

$$W_n - \mu_W \approx N(0, \sigma^2_W/n).$$

Suppose that S^2_W is a consistent estimator of σ^2_W so that the (asymptotic) *standard error* of W_n is SE$(W_n) = S_W/\sqrt{n}$. Let z_α be the α percentile of the N(0,1) distribution. Hence $P(Z \le z_\alpha) = \alpha$ if $Z \sim N(0,1)$. Then

$$1 - \alpha \approx P(-z_{1-\alpha/2} \le \frac{W_n - \mu_W}{SE(W_n)} \le z_{1-\alpha/2}),$$

and an approximate or large sample $100(1-\alpha)\%$ CI for μ_W is given by

$$(W_n - z_{1-\alpha/2} SE(W_n), W_n + z_{1-\alpha/2} SE(W_n)).$$

Three common approximate level α tests of hypotheses all use the *null hypothesis* $H_o : \mu_W = \mu_o$. A right tailed test uses the *alternative hypothesis* $H_A : \mu_W > \mu_o$, a left tailed test uses $H_A : \mu_W < \mu_o$, and a two tail test uses $H_A : \mu_W \neq \mu_o$. The test statistic is

$$t_o = \frac{W_n - \mu_o}{SE(W_n)},$$

and the (approximate) *p-values* are $P(Z > t_o)$ for a right tail test, $P(Z < t_o)$ for a left tail test, and $2P(Z > |t_o|) = 2P(Z < -|t_o|)$ for a two tail test. The null hypothesis H_o is rejected if the p-value $< \alpha$.

Remark 2.1. Frequently the large sample CIs and tests can be improved for smaller samples by substituting a t distribution with p degrees of freedom for the standard normal distribution Z where $p \equiv p_n$ is some increasing function of the sample size n. Then the $100(1-\alpha)\%$ CI for μ_W is given by

$$(W_n - t_{p,1-\alpha/2} SE(W_n), W_n + t_{p,1-\alpha/2} SE(W_n)).$$

The test statistic rarely has an exact t_p distribution, but the approximation tends to make the CIs and tests more *conservative;* ie, the CIs are longer and

H_o is less likely to be rejected. This book will typically use very simple rules for p and not investigate the exact distribution of the test statistic.

Paired and two sample procedures can be obtained directly from the one sample procedures. Suppose there are two samples $Y_1, ..., Y_n$ and $X_1, ..., X_m$. If $n = m$ and it is known that (Y_i, X_i) match up in correlated pairs, then *paired* CIs and tests apply the one sample procedures to the differences $D_i = Y_i - X_i$. Otherwise, assume the two samples are independent, that n and m are large, and that

$$\left(\begin{array}{c} \sqrt{n}(W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ \sqrt{m}(W_m(\boldsymbol{X}) - \mu_W(X)) \end{array} \right) \xrightarrow{D} N_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \ \left(\begin{array}{cc} \sigma^2_W(Y) & 0 \\ 0 & \sigma^2_W(X) \end{array} \right) \right).$$

Then

$$\left(\begin{array}{c} (W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ (W_m(\boldsymbol{X}) - \mu_W(X)) \end{array} \right) \approx N_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \ \left(\begin{array}{cc} \sigma^2_W(Y)/n & 0 \\ 0 & \sigma^2_W(X)/m \end{array} \right) \right),$$

and

$$W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}) - (\mu_W(Y) - \mu_W(X)) \approx N(0, \frac{\sigma^2_W(Y)}{n} + \frac{\sigma^2_W(X)}{m}).$$

Hence

$$SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) = \sqrt{\frac{S^2_W(\boldsymbol{Y})}{n} + \frac{S^2_W(\boldsymbol{X})}{m}},$$

and the large sample $100(1-\alpha)\%$ CI for $\mu_W(Y) - \mu_W(X)$ is given by

$$(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) \pm z_{1-\alpha/2} SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})).$$

Often approximate level α tests of hypotheses use the *null hypothesis* $H_o : \mu_W(Y) = \mu_W(X)$. A right tailed test uses the *alternative hypothesis* $H_A : \mu_W(Y) > \mu_W(X)$, a left tailed test uses $H_A : \mu_W(Y) < \mu_W(X)$, and a two tail test uses $H_A : \mu_W(Y) \neq \mu_W(X)$. The test statistic is

$$t_o = \frac{W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})}{SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}))},$$

and the (approximate) *p-values* are $P(Z > t_o)$ for a right tail test, $P(Z < t_o)$ for a left tail test, and $2P(Z > |t_o|) = 2P(Z < -|t_o|)$ for a two tail test. The null hypothesis H_o is rejected if the p-value $< \alpha$.

Remark 2.2. Again a t_p distribution will often be used instead of the N(0,1) distribution. If p_n is the degrees of freedom used for a single sample procedure when the sample size is n, use $p = \min(p_n, p_m)$ for the two sample procedure. These CIs are known as *Welch intervals.* See Welch (1937) and Yuen (1974).

Example 2.10. Consider the single sample procedures where $W_n = \overline{Y}_n$. Then $\mu_W = E(Y)$, $\sigma^2_W = \text{VAR}(Y)$, $S_W = S_n$, and $p = n - 1$. Let t_p denote a random variable with a t distribution with p degrees of freedom and let the α percentile $t_{p,\alpha}$ satisfy $P(t_p \leq t_{p,\alpha}) = \alpha$. Then the classical *t-interval* for $\mu \equiv E(Y)$ is

$$\overline{Y}_n \pm t_{n-1,1-\alpha/2} \frac{S_n}{\sqrt{n}}$$

and the *t-test statistic* is

$$t_o = \frac{\overline{Y} - \mu_o}{S_n/\sqrt{n}}.$$

The right tailed p-value is given by $P(t_{n-1} > t_o)$.

Now suppose that there are two samples where $W_n(\boldsymbol{Y}) = \overline{Y}_n$ and $W_m(\boldsymbol{X}) = \overline{X}_m$. Then $\mu_W(Y) = E(Y) \equiv \mu_Y$, $\mu_W(X) = E(X) \equiv \mu_X$, $\sigma^2_W(Y) = \text{VAR}(Y) \equiv \sigma^2_Y$, $\sigma^2_W(X) = \text{VAR}(X) \equiv \sigma^2_X$, and $p_n = n - 1$. Let $p = \min(n-1, m-1)$. Since

$$SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) = \sqrt{\frac{S^2_n(\boldsymbol{Y})}{n} + \frac{S^2_m(\boldsymbol{X})}{m}},$$

the *two sample t-interval* for $\mu_Y - \mu_X$

$$(\overline{Y}_n - \overline{X}_m) \pm t_{p,1-\alpha/2} \sqrt{\frac{S^2_n(\boldsymbol{Y})}{n} + \frac{S^2_m(\boldsymbol{X})}{m}}$$

and *two sample t-test statistic*

$$t_o = \frac{\overline{Y}_n - \overline{X}_m}{\sqrt{\frac{S^2_n(\boldsymbol{Y})}{n} + \frac{S^2_m(\boldsymbol{X})}{m}}}.$$

The right tailed p-value is given by $P(t_p > t_o)$. For sample means, values of the degrees of freedom that are more accurate than $p = \min(n-1, m-1)$ can be computed. See Moore (2007, p. 474).

2.6 Some Two Stage Trimmed Means

Robust estimators are often obtained by applying the sample mean to a sequence of consecutive order statistics. The sample median, trimmed mean, metrically trimmed mean, and two stage trimmed means are examples. For the trimmed mean given in Definition 2.10 and for the Winsorized mean, defined below, the proportion of cases trimmed and the proportion of cases covered are fixed.

Definition 2.11. Using the same notation as in Definition 2.10, the *Winsorized mean*

$$W_n = W_n(L_n, U_n) = \frac{1}{n}[L_n Y_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y_{(i)} + (n - U_n) Y_{(U_n)}]. \tag{2.13}$$

Definition 2.12. A *randomly trimmed mean*

$$R_n = R_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{2.14}$$

where $L_n < U_n$ are integer valued random variables. $U_n - L_n$ of the cases are *covered* by the randomly trimmed mean while $n - U_n + L_n$ of the cases are trimmed.

Definition 2.13. The *metrically trimmed mean* (also called the Huber type skipped mean) M_n is the sample mean of the cases inside the interval

$$[\hat{\theta}_n - k_1 D_n,\ \hat{\theta}_n + k_2 D_n]$$

where $\hat{\theta}_n$ is a location estimator, D_n is a scale estimator, $k_1 \geq 1$, and $k_2 \geq 1$.

The proportions of cases covered and trimmed by randomly trimmed means such as the metrically trimmed mean are now random. Typically the sample median MED(n) and the sample mad MAD(n) are used for $\hat{\theta}_n$ and D_n, respectively. The amount of trimming will depend on the distribution of the data. For example, if M_n uses $k_1 = k_2 = 5.2$ and the data is normal (Gaussian), about 1% of the data will be trimmed while if the data is Cauchy, about 12% of the data will be trimmed. Hence the upper and lower trimming points estimate lower and upper population percentiles $L(F)$ and $U(F)$ and change with the distribution F.

Two stage estimators are frequently used in robust statistics. Often the initial estimator used in the first stage has good resistance properties but has a low asymptotic relative efficiency or no convenient formula for the SE. Ideally, the estimator in the second stage will have resistance similar to the initial estimator but will be efficient and easy to use. The metrically trimmed mean M_n with tuning parameter $k_1 = k_2 \equiv k = 6$ will often be the initial estimator for the two stage trimmed means. That is, retain the cases that fall in the interval

$$[\text{MED}(n) - 6\text{MAD}(n), \text{MED}(n) + 6\text{MAD}(n)].$$

Let $L(M_n)$ be the number of observations that fall to the left of $\text{MED}(n) - k_1 \text{ MAD}(n)$ and let $n - U(M_n)$ be the number of observations that fall to the right of $\text{MED}(n) + k_2 \text{ MAD}(n)$. When $k_1 = k_2 \equiv k \geq 1$, at least half of the cases will be covered. Consider the set of 51 trimming proportions in the set $C = \{0, 0.01, 0.02, ..., 0.49, 0.50\}$. Alternatively, the coarser set of 6 trimming proportions $C = \{0, 0.01, 0.1, 0.25, 0.40, 0.49\}$ may be of interest. The greatest integer function (eg $\lfloor 7.7 \rfloor = 7$) is used in the following definitions.

Definition 2.14. Consider the smallest proportion $\alpha_{o,n} \in C$ such that $\alpha_{o,n} \geq L(M_n)/n$ and the smallest proportion $1 - \beta_{o,n} \in C$ such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Let $\alpha_{M,n} = \max(\alpha_{o,n}, 1 - \beta_{o,n})$. Then the *two stage symmetrically trimmed mean* $T_{S,n}$ is the $\alpha_{M,n}$ trimmed mean. Hence $T_{S,n}$ is a randomly trimmed mean with $L_n = \lfloor n \quad \alpha_{M,n} \rfloor$ and $U_n = n - L_n$. If $\alpha_{M,n} = 0.50$, then use $T_{S,n} = \text{MED}(n)$.

Definition 2.15. As in the previous definition, consider the smallest proportion $\alpha_{o,n} \in C$ such that $\alpha_{o,n} \geq L(M_n)/n$ and the smallest proportion $1 - \beta_{o,n} \in C$ such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Then the *two stage asymmetrically trimmed mean* $T_{A,n}$ is the $(\alpha_{o,n}, 1 - \beta_{o,n})$ trimmed mean. Hence $T_{A,n}$ is a randomly trimmed mean with $L_n = \lfloor n \quad \alpha_{o,n} \rfloor$ and $U_n = \lfloor n \quad \beta_{o,n} \rfloor$. If $\alpha_{o,n} = 1 - \beta_{o,n} = 0.5$, then use $T_{A,n} = \text{MED}(n)$.

Example 2.11. These two stage trimmed means are almost as easy to compute as the classical trimmed mean, and no knowledge of the unknown parameters is needed to do inference. First, order the data and find the number of cases $L(M_n)$ less than $\text{MED}(n) - k_1\text{MAD}(n)$ and the number of cases $n - U(M_n)$ greater than $\text{MED}(n) + k_2\text{MAD}(n)$. (These are the cases trimmed by the metrically trimmed mean M_n, but M_n need not be computed.) Next, convert these two numbers into percentages and round both percentages up to the nearest integer. For $T_{S,n}$ find the maximum of the two percentages. For example, suppose that there are $n = 205$ cases and M_n trims the smallest 15 cases and the largest 20 cases. Then $L(M_n)/n = 0.073$ and $1 - (U(M_n)/n) = 0.0976$. Hence M_n trimmed the 7.3% smallest cases and the 9.76% largest cases, and $T_{S,n}$ is the 10% trimmed mean while $T_{A,n}$ is the (0.08, 0.10) trimmed mean.

Definition 2.16. The standard error SE$_{RM}$ for the two stage trimmed means given in Definitions 2.10, 2.14 and 2.15 is

$$SE_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$$

where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

$$\frac{[L_n Y^2_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y^2_{(i)} + (n - U_n)Y^2_{(U_n)}] - \ n\ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}. \tag{2.15}$$

Remark 2.3. A simple method for computing $V_{SW}(L_n, U_n)$ has the following steps. First, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \le L_n \\ Y_{(i)}, & L_n + 1 \le i \le U_n \\ Y_{(U_n)}, & i \ge U_n + 1. \end{cases}$$

Then the Winsorized variance is the sample variance $S^2_n(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S^2_n(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \tag{2.16}$$

Notice that the SE given in Definition 2.16 is the SE for the δ trimmed mean where L_n and U_n are fixed constants rather than random.

Application 2.4. Let T_n be the two stage (symmetrically or) asymmetrically trimmed mean that trims the L_n smallest cases and the $n - U_n$ largest cases. Then for the one and two sample procedures described in Section 2.5, use the one sample standard error SE$_{RM}(L_n, U_n)$ given in Definition 2.16 and the t_p distribution where the degrees of freedom $p = U_n - L_n - 1$.

The CIs and tests for the δ trimmed mean and two stage trimmed means given by Applications 2.3 and 2.4 are very similar once L_n has been computed. For example, a large sample 100 $(1 - \alpha)$% confidence interval (CI) for μ_T is

$$(T_n - t_{U_n - L_n - 1, 1 - \frac{\alpha}{2}} SE_{RM}(L_n, U_n), T_n + t_{U_n - L_n - 1, 1 - \frac{\alpha}{2}} SE_{RM}(L_n, U_n)) \quad (2.17)$$

where $P(t_p \le t_{p, 1 - \frac{\alpha}{2}}) = 1 - \alpha/2$ if t_p is from a t distribution with p degrees of freedom. Section 2.7 provides the asymptotic theory for the δ and two stage trimmed means and shows that μ_T is the mean of a truncated distribution. Chapter 3 gives suggestions for k_1 and k_2 while Chapter 4 provides a simulation study comparing the robust and classical point estimators and intervals. Next Examples 2.7, 2.8 and 2.9 are repeated using the intervals based on the two stage trimmed means instead of the median.

Example 2.12. In 1979 a student received the following scores for the nonverbal, verbal, reading, English, math, science, social studies, and problem solving sections of a standardized test:
6, 9, 9, 7, 8, 9, 9, 7.
Assume that if this student took the exam many times, then these scores would be well approximated by a symmetric distribution with mean μ. Find a 95% CI for μ.

Solution. If $T_{A,n}$ or $T_{S,n}$ is used with the metrically trimmed mean that uses $k = k_1 = k_2$, eg $k = 6$, then $\mu_T(a, b) = \mu$. When computing small examples by hand, it is convenient to sort the data:
6, 7, 7, 8, 9, 9, 9, 9.
Thus MED$(n) = (8 + 9)/2 = 8.5$. The ordered residuals $Y_{(i)} - \text{MED}(n)$ are
-2.5, -1.5, -1.5, 0.5, 0.5, 0.5, 0.5, 0.5.
Find the absolute values and sort them to get
0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 2.5.
Then MAD$(n) = 0.5$, MED$(n) - 6MAD(n) = 5.5$, and MED$(n) + 6MAD(n)$ $= 11.5$. Hence no cases are trimmed by the metrically trimmed mean, ie $L(M_n) = 0$ and $U(M_n) = n = 8$. Thus $L_n = \lfloor 8(0) \rfloor = 0$, and $U_n = n - L_n = 8$. Since no cases are trimmed by the two stage trimmed means, the robust interval will have the same endpoints as the classical t–interval. To see this, note that $M_n = T_{S,n} = T_{A,n} = \overline{Y} = (6 + 7 + 7 + 8 + 9 + 9 + 9 + 9)/8 = 8 = W_n(L_n, U_n)$. Now $V_{SW}(L_n, U_n) = (1/7)[\sum_{i=1}^n Y_{(i)}^2 - 8(8^2)]/[8/8]^2$ $= (1/7)[(522 - 8(64)] = 10/7 \approx 1.4286$, and $t_{7,0.975} \approx 2.365$. Hence the 95% CI for μ is $8 \pm 2.365(\sqrt{1.4286/8}) = (7.001, 8.999)$.

Example 2.13. In the last example, what happens if a 6 becomes 66 and a 9 becomes 99? Use $k = 6$ and $T_{A,n}$. Then the ordered data are
7, 7, 8, 9, 9, 9, 66, 99.
Thus MED$(n) = 9$ and MAD$(n) = 1.5$. With $k = 6$, the metrically trimmed mean M_n trims the two values 66 and 99. Hence the left and right trimming proportions of the metrically trimmed mean are 0.0 and $0.25 = 2/8$, respectively. These numbers are also the left and right trimming proportions of $T_{A,n}$ since after converting these proportions into percentages, both percentages are integers. Thus $L_n = \lfloor 0 \rfloor = 0$, $U_n = \lfloor 0.75(8) \rfloor = 6$ and the two stage asymmetrically trimmed mean trims 66 and 99. So $T_{A,n} = 49/6 \approx 8.1667$. To compute the scaled Winsorized variance, use Remark 2.3 to find that the d_i's are
7, 7, 8, 9, 9, 9, 9, 9
and

$$V_{SW} = \frac{S_n^2(d_1, ..., d_8)}{[(6-0)/8]^2} \approx \frac{0.8393}{.5625} \approx 1.4921.$$

Hence the robust confidence interval is $8.1667 \pm t_{5,0.975}\sqrt{1.4921/8} \approx 8.1667 \pm 1.1102 \approx (7.057, 9.277)$. The classical confidence interval $\overline{Y} \pm t_{n-1,0.975} S/\sqrt{n}$ blows up and is equal to $(-2.955, 56.455)$.

Example 2.14. Use $k = 6$ and $T_{A,n}$ to compute a robust CI using the 87 heights from the Buxton (1920) data that includes 5 outliers. The mean height is $\overline{Y} = 1598.862$ while $T_{A,n} = 1695.22$. The classical 95% CI is (1514.206,1683.518) and is more than five times as long as the robust 95% CI which is (1679.907,1710.532). In this example the five outliers can be corrected. For the corrected data, no cases are trimmed and the robust and classical estimators have the same values. The results are $\overline{Y} = 1692.356 = T_{A,n}$ and the robust and classical 95% CIs are both (1678.595,1706.118). Note that the outliers did not have much affect on the robust confidence interval.

2.7 Asymptotics for Two Stage Trimmed Means

Large sample or asymptotic theory is very important for understanding robust statistics. Convergence in distribution, convergence in probability, almost everywhere (sure) convergence, and tightness (bounded in probability) are reviewed in the following remark.

Remark 2.4. Let $X_1, X_2, ...$ be random variables with corresponding cdfs $F_1, F_2, ...$ Let X be a random variable with cdf F. Then X_n *converges in distribution to* X if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at each continuity point t of F. If $X_1, X_2, ...$ and X share a common probability space, then X_n *converges in probability to* X if

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1,$$

for every $\epsilon > 0$, and X_n *converges almost everywhere* (or *almost surely*, or *with probability 1*) to X if

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

The three types of convergence will be denoted by

$$X_n \xrightarrow{D} X, \ X_n \xrightarrow{P} X, \text{ and } X_n \xrightarrow{ae} X,$$

respectively. Notation such as "X_n converges to X ae" will also be used. Serfling (1980, p. 8-9) defines W_n to be *bounded in probability,* $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| > D_\epsilon) < \epsilon$$

for all $n \geq N_\epsilon$, and $W_n = O_P(n^{-\delta})$ if $n^\delta W_n = O_P(1)$. The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

$$n^\delta W_n \xrightarrow{P} 0.$$

Truncated and Winsorized random variables are important because they simplify the asymptotic theory of robust estimators. Let Y be a random variable with continuous cdf F and let $\alpha = F(a) < F(b) = \beta$. Thus α is the *left trimming proportion* and $1 - \beta$ is the *right trimming proportion.* Let $F(a-) = P(Y < a)$. (Refer to Proposition 4.1 for the notation used below.)

Definition 2.17. The *truncated random variable* $Y_T \equiv Y_T(a, b)$ with *truncation points* a and b has cdf

$$F_{Y_T}(y|a, b) = G(y) = \frac{F(y) - F(a-)}{F(b) - F(a-)} \tag{2.18}$$

for $a \leq y \leq b$. Also G is 0 for $y < a$ and G is 1 for $y > b$. The mean and variance of Y_T are

$$\mu_T = \mu_T(a,b) = \int_{-\infty}^{\infty} y dG(y) = \frac{\int_a^b y dF(y)}{\beta - \alpha} \tag{2.19}$$

and

$$\sigma_T^2 = \sigma_T^2(a,b) = \int_{-\infty}^{\infty} (y - \mu_T)^2 dG(y) = \frac{\int_a^b y^2 dF(y)}{\beta - \alpha} - \mu_T^2.$$

See Cramér (1946, p. 247).

Definition 2.18. The *Winsorized random variable*

$$Y_W = Y_W(a,b) = \begin{cases} a, & Y \leq a \\ Y, & a \leq Y \leq b \\ b, & Y \geq b. \end{cases}$$

If the cdf of $Y_W(a,b) = Y_W$ is F_W, then

$$F_W(y) = \begin{cases} 0, & y < a \\ F(a), & y = a \\ F(y), & a < y < b \\ 1, & y \geq b. \end{cases}$$

Since Y_W is a mixture distribution with a point mass at a and at b, the mean and variance of Y_W are

$$\mu_W = \mu_W(a,b) = \alpha a + (1-\beta) b + \int_a^b y dF(y)$$

and

$$\sigma_W^2 = \sigma_W^2(a,b) = \alpha a^2 + (1-\beta) b^2 + \int_a^b y^2 dF(y) - \mu_W^2.$$

Definition 2.19. The *quantile function*

$$F_Q^{-1}(t) = Q(t) = \inf\{y : F(y) \geq t\}. \tag{2.20}$$

Note that $Q(t)$ is the left continuous inverse of F and if F is strictly increasing and continuous, then F has an inverse F^{-1} and $F^{-1}(t) = Q(t)$. The following conditions on the cdf are used.

Regularity Conditions. (R1) Let $Y_1, \ldots, Y_n$ be iid with cdf F.
(R2) Let F be continuous and strictly increasing at $a = Q(\alpha)$ and $b = Q(\beta)$.

The following theorem is proved in Bickel (1965), Stigler (1973a), and Shorack and Wellner (1986, p. 678-679). The α trimmed mean is asymptotically equivalent to the $(\alpha, 1-\alpha)$ trimmed mean. Let T_n be the $(\alpha, 1-\beta)$ trimmed mean. Lemma 2.3 shows that the standard error SE_{RM} given in the previous section is estimating the appropriate asymptotic standard deviation of T_n.

Theorem 2.2. If conditions (R1) and (R2) hold and if $0 < \alpha < \beta < 1$, then

$$\sqrt{n}(T_n - \mu_T(a,b)) \xrightarrow{D} N[0, \frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2}]. \tag{2.21}$$

Lemma 2.3: Shorack and Wellner (1986, p. 680). Assume that regularity conditions (R1) and (R2) hold and that

$$\frac{L_n}{n} \xrightarrow{P} \alpha \text{ and } \frac{U_n}{n} \xrightarrow{P} \beta. \tag{2.22}$$

Then

$$V_{SW}(L_n, U_n) \xrightarrow{P} \frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2}.$$

Since $L_n = \lfloor n\alpha \rfloor$ and $U_n = n - L_n$ (or $L_n = \lfloor n\alpha \rfloor$ and $U_n = \lfloor n\beta \rfloor$) satisfy the above lemma, the standard error SE_{RM} can be used for both trimmed means and two stage trimmed means: $\text{SE}_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$ where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

$$\frac{[L_n Y^2_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y^2_{(i)} + (n - U_n) Y^2_{(U_n)}] - \ n\ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}.$$

Again L_n is the number of cases trimmed to the left and $n - U_n$ is the number of cases trimmed to the right by the trimmed mean.

The following notation will be useful for finding the asymptotic distribution of the two stage trimmed means. Let $a = \text{MED}(Y) - k\text{MAD}(Y)$ and $b = \text{MED}(Y) + k\text{MAD}(Y)$ where $\text{MED}(Y)$ and $\text{MAD}(Y)$ are the population median and median absolute deviation respectively. Let $\alpha = F(a-) =$

$P(Y < a)$ and let $\alpha_o \in C = \{0, 0.01, 0.02, ..., 0.49, 0.50\}$ be the smallest value in C such that $\alpha_o \geq \alpha$. Similarly, let $\beta = F(b)$ and let $1-\beta_o \in C$ be the smallest value in the index set C such that $1 - \beta_o \geq 1 - \beta$. Let $\alpha_o = F(a_o-)$, and let $\beta_o = F(b_o)$. Recall that $L(M_n)$ is the number of cases trimmed to the left and that $n - U(M_n)$ is the number of cases trimmed to the right by the metrically trimmed mean M_n. Let $\alpha_{o,n} \equiv \hat{\alpha}_o$ be the smallest value in C such that $\alpha_{o,n} \geq L(M_n)/n$, and let $1-\beta_{o,n} \equiv 1-\hat{\beta}_o$ be the smallest value in C such that $1-\beta_{o,n} \geq 1-(U(M_n)/n)$. Then the robust estimator $T_{A,n}$ is the $(\alpha_{o,n}, 1-\beta_{o,n})$ trimmed mean while $T_{S,n}$ is the $\max(\alpha_{o,n}, 1 - \beta_{o,n})100\%$ trimmed mean. The following lemma is useful for showing that $T_{A,n}$ is asymptotically equivalent to the $(\alpha_o, 1 - \beta_o)$ trimmed mean and that $T_{S,n}$ is asymptotically equivalent to the $\max(\alpha_o, 1 - \beta_o)$ trimmed mean.

Lemma 2.4: Shorack and Wellner (1986, p. 682-683). Let F have a strictly positive and continuous derivative in some neighborhood of $\text{MED}(Y) \pm k\text{MAD}(Y)$. Assume that

$$\sqrt{n}(MED(n) - MED(Y)) = O_P(1) \tag{2.23}$$

and

$$\sqrt{n}(MAD(n) - MAD(X)) = O_P(1). \tag{2.24}$$

Then

$$\sqrt{n}(\frac{L(M_n)}{n} - \alpha) = O_P(1) \tag{2.25}$$

and

$$\sqrt{n}(\frac{U(M_n)}{n} - \beta) = O_P(1). \tag{2.26}$$

Corollary 2.5. Let $Y_1, ..., Y_n$ be iid from a distribution with cdf F that has a strictly positive and continuous pdf f on its support. Let $\alpha_M = \max(\alpha_o, 1 - \beta_o) \leq 0.49$, $\beta_M = 1 - \alpha_M$, $a_M = F^{-1}(\alpha_M)$, and $b_M = F^{-1}(\beta_M)$. Assume that α and $1 - \beta$ are not elements of $C = \{0, 0.01, 0.02, ..., 0.50\}$. Then

$$\sqrt{n}[T_{A,n} - \mu_T(a_o, b_o)] \xrightarrow{D} N(0, \frac{\sigma^2_W(a_o, b_o)}{(\beta_o - \alpha_o)^2}),$$

and

$$\sqrt{n}[T_{S,n} - \mu_T(a_M, b_M)] \xrightarrow{D} N(0, \frac{\sigma^2_W(a_M, b_M)}{(\beta_M - \alpha_M)^2}).$$

Proof. The first result follows from Theorem 2.2 if the probability that $T_{A,n}$ is the $(\alpha_o, 1-\beta_o)$ trimmed mean goes to one as n tends to infinity. This condition holds if $L(M_n)/n \xrightarrow{D} \alpha$ and $U(M_n)/n \xrightarrow{D} \beta$. But these conditions follow from Lemma 2.4. The proof for $T_{S,n}$ is similar. QED

2.8 L, R, and M Estimators

Definition 2.20. An *L-estimator* is a linear combination of order statistics.

$$T_{L,n} = \sum_{i=1}^{n} c_{n,i} Y_{(i)}$$

for some choice of constants $c_{n,i}$.

The sample mean, median and trimmed mean are L-estimators. Often only a fixed number of the $c_{n,i}$ are nonzero. Examples include the max $= Y_{(n)}$, the min $= Y_{(1)}$, the range $= Y_{(n)} - Y_{(1)}$, and the midrange $= (Y_{(n)} + Y_{(1)})/2$. The following definition and theorem are useful for L-estimators such as the interquartile range and median that use a fixed linear combination of sample quantiles. Recall that the smallest integer function $\lceil x \rceil$ rounds up, eg $\lceil 7.7 \rceil = 8$.

Definition 2.21. The *sample α quantile* $\hat{\xi}_{n,\alpha} = Y_{(\lceil n\alpha \rceil)}$. The *population quantile* $\xi_\alpha = Q(\alpha) = \inf\{y : F(y) \geq \alpha\}$.

Theorem 2.6: Serfling (1980, p. 80). Let $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_k < 1$. Suppose that F has a density f that is positive and continuous in neighborhoods of $\xi_{\alpha_1}, ..., \xi_{\alpha_k}$. Then

$$\sqrt{n}[(\hat{\xi}_{n,\alpha_1}, ..., \hat{\xi}_{n,\alpha_k})^T - (\xi_{\alpha_1}, ..., \xi_{\alpha_k})^T] \xrightarrow{D} N_k(\mathbf{0}, \mathbf{\Sigma})$$

where $\mathbf{\Sigma} = (\sigma_{ij})$ and

$$\sigma_{ij} = \frac{\alpha_i(1-\alpha_j)}{f(\xi_{\alpha_i})f(\xi_{\alpha_j})}$$

for $i \leq j$ and $\sigma_{ij} = \sigma_{ji}$ for $i > j$.

R-estimators are derived from rank tests and include the sample mean and median. See Hettmansperger and McKean (1998).

Definition 2.22. An *M-estimator* of location T with preliminary estimator of scale MAD(n) is computed with at least one Newton step

$$T^{(m+1)} = T^{(m)} + \text{MAD}(n) \; \frac{\sum_{i=1}^n \psi(\frac{Y_i - T^{(m)}}{\text{MAD}(n)})}{\sum_{i=1}^n \psi'(\frac{Y_i - T^{(m)}}{\text{MAD}(n)})}$$

where $T^{(0)} = \text{MED}(n)$. In particular, the *one step M-estimator*

$$T^{(1)} = \text{MED}(n) + \text{MAD}(n) \; \frac{\sum_{i=1}^n \psi(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)})}{\sum_{i=1}^n \psi'(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)})}.$$

The key to M-estimation is finding a good ψ. The sample mean and sample median are M-estimators. Recall that *Newton's method* is an iterative procedure for finding the solution T to the equation $h(T) = 0$ where M-estimators use

$$h(T) = \sum_{i=1}^n \psi(\frac{Y_i - T}{S}).$$

Thus

$$h'(T) = \frac{d}{dT} h(T) = \sum_{i=1}^n \psi'(\frac{Y_i - T}{S})(\frac{-1}{S})$$

where $S = \text{MAD}(n)$ and

$$\psi'(\frac{Y_i - T}{S}) = \frac{d}{dy}\psi(y)$$

evaluated at $y = (Y_i - T)/S$. Beginning with an initial guess $T^{(0)}$, successive terms are generated from the formula $T^{(m+1)} = T^{(m)} - h(T^{(m)})/h'(T^{(m)})$. Often the iteration is stopped if $|T^{(m+1)} - T^{(m)}| < \epsilon$ where ϵ is a small constant. However, one step M-estimators often have the same asymptotic properties as the fully iterated versions. The following example may help clarify notation.

Example 2.15. Huber's M-estimator uses

$$\psi_k(y) = \begin{cases} -k, & y < -k \\ y, & -k \le y \le k \\ k, & y > k. \end{cases}$$

Now

$$\psi'_k(\frac{Y-T}{S}) = 1$$

if $T - kS \leq Y \leq T + kS$ and is zero otherwise (technically the derivative is undefined at $y = \pm\ k$, but assume that Y is a continuous random variable so that the probability of a value occuring on a "corner" of the ψ function is zero). Let L_n count the number of observations $Y_i < \text{MED}(n) - k\text{MAD}(n)$, and let $n - U_n$ count the number of observations $Y_i > \text{MED}(n) + k\text{MAD}(n)$. Set $T^{(0)} = \text{MED}(n)$ and $S = \text{MAD}(n)$. Then

$$\sum_{i=1}^{n} \psi'_k(\frac{Y_i - T^{(0)}}{S}) = U_n - L_n.$$

Since

$$\psi_k(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)}) =$$

$$\begin{cases} -k, & Y_i < \text{MED}(n) - k\text{MAD}(n) \\ \tilde{Y}_i, & \text{MED}(n) - k\text{MAD}(n) \leq Y_i \leq \text{MED}(n) + k\text{MAD}(n) \\ k, & Y_i > \text{MED}(n) + k\text{MAD}(n), \end{cases}$$

where $\tilde{Y}_i = (Y_i - \text{MED}(n))/\text{MAD}(n)$,

$$\sum_{i=1}^{n} \psi_k(\frac{Y_{(i)} - T^{(0)}}{S}) = -kL_n + k(n - U_n) + \sum_{i=L_n+1}^{U_n} \frac{Y_{(i)} - T^{(0)}}{S}.$$

Hence

$$\text{MED}(n) + S\ \frac{\sum_{i=1}^{n} \psi_k(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)})}{\sum_{i=1}^{n} \psi'_k(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)})}$$

$$= \text{MED}(n) + \frac{k\text{MAD}(n)(n - U_n - L_n) + \sum_{i=L_n+1}^{U_n}[Y_{(i)} - \text{MED}(n)]}{U_n - L_n},$$

and Huber's one step M-estimator

$$H_{1,n} = \frac{k\text{MAD}(n)(n - U_n - L_n) + \sum_{i=L_n+1}^{U_n} Y_{(i)}}{U_n - L_n}.$$

2.9 Asymptotic Theory for the MAD

Let MD(n) = MED$(|Y_i - \text{MED}(Y)|,\ i = 1, \ldots, n)$. Since MD$(n)$ is a median and convergence results for the median are well known, see for example Serfling (1980, p. 74-77) or Theorem 2.6 from the previous section, it is simple to prove convergence results for MAD(n). Typically MED(n) = MED$(Y)+O_P(n^{-1/2})$ and MAD(n) = MAD$(Y)+O_P(n^{-1/2})$. Equation (2.27) in the proof of the following lemma implies that if MED(n) converges to MED(Y) ae and MD(n) converges to MAD(Y) ae, then MAD(n) converges to MAD(Y) ae.

Lemma 2.7. If MED(n) = MED$(Y) + O_P(n^{-\delta})$ and MD(n) = MAD$(Y) + O_P(n^{-\delta})$, then MAD(n) = MAD$(Y) + O_P(n^{-\delta})$.

Proof. Let $W_i = |Y_i - \text{MED}(n)|$ and let $V_i = |Y_i - \text{MED}(Y)|$. Then

$$W_i = |Y_i - \text{MED}(Y) + \text{MED}(Y) - \text{MED}(n)| \leq V_i + |\text{MED}(Y) - \text{MED}(n)|,$$

and

$$\text{MAD}(n) = \text{MED}(W_1, \ldots, W_n) \leq \text{MED}(V_1, \ldots, V_n) + |\text{MED}(Y) - \text{MED}(n)|.$$

Similarly

$$V_i = |Y_i - \text{MED}(n) + \text{MED}(n) - \text{MED}(Y)| \leq W_i + |\text{MED}(n) - \text{MED}(Y)|$$

and thus

$$\text{MD}(n) = \text{MED}(V_1, \ldots, V_n) \leq \text{MED}(W_1, \ldots, W_n) + |\text{MED}(Y) - \text{MED}(n)|.$$

Combining the two inequalities shows that

$$\text{MD}(n) - |\text{MED}(Y) - \text{MED}(n)| \leq \text{MAD}(n) \leq \text{MD}(n) + |\text{MED}(Y) - \text{MED}(n)|,$$

or

$$|\text{MAD}(n) - \text{MD}(n)| \leq |\text{MED}(n) - \text{MED}(Y)|. \qquad (2.27)$$

Adding and subtracting MAD(Y) to the left hand side shows that

$$|\text{MAD}(n) - \text{MAD}(Y) - O_P(n^{-\delta})| = O_P(n^{-\delta}) \qquad (2.28)$$

and the result follows. QED

The main point of the following theorem is that the joint distribution of MED(n) and MAD(n) is asymptotically normal. Hence the limiting distribution of MED$(n) + k$MAD(n) is also asymptotically normal for any constant k. The parameters of the covariance matrix are quite complex and hard to estimate. The assumptions of f used in Theorem 2.8 guarantee that MED(Y) and MAD(Y) are unique.

Theorem 2.8: Falk (1997). Let the cdf F of Y be continuous near and differentiable at MED$(Y) = F^{-1}(1/2)$ and MED$(Y)\pm$MAD(Y). Assume that $f = F'$, $f(F^{-1}(1/2)) > 0$, and $A \equiv f(F^{-1}(1/2) - \text{MAD}(Y)) + f(F^{-1}(1/2) + \text{MAD}(Y)) > 0$. Let $C \equiv f(F^{-1}(1/2) - \text{MAD}(Y)) - f(F^{-1}(1/2) + \text{MAD}(Y))$, and let $B \equiv C^2 + 4Cf(F^{-1}(1/2))[1 - F(F^{-1}(1/2) - \text{MAD}(Y)) - F(F^{-1}(1/2) + \text{MAD}(Y))]$. Then

$$\sqrt{n}\left(\begin{pmatrix} \text{MED}(n) \\ \text{MAD}(n) \end{pmatrix} - \begin{pmatrix} \text{MED}(Y) \\ \text{MAD}(Y) \end{pmatrix} \right) \overset{D}{\rightarrow}$$

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_M & \sigma_{M,D} \\ \sigma_{M,D} & \sigma^2_D \end{pmatrix} \right) \tag{2.29}$$

where

$$\sigma^2_M = \frac{1}{4f^2(F^{-1}(\frac{1}{2}))}, \ \sigma^2_D = \frac{1}{4A^2}(1 + \frac{B}{f^2(F^{-1}(\frac{1}{2}))}),$$

and

$$\sigma_{M,D} = \frac{1}{4Af(F^{-1}(\frac{1}{2}))}(1 - 4F(F^{-1}(\frac{1}{2}) + \text{MAD}(Y)) + \frac{C}{f(F^{-1}(\frac{1}{2}))}).$$

Determining whether the population median and mad are unique can be useful. Recall that $F(y) = P(Y \le y)$ and $F(y-) = P(Y < y)$. The median is unique unless there is a flat spot at $F^{-1}(0.5)$, that is, unless there exist a and b with $a < b$ such that $F(a) = F(b) = 0.5$. MAD(Y) may be unique even if MED(Y) is not, see Problem 2.7. If MED(Y) is unique, then MAD(Y) is unique unless F has flat spots at both $F^{-1}(\text{MED}(Y) - \text{MAD}(Y))$ and $F^{-1}(\text{MED}(Y) + \text{MAD}(Y))$. Moreover, MAD$(Y)$ is unique unless there exist $a_1 < a_2$ and $b_1 < b_2$ such that $F(a_1) = F(a_2)$, $F(b_1) = F(b_2)$,

$$P(a_i \le Y \le b_i) = F(b_i) - F(a_i-) \ge 0.5,$$

and

$$P(Y \le a_i) + P(Y \ge b_i) = F(a_i) + 1 - F(b_i-) \ge 0.5$$

for $i = 1, 2$. The following lemma gives some simple bounds for MAD(Y).

Lemma 2.9. Assume MED(Y) and MAD(Y) are unique. a) Then

$$\min\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\} \leq \text{MAD}(Y) \leq$$
$$\max\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\}. \qquad (2.30)$$

b) If Y is symmetric about $\mu = F^{-1}(0.5)$, then the three terms in a) are equal.
c) If the distribution is symmetric about zero, then MAD(Y) = $F^{-1}(0.75)$.
d) If Y is symmetric and continuous with a finite second moment, then

$$\text{MAD}(Y) \leq \sqrt{2\text{VAR}(Y)}.$$

e) Suppose $Y \in [a, b]$. Then

$$0 \leq \text{MAD}(Y) \leq m = \min\{\text{MED}(Y) - a, b - \text{MED}(Y)\} \leq (b - a)/2,$$

and the inequalities are sharp.

Proof. a) This result follows since half the mass is between the upper and lower quartiles and the median is between the two quartiles.

b) and c) are corollaries of a).

d) This inequality holds by Chebyshev's inequality, since

$$P(\ |Y - E(Y)| \geq \text{MAD}(Y)\ ) = 0.5 \geq P(\ |Y - E(Y)| \geq \sqrt{2\text{VAR}(Y)}\ ),$$

and $E(Y) = \text{MED}(Y)$ for symmetric distributions with finite second moments.

e) Note that if MAD(Y) $> m$, then either MED(Y) − MAD(Y) $< a$ or MED(Y) + MAD(Y) $> b$. Since at least half of the mass is between a and MED(Y) and between MED(Y) and b, this contradicts the definition of MAD(Y). To see that the inequalities are sharp, note that if at least half of the mass is at some point $c \in [a, b]$, than MED(Y) = c and MAD(Y) = 0. If each of the points a, b, and c has 1/3 of the mass where $a < c < b$, then MED(Y) = c and MAD(Y) = m. QED

Many other results for MAD(Y) and MAD(n) are possible. For example, note that Lemma 2.9 b) implies that when Y is symmetric, MAD(Y) = $F^{-1}(3/4) - \mu$ and $F(\mu + \text{MAD}(Y)) = 3/4$. Also note that MAD($Y$) and the interquartile range IQR(Y) are related by

$$2\text{MAD}(Y) = \text{IQR}(Y) \equiv F^{-1}(0.75) - F^{-1}(0.25)$$

when Y is symmetric. Moreover, results similar to those in Lemma 2.9 hold for MAD(n) with quantiles replaced by order statistics. One way to see this is to note that the distribution with a point mass of $1/n$ at each observation $Y_1, \ldots, Y_n$ will have a population median equal to MED(n). To illustrate the outlier resistance of MAD(n) and MED(n), consider the following lemma.

Lemma 2.10. If $Y_1, \ldots, Y_n$ are n fixed points, and if $m \leq n-1$ arbitrary points $W_1, \ldots, W_m$ are added to form a sample of size $n+m$, then

$$\text{MED}(n+m) \in [Y_{(1)}, Y_{(n)}] \text{ and } 0 \leq \text{MAD}(n+m) \leq Y_{(n)} - Y_{(1)}. \qquad (2.31)$$

Proof. Let the order statistics of $Y_1, \ldots, Y_n$ be $Y_{(1)} \leq \cdots \leq Y_{(n)}$. By adding a single point W, we can cause the median to shift by half an order statistic, but since at least half of the observations are to each side of the sample median, we need to add at least $m = n-1$ points to move MED($n+m$) to $Y_{(1)}$ or to $Y_{(n)}$. Hence if $m \leq n-1$ points are added, $[\text{MED}(n+m) - (Y_{(n)} - Y_{(1)}), \text{MED}(n+m) + (Y_{(n)} - Y_{(1)})]$ contains at least half of the observations and $\text{MAD}(n+m) \leq Y_{(n)} - Y_{(1)}$. QED

Hence if $Y_1, \ldots, Y_n$ are a random sample with cdf F and if $W_1, \ldots, W_{n-1}$ are arbitrary, then the sample median and mad of the combined sample, MED($n+n-1$) and MAD($n+n-1$), are bounded by quantities from the random sample from F.

2.10 Summary

1) Given a small data set, recall that

$$\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

and the *sample variance*

$$S^2 = S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1},$$

and the *sample standard deviation* (SD)

$$S = S_n = \sqrt{S_n^2}.$$

If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then the $Y_{(i)}$'s are called the *order statistics*. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used. To find the sample median, sort the data from smallest to largest and find the middle value or values.

The *sample median absolute deviation*

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|,\ i = 1, \ldots, n).$$

To find $\text{MAD}(n)$, find $D_i = |Y_i - \text{MED}(n)|$, then find the sample median of the D_i by ordering them from smallest to largest and finding the middle value or values.

2) Find the population median $M = \text{MED}(Y)$ by solving the equation $F(M) = 0.5$ for M where the cdf $F(y) = P(Y \leq y)$. If Y has a pdf $f(y)$ that is symmetric about μ, then $M = \mu$. If $W = a + bY$, then $\text{MED}(W) = a + b\text{MED}(Y)$. Often $a = \mu$ and $b = \sigma$.

3) To find the population median absolute deviation $D = \text{MAD}(Y)$, first find $M = \text{MED}(Y)$ as in 2) above.
a) Then solve $F(M + D) - F(M - D) = 0.5$ for D.
b) If Y has a pdf that is symmetric about μ, then let $U = y_{0.75}$ where $P(Y \leq y_\alpha) = \alpha$, and y_α is the 100αth percentile of Y for $0 < \alpha < 1$. Hence $M = y_{0.5}$ is the 50th percentile and U is the 75th percentile. Solve $F(U) = 0.75$ for U. Then $D = U - M$.
c) If $W = a + bY$, then $\text{MAD}(W) = |b|\text{MAD}(Y)$.

$\text{MED}(Y)$ and $\text{MAD}(Y)$ need not be unique, but for "brand name" continuous random variables, they are unique.

4) A large sample 100 $(1 - \alpha)$% confidence interval (CI) for θ is

$$\hat{\theta} \pm t_{p,1-\frac{\alpha}{2}} SE(\hat{\theta})$$

where $P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ if t_p is from a t distribution with p degrees of freedom. We will use 95% CIs so $\alpha = 0.05$ and $t_{p,1-\frac{\alpha}{2}} = t_{p,0.975} \approx 1.96$ for $p > 20$. Be able to find $\hat{\theta}$, p and $SE(\hat{\theta})$ for the following three estimators.

a) The **classical CI for the population mean** $\theta = \mu$ uses $\hat{\theta} = \overline{Y}$, $p = n - 1$ and $SE(\overline{Y}) = S/\sqrt{n}$.

Let $\lfloor x \rfloor$ denote the "greatest integer function". Then $\lfloor x \rfloor$ is the largest integer less than or equal to x (eg, $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (eg, $\lceil 7.7 \rceil = 8$).

b) Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\ \rceil$. Then the **CI for the population median** $\theta = \text{MED}(Y)$ uses $\hat{\theta} = \text{MED}(n)$, $p = U_n - L_n - 1$ and

$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}).$$

c) The 25% trimmed mean

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)}$$

where $L_n = \lfloor n/4 \rfloor$ and $U_n = n - L_n$. That is, order the data, delete the L_n smallest cases and the L_n largest cases and take the sample mean of the remaining $U_n - L_n$ cases. The 25% trimmed mean is estimating the population truncated mean

$$\mu_T = \int_{y_{0.25}}^{y_{0.75}} 2y f_Y(y) dy.$$

To perform inference, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \leq L_n \\ Y_{(i)}, & L_n + 1 \leq i \leq U_n \\ Y_{(U_n)}, & i \geq U_n + 1. \end{cases}$$

(The "half set" of retained cases is not changed, but replace the L_n smallest deleted cases by the smallest retained case $Y_{(L_n+1)}$ and replace the L_n largest deleted cases by the largest retained case $Y_{(U_n)}$.) Then the Winsorized variance is the sample variance $S_n^2(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}.$$

Then the **CI for the population truncated mean** $\theta = \mu_T$ uses $\hat{\theta} = T_n$, $p = U_n - L_n - 1$ and

$$SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n}.$$

2.11 Complements

Chambers, Cleveland, Kleiner and Tukey (1983) is an excellent source for graphical procedures such as quantile plots, QQ-plots, and box plots.

The confidence intervals and tests for the sample median and 25% trimmed mean can be modified for certain types of **censored data** as can the robust point estimators based on MED(n) and MAD(n). Suppose that in a reliability study the Y_i are failure times and the study lasts for T hours. Let $Y_{(R)} < T$ but $T < Y_{(R+1)} < \cdots < Y_{(n)}$ so that only the first R failure times are known and the last $n - R$ failure times are unknown but greater than T (similar results hold if the first L failure times are less than T but unknown while the failure times $T < Y_{(L+1)} < \cdots < Y_{(n)}$ are known). Then create a pseudo sample $Z_{(i)} = Y_{(R)}$ for $i > R$ and $Z_{(i)} = Y_{(i)}$ for $i \leq R$. Then compute the robust estimators based $Z_1, ..., Z_n$. These estimators will be identical to the estimators based on $Y_1, ..., Y_n$ (no censoring) if the amount of right censoring is moderate. For a one parameter family, nearly half of the data can be right censored if the estimator is based on the median. If the sample median and MAD are used for a two parameter family, the proportion of right censored data depends on the skewness of the distribution. Symmetric data can tolerate nearly 25% right censoring, right skewed data a larger percentage, and left skewed data a smaller percentage. See Olive (2006). He and Fung (1999) present an alternative robust method that also works well for this type of censored data.

Huber (1981, p. 74-75) and Chen (1998) show that the sample median minimizes the asymptotic bias for estimating MED(Y) for the family of symmetric contaminated distributions, and Huber (1981) concludes that since the asymptotic variance is going to zero for reasonable estimators, MED(n) is the estimator of choice for large n. Hampel, Ronchetti, Rousseeuw, and Stahel (1986, p. 133-134, 142-143) contains some other optimality properties of MED(n) and MAD(n). Larocque and Randles (2008), McKean and Schrader (1984) and Bloch and Gastwirth (1968) are useful references for estimating the SE of the sample median.

Section 2.4 is based on Olive (2005b). Several other approximations for the standard error of the sample median $SE(\text{MED}(n))$ could be used.

a) McKean and Schrader (1984) proposed

$$SE(\text{MED}(n)) = \frac{Y_{(n-c+1)} - Y_{(c)}}{2z_{1-\frac{\alpha}{2}}}$$

where $c = (n+1)/2 \ - \ z_{1-\alpha/2}\sqrt{n/4}$ is rounded up to the nearest integer. This estimator was based on the half length of a distribution free 100 $(1-\alpha)$% CI $(Y_{(c)}, Y_{(n-c+1)})$ for MED(Y). Use the t_p approximation with $p = \lfloor 2\sqrt{n} \rfloor - 1$.

b) This proposal is also due to Bloch and Gastwirth (1968). Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil 0.5n^{0.8} \ \rceil$ and use

$$SE(\text{MED}(n)) = \frac{Y_{(U_n)} - Y_{(L_n+1)}}{2n^{0.3}}.$$

Use the t_p approximation with $p = U_n - L_n - 1$.

c) MED(n) is the 50% trimmed mean, so trimmed means with trimming proportions close to 50% should have an asymptotic variance close to that of the sample median. Hence an ad hoc estimator is

$$SE(\text{MED}(n)) = SE_{RM}(L_n, U_n)$$

where $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4} \ \rceil$ and $SE_{RM}(L_n, U_n)$ is given by Definition 2.16 on p. 46. Use the t_p approximation with $p = U_n - L_n - 1$.

In a small simulation study (see Section 4.6), the proposal in Application 2.2 using $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4} \ \rceil$ seemed to work best. Using $L_n = \lfloor n/2 \rfloor - \lceil 0.5n^{0.8} \ \rceil$ gave better coverages for symmetric data but is vulnerable to a single cluster of shift outliers if $n \leq 100$.

An enormous number of procedures have been proposed that have better robustness or asymptotic properties than the classical procedures when outliers are present. Huber (1981), Hampel, Ronchetti, Rousseeuw, and Stahel (1986) and Staudte and Sheather (1990) are standard references. **For location–scale families, we recommend using the robust estimators from Application 2.1 to create a highly robust asymptotically efficient cross checking estimator.** See Olive (2006) and He and Fung (1999). Joiner and Hall (1983) compare and contrast L, R, and M-estimators while Jureckova and Sen (1996) derive the corresponding asymptotic theory. Mosteller (1946) is an early reference for L-estimators. Bickel (1965), Dixon and Tukey (1968), Stigler (1973a), Tukey and McLaughlin (1963) and Yuen (1974) discuss trimmed and Winsorized means while Prescott (1978) examines adaptive methods of trimming. Bickel (1975) examines one-step M-estimators, and Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972) present a simulation study comparing trimmed means and M-estimators. A robust method for massive data sets is given in Rousseeuw and Bassett (1990).

Hampel (1985) considers metrically trimmed means. Shorack (1974) and Shorack and Wellner (1986, section 19.3) derive the asymptotic theory for a large class of robust procedures for the iid location model. Special cases include trimmed, Winsorized, metrically trimmed, and Huber type skipped means. Also see Kim (1992) and papers in Hahn, Mason, and Weiner (1991). Olive (2001) considers two stage trimmed means.

Shorack and Wellner (1986, p. 3) and Parzen (1979) discuss the quantile function while Stigler (1973b) gives historic references to trimming techniques, M-estimators, and to the asymptotic theory of the median. David (1995, 1998), Field (1985), and Sheynin (1997) also contain historical references.

Scale estimators are essential for testing and are discussed in Falk (1997), Hall and Welsh (1985), Lax (1985), Rousseeuw and Croux (1992, 1993), and Simonoff (1987b). There are many alternative approaches for testing and confidence intervals. Guenther (1969) discusses classical confidence intervals while Gross (1976) considers robust confidence intervals for symmetric distributions. Basically all of the methods which truncate or Winsorize the tails worked. Wilcox (2005) uses trimmed means for testing while Kafadar (1982) uses the biweight M-estimator. Also see Horn (1983). Hettmansperger and McKean (1998) consider rank procedures.

Wilcox (2005) gives an excellent discussion of the problems that outliers and skewness can cause for the one and two sample t–intervals, the t–test, tests for comparing 2 groups and the ANOVA F test. Wilcox (2005) replaces ordinary population means by truncated population means and uses trimmed means to create analogs of one, two, and three way anova, multiple comparisons, and split plot designs.

Often a large class of estimators is defined and picking out good members from the class can be difficult. Freedman and Diaconis (1982) and Clarke

(1986) illustrate some potential problems for M-estimators. Jureckova and Sen (1996, p. 208) show that under symmetry a large class of M-estimators is asymptotically normal, but the asymptotic theory is greatly complicated when symmetry is not present. Stigler (1977) is a very interesting paper and suggests that Winsorized means (which are often called “trimmed means” when the trimmed means from Definition 2.10 do not appear in the paper) are adequate for finding outliers.

2.12 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

2.1. Write the location model in matrix form.

2.2. Let $f_Y(y)$ be the pdf of Y. If $W = \mu + Y$ where $-\infty < \mu < \infty$, show that the pdf of W is $f_W(w) = f_Y(w - \mu)$.

2.3. Let $f_Y(y)$ be the pdf of Y. If $W = \sigma Y$ where $\sigma > 0$, show that the pdf of W is $f_W(w) = (1/\sigma) f_Y(w/\sigma)$.

2.4. Let $f_Y(y)$ be the pdf of Y. If $W = \mu + \sigma Y$ where $-\infty < \mu < \infty$ and $\sigma > 0$, show that the pdf of W is $f_W(w) = (1/\sigma) f_Y((w - \mu)/\sigma)$.

2.5. Use Theorem 2.8 to find the limiting distribution of $\sqrt{n}(\text{MED}(n) - \text{MED}(Y))$.

2.6. The interquartile range $\text{IQR}(n) = \hat{\xi}_{n,0.75} - \hat{\xi}_{n,0.25}$ and is a popular estimator of scale. Use Theorem 2.6 to show that

$$\sqrt{n}\frac{1}{2}(\text{IQR}(n) - \text{IQR}(Y)) \xrightarrow{D} N(0, \sigma_A^2)$$

where

$$\sigma_A^2 = \frac{1}{64}\left[\frac{3}{[f(\xi_{3/4})]^2} - \frac{2}{f(\xi_{3/4})f(\xi_{1/4})} + \frac{3}{[f(\xi_{1/4})]^2}\right].$$

2.7. Let the pdf of Y be $f(y) = 1$ if $0 < y < 0.5$ or if $1 < y < 1.5$. Assume that $f(y) = 0$, otherwise. Then Y is a mixture of two uniforms, one $U(0, 0.5)$ and the other $U(1, 1.5)$. Show that the population median MED(Y) is not unique but the population mad MAD(Y) is unique.

2.8. a) Let $L_n = 0$ and $U_n = n$. Prove that $\text{SE}_{RM}(0, n) = S/\sqrt{n}$. In other words, the SE given by Definition 2.16 reduces to the SE for the sample mean if there is no trimming.

b) Prove Remark 2.3:

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{[(U_n - L_n)/n]^2}.$$

2.9. Find a 95% CI for μ_T based on the 25% trimmed mean for the following data sets. Follow Examples 2.12 and 2.13 closely with $L_n = \lfloor 0.25n \rfloor$ and $U_n = n - L_n$.

a) 6, 9, 9, 7, 8, 9, 9, 7

b) 66, 99, 9, 7, 8, 9, 9, 7

2.10. Consider the data set 6, 3, 8, 5, and 2. Show work.

a) Find the sample mean $\overline{Y}$.

b) Find the standard deviation S

c) Find the sample median MED(n).

d) Find the sample median absolute deviation MAD(n).

2.11*. The Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) is listed below.

```
1.2  2.4  1.3  1.3  0.0  1.0  1.8  0.8  4.6  1.4
```

a) Find the sample mean $\overline{Y}$.

b) Find the sample standard deviation S.

c) Find the sample median MED(n).

d) Find the sample median absolute deviation MAD(n).

e) Plot the data. Are any observations unusually large or unusually small?

2.12*. Consider the following data set on Spring 2004 Math 580 homework scores.

```
66.7  76.0  89.7  90.0  94.0  94.0  95.0  95.3  97.0  97.7
```

Then $\overline{Y} = 89.54$ and $S^2 = 103.3604$.

a) Find SE($\overline{Y}$).

b) Find the degrees of freedom p for the classical CI based on $\overline{Y}$.

Parts c)-g) refer to the CI based on MED(n).

c) Find the sample median MED(n).

d) Find L_n.

e) Find U_n.

f) Find the degrees of freedom p.

g) Find SE(MED(n)).

2.13*. Consider the following data set on Spring 2004 Math 580 homework scores.

```
66.7  76.0  89.7  90.0  94.0  94.0  95.0  95.3  97.0  97.7
```

Consider the CI based on the 25% trimmed mean.

a) Find L_n.

b) Find U_n.

c) Find the degrees of freedom p.

d) Find the 25% trimmed mean T_n.

e) Find $d_1, ..., d_{10}$.

f) Find $\overline{d}$.

g) Find $S^2(d_1, ..., d_{10})$.

e) Find SE(T_n).

2.14. Consider the data set 6, 3, 8, 5, and 2.

a) Referring to Application 2.2 on p. 37, find L_n, U_n, p and SE(MED(n)).

b) Referring to Application 2.3 on p. 38, let T_n be the 25% trimmed mean. Find L_n, U_n, p, T_n and SE(T_n).

R/Splus problems

2.15*. Use the commands

```
height <- rnorm(87, mean=1692, sd = 65)
height[61:65] <- 19.0
```

to simulate data similar to the Buxton heights. Make a plot similar to Figure 2.1 using the following *R/Splus* commands.

```
> par(mfrow=c(2,2))
> plot(height)
> title("a) Dot plot of heights")
> hist(height)
> title("b) Histogram of heights")
> length(height)
[1] 87
> val <- quantile(height)[4] - quantile(height)[2]
> val
   75%
 103.5
> wid <- 4*1.06*min(sqrt(var(height)),val/1.34)*(87^(-1/5))
> wid
[1] 134.0595
> dens<- density(height,width=wid)
> plot(dens$x,dens$y)
> lines(dens$x,dens$y)
> title("c) Density of heights")
> boxplot(height)
> title("d) Boxplot of heights")
```

2.16***.** The following command computes MAD(n).

```
mad(y, constant=1)
```

a) Let $Y \sim N(0,1)$. Estimate MAD(Y) with the following commands.

```
y <- rnorm(10000)
mad(y, constant=1)
```

b) Let $Y \sim$ EXP(1). Estimate MAD(Y) with the following commands.

```
y <- rexp(10000)
mad(y, constant=1)
```

2.17***.** The following commands computes the α trimmed mean. The default uses $tp = 0.25$ and gives the 25% trimmed mean.

```
 tmn <-
function(x, tp = 0.25)
```

```
{
mean(x, trim = tp)
}
```

a) Compute the 25% trimmed mean of 10000 simulated $N(0,1)$ random variables with the following commands.

```
y <- rnorm(10000)
tmn(y)
```

b) Compute the mean and 25% trimmed mean of 10000 simulated EXP(1) random variables with the following commands.

```
y <- rexp(10000)
mean(y)
tmn(y)
```

2.18. The following *R/Splus* function computes the metrically trimmed mean.

```
metmn <-
function(x, k = 6)
{
madd <- mad(x, constant = 1)
med <- median(x)
mean(x[(x >= med - k * madd) & (x <= med + k * madd)])
}
```

Compute the metrically trimmed mean of 10000 simulated $N(0,1)$ random variables with the following commands.

```
y <- rnorm(10000)
metmn(y)
```

Warning: For the following problems, use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *ratmn*, will display the code for the function. Use the `args` command, eg *args(ratmn)*, to display the needed arguments for the function.

2.19. Download the *R/Splus* function `ratmn` that computes the two stage asymmetrically trimmed mean $T_{A,n}$. Compute the $T_{A,n}$ for 10000 simulated $N(0,1)$ random variables with the following commands.

```
y <- rnorm(10000)
ratmn(y)
```

2.20. Download the *R/Splus* function `rstmn` that computes the two stage symmetrically trimmed mean $T_{S,n}$. Compute the $T_{S,n}$ for 10000 simulated $N(0,1)$ random variables with the following commands.

```
y <- rnorm(10000)
rstmn(y)
```

2.21*. a) Download the `cci` function which produces a classical CI. The default is a 95% CI.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *cci(height)*.

2.22*. a) Download the *R/Splus* function `medci` that produces a CI using the median and the Bloch and Gastwirth SE.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *medci(height)*.

2.23*. a) Download the *R/Splus* function `tmci` that produces a CI using the 25% trimmed mean as a default.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *tmci(height)*.

2.24. a) Download the *R/Splus* function `atmci` that produces a CI using $T_{A,n}$.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *atmci(height)*.

2.25. a) Download the *R/Splus* function `stmci` that produces a CI using $T_{S,n}$.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *stmci(height)*.

2.26. a) Download the *R/Splus* function `med2ci` that produces a CI using the median and $SE_{RM}(L_n, U_n)$.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *med2ci(height)*.

2.27. a) Download the *R/Splus* function `cgci` that produces a CI using $T_{S,n}$ and the coarse grid $C = \{0, 0.01, 0.1, 0.25, 0.40, 0.49\}$.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *cgci(height)*.

2.28. a) Bloch and Gastwirth (1968) suggest using

$$SE(\text{MED}(n)) = \frac{\sqrt{n}}{4m}[Y_{([n/2]+m)} - Y_{([n/2]-m)}]$$

where $m \to \infty$ but $n/m \to 0$ as $n \to \infty$. Taking $m = 0.5n^{0.8}$ is optimal in some sense, but not as resistant as the choice $m = \sqrt{n/4}$. Download the *R/Splus* function `bg2ci` that is used to simulate the CI that uses MED(n) and the "optimal" BG SE.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *bg2ci(height)*.

2.29. a) Enter the following commands to create a function that produces a Q plot.

```
qplot<-
function(y)
{ plot(sort(y), ppoints(y))
title("QPLOT")}
```

b) Make a Q plot of the height data from Problem 2.15 with the following command.

```
qplot(height)
```

c) Make a Q plot for $N(0, 1)$ data with the following commands.

```
Y <- rnorm(1000)
qplot(y)
```