Chapter 2
The Location Model

The location model is used when there is one variable Y, such as height, of interest. The location model is a special case of the multivariate location and dispersion model, where there are p variables $x_1, ..., x_p$ of interest, such as height and weight if $p = 2$. See Chapter 3.

The *location model* is

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

where $e_1, ..., e_n$ are error random variables, often independent and identically distributed (iid) with zero mean. For example, if the Y_i are iid from a normal distribution with mean μ and variance σ^2, written $Y_i \sim N(\mu, \sigma^2)$, then the e_i are iid with $e_i \sim N(0, \sigma^2)$. The location model 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 cumulative distribution function (cdf) F, median MED(Y), mean $E(Y)$, and variance $V(Y)$ if they exist. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$. The ith *case* is Y_i.

An 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 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

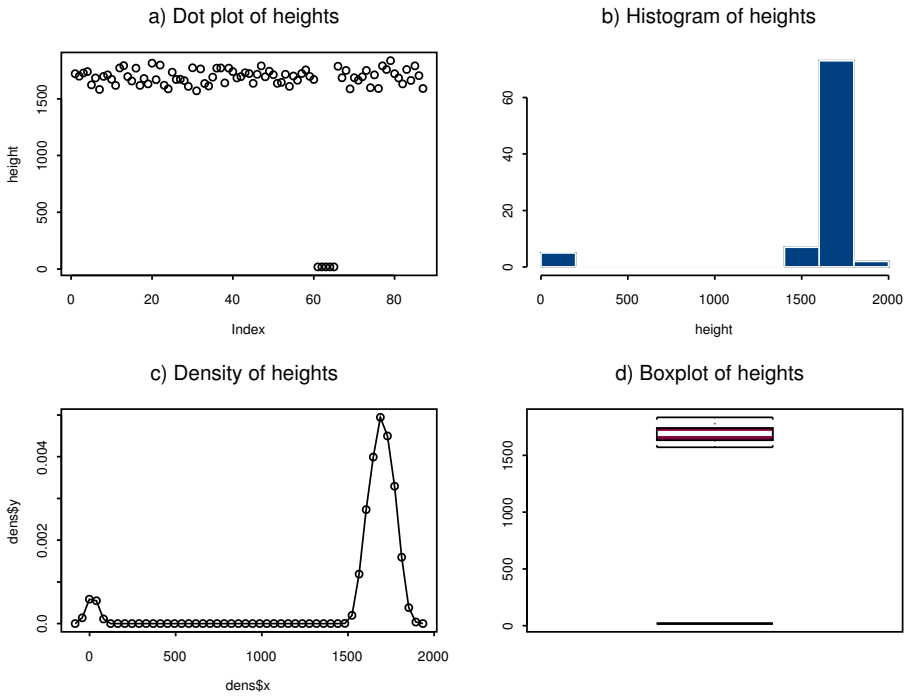

Fig. 2.1 Dot plot, histogram, density estimate, and box plot for heights from Buxton (1920).

that the presence of outliers can be detected in all four plots, but the dot plot of case index versus Y may be easiest to use. Problem 2.22 shows how to make a similar figure.

2.1 Four Essential Statistics

Point estimation is one of the oldest problems in statistics and four 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; i.e., 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*

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

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

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

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 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 these estimators are nonparametric estimators of the corresponding population quantities, they are useful for a very wide range of distributions. Since MAD(n) is the median of n distances, at least half of the observations are within a distance MAD(n) of MED(n) and at least half of the observations are a distance of MAD(n) or more away from MED(n). For small data sets, sort the data. Then the median is the middle observation if n is odd, and the average of the two middle observations if n is even.

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

2.2 A Note on Notation

Table 2.1 Some commonly used notation.

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

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 MED(n), MAD(n), MED$(Y_i, i = 1, ..., n)$, or 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 med(n), 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 [MED(Y) − MAD(Y), MED(Y) + MAD(Y)] and at least half of the mass of the distribution is outside of the interval (MED(Y) − MAD(Y), MED(Y) + MAD(Y)). In other words, MAD(Y) is any value such that

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

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

Warning. There is often no simple formula for MAD(Y). For example, if $Y \sim \text{Gamma}(\nu, \lambda)$, then 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.

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 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 mad and median for some "brand name" distributions. The distributions are location–scale families except for the ex-

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

NAME	Section	MED(Y)	MAD(Y)
Cauchy C(μ, σ)	11.4.3	μ	σ
double exponential DE(θ, λ)	11.4.6	θ	0.6931λ
exponential EXP(λ)	11.4.7	0.6931λ	$\lambda/2.0781$
two parameter exponential EXP(θ, λ)	11.4.8	$\theta + 0.6931\lambda$	$\lambda/2.0781$
half normal HN(μ, σ)	11.4.12	$\mu + 0.6745\sigma$	$0.3991\ \sigma$
largest extreme value LEV(θ, σ)	11.4.13	$\theta + 0.3665\sigma$	0.7670σ
logistic L(μ, σ)	11.4.14	μ	$1.0986\ \sigma$
normal N(μ, σ^2)	11.4.19	μ	0.6745σ
Rayleigh R(μ, σ)	11.4.23	$\mu + 1.1774\sigma$	0.4485σ
smallest extreme value SEV(θ, σ)	11.4.24	$\theta - 0.3665\sigma$	0.7670σ
t_p	11.4.25	0	$t_{p,3/4}$
uniform U(θ_1, θ_2)	11.4.27	$(\theta_1 + \theta_2)/2$	$(\theta_2 - \theta_1)/4$

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

Name	Section	MED(Y)	MAD(Y)
binomial BIN(k,ρ)	11.4.1	$k\rho$	$0.6745\sqrt{k\rho(1-\rho)}$
chi-square χ^2_p	11.4.5	$p - 2/3$	$0.9536\sqrt{p}$
gamma G(ν, λ)	11.4.9	$\lambda(\nu - 1/3)$	$\lambda\sqrt{\nu}/1.483$

ponential and t_p distributions. The notation t_p denotes a t distribution with p degrees of freedom while $t_{p,\delta}$ is the δ quantile of the t_p distribution, i.e. $P(t_p \leq t_{p,\delta}) = \delta$. Hence $t_{p,0.5} = 0$ is the population median. The second column of Table 2.2 gives the subsection of Chapter 11 where the random variable is described further. For example, the exponential (λ) random variable is described in Section 11.4.7. Table 2.3 presents approximations for the binomial, chi-square and gamma distributions.

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

Theorem 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

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

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

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

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 $\text{MAD}(Y)$ is to generate a simulated random sample $Y_1, ..., Y_n$ for $n \approx 10000$ and then compute $\text{MAD}(n)$. 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 $\text{MED}(Z) = 0$ and $\text{MED}(W) = \mu + \sigma\text{MED}(Z) = \mu$. 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. Theorem 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 command.

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

Hence $\text{MAD}(W) \approx 0.6745\sigma$.

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 $\exp(\log(1/2)) = \exp(-\log(2)) = 0.5$, $\text{MED}(W) = \log(2)\lambda$. Since the exponential distribution is a scale family with scale parameter λ, $\text{MAD}(W) = D\lambda$ 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 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(0.481)
[1] 0.9995264
> tem(0.482)
[1] 1.001763
> 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 $\Phi(z_\delta) = \delta$. Note that $z_\delta = \Phi^{-1}(\delta)$. Suppose Y has a mixture distribution with cdf $F_Y(y) = (1 - \gamma)F_W(y) + \gamma F_C(y)$ where $W \sim N(\mu, \sigma^2)$ and C is a random variable far to the right of μ. See Remark 11.1. 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, $F_C(MED(Y)) \approx 0$ and

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

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, $F_C(\text{MED}(Y) + \text{MAD}(Y)) \approx 0$ and

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

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 $\text{MED}(n) = \text{MED}(Y)$ and $\text{MAD}(n) = \text{MAD}(Y)$. In particular, the location and scale parameters of a location–scale family can often be estimated robustly using $c_1\text{MED}(n)$ and $c_2\text{MAD}(n)$ where c_1 and c_2 are appropriate constants. Table 2.4 shows some of the point estimators and Chapter 11 has additional examples. 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, $\text{MED}(Y) = \mu$ and $\text{MAD}(Y) \approx 0.6745\sigma$. Hence $\hat{\mu} = \text{MED}(n)$ and $\hat{\sigma} \approx \text{MAD}(n)/0.6745 \approx 1.483\text{MAD}(n)$.

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

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

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)	$\hat{\theta_1} = \text{MED}(n) - 2\text{MAD}(n)$	$\hat{\theta_2} = \text{MED}(n) + 2\text{MAD}(n)$

for large ν. If X is $N(\mu, \sigma^2)$ then $\text{MAD}(X) \approx \sigma/1.483$. Hence $\text{MAD}(Y) \approx \lambda\sqrt{\nu}/1.483$. Assuming that ν is large, solve $\text{MED}(n) = \lambda\nu$ and $\text{MAD}(n) = \lambda\sqrt{\nu}/1.483$ 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 $\text{MED}(n)$ and $\text{MAD}(n)$ 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 2.11. A special case of this result follows. Let ξ_δ be the δ quantile of Y. Thus $P(Y \leq \xi_\delta) = \delta$. 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 Prediction Intervals and the Shorth

Prediction intervals are important. Applying certain prediction intervals or prediction regions to the bootstrap sample will result in confidence intervals or confidence regions. The prediction intervals and regions are based on samples of size n, while the bootstrap sample size is $B = B_n$. Hence this section and the following section are important.

Definition 2.10. Consider predicting a future test value Y_f given a training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the *population shorth*: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

If Y_f has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1-\delta$ as $n \to \infty$. The interpretation of a 100 $(1-\delta)\%$ PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where the k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI, then the probability that $Y_{fi} \in PI_i$ for j of the PIs approximately follows a binomial$(k, \rho = 1-\delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number J, say. Secondly, many confidence intervals work well for large classes of distributions while many prediction intervals assume that the distribution of the data is known up to some unknown parameters. Usually the $N(\mu, \sigma^2)$ distribution is assumed, and the parametric PI may not perform well if the normality assumption is violated.

The following two nonparametric PIs often work well if the Y_i are iid and $n \geq 50$. Consider the location model, $Y_i = \mu + e_i$, where $Y_1, ..., Y_n, Y_f$ are iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of n iid random variables $Y_1, ..., Y_n$ that make up the training data. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1-\delta/2) \rceil$ where $\lceil x \rceil$ is the smallest integer $\geq x$. For example, $\lceil 7.7 \rceil = 8$. See Frey (2013) for references for the following PI.

Definition 2.11. The large sample $100(1-\delta)\%$ *nonparametric prediction interval* for Y_f is

$$[Y_{(k_1)}, Y_{(k_2)}] \tag{2.8}$$

where $0 < \delta < 1$.

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals. With the Y_i and $Y_{(i)}$ as in the above paragraph above Definition 2.11, let the shortest closed interval containing at least c of the Y_i be

$$\text{shorth(c)} = [\mathrm{Y}_{(\mathrm{s})}, \mathrm{Y}_{(\mathrm{s+c-1})}]. \tag{2.9}$$

Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{2.10}$$

Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$. An interesting fact is that the maximum undercoverage occurs for the family of uniform $U(\theta_1, \theta_2)$ distributions. See Section 11.4.27. Frey (2013) used the following shorth PI.

Definition 2.12. The large sample $100(1-\delta)\%$ *shorth PI* is

$$[Y_{(s)}, Y_{(s+c-1)}] \text{ where } \mathrm{c} = \min(\mathrm{n}, \lceil \mathrm{n}[1 - \delta + 1.12\sqrt{\delta/\mathrm{n}}\ ]\ \rceil). \tag{2.11}$$

A problem with the prediction intervals that cover $\approx 100(1-\delta)\%$ of the training data cases Y_i, such as (2.11), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data than on training data. For iid data, Frey (2013) used (2.11) to correct for undercoverage.

Example 2.7. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76

order data: 76 78 89 111 778

                  13 = 89 - 76
```

```
                        33 = 111 - 78

                              689 = 778 - 89
shorth(3) = [76,89]
```

Remark 2.1. The sample shorth converges to the population shorth rather slowly. Grübel (1988) shows that under regularity conditions for iid data, the length and center of the shorth($k_n = \lceil n(1-\delta) \rceil$) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval.

Remark 2.2. The large sample $100(1-\delta)\%$ shorth PI (2.11) may or may not be asymptotically optimal if the $100(1-\delta)\%$ population shorth is $[L_s, U_s]$ and $F(x)$ is not strictly increasing in intervals $(L_s - \delta, L_s + \delta)$ and $(U_s - \delta, U_s + \delta)$ for some $\delta > 0$. To see the issue, suppose Y has probability mass function (pmf) $p(0) = 0.4,\ p(1) = 0.3,\ p(2) = 0.2,\ p(3) = 0.06$, and $p(4) = 0.04$. Then the 90% population shorth is [0,2] and the $100(1-\delta)\%$ population shorth is [0,3] for $(1-\delta) \in (0.9, 0.96]$. Let $W_i = I(Y_i \leq x) = 1$ if $Y_i \leq x$ and 0, otherwise. The empirical cdf

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_{(i)} \leq x)$$

is the sample proportion of $Y_i \leq x$. If $Y_1, ..., Y_n$ are iid, then for fixed x, $n\hat{F}_n(x) \sim binomial(n, F(x))$. Thus $\hat{F}_n(x) \sim AN(F(x), F(x)(1-F(x))/n)$. For the Y with the above pmf, $\hat{F}_n(2) \xrightarrow{P} 0.9$ as $n \to \infty$ with $P(\hat{F}_n(2) < 0.9) \to 0.5$ and $P(\hat{F}_n(2) \geq 0.9) \to 0.5$ as $n \to \infty$. Hence the large sample 90% PI (2.11) will be [0,2] or [0,3] with probabilities $\to 0.5$ as $n \to \infty$ with expected asymptotic length of 2.5 and expected asymptotic coverage converging to 0.93. However, the large sample $100(1-\delta)\%$ PI (2.11) converges to [0,3] and is asymptotically optimal with asymptotic coverage 0.96 for $(1-\delta) \in (0.9, 0.96)$.

For a random variable Y, the $100(1-\delta)\%$ *highest density region* is a union of $k \geq 1$ disjoint intervals such that the mass within the intervals $\geq 1-\delta$ and the sum of the k interval lengths is as small as possible. Suppose that $f(z)$ is a unimodal pdf that has interval support, and that the pdf $f(z)$ of Y decreases rapidly as z moves away from the mode. Let $[a, b]$ be the shortest interval such that $F_Y(b) - F_Y(a) = 1 - \delta$ where the cdf $F_Y(z) = P(Y \leq z)$. Then the interval $[a, b]$ is the $100(1-\delta)$ highest density region. To find the $100(1-\delta)\%$ highest density region of a pdf, move a horizontal line down from the top of the pdf. The line will intersect the pdf or the boundaries of the support of the pdf at $[a_1, b_1], ..., [a_k, b_k]$ for some $k \geq 1$. Stop moving the line when the areas under the pdf corresponding to the intervals is equal to $1-\delta$. As an example, let $f(z) = e^{-z}$ for $z > 0$. See Figure 2.2 where the area under the pdf from 0 to 1 is 0.368. Hence [0,1] is the 36.8% highest

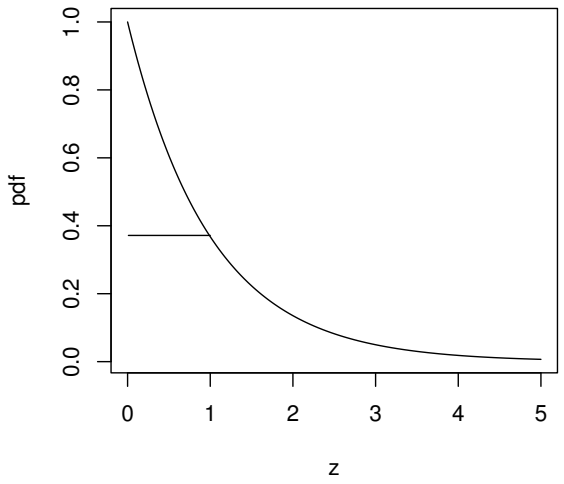

Fig. 2.2 The 36.8% Highest Density Region is [0,1]

density region. The shorth PI estimates the highest density interval which is the highest density region for a distribution with a unimodal pdf. Often the highest density region is an interval $[a, b]$ where $f(a) = f(b)$, especially if the support where $f(z) > 0$ is $(-\infty, \infty)$.

Applications 2.2. Variants of the shorth PI have many applications. The shorth PI tends to be asymptotically optimal for iid data. A shorth PI for multiple linear regression was given by Olive (2007); for the additive error regression model, including multiple linear regression, by Olive (2013a) and Pelawa Watagoda and Olive (2020); for many parametric regression models, including GLMs, GAMs and some survival regression models, by Olive et al. (2020); and for some time series models and renewal processes by Haile and Olive (2021). The following section shows that under regularity conditions, applying the shorth PI on a bootstrap sample results in a confidence interval. For Bayesian statistics, generate random variables from the the posterior distribution and apply the shorth PI to estimate the *highest density Bayesian credible interval*. See Olive (2014, p. 364) and Chen and Shao (1999).

Prediction intervals are closely related to percentiles or quantiles. The 95th percentile is the 0.95 quantile. The $100p$th percentile π_p satisfies $F(\pi_p) = P(X \leq \pi_p) = p$ if X is a continuous RV with increasing $F(x)$. Then to find π_p, let $\pi = \pi_p$ and solve $F(\pi) \stackrel{\text{set}}{=} p$ for π. In the literature, often the terms "quantiles" and "percentiles" are used interchangeably.

For a general RV X, π_p satisfies $F(\pi_p-) = P(X < \pi_p) \leq p \leq F(\pi_p) = P(X \leq \pi_p)$. So $F(\pi_p-) \leq p$ and $F(\pi_p) \geq p$. Then graphing $F(x)$ can be useful for finding π_p. The population median is the 50th percentile and 0.5 quantile. For iid data from a symmetric distribution, $\text{MED}(n) + \text{MAD}(n)$ estimates the 75th percentile while $\text{MED}(n) - \text{MAD}(n)$ estimates the 25th percentile.

Definition 2.13. The *sample* ρ *quantile* $\hat{\xi}_{n,\rho} = Y_{(\lceil n\rho \rceil)}$. The *population quantile* $y_\rho = \pi_\rho = \xi_\rho = Q(\rho) = \inf\{y \in \mathbb{R} : F(y) \geq \rho\}$ where Q is the *quantile function* and $0 < \rho < 1$.

For a random variable Y, we may use $Y_\delta, y_\delta, \pi_\delta$, or ξ_δ to denote the 100δth percentile with $P(Y \leq y_\delta) = F(y_\delta) = \delta$ if Y is from a continuous distribution with strictly increasing cdf. If the cdf has flat spots, e.g. if Y has a pmf, the following definition for a population quantile is often used. If F is continuous and strictly increasing, then $Q = F^{-1}$. The quantile function satisfies $Q(\rho) \leq y$ iff $F(y) \leq \rho$. For large sample theory and convergence in distribution, see Chapter 11. For the multivariate normal distribution, see Chapter 3.

Theorem 2.2: Serfling (1980, p. 80). Let $0 < \rho_1 < \rho_2 < \cdots < \rho_k < 1$. Suppose that F has a pdf f that is positive and continuous in neighborhoods of $\xi_{\rho_1}, ..., \xi_{\rho_k}$. Then

$$\sqrt{n}[(\hat{\xi}_{n,\rho_1}, ..., \hat{\xi}_{n,\rho_k})^T - (\xi_{\rho_1}, ..., \xi_{\rho_k})^T] \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{\Sigma})$$

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

$$\sigma_{ij} = \frac{\rho_i(1-\rho_j)}{f(\xi_{\rho_i})f(\xi_{\rho_j})}$$

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

Warning: Software often uses a slightly different definition of the sample quantile then the one given in Definition 2.13. Next we give an alternative estimator. See Klugman et al. (2008, p. 377). Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}$ be the order statistics of $X_1, ..., X_n$. Let the greatest integer function $\lfloor x \rfloor$ = the greatest integer $\leq x$, i.e. $\lfloor 7.7 \rfloor = 7$. The *smoothed empirical estimator of a percentile* π_p is $\hat{\pi}_p = X_{(j)}$ if $j = (n+1)p$ is an integer, and $\hat{\pi}_p = (1-h)X_{(j)} + hX_{(j+1)}$ if $(n+1)p$ is not an integer where $j = \lfloor (n+1)p \rfloor$ and $h = (n+1)p - j$. Here $\hat{\pi}_p$ is undefined if $j = 0$ or $j = n+1$, equivalently, $\hat{\pi}_p$ is undefined if $0 \leq p < 1/(n+1)$ or if $p = 1$.

Remark 2.3. If the data $z_1, ..., z_n$ are not iid, but the sample percentiles applied to the data give consistent estimators of the population percentiles, then typically the shorth interval applied to the data estimates the population shorth. As an example, assume that the sample percentiles of the residuals r_i converge to the population percentiles of the iid unimodal errors e_i: $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$. Also assume that the population shorth $[\xi_{\delta_1}, \xi_{1-\delta_2}]$ is unique and has length L. We want to show that the shorth of the residuals converges to the population

shorth of the e_i: $[\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}] \xrightarrow{P} [\xi_{\delta_1}, \xi_{1-\delta_2}]$. Let L_n be the length of $[\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Let $0 < \tau < 1$ and $0 < \epsilon < L$ be arbitrary. Assume n is large enough so that the correction factor is negligible. Then $P(L_n > L + \epsilon) \to 0$ since $[\hat{\xi}_{\delta_1}, \hat{\xi}_{1-\delta_2}]$ covers 100 $(1-\delta)\%$ of the data and $L_n = \tilde{\xi}_{1-\delta_2} - \tilde{\xi}_{\delta_1} \leq \hat{\xi}_{1-\delta_2} - \hat{\xi}_{\delta_1} \xrightarrow{P} L$ as $n \to \infty$ since the sample percentiles are consistent and the shorth is the shortest interval covering 100 $(1-\delta)\%$ of the data. If $P(L_n < L - \epsilon) > \tau$ eventually, then the shorth is an interval covering 100 $(1-\delta)\%$ of the cases that is shorter than the population shorth with positive probability τ. Hence at least one of $\hat{\xi}_{1-\delta_2}$ or $\hat{\xi}_{\delta_1}$ would not converge, a contradiction. Since ϵ and τ were arbitrary, $L_n \xrightarrow{P} L$. If $P(\tilde{\xi}_{\delta_1} < \xi_{\delta_1} - \epsilon) > \tau$ eventually, then $P(\tilde{\xi}_{1-\delta_2} < \xi_{1-\delta_2} - \epsilon/2) > \tau$ eventually since $L_n = \tilde{\xi}_{1-\delta_2} - \tilde{\xi}_{\delta_1} \xrightarrow{P} L = \xi_{1-\delta_2} - \xi_{\delta_1}$. But such an interval (of length going to L in probability with left endpoint less than $\xi_{\delta_1} - \epsilon$ and right endpoint less than $\xi_{1-\delta_2} - \epsilon/2$) contains more than $100(1-\delta)\%$ of the cases with probability going to one since the population shorth is the unique shortest interval covering $100(1-\delta)\%$ of the mass. Hence there is an interval covering $100(1-\delta)\%$ of the cases that is shorter than the shorth, with probability going to one, a contradiction. The case $P(\tilde{\xi}_{\delta_1} > \xi_{\delta_1} + \epsilon) > \tau$ can be handled similarly. Since ϵ and τ were arbitrary, $\tilde{\xi}_{\delta_1} \xrightarrow{P} \xi_{\delta_1}$. The proof that $\tilde{\xi}_{1-\delta_2} \xrightarrow{P} \xi_{1-\delta_2}$ is similar.

2.5 Bootstrap Confidence Intervals and Tests

Bootstrap tests and bootstrap confidence intervals are resampling algorithms used to provide information about the sampling distribution of a statistic $T_n \equiv T_n(\boldsymbol{Y}_n)$ where $\boldsymbol{Y}_n = (Y_1, ..., Y_n)^T$ and the Y_i are iid from a distribution with cdf $F(y) = P(Y \leq y)$. Then T_n has a cdf $H_n(y) = P(T_n \leq y)$. If $F(y)$ is known, then B independent samples $\boldsymbol{Y}^*_{j,n} = (Y^*_{j,1}, ..., Y^*_{j,n})^T$ of size n could be generated, where the $Y^*_{j,k}$ are iid from a distribution with cdf F and $j = 1, ..., B$. Then the statistic T_n is computed for each sample, resulting in B statistics $T^*_{1,n}(F), ..., T^*_{B,n}(F)$ which are iid from a distribution with cdf $H_n(y)$. The sample size n is often suppressed. This resampling scheme is a special case of the parametric bootstrap where the distribution is known. Usually the parametric bootstrap estimators the parameters of the parametric distribution that is known up to the unknown parameters. For example, if the Y_i are iid $N(\mu, \sigma^2)$, generate n iid $Y^*_i \sim N(\overline{Y}, S^2_n)$ to produce $\boldsymbol{Y}^*_{j,n}$ for $j = 1, ..., B$ where S^2_n is the sample variance of $Y_1, ..., Y_n$. We will discuss the nonparametric bootstrap below. Chapter 3 will discuss the bootstrap for statistics that are random vectors. Several bootstrap methods will be used throughout the text.

Definition 2.14. Suppose that data $y_1, ..., y_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the y_i are the

possible values, and each value is equally likely. If W is a random variable having the empirical distribution, then $p_i = P(W = y_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 2.8. Let W be a random variable having the empirical distribution given by Definition 2.14. Show that $E(W) = \overline{y} \equiv \overline{y}_n$ and $V(W) = \frac{n-1}{n} S_n^2$.

Solution: Recall that for a discrete random vector, the population expected value $E(W) = \sum y_i p_i$ where y_i are the values that W takes with positive probability p_i. Similarly, the population variance

$$V(W) = E[(W - E(W))^2] = \sum (y_i - E(W))^2 p_i.$$

Hence

$$E(W) = \sum_{i=1}^{n} y_i \frac{1}{n} = \overline{y},$$

and

$$V(W) = \sum_{i=1}^{n} (y_i - \overline{y})^2 \frac{1}{n} = \frac{n-1}{n} S_n^2. \quad \square$$

Example 2.9. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

The following notation is useful for the next definition. Suppose there is data $y_1, ..., y_n$ collected into an $n \times 1$ vector $\boldsymbol{y}$. Let the statistic $T_n = t(\boldsymbol{y}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\theta = T(F)$, and let $t(\boldsymbol{y}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $y_1^*, ..., y_n^*$ of size n was drawn with replacement from the observed sample $y_1, ..., y_n$. Let $T_j^* = t(\boldsymbol{y}_j^*)$ where $\boldsymbol{y}_j^* = (y_{1j}^*, ..., y_{nj}^*)^T$ corresponds to the jth sample. The B samples are drawn independently. Hence $\boldsymbol{y}_1^*, ..., \boldsymbol{y}_B^*$ are iid with respect to the bootstrap distribution.

Definition 2.15. The *empirical bootstrap* or **nonparametric bootstrap** or *naive bootstrap* draws B samples of size n with replacement from the observed sample $y_1, ..., y_n$. Then $T_j^* = T_{jn}^* = t(\boldsymbol{y}_j^*)$ is computed from the jth bootstrap sample for $j = 1, ..., B$. Then $T_1^*, ..., T_B^*$ is the *bootstrap sample* produced by the nonparametric bootstrap.

Example 2.10. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ samples (of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

Under regularity conditions, applying three prediction intervals to the bootstrap sample results in a confidence interval. Theory for bootstrap confidence regions will be given in Section 3.7, and a confidence interval is a special case of a confidence region. When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations $(2SD(\overline{Y}_n) = 2\sigma/\sqrt{n})$ of μ is about 95%. Hence the probability that μ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\mu - 1.96S/\sqrt{n}, \mu + 1.96S/\sqrt{n}\ ]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if μ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}\ ]$ is a large sample 95% confidence interval for the population mean μ. Note that the lengths of the two intervals are the same. Where the interval is centered determines whether the interval is a confidence or a prediction interval.

For a confidence interval, we often want the following probability to converge to $1 - \delta$ if the confidence interval is based on a statistic with an asymptotic distribution that has a probability density function. For a large sample level δ test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, reject H_0 if θ_0 is not in the large sample $100(1 - \delta)\%$ confidence interval (CI) for θ.

Definition 2.16. The interval $[\hat{L}_n, \hat{U}_n]$ is a large sample $100(1 - \delta)\%$ *confidence interval* for θ if $P(\hat{L}_n \leq \theta \leq \hat{U}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

Next we discuss bootstrap confidence intervals (2.12) and (2.13) that are obtained by applying prediction intervals (2.8) and (2.11) to the bootstrap sample with B used instead of n. See Efron (1982) and Chen (2016) for the percentile method CI. Let T_n be an estimator of a parameter θ such as $T_n = \overline{Z} = \sum_{i=1}^{n} Z_i/n$ with $\theta = E(Z_1)$. Let $T_1^*, ..., T_B^*$ be a bootstrap sample for T_n. Let $T_{(1)}^*, ..., T_{(B)}^*$ be the order statistics of the the bootstrap sample.

Definition 2.17. The bootstrap large sample $100(1-\delta)\%$ percentile confidence interval for θ is an interval $[T_{(k_L)}^*, T_{(K_U)}^*]$ containing $\approx \lceil B(1-\delta) \rceil$ of the T_i^*. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. A common choice is

$$[T_{(k_1)}^*, T_{(k_2)}^*]. \tag{2.12}$$

Definition 2.18. The large sample $100(1-\delta)\%$ *shorth*(c) *CI*

$$[T_{(s)}^*, T_{(s+c-1)}^*] \tag{2.13}$$

uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length. Here

$$c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{2.14}$$

The shorth CI can be regarded as the shortest percentile method confidence interval, asymptotically. Hence the shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples. Olive (2014: p. 238, 2017b: p. 168, 2018) recommended using the shorth CI for the percentile CI.

The following correction factor is useful for the next three bootstrap CIs. Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + 1/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta/B), \quad \text{otherwise.} \tag{2.15}$$

If $1-\delta < 0.999$ and $q_B < 1-\delta+0.001$, set $q_B = 1-\delta$. Let $a_{(U_B)}$ be the $100q_B$th sample quantile of the $a_i = |T_i^* - \overline{T}^*|$. Let $b_{(U_B,T)}$ be the $100q_B$th sample quantile of the $b_i = |T_i^* - T_n|$. Equation (2.15) is often useful for getting good coverage when $B \geq 200$. Undercoverage could occur without the correction factor. This result is useful because the bootstrap confidence intervals can be slow to simulate. Hence we want to use small values of $B \geq 200$.

The percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T_i^*. Let $a_i = |T_i^* - \overline{T}^*|$. The following three CIs are the special cases of the prediction region method confidence region, modified Bickel and Ren confidence region, and hybrid confidence region for a $g \times 1$ parameter vector $\boldsymbol{\theta}$ when $g = 1$. See Section 3.4. The sample mean of the bootstrap sample $\overline{T}^* = \dfrac{1}{B}\sum_{i=1}^{B} T_i^*$ is the *bagging estimator*.

Definition 2.19. a) The large sample $100(1-\delta)\%$ *prediction region method CI* is

$$[\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}], \tag{2.16}$$

which is a closed interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*.

b) The large sample $100(1-\delta)\%$ *modified Bickel and Ren CI* is

$$[T_n - b_{(U_B,T)}, T_n + b_{(U_B,T)}], \tag{2.17}$$

which is a closed interval centered at T_n just long enough to cover "U_B, T" of the T_i^*.

c) The large sample $100(1-\delta)\%$ *hybrid CI* is

$$[T_n - a_{(U_B)}, T_n + a_{(U_B)}]. \tag{2.18}$$

This CI is the prediction region method CI shifted to have center T_n instead of $\overline{T}^*$.

Both CIs (2.16) and (2.17) are special cases of the percentile method of Definition 2.17. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal.

Remark 2.4. The shorth(c) CI (2.13) is often very short, but sometimes needs larger sample sizes for good coverage than the percentile CI (2.12), the prediction region method CI (2.16) or the modified Bickel and Ren CI (2.17). The hybrid CI has the same length as the prediction region method CI and is usually shorter than the modified Bickel and Ren CI since the T_i^* tend to be closer, on average, to $\overline{T}^*$ than to T_n. The hybrid CI was more prone to undercoverage than CIs (2.16) and (2.17).

Application 2.3. We recommend using using the percentile CI (2.12), the prediction region method CI (2.16), the modified Bickel and Ren CI (2.17), and possibly the shorth CI (2.13) for robust statistics with good large sample theory and good bootstrap theory, but with a standard error that is difficult to estimate. The sample median is such a statistic. In the next section, CI (2.19) for the population median is useful for hand calculations, but likely needs a larger sample size n than CIs (2.12), (2.16), and (2.17) for good coverage.

Remark 2.5, Pelawa Watagoda and Olive (2019). If $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$, and if $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$ where U has a unimodal probability density function symmetric about zero with $E(U) = 0$, then the confidence intervals from the (2.16), (2.17), (2.18), the shorth confidence interval (2.13), and the "usual" percentile method confidence interval (2.12) are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically). See Section 3.5.

2.6 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" (e.g., $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (e.g., $\lceil 7.7 \rceil = 8$).

Application 2.4: 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

$$\text{MED}(n) \pm t_{p,1-\alpha/2} SE(\text{MED}(n)). \tag{2.19}$$

Warning. This CI is easy to compute by hand, but tends to be long with undercoverage if $n < 100$. See Baszczyńska and Pekasiewicz (2010) for two competitors that work better. We recommend bootstrap confidence intervals in Application 2.3 from the last Section for the population median.

Definition 2.20. The symmetrically trimmed mean or the α *trimmed mean*

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{2.20}$$

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

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

The trimmed mean is estimating a truncated mean μ_T. See Section 11.5 for truncated distributions. Assume that Y has a probability density function $f_Y(y)$ that is continuous and positive on its support. Let y_α be the quantile satisfying $P(Y \leq y_\alpha) = \alpha$. Then

$$\mu_T = \frac{1}{1-2\alpha} \int_{y_\alpha}^{y_{1-\alpha}} y f_Y(y) dy. \tag{2.21}$$

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

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

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

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \tag{2.22}$$

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

Application 2.5: inference with the α trimmed mean. A large sample 100 $(1 - \delta)$% confidence interval (CI) for μ_T is

$$T_n \pm t_{p,1-\frac{\delta}{2}} SE(T_n) \tag{2.23}$$

where $P(t_p \leq t_{p,1-\frac{\delta}{2}}) = 1 - \delta/2$ 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 $\alpha = 0$, but $\alpha = 0.25$ gives a robust CI.

Example 2.11. Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean μ, and 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 $t_{p,1-\delta/2}$, 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 $SE($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 $SE(\overline{Y}) < SE($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 2.12. 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 2.13. 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 $[1672.096, 1707.282]$. See Problems 2.28, 2.29 and 2.30 for *rpack* software.

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, 2.7, 2.8, 2.9, and 2.15 provide additional information on CIs and tests.

2.7 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

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

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

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

See Section 11.6 for more information on large sample theory and convergence in distribution. Suppose that S_W^2 is a consistent estimator of σ_W^2 so that the (asymptotic) *standard error* of W_n is $\mathrm{SE}(W_n) = S_W/\sqrt{n}$. Let z_δ be the δ quantile of the N(0,1) distribution. Hence $P(Z \leq z_\delta) = \delta$ if $Z \sim N(0,1)$. Then

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

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

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

Three common approximate level δ tests of hypotheses all use the *null hypothesis* $H_0 : \mu_W = \mu_0$. A right tailed test uses the *alternative hypothesis* $H_A : \mu_W > \mu_0$, a left tailed test uses $H_A : \mu_W < \mu_0$, and a two tail test uses $H_A : \mu_W \neq \mu_o$. The test statistic is

$$t_0 = \frac{W_n - \mu_0}{SE(W_n)},$$

and the (approximate) *p-values* are $P(Z > t_0)$ for a right tail test, $P(Z < t_0)$ for a left tail test, and $2P(Z > |t_0|) = 2P(Z < -|t_0|)$ for a two tail test. The null hypothesis H_0 is rejected if the p-value $< \delta$.

Remark 2.6. 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 $100(1-\delta)\%$ CI for μ_W is given by

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

The test statistic rarely has an exact t_p distribution, but the approximation tends to make the CIs and tests more *conservative;* i.e., the CIs are longer and H_0 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

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

Then

$$\begin{pmatrix} (W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ (W_m(\boldsymbol{X}) - \mu_W(X)) \end{pmatrix} \approx N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_W^2(Y)/n & 0 \\ 0 & \sigma_W^2(X)/m \end{pmatrix} \right),$$

and

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

Hence $SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) =$

$$\sqrt{\frac{S_W^2(\boldsymbol{Y})}{n} + \frac{S_W^2(\boldsymbol{X})}{m}} = \sqrt{[SE(W_n(\boldsymbol{Y}))]^2 + [SE(W_m(\boldsymbol{X}))]^2},$$

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

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

Often approximate level δ tests of hypotheses use the *null hypothesis* $H_0 : \mu_W(Y) = \mu_W(X)$. 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

$$t_0 = \frac{W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})}{SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}))},$$

and the (approximate) *p-values* are $P(Z > t_0)$ for a right tail test, $P(Z < t_0)$ for a left tail test, and $2P(Z > |t_0|) = 2P(Z < -|t_0|)$ for a two tail test. The null hypothesis H_0 is rejected if the p-value $< \delta$.

Remark 2.7. 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 if a better formula is not given. These CIs are known as *Welch intervals*. See Welch (1937) and Yuen (1974).

Example 2.14. Consider the single sample procedures where $W_n = \overline{Y}_n$. Then $\mu_W = E(Y)$, $\sigma_W^2 = \text{VAR}(Y)$, $S_W = S_n$, 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 $t_{p,\delta}$ satisfy $P(t_p \leq t_{p,\delta}) = \delta$. Then the classical *t-interval* for $\mu \equiv E(Y)$ is

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

and the *t-test statistic* is

$$t_0 = \frac{\overline{Y} - \mu_0}{S_n/\sqrt{n}}.$$

The right tailed p-value is given by $P(t_{n-1} > t_0)$.

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 $\mu_Y - \mu_X$ is

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

and *two sample t-test statistic* is

$$t_0 = \frac{\overline{Y}_n - \overline{X}_m}{\sqrt{\frac{S^2_n(\boldsymbol{Y})}{n} + \frac{S^2_m(\boldsymbol{X})}{m}}}.$$

The right tailed p-value is given by $P(t_p > t_0)$. 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).

2.8 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 2.20 and for the Winsorized mean, defined below, the proportion of cases trimmed and the proportion of cases covered are fixed.

Definition 2.21. Using the same notation as in Definition 2.20, the *Winsorized mean*

$$W_n = W_n(L_n, U_n) = \frac{1}{n}[L_n Y_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y_{(i)} + (n - U_n)Y_{(U_n)}]. \quad (2.24)$$

Definition 2.22. A *randomly trimmed mean*

$$R_n = R_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \quad (2.25)$$

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 2.23. 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 MED(n) $- k_1$ MAD(n) and let $n - U(M_n)$ be the number of observations that fall to the right of MED(n) $+ k_2$ 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 (e.g. $\lfloor 7.7 \rfloor = 7$) is used in the following definitions.

Definition 2.24. 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 $L_n = \lfloor n\ \ \alpha_{M,n} \rfloor$ and $U_n = n - L_n$. If $\alpha_{M,n} = 0.50$, then use $T_{S,n} = \text{MED}(n)$.

Definition 2.25. 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 $L_n = \lfloor n \ \ \alpha_{o,n} \rfloor$ and $U_n = \lfloor n \ \ \beta_{o,n} \rfloor$. If $\alpha_{o,n} = 1-\beta_{o,n} = 0.5$, then use $T_{A,n} = \text{MED}(n)$.

Example 2.15. 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 2.26. The standard error SE_{RM} for the two stage trimmed means given in Definitions 2.20, 2.24 and 2.25 is

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

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

$$\frac{[L_n Y_{(L_n+1)}^2 + \sum_{i=L_n+1}^{U_n} Y_{(i)}^2 + (n-U_n)Y_{(U_n)}^2] - \ n\ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}. \quad (2.26)$$

Remark 2.8. A simple method for computing $V_{SW}(L_n, U_n)$ has the following steps. First, find $d_1, ..., d_n$ where

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

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

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \quad (2.27)$$

Notice that the SE given in Definition 2.26 is the SE for the δ trimmed mean where L_n and U_n are fixed constants rather than random.

Application 2.6. 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 2.7, use the one sample standard error $\mathrm{SE}_{RM}(L_n, U_n)$ given in Definition 2.26 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 2.5 and 2.6 are very similar once L_n has been computed. For example, a large sample 100 $(1 - \alpha)\%$ confidence interval (CI) for μ_T is

$$(T_n - t_{U_n - L_n - 1, 1-\frac{\alpha}{2}} SE_{RM}(L_n, U_n), T_n + t_{U_n - L_n - 1, 1-\frac{\alpha}{2}} SE_{RM}(L_n, U_n)) \tag{2.28}$$

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. Section 2.9 provides the asymptotic theory for the δ and two stage trimmed means and shows that μ_T is the mean of a truncated distribution. Section 11.4 gives suggestions for k_1 and k_2 while Section 2.15 provides a simulation study comparing the robust and classical point estimators and intervals. Next Examples 2.11, 2.12, and 2.13 are repeated using the intervals based on the two stage trimmed means instead of the median.

Example 2.16. Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean μ, and find a 95% CI for μ.

Solution. If $T_{A,n}$ or $T_{S,n}$ is used with the metrically trimmed mean that uses $k = k_1 = k_2$, e.g. $k = 6$, 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 $\mathrm{MED}(n) = (8 + 9)/2 = 8.5$. The ordered residuals $Y_{(i)} - \mathrm{MED}(n)$ 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 $\mathrm{MAD}(n) = 0.5$, $\mathrm{MED}(n) - 6MAD(n) = 5.5$, and $\mathrm{MED}(n) + 6MAD(n) = 11.5$. Hence no cases are trimmed by the metrically trimmed mean, i.e. $L(M_n) = 0$ 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 2.17. 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 $\mathrm{MED}(n) = 9$ and $\mathrm{MAD}(n) = 1.5$. 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 $0.25 = 2/8$, respec-

tively. 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 2.8 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 2.18. 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.

2.9 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 covered in Section 11.6.

Truncated and Winsorized random variables are important because they simplify the asymptotic theory of robust estimators. See Section 11.5. 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 Theorem 11.1 for the notation used below.)

Definition 2.27. The *truncated random variable* $Y_T \equiv Y_T(a, b)$ with *truncation points* a and b has cdf

$$F_{Y_T}(y|a, b) = G(y) = \frac{F(y) - F(a-)}{F(b) - F(a-)} \tag{2.29}$$

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

$$\mu_T = \mu_T(a,b) = \int_{-\infty}^{\infty} y dG(y) = \frac{\int_a^b y dF(y)}{\beta - \alpha} \tag{2.30}$$

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 2.28. The *Winsorized random variable*

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

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

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

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

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

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 $b = Q(\beta)$. (See Definition 2.13 for the quantile function Q.)

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. Theorem 2.4 shows that the standard error SE_{RM} given in the previous section is estimating the appropriate asymptotic standard deviation of T_n.

Theorem 2.3. If conditions (R1) and (R2) hold and if $0 < \alpha < \beta < 1$, then

$$\sqrt{n}(T_n - \mu_T(a,b)) \xrightarrow{D} N[0, \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2}]. \quad (2.31)$$

Theorem 2.4: Shorack and Wellner (1986, p. 680). Assume that regularity conditions (R1) and (R2) hold and that

$$\frac{L_n}{n} \xrightarrow{P} \alpha \text{ and } \frac{U_n}{n} \xrightarrow{P} \beta. \quad (2.32)$$

Then

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

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: $\mathrm{SE}_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$ where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

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

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 $a = \mathrm{MED}(Y) - k\mathrm{MAD}(Y)$ and $b = \mathrm{MED}(Y) + k\mathrm{MAD}(Y)$ where $\mathrm{MED}(Y)$ and $\mathrm{MAD}(Y)$ 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.

Theorem 2.5: Shorack and Wellner (1986, p. 682-683). Let F have a strictly positive and continuous derivative in some neighborhood of $\mathrm{MED}(Y) \pm k\mathrm{MAD}(Y)$. Assume that

$$\sqrt{n}(MED(n) - MED(Y)) = O_P(1) \quad (2.33)$$

and

$$\sqrt{n}(MAD(n) - MAD(X)) = O_P(1). \tag{2.34}$$

Then

$$\sqrt{n}(\frac{L(M_n)}{n} - \alpha) = O_P(1) \tag{2.35}$$

and

$$\sqrt{n}(\frac{U(M_n)}{n} - \beta) = O_P(1). \tag{2.36}$$

Theorem 2.6. 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 2.3 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 Theorem 2.5. The proof for $T_{S,n}$ is similar. □

2.10 L, R, and M Estimators

Definition 2.29. 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. Other examples include the max $= Y_{(n)}$, the min $= Y_{(1)}$, the range $= Y_{(n)} - Y_{(1)}$, and the midrange $= (Y_{(n)} + Y_{(1)})/2$. Definition 2.13 and Theorem 2.2 are useful for L-estimators such as the interquartile range and median that use a fixed linear combination of sample quantiles.

R-estimators are derived from rank tests and include the sample mean and median. See Hettmansperger and McKean (2010).

Definition 2.30. An *M-estimator* of location T with preliminary estimator of scale $\mathrm{MAD}(n)$ is computed with at least one Newton step

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

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

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

The key to M-estimation is finding a good ψ. The sample mean and sample median are M-estimators. *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 $S = \mathrm{MAD}(n)$ 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 2.19. 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 $T - kS \le Y \le T + kS$ 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 occurring 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

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

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

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

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

2.11 Asymptotic Theory for the MAD

Let $\text{MD}(n) = \text{MED}(|Y_i - \text{MED}(Y)|,\ i = 1, \ldots, n)$. Since $\text{MD}(n)$ is a median and convergence results for the median are well known, see for example Serfling (1980, p. 74-77) or Theorem 2.2 from Section 2.4, it is simple to prove convergence results for $\text{MAD}(n)$. Typically $\text{MED}(n) = \text{MED}(Y) + O_P(n^{-1/2})$ and $\text{MAD}(n) = \text{MAD}(Y) + O_P(n^{-1/2})$. 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.

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

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

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

and

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

Similarly

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

and thus

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

Combining the two inequalities shows that

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

or

$$|\text{MAD}(n) - \text{MD}(n)| \le |\text{MED}(n) - \text{MED}(Y)|. \tag{2.37}$$

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

$$|\text{MAD}(n) - \text{MAD}(Y) - O_P(n^{-\delta})| = O_P(n^{-\delta}) \tag{2.38}$$

and the result follows. □

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 $\text{MED}(Y) = F^{-1}(1/2)$ and $\text{MED}(Y) \pm \text{MAD}(Y)$. 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

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

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_M^2 & \sigma_{M,D}\\ \sigma_{M,D} & \sigma_D^2\end{pmatrix}\right) \tag{2.39}$$

where

$$\sigma_M^2 = \frac{1}{4f^2(F^{-1}(\frac{1}{2}))},\ \sigma_D^2 = \frac{1}{4A^2}(1 + \frac{B}{f^2(F^{-1}(\frac{1}{2}))}),$$

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 $F(y) = P(Y \leq y)$ 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 $F(a_1) = F(a_2)$, $F(b_1) = F(b_2)$,

$$P(a_i \leq Y \leq b_i) = F(b_i) - F(a_i-) \geq 0.5,$$

and

$$P(Y \leq a_i) + P(Y \geq b_i) = F(a_i) + 1 - F(b_i-) \geq 0.5$$

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

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

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

$$\max\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\}. \tag{2.40}$$

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 $\text{MAD}(Y) > m$, then either $\text{MED}(Y) - \text{MAD}(Y) < a$ or $\text{MED}(Y) + \text{MAD}(Y) > b$. Since at least half of the mass is between a and $\text{MED}(Y)$ and between $\text{MED}(Y)$ and b, this contradicts the definition of $\text{MAD}(Y)$. 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 $\text{MED}(Y) = c$ and $\text{MAD}(Y) = 0$. If each of the points a, b, and c has $1/3$ of the mass where $a < c < b$, then $\text{MED}(Y) = c$ and $\text{MAD}(Y) = m$. □

Many other results for $\text{MAD}(Y)$ and $\text{MAD}(n)$ are possible. For example, note that Theorem 2.9 b) implies that when Y is symmetric, $\text{MAD}(Y) = F^{-1}(3/4) - \mu$ and $F(\mu + \text{MAD}(Y)) = 3/4$. Also note that $\text{MAD}(Y)$ and the interquartile range $\text{IQR}(Y)$ 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 Theorem 2.9 hold for $\text{MAD}(n)$ 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 $\text{MED}(n)$. To illustrate the outlier resistance of $\text{MAD}(n)$ and $\text{MED}(n)$, consider the following lemma.

Theorem 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

$$\text{MED}(n+m) \in [Y_{(1)}, Y_{(n)}] \text{ and } 0 \leq \text{MAD}(n+m) \leq Y_{(n)} - Y_{(1)}. \quad (2.41)$$

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 $\text{MED}(n+m)$ 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 $\text{MAD}(n+m) \leq Y_{(n)} - Y_{(1)}$. □

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, $\text{MED}(n+n-1)$ and $\text{MAD}(n+n-1)$, are bounded by quantities from the random sample from F.

2.12 Some Other Estimators

2.12.1 The Median of Estimators Estimator

The machine learning literature has estimators like the following. Let $n = Km+J$ with $0 \leq J < K$. Let $X_1, ..., X_n$ be iid data and let statistic T, such as the sample mean, be a function of the data that is a consistent estimator of θ. Randomly divide the data into K blocks of equal size n (omit the remaining J cases if $J \neq 0$. Let T_i be the statistic computed from the m cases in block i. Then $T_1, ..., T_K$ are iid. The *median of estimators* $MED(K)$ is the sample median of the T_i.

The above procedure gives a point estimator of θ with some outlier resistance, but is is hard to get confidence intervals for general T since the population median $\theta_{K,n}$ of the T_i depends on K and n. Typically $\sqrt{n}(\theta_{K,n} - \theta) = O_P(1)$ but not $o_p(1)$. Hence we can not use the confidence interval (2.19) for θ. There is a clever way to get a confidence interval for the median of means where T is the sample mean. See Laforgue et al. (2019) for references. Roughly half of the $K/2$ blocks need bad contamination for the median of estimators estimator to be arbitrarily bad.

2.12.2 LMS, LTA, LTS

The location model is a special case of the multiple linear regression model and of the multivariate location and dispersion model where $p = 1$. Truncated distributions are useful for explaining what is being estimated in the location model. See Section 11.5. The LMS, LTS, and LTA regression estimators can be computed for the location model.

Definition 2.31. Consider intervals that contain c_n cases: $[Y_{(1)}, Y_{(c_n)}]$, $[Y_{(2)}, Y_{(c_n+1)}], ..., [Y_{(n-c_n+1)}, Y_{(n)}]$. Denote the set of c_n cases in the ith interval by J_i, for $i = 1, 2, ..., n - c_n + 1$. Often $c_n = \lfloor n/2 \rfloor + 1$.

i) Let the shorth(c_n) estimator $= [Y_{(s)}, Y_{(s+c_n-1)}]$ be the shortest such interval. Then the *least median of squares estimator* LMS(c_n) is $(Y_{(s)} + Y_{(s+c_n-1)})/2$, the midpoint of the shorth(c_n) interval. The LMS estimator is also called the *least quantile of squares estimator* LQS(c_n).

ii) Compute the sample mean and sample variance $(\overline{Y}_{J_i}, S^2_{J_i})$ of the c_n cases in the ith interval. The *minimum covariance determinant* estimator MCD(c_n) estimator $(\overline{Y}_{MCD}, S^2_{MCD})$ is equal to the $(\overline{Y}_{J_j}, S^2_{J_j})$ with the smallest $S^2_{J_i}$. The *least trimmed sum of squares estimator* is LTS$(c_n) = \overline{Y}_{MCD}$.

iii) Compute the sample median M_{J_i} of the c_n cases in the ith interval. Let $Q_{LTA}(M_{J_i}) = \sum_{j \in J_i} |y_j - M_{J_i}|$. The *least trimmed sum of absolute deviations estimator* LTA(c_n) is equal to the M_{J_j} with the smallest $Q_{LTA}(M_{J_i})$.

Definition 2.32. In a location model *concentration algorithm*, let the jth *start* be $(T_{-1,j}, C_{-1,j})$, an estimator of location and dispersion. Then the classical estimator $(T_{0,j}, C_{0,j}) = (\overline{Y}_{0,j}, S^2_{0,j})$ is computed from the c_n cases closest to $T_{-1,j}$. This iteration can be continued for k steps resulting in the sequence of estimators $(T_{-1,j}, C_{-1,j}), (\overline{Y}_{0,j}, S^2_{0,j}), ..., (\overline{Y}_{k,j}, S^2_{k,j})$. The result of the iteration $(\overline{Y}_{k,j}, S^2_{k,j})$ is called the jth *attractor*. If K_n starts are used, then $j = 1, ..., K_n$. The *concentration attractor*, $(\overline{Y}_A, S^2_A)$, is the attractor chosen by the algorithm. The attractor is used to obtain the final estimator. The FLTS and FMCD algorithms choose the attractor with the smallest $S^2_{k,j}$.

In a location concentration algorithm that uses k steps for each start, the dispersion estimators do not need to be computed since the c_n cases closest to the location estimator $T_{-1,j}$ or $\overline{Y}_{i,j}$ are used in the concentration step for $i = 0, 1, ..., k-1$. Attractors in a concentration algorithm can also be obtained by iterating to convergence. In this case the number of concentration steps k is not fixed and is unknown, but convergence is typically very fast for the location model. As notation, $(\overline{Y}_{\infty,j}, S^2_{\infty,j})$ is the jth attractor that results when the algorithm is iterated to convergence.

Theorem 2.11 Rousseeuw and van Driessen (1999): $S^2_{i+1,j} \leq S^2_{i,j}$, and the attractor converges when equality is obtained.

Definition 2.33. i) For the elemental FLTS concentration algorithm, $C_{-1,j} = 1$ while $T_{-1,j} = Y^*_j$ where Y^*_j is a randomly selected case. $K_n = 500$ starts are used.

ii) For the elemental FMCD concentration algorithm, randomly select two cases. Then $(T_{-1,j}, C_{-1,j})$ is the sample mean and variance of these two cases. $K_n = 500$ starts are used.

iii) The MB estimator uses $(T_{-1,1}, C_{-1,1}) = (\text{MED}(n), 1)$ as the only start. Hence the start uses the sample median as the location estimator.

iv) The DGK estimator uses the sample mean and variance of all n cases, $(T_{-1,1}, C_{-1,1}) = (\overline{Y}, S^2)$, as the only start.

Concentration algorithm estimators can have problems if the distribution is not unimodal. For example, the population shorth is not unique for the uniform distribution. Outliers can easily make the distribution multimodal.

Remark 2.9. Let $[Y_{(d)}, Y_{(d+c_n-1)}]$ be the LTS interval and $[Y_{(a)}, Y_{(a+c_n-1)}]$ be the LTA interval. The population quantities are $[a_{LTS}, b_{LTS}]$ and $[a_{LTA}, b_{LTA}]$. Take $c = c_n$ given by Equation (2.12). Then the two above intervals should be useful large sample $100(1-\delta)\%$ PIs, and the population quantities will equal the population shorth for many distributions. Among intervals that contain c_n observations, the coverage should be the worst for the shortest and longest intervals for clean data (with no outliers). The shortest interval behaves well by Frey (2013). The longest interval is not outlier resistant. It is

possible that the LTS and LTA PIs converge at $\sqrt{n}$ rate instead of the slower rate for the shorth interval given by Remark 2.1.

Definition 2.34. Let $\boldsymbol{W} = (Y_1, ..., Y_n)^T$ be the *clean data*, and $\boldsymbol{W}_d^n = (W_1, ..., W_n)^T$ be the contaminated data after d_n of the Y_i have been replaced by arbitrarily bad cases.. The *breakdown value* of a location estimator T_n is

$$B(T, \boldsymbol{W}) = \min\{\frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} |T(\boldsymbol{W}_d^n)| = \infty\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d_n \leq n$. The *breakdown value* of a dispersion estimator C_n is

$$B(C_n, \boldsymbol{W}) = \min\{\frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \max(|C_n(\boldsymbol{W}_d^n)|, |1/C_n(\boldsymbol{W}_d^n)|) = \infty\}.$$

Since the sup is used, there exists a real numbers M_1 and $0 < m < M_2$ that depend on the estimator and the clean data $Y_1, ..., Y_n$ but not on the outliers such that $0 \leq |T_n| < M_1$ and $0 < m < |C_n| < M_2$ if the number of outliers d_n is less than the breakdown value. For MED(n), $M_1 = \max(|Y_{(1)}|, |Y_{(n)}|)$.

Suppose $c_n \approx n/2$. For the MCD(c_n) and MB estimators, the breakdown value $d_n/n \to 0.5$ for both the location and dispersion estimators if the Y_i are distinct. Such estimators are called high breakdown estimators. See Chapter 3. LTS(c_n) is also a high breakdown estimator. The sample mean and variance both have breakdown value $1/n$. The sample mean and variance applied to a randomly selected elemental set of two randomly selected cases also has breakdown value $1/n$. A concentration algorithm that has K_n randomly selected elemental sets can be made to breakdown by changing 1 case in each elemental set. Hence the elemental concentration algorithm has breakdown value $\leq K_n/n \to 0$ as $n \to \infty$. Hence the FLTS and FMCD estimators can not produce the high breakdown LTS and MCD estimators.

Consider the attractor of a concentration algorithm. If 26% of the cases are large positive outliers, and the start $T_{-1,j}$ is closer in distance to the outliers than to the bulk of the data, then the sample mean of the $c_n \approx n/2$ cases closest to $T_{-1,j}$ is closer to the outliers than to the bulk of the data. Hence the location estimator of the attractor, $T_{k,j}$ or $T_{\infty,j}$, is the sample mean of the c_n largest order statistics. Hence the attractor is not the MCD(c_n) estimator.

Next we give a theorem for the metrically trimmed mean M_n. Lopuhaä (1999) shows the following result. Suppose $(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n)$ is an estimator of multivariate location and dispersion. Suppose that the iid data follow an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. Let $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ be the classical estimator applied to the set J of cases with squared Mahalanobis distances $D_i^2(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n) \leq k^2$. Under regularity conditions, if $(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n) \xrightarrow{P} (\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $0 < \delta \leq 0.5$, then $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J) \xrightarrow{P} (\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ with the same rate n^δ

where $s > 0$ and $d > 0$ are some constants. *See Chapter 3 for discussion of the above quantities.*

In the univariate setting with $p = 1$, let $\hat{\theta}_n = \hat{\mu}_n$ and let $D_n^2 = \boldsymbol{C}_n$ where D_n is an estimator of scale. Suppose the classical estimator $(\overline{Y}_J, S_J^2) \equiv (\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ is applied to the set J of cases with $\hat{\theta}_n - kD_n \leq Y_i \leq \hat{\theta}_n + kD_n$. Hence $\overline{Y}_J$ is the metrically trimmed mean M_n with $k_1 = k_2 \equiv k$. See Definition 2.23.

The population quantity estimated by $(\overline{Y}_J, S_J^2)$ is the truncated mean and variance $(\mu_T(a,b), \sigma_T^2(a,b))$ of Definition 2.27 where $\hat{\theta}_n - kD_n \xrightarrow{P} a$ and $\hat{\theta}_n + kD_n \xrightarrow{P} b$. In the theorem below, the pdf corresponds to an elliptically contoured distribution with $p = 1$ and $\boldsymbol{\Sigma} = \tau^2$. Each pdf corresponds to a location scale family with location parameter μ and scale parameter τ. Note that $(\hat{\theta}_n, D_n) = (\text{MED}(n), \text{MAD}(n))$ results in a $\sqrt{n}$ consistent estimator (M_n, S_J^2).

Assumption E1: Suppose $Y_1, ..., Y_n$ are iid from an $EC_1(\mu, \tau^2, g)$ distribution with pdf

$$f(y) = \frac{c}{\tau} \; g\left[\left(\frac{y-\mu}{\tau}\right)^2\right]$$

where g is continuously differentiable with finite 4th moment $\int y^4 g(y^2)dy < \infty$, $c > 0$ is some constant, $\tau > 0$ where y and μ are real.

Theorem 2.12. Let M_n be the metrically trimmed mean with $k_1 = k_2 \equiv k$. Assume (E1) holds. If $(\hat{\theta}_n, D_n^2) \xrightarrow{P} (\mu, s\tau^2)$ with rate n^δ for some constant $s > 0$ where $0 < \delta \leq 0.5$, then $(M_n, S_J^2) \xrightarrow{P} (\mu, \sigma_T^2(a,b))$ with the same rate n^δ.

Proof. The result is a special case of Lopuhaä (1999) which shows that $(M_n, S_J^2) \xrightarrow{P} (\mu, d\tau^2)$ with rate n^δ. Since $k_1 = k_2 = k$, $d\tau^2 = \sigma_T^2(a,b)$. □

Note that the classical estimator applied to the set $\tilde{J}$ of cases Y_i between a and b is a $\sqrt{n}$ consistent estimator of $(\mu_T(a,b), \sigma_T^2(a,b))$. Consider the set J of cases with $\text{MED}(n) - k\text{MAD}(n) \leq Y_i \leq \text{MED}(N) + k\text{MAD}(n)$. By Lemma 2.4 sets $\tilde{J}$ and J differ primarily in neighborhoods of a and b. This result leads to the following conjecture.

Conjecture 2.1. If $Y_1, ..., Y_n$ are iid from a distribution with a pdf that is positive in neighborhoods of a and b, and if $\hat{\theta}_n - k_1 D_n \xrightarrow{P} a$ and $\hat{\theta}_n + k_2 D_n \xrightarrow{P} b$ at rate $n^{0.5}$, then $(M_n, S_J^2) \xrightarrow{P} (\mu_T(a,b), \sigma_T^2(a,b))$ with rate $n^{0.5}$.

The following result follows from Theorem 3.14b applied to the location model.

Theorem 2.13. Let $(\overline{Y}_A, S_A^2)$ be the DGK or MB estimator that uses k concentration steps with $c_n \approx n/2$. Assume (E1) holds and let $[a, b]$ be

the highest density region containing half of the mass. Then $(\overline{Y}_A, S_A^2) \xrightarrow{P} (\mu, \sigma_T^2(a,b))$ with rate n^δ.

2.13 Asymptotic Variances for Trimmed Means

The truncated distributions will be useful for finding the asymptotic variances of trimmed and two stage trimmed means. Assume that Y is from a symmetric location–scale family with parameters μ and σ and that the truncation points are $a = \mu - z\sigma$ and $b = \mu + z\sigma$. Recall that for the trimmed mean T_n,

$$\sqrt{n}(T_n - \mu_T(a,b)) \xrightarrow{D} N\left(0, \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2}\right).$$

Since the family is symmetric and the truncation is symmetric, $\alpha = F(a) = 1 - \beta$ and $\mu_T(a,b) = \mu$.

Definition 2.35. Let $Y_1, ..., Y_n$ be iid random variables and let $D_n \equiv D_n(Y_1, ..., Y_n)$ be an estimator of a parameter μ_D such that

$$\sqrt{n}(D_n - \mu_D) \xrightarrow{D} N(0, \sigma_D^2).$$

Then the *asymptotic variance* of $\sqrt{n}(D_n - \mu_D)$ is σ_D^2 and the *asymptotic variance* (AV) of D_n is σ_D^2/n. If S_D^2 is a consistent estimator of σ_D^2, then the (asymptotic) *standard error* (SE) of D_n is $S_D/\sqrt{n}$.

Remark 2.10. In the literature, usually either σ_D^2 or σ_D^2/n is called the asymptotic variance of D_n. The parameter σ_D^2 is a function of both the estimator D_n and the underlying distribution F of Y_1. Frequently $n\text{VAR}(D_n)$ converges in distribution to σ_D^2, but not always. See Staudte and Sheather (1990, p. 51) and Lehmann (1999, p. 232).

Example 2.20. If $Y_1, ..., Y_n$ are iid from a distribution with mean μ and variance σ^2, then by the central limit theorem,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Recall that $\text{VAR}(\overline{Y}_n) = \sigma^2/n = AV(\overline{Y}_n)$ and that the standard error $SE(\overline{Y}_n) = S_n/\sqrt{n}$ where S_n^2 is the sample variance.

Remark 2.11. Returning to the trimmed mean T_n where Y is from a symmetric location–scale family, take $\mu = 0$ since the asymptotic variance does not depend on μ. Then

$$n\ AV(T_n) = \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2} = \frac{\sigma_T^2(a,b)}{1-2\alpha} + \frac{2\alpha(F^{-1}(\alpha))^2}{(1-2\alpha)^2}.$$

See, for example, Bickel (1965). This formula is useful since the variance of the truncated distribution $\sigma_T^2(a,b)$ has been computed for several distributions in Section 11.5.

Definition 2.36. An estimator D_n is a *location and scale equivariant estimator* if $D_n(\alpha+\beta Y_1, ..., \alpha+\beta Y_n) = \alpha + \beta D_n(Y_1, ..., Y_n)$ where α and β are arbitrary real constants.

Remark 2.12. Many location estimators such as the sample mean, sample median, trimmed mean, metrically trimmed mean, and two stage trimmed means are equivariant. Let $Y_1, ..., Y_n$ be iid from a distribution with cdf $F_Y(y)$ and suppose that D_n is an equivariant estimator of $\mu_D \equiv \mu_D(F_Y) \equiv \mu_D(F_Y(y))$. If $X_i = \alpha + \beta Y_i$ where $\beta \neq 0$, then the cdf of X is $F_X(y) = F_Y((y-\alpha)/\beta)$. Suppose that

$$\mu_D(F_X) \equiv \mu_D[F_Y(\frac{y-\alpha}{\beta})] = \alpha + \beta \mu_D[F_Y(y)]. \tag{2.42}$$

Let $D_n(\boldsymbol{Y}) \equiv D_n(Y_1, ..., Y_n)$. If $\sqrt{n}[D_n(\boldsymbol{Y}) - \mu_D(F_Y(y))] \xrightarrow{D} N(0, \sigma_D^2)$, then

$$\sqrt{n}[D_n(\boldsymbol{X}) - \mu_D(F_X)] = \sqrt{n}[\alpha + \beta D_n(\boldsymbol{Y}) - (\alpha + \beta \mu_D(F_Y))] \xrightarrow{D} N(0, \beta^2 \sigma_D^2).$$

This result is especially useful when F is a cdf from a location–scale family with parameters μ and σ. In this case, Equation (2.42) holds when μ_D is the population mean, population median, and the population truncated mean with truncation points $a = \mu - z_1\sigma$ and $b = \mu + z_2\sigma$ (the parameter estimated by trimmed and two stage trimmed means).

Refer to the notation for two stage trimmed means below Theorem 2.4. Then from Theorem 2.6,

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

and

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

If the distribution of Y is symmetric then $T_{A,n}$ and $T_{S,n}$ are asymptotically equivalent. It is important to note that no knowledge of the unknown distribution and parameters is needed to compute the two stage trimmed means and their standard errors.

The next three lemmas find the asymptotic variance for trimmed and two stage trimmed means when the underlying distribution is normal, double exponential and Cauchy, respectively. Assume $a = \text{MED}(Y) - k\text{MAD}(Y)$ and $b = \text{MED}(Y) + k\text{MAD}(Y)$.

Theorem 2.14. Suppose that Y comes from a normal $N(\mu, \sigma^2)$ distribution. Let $\Phi(x)$ be the cdf and let $\phi(x)$ be the density of the standard normal. Then for the α trimmed mean,

$$n\ AV = \left(\frac{1 - \frac{2z\phi(z)}{2\Phi(z)-1}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \right) \sigma^2 \tag{2.43}$$

where $\alpha = \Phi(-z)$, and $z = k\Phi^{-1}(0.75)$. For the two stage estimators, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and $z_J = -\Phi^{-1}(\alpha_J)$ in Equation (2.43).

Proof. If Y follows the normal $N(\mu, \sigma^2)$ distribution, then $a = \mu - k\text{MAD}(Y)$ and $b = \mu + k\text{MAD}(Y)$ where $\text{MAD}(Y) = \Phi^{-1}(0.75)\sigma$. It is enough to consider the standard N(0,1) distribution since $n\ AV(T_n, N(\mu, \sigma^2)) = \sigma^2\ n\ AV(T_n, N(0,1))$. If $a = -z$ and $b = z$, then by Theorem 11.6,

$$\sigma_T^2(a,b) = 1 - \frac{2z\phi(z)}{2\Phi(z) - 1}.$$

Use Remark 2.11 with $z = k\Phi^{-1}(0.75)$, and $\alpha = \Phi(-z)$ to get Equation (2.43). □

Theorem 2.15. Suppose that Y comes from a double exponential DE(0,1) distribution. Then for the α trimmed mean,

$$n\ AV = \frac{\frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \tag{2.44}$$

where $z = k\log(2)$ and $\alpha = 0.5\exp(-z)$. For the two stage estimators, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and let $z_J = -\log(2\alpha_J)$.

Proof Sketch. For the $DE(0,1)$ distribution, $\text{MAD}(Y) = \log(2)$. If the DE(0,1) distribution is truncated at $-z$ and z, then use Remark 2.11 with

$$\sigma_T^2(-z,z) = \frac{2 - (z^2 + 2z + 2)e^{-z}}{1 - e^{-z}}.$$

Theorem 2.16. Suppose that Y comes from a Cauchy (0,1) distribution. Then for the α trimmed mean,

$$n\ AV = \frac{z - \tan^{-1}(z)}{(1-2\alpha)\tan^{-1}(z)} + \frac{2\alpha(\tan[\pi(\alpha - \frac{1}{2})])^2}{(1-2\alpha)^2} \tag{2.45}$$

where $z = k$ and

$$\alpha = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(z).$$

For the two stage estimators, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and let $z_J = \tan[\pi(\alpha_J - 0.5)]$.

Proof Sketch. For the $C(0,1)$ distribution, $\text{MAD}(Y) = 1$. If the C(0,1) distribution is truncated at $-z$ and z, then use Remark 2.11 with

$$\sigma_T^2(-z,z) = \frac{z - \tan^{-1}(z)}{\tan^{-1}(z)}.$$

2.14 Simulation

In statistics, *simulation* uses computer generated pseudo-random variables in place of real data. This artificial data can be used just like real data to produce histograms and confidence intervals and to compare estimators. Since the artificial data is under the investigator's control, often the theoretical behavior of the statistic is known. This knowledge can be used to estimate population quantities (such as $\text{MAD}(Y)$) that are otherwise hard to compute and to check whether software is running correctly.

Example 2.21. The *R* software is especially useful for generating random variables. The command

```
Y <- rnorm(100)
```

creates a vector Y that contains 100 pseudo iid N(0,1) variables. More generally, the command

```
Y <- rnorm(100,10,sd=4)
```

creates a vector Y that contains 100 pseudo iid $N(10, 16)$ variables since $4^2 = 16$. To study the sampling distribution of $\overline{Y}_n$, we could generate K $N(0,1)$ samples of size n, and compute $\overline{Y}_{n,1}, ..., \overline{Y}_{n,K}$ where the notation $\overline{Y}_{n,j}$ denotes the sample mean of the n pseudo-variates from the jth sample. The command

```
M <- matrix(rnorm(1000),nrow=100,ncol=10)
```

creates a 100×10 matrix containing 100 samples of size 10. (Note that 100(10) = 1000.) The command

```
M10 <- apply(M,1,mean)
```

creates the vector M10 of length 100 which contains $\overline{Y}_{n,1}, ..., \overline{Y}_{n,K}$ where $K = 100$ and $n = 10$. A histogram from this vector should resemble the pdf of a $N(0, 0.1)$ random variable. The sample mean and variance of the 100 vector entries should be close to 0 and 0.1, respectively.

Example 2.22. Similarly the command

```
M <- matrix(rexp(1000),nrow=100,ncol=10)
```

creates a 100×10 matrix containing 100 samples of size 10 exponential(1) (pseudo) variates. (Note that $100(10) = 1000$.) The command

```
M10 <- apply(M,1,mean)
```

gets the sample mean for each (row) sample of 10 observations. The command

```
M <- matrix(rexp(10000),nrow=100,ncol=100)
```

creates a 100×100 matrix containing 100 samples of size 100 exponential(1) (pseudo) variates. (Note that $100(100) = 10000$.) The command

```
M100 <- apply(M,1,mean)
```

gets the sample mean for each (row) sample of 100 observations. The commands

```
hist(M10) and hist(M100)
```

will make histograms of the 100 sample means. The first histogram should be more skewed than the second, illustrating the central limit theorem.

Example 2.23. As a slightly more complicated example, suppose that it is desired to approximate the value of MAD(Y) when Y is the mixture distribution with cdf $F(y) = 0.95\Phi(y) + 0.05\Phi(y/3)$. That is, roughly 95% of the variates come from a $N(0, 1)$ distribution and 5% from a $N(0, 9)$ distribution. Since MAD(n) is a good estimator of MAD(Y), the following R commands can be used to approximate MAD(Y).

```
contam <- rnorm(10000,0,(1+2*rbinom(10000,1,0.05)))
mad(contam,constant=1)
```

Running these commands suggests that MAD(Y) ≈ 0.70. Now $F(MAD(Y)) = 0.75$. To find $F(0.7)$, use the command

```
0.95*pnorm(.7) + 0.05*pnorm(.7/3)
```

which gives the value 0.749747. Hence the approximation was quite good.

Definition 2.37. Let $T_{1,n}$ and $T_{2,n}$ be two estimators of a parameter τ such that

$$n^{\delta}(T_{1,n} - \tau) \xrightarrow{D} N(0, \sigma_1^2(F))$$

and

$$n^{\delta}(T_{2,n} - \tau) \xrightarrow{D} N(0, \sigma_2^2(F)),$$

then the *asymptotic relative efficiency* of $T_{1,n}$ with respect to $T_{2,n}$ is

$$ARE(T_{1,n}, T_{2,n}) = \frac{\sigma_2^2(F)}{\sigma_1^2(F)} = \frac{AV(T_{2,n})}{AV(T_{1,n})}.$$

This definition brings up several issues. First, both estimators must have the same convergence rate n^{δ}. Usually $\delta = 0.5$. If $T_{i,n}$ has convergence rate

n^{δ_i}, then estimator $T_{1,n}$ is judged to be better than $T_{2,n}$ if $\delta_1 > \delta_2$. Secondly, the two estimators need to estimate the same parameter τ. This condition will often not hold unless the distribution is symmetric about μ. Then $\tau = \mu$ is a natural choice. Thirdly, robust estimators are often judged by their Gaussian efficiency with respect to the sample mean (thus F is the normal distribution). Since the normal distribution is a location–scale family, it is often enough to compute the ARE for the standard normal distribution. If the data come from a distribution F and the ARE can be computed, then $T_{1,n}$ is judged to be a better estimator at the data than $T_{2,n}$ if the $ARE > 1$.

In simulation studies, typically the underlying distribution F belongs to a symmetric location–scale family. There are at least two reasons for using such distributions. First, if the distribution is symmetric, then the population median MED(Y) is the point of symmetry and the natural parameter to estimate. Under the symmetry assumption, there are many estimators of MED(Y) that can be compared via their ARE with respect to the sample mean or maximum likelihood estimator (MLE). Secondly, once the ARE is obtained for one member of the family, it is typically obtained for *all members of the location–scale family.* That is, suppose that $Y_1, ..., Y_n$ are iid from a location–scale family with parameters μ and σ. Then $Y_i = \mu + \sigma Z_i$ where the Z_i are iid from the same family with $\mu = 0$ and $\sigma = 1$. Typically

$$AV[T_{i,n}(\boldsymbol{Y})] = \sigma^2 AV[T_{i,n}(\boldsymbol{Z})], \quad \text{so}$$

$$ARE[T_{1,n}(\boldsymbol{Y}), T_{2,n}(\boldsymbol{Y})] = ARE[T_{1,n}(\boldsymbol{Z}), T_{2,n}(\boldsymbol{Z})].$$

Example 2.24. If $T_{2,n} = \overline{Y}$, then by the central limit theorem $\sigma_2^2(F) = \sigma^2$ when F is the $N(\mu, \sigma^2)$ distribution. Then $ARE(T_{A,n}, \overline{Y}_n) = \sigma^2/(nAV)$ where nAV is given by Equation (2.43). Note that the ARE does not depend on σ^2. If $k \in [5, 6]$, then $J = 1$, and $ARE(T_{A,n}, \overline{Y}_n) \approx 0.996$. Hence $T_{S,n}$ and $T_{A,n}$ are asymptotically equivalent to the 1% trimmed mean and are almost as good as the optimal sample mean at Gaussian data.

Warning: Claiming superefficiency of robust estimators at the normal distribution due to simulation and without any theory, as done by Zuo (2010), is unwise. The 1% trimmed mean, $T_{S,n}$ and $T_{A,n}$ (both with $k_1 = k_2 = 6$) often had simulated variances that beat $\overline{Y}$ for "normal" data. This simulation result happens since these three robust estimators are nearly as efficient as $\overline{Y}$ (though certainly not superefficient) at normal data, and pseudo–normal data is used instead of genuine normal data. The following R output illustrates the phenomenon. For $n = 500$ and 100 runs, only the sample median had a smaller simulated variance than $\overline{Y}$ at N(0,1) data. Here trmn is the 1% trimmed mean, rstmn = $T_{S,n}$ and ratmn = $T_{A,n}$. Let $\overline{T}_i$ be the value of the robust point estimator for the ith sample for $i = 1, ..., 100$. Let $S^2(T)$ be the sample variance of $T_1, ..., T_{100}$. Then $nS^2(T)$ is shown by the "vars" line. For $\overline{Y}$ the value 1.1359 estimates $n\sigma^2/n = 1.0$.

```
locsim(n=500) #from rpack
[1] "mean,median,trmn,rstmn,ratmn"
$vars:
[1] 1.135908 1.616481 1.125468 1.135834 1.125910
```

Example 2.25. If F is the $DE(0,1)$ cdf, then the asymptotic efficiency of $T_{A,n}$ with respect to the mean is $ARE = 2/(nAV)$ where nAV is given by Equation (2.44). If $k = 5$, then $J = 2$, and $ARE(T_{A,n}, \overline{Y}_n) \approx 1.108$. Hence $T_{S,n}$ and $T_{A,n}$ are asymptotically equivalent to the 2% trimmed mean and perform better than the sample mean. If $k = 6$, then $J = 1$, and $ARE(T_{A,n}, \overline{Y}_n) \approx 1.065$.

The results from a small simulation are presented in Table 2.5. For each sample size n, 500 samples were generated. The sample mean $\overline{Y}$, sample median, 1% trimmed mean, and $T_{S,n}$ were computed. The latter estimator was computed using the trimming parameter $k = 5$. Next the sample variance $S^2(T)$ of the 500 values $T_1, ..., T_{500}$ was computed where T is one of the four estimators. The value in the table is $nS^2(T)$. These numbers estimate n times the actual variance of the estimators. Suppose that for $n \geq N$, the tabled numbers divided by n are close to the asymptotic variance. Then the asymptotic theory may be useful if the sample size $n \geq N$ and if the distribution corresponding to F is a reasonable approximation to the data (but see Lehmann 1999, p. 74). The scaled asymptotic variance σ_D^2 is reported in the rows $n = \infty$. The simulations were performed for normal and double exponential data, and the simulated values are close to the theoretical values.

Table 2.5 Simulated Scaled Variance, 500 Runs, k = 5

F	n	$\overline{Y}$	MED(n)	1% TM	$T_{S,n}$
N(0,1)	10	1.116	1.454	1.116	1.166
N(0,1)	50	0.973	1.556	0.973	0.974
N(0,1)	100	1.040	1.625	1.048	1.044
N(0,1)	1000	1.006	1.558	1.008	1.010
N(0,1)	∞	1.000	1.571	1.004	1.004
DE(0,1)	10	1.919	1.403	1.919	1.646
DE(0,1)	50	2.003	1.400	2.003	1.777
DE(0,1)	100	1.894	0.979	1.766	1.595
DE(0,1)	1000	2.080	1.056	1.977	1.886
DE(0,1)	∞	2.000	1.000	1.878	1.804

A small simulation study was used to compare some simple randomly trimmed means. The $N(0,1)$, $0.75N(0,1) + 0.25N(100,1)$ (shift), C(0,1), DE(0,1) and exponential(1) distributions were considered. For each distribution $K = 500$ samples of size $n = 10$, 50, 100, and 1000 were generated. See Problem 2.37.

Six different CIs

$$D_n \pm t_{d,0.975} SE(D_n)$$

were used. The degrees of freedom $d = U_n - L_n - 1$, and usually $SE(D_n) = SE_{RM}(L_n, U_n)$. See Definition 2.26.
(i) The classical interval used $D_n = \overline{Y}$, $d = n-1$ and SE $= S/\sqrt{n}$. Note that $\overline{Y}$ is a 0% trimmed mean that uses $L_n = 0, U_n = n$ and $SE_{RM}(0, n) = S/\sqrt{n}$.
(ii) This robust interval used $D_n = T_{A,n}$ with $k_1 = k_2 = 6$ and $SE = SE_{RM}(L_n, U_n)$ where U_n and L_n are given by Definition 2.25.
(iii) This resistant interval used $D_n = T_{S,n}$ with $k_1 = k_2 = 3.5$, and $SE = SE_{RM}(L_n, U_n)$ where U_n and L_n are given by Definition 2.24.
(iv) This resistant interval used $D_n = \text{MED}(n)$ with $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\ \rceil$. Note that $d = U_n - L_n - 1 \approx \sqrt{n}$. Following Application 2.4, $SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)})$.
(v) This resistant interval again used $D_n = \text{MED}(n)$ with $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\ \rceil$, but $SE(\text{MED}(n)) = SE_{RM}(L_n, U_n)$ was used. Note that MED(n) is the 50% trimmed mean and that the percentage of cases used to compute the SE goes to 0 as $n \to \infty$.
(vi) This resistant interval used the 25% trimmed mean for D_n and $SE = SE_{RM}(L_n, U_n)$ where U_n and L_n are given by $L_n = \lfloor 0.25n \rfloor$ and $U_n = n - L_n$.

Table 2.6 Simulated 95% CI Coverages, 500 Runs

F and n		$\overline{Y}$	$T_{A,n}$	$T_{S,n}$	MED	(v)	25% TM
N(0,1)	10	0.960	0.942	0.926	0.948	0.900	0.938
N(0,1)	50	0.948	0.946	0.930	0.936	0.890	0.926
N(0,1)	100	0.932	0.932	0.932	0.900	0.898	0.938
N(0,1)	1000	0.942	0.934	0.936	0.940	0.940	0.936
DE(0,1)	10	0.966	0.954	0.950	0.970	0.944	0.968
DE(0,1)	50	0.948	0.956	0.958	0.958	0.932	0.954
DE(0,1)	100	0.956	0.940	0.948	0.940	0.938	0.938
DE(0,1)	1000	0.948	0.940	0.942	0.936	0.930	0.944
C(0,1)	10	0.974	0.968	0.964	0.980	0.946	0.962
C(0,1)	50	0.984	0.982	0.960	0.960	0.932	0.966
C(0,1)	100	0.970	0.996	0.974	0.940	0.938	0.968
C(0,1)	1000	0.978	0.992	0.962	0.952	0.942	0.950
EXP(1)	10	0.892	0.816	0.838	0.948	0.912	0.916
EXP(1)	50	0.938	0.886	0.892	0.940	0.922	0.950
EXP(1)	100	0.938	0.878	0.924	0.930	0.920	0.954
EXP(1)	1000	0.952	0.848	0.896	0.926	0.922	0.936
SHIFT	10	0.796	0.904	0.850	0.940	0.910	0.948
SHIFT	50	0.000	0.986	0.620	0.740	0.646	0.820
SHIFT	100	0.000	0.988	0.240	0.376	0.354	0.610
SHIFT	1000	0.000	0.992	0.000	0.000	0.000	0.442

In order for a location estimator to be used for inference, there must exist a useful SE and a useful cutoff value t_d where the degrees of freedom d is a function of n. Two criteria will be used to evaluate the CIs. First, the

observed coverage is the proportion of the $K = 500$ runs for which the CI contained the parameter estimated by D_n. This proportion should be near the nominal coverage 0.95. Notice that if W is the proportion of runs where the CI contains the parameter, then KW is a binomial random variable. Hence the SE of W is $\sqrt{\hat{p}(1-\hat{p})/K} \approx 0.013$ for the observed proportion $\hat{p} \in [0.9, 0.95]$, and an observed coverage between 0.92 and 0.98 suggests that the observed coverage is close to the nominal coverage of 0.95.

The second criterion is the scaled length of the CI $= \sqrt{n}$ CI length $=$

$$\sqrt{n}(2)(t_{d,0.975})(SE(D_n)) \approx 2(1.96)(\sigma_D)$$

where the approximation holds if $d > 30$, if $\sqrt{n}(D_n - \mu_D) \xrightarrow{D} N(0, \sigma_D^2)$, and if $SE(D_n)$ is a good estimator of $\sigma_D/\sqrt{n}$ for the given value of n.

Table 2.7 Simulated Scaled CI Lengths, 500 Runs

F and n		$\overline{Y}$	$T_{A,n}$	$T_{S,n}$	MED	(v)	25% TM
N(0,1)	10	4.467	4.393	4.294	7.803	6.030	5.156
N(0,1)	50	4.0135	4.009	3.981	5.891	5.047	4.419
N(0,1)	100	3.957	3.954	3.944	5.075	4.961	4.351
N(0,1)	1000	3.930	3.930	3.940	5.035	4.928	4.290
N(0,1)	∞	3.920	3.928	3.928	4.913	4.913	4.285
DE(0,1)	10	6.064	5.534	5.078	7.942	6.120	5.742
DE(0,1)	50	5.591	5.294	4.971	5.360	4.586	4.594
DE(0,1)	100	5.587	5.324	4.978	4.336	4.240	4.404
DE(0,1)	1000	5.536	5.330	5.006	4.109	4.021	4.348
DE(0,1)	∞	5.544	5.372	5.041	3.920	3.920	4.343
C(0,1)	10	54.590	10.482	9.211	12.682	9.794	9.858
C(0,1)	50	94.926	10.511	8.393	7.734	6.618	6.794
C(0,1)	100	243.4	10.782	8.474	6.542	6.395	6.486
C(0,1)	1000	515.9	10.873	8.640	6.243	6.111	6.276
C(0,1)	∞	∞	10.686	8.948	6.157	6.157	6.255
EXP(1)	10	4.084	3.359	3.336	6.012	4.648	3.949
EXP(1)	50	3.984	3.524	3.498	4.790	4.105	3.622
EXP(1)	100	3.924	3.527	3.503	4.168	4.075	3.571
EXP(1)	1000	3.914	3.554	3.524	3.989	3.904	3.517
SHIFT	10	184.3	18.529	24.203	203.5	166.2	189.4
SHIFT	50	174.1	7.285	9.245	18.686	16.311	180.1
SHIFT	100	171.9	7.191	29.221	7.651	7.481	177.5
SHIFT	1000	169.7	7.388	9.453	7.278	7.123	160.6

Tables 2.6 and 2.7 can be used to examine the six different interval estimators. A good estimator should have an observed coverage $\hat{p} \in [.92, .98]$, and a small scaled length. In Table 2.6, coverages were good for $N(0,1)$ data, except the interval (v) where $SE_{RM}(L_n, U_n)$ is slightly too small for $n \leq 100$. The coverages for the C(0,1) and DE(0,1) data were all good even for $n = 10$.

For the mixture $0.75N(0,1) + 0.25N(100,1)$, the "coverage" counted the number of times 0 was contained in the interval and divided the result by 500.

These rows do not give a genuine coverage since the parameter μ_D estimated by D_n is not 0 for any of these estimators. For example $\overline{Y}$ estimates $\mu = 25$. Since the median, 25% trimmed mean, and $T_{S,n}$ trim the same proportion of cases to the left as to the right, MED(n) is estimating MED$(Y) \approx \Phi^{-1}(2/3) \approx 0.43$ while the parameter estimated by $T_{S,n}$ is approximately the mean of a truncated standard normal random variable where the truncation points are $\Phi^{-1}(.25)$ and ∞. The 25% trimmed mean also has trouble since the number of outliers is a binomial$(n, 0.25)$ random variable. Hence approximately half of the samples have more than 25% outliers and approximately half of the samples have less than 25% outliers. This fact causes the 25% trimmed mean to have great variability. The parameter estimated by $T_{A,n}$ is zero to several decimal places. Hence the coverage of the $T_{A,n}$ interval is quite high.

The exponential(1) distribution is skewed, so the central limit theorem is not a good approximation for $n = 10$. The estimators $\overline{Y}, T_{A,n}, T_{S,n}$, MED$(n)$ and the 25% trimmed mean are estimating the parameters 1, 0.89155, 0.83071, log(2) and 0.73838 respectively. Now the coverages of $T_{A,n}$ and $T_{S,n}$ are slightly too small. For example, $T_{S,n}$ is asymptotically equivalent to the 10% trimmed mean since the metrically trimmed mean truncates the largest 9.3% of the cases, asymptotically. For small n, the trimming proportion will be quite variable and the mean of a truncated exponential distribution with the largest γ percent of cases trimmed varies with γ. This variability of the truncated mean does not occur for symmetric distributions if the trimming is symmetric since then the truncated mean μ_T is the point of symmetry regardless of the amount of truncation.

Examining Table 2.7 for N(0,1) data shows that the scaled lengths of the first 3 intervals are about the same. The rows labeled ∞ give the scaled length $2(1.96)(\sigma_D)$ expected if $\sqrt{n}SE$ is a good estimator of σ_D. The median interval and 25% trimmed mean interval are noticeably larger than the classical interval. Since the degrees of freedom $d \approx \sqrt{n}$ for the median intervals, $t_{d,0.975}$ is considerably larger than $1.96 = z_{0.975}$ for $n \leq 100$.

The intervals for the C(0,1) and DE(0,1) data behave about as expected. The classical interval is very long at C(0,1) data since the first moment of C(0,1) data does not exist. Notice that for $n \geq 50$, all of the resistant intervals are shorter on average than the classical intervals for DE(0,1) data.

For the mixture distribution, examining the length of the interval should be fairer than examining the "coverage." The length of the 25% trimmed mean is long since about half of the time the trimmed data contains no outliers while half of the time the trimmed data does contain outliers. When $n = 100$, the length of the $T_{S,n}$ interval is quite long. This occurs because the $T_{S,n}$ will usually trim all outliers, but the actual proportion of outliers is binomial(100, 0.25). Hence $T_{S,n}$ is sometimes the 20% trimmed mean and sometimes the 30% trimmed mean. But the parameter μ_T estimated by the γ % trimmed mean varies quite a bit with γ. When $n = 1000$, the trimming proportion is much less variable, and the CI length is shorter.

For exponential(1) data, $2(1.96)(\sigma_D) = 3.9199$ for $\overline{Y}$ and MED(n). The 25% trimmed mean appears to be the best of the six intervals since the scaled length is the smallest while the coverage is good.

2.15 Sequential Analysis

This section is not yet written. See Huber and Ronchetti (2009, pp. 267-268), Olive (1998), and Quang (1985).

2.16 Summary

1) Given a small data set, $\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}$ 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* MED(n) $= Y_{((n+1)/2)}$ if n is odd, MED(n) $= \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2}$ if n is even. The notation MED(n) = 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 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 MED(W) = $a + b$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 $P(Y \leq y_\delta) = \delta$, and y_δ is the 100δth 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 $\text{MAD}(W) = |b|\text{MAD}(Y)$.

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

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

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

where $P(t_p \leq t_{p,1-\frac{\delta}{2}}) = 1 - \alpha/2$ if t_p is from a t distribution with p degrees of freedom. We will use 95% CIs so $\delta = 0.05$ and $t_{p,1-\frac{\delta}{2}} = t_{p,0.975} \approx 1.96$ 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 (e.g., $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (e.g., $\lceil 7.7 \rceil = 8$).

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 $T_n = T_n(L_n, U_n) = \dfrac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)}$ 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 $V_{SW}(L_n, U_n) = \dfrac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}$.

Then the **CI for the population truncated mean** $\theta = \mu_T$ uses $\hat{\theta} = T_n$, $p = U_n - L_n - 1$ and $SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n}$.

5) The δ quantile or 100δth percentile $y_\delta = \pi_\delta = \xi_\delta$ satisfies $P(Y \leq y_\delta) = \delta$. The *sample* δ *quantile* or sample 100δth percentile $\hat{\xi}_{n,\rho} = Y_{(\lceil n\delta \rceil)}$. Software often uses $\tilde{\xi}_{n,\rho} = \gamma_n Y_{(\lceil n\delta \rceil)} + (1-\gamma_n) Y_{(\lfloor n\delta \rfloor)}$ for some $0 \leq \gamma_n \leq 1$.

6) Consider intervals that contain c cases $[Y_{(1)}, Y_{(c)}], [Y_{(2)}, Y_{(c+1)}], ..., [Y_{(n-c+1)}, Y_{(n)}]$. Compute $Y_{(c)} - Y_{(1)}, Y_{(c+1)} - Y_{(2)}, ..., Y_{(n)} - Y_{(n-c+1)}$. Then the estimator shorth$(c) = [Y_{(s)}, Y_{(s+c-1)}]$ is the interval with the shortest length. The shorth(c) interval is a large sample $100(1-\delta)\%$ PI if $c/n \to 1-\delta$ as $n \to \infty$ that estimates the population shorth. Hence the shorth PI is often asymptotically optimal.

7) A large sample $100(1-\delta)\%$ prediction interval (PI) $[\hat{L}_n, \hat{U}_n]$ is such that $P(Y_f \in [\hat{L}_n, \hat{U}_n])$ is eventually bounded below by $1-\delta$ as $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the *population shorth*: the shortest interval covering at least $100(1-\delta)\%$ of the mass. So $F(U_s) - F(L_s-) \geq 1-\delta$, and if $F(b) - F(a-) \geq 1-\delta$, then $b - a \geq U_s - L_s$. The population shorth need not be unique, but the length of the population shorth is unique.

8) The interval $[\hat{L}_n, \hat{U}_n]$ is a large sample $100(1-\delta)\%$ *confidence interval* for θ if $P(\hat{L}_n \leq \theta \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

9) Given B samples drawn with replacement from the cases (nonparametric bootstrap), be able to compute simple statistics T_j^* from the jth sample such as the sample mean, the sample median, the max, the min, the range = max $-$ min. See Example 2.10. The bagging estimator is $\overline{T}^* = \dfrac{1}{B}\sum_{j=1}^{B} T_j^*$.

10) The bootstrap sample is $T_1^*, ..., T_B^*$. Often B is a fixed number such as $B = 1000$, but using $B = \max(1000, \lceil n \log(n) \rceil)$ works better if you want the coverage of the bootstrap CI to converge to $1-\delta$ as $n \to \infty$.

11) Given a bootstrap sample $T_1^*, ..., T_B^*$, let the order statistics be $T_{(1)}^*, ..., T_{(B)}^*$. Applying certain PIs to the bootstrap sample results in CIs. The shorth(c) CI is found as in 6). The prediction region method CI is $[\overline{T}^* - a, \overline{T}^* + a]$, which is the interval centered at $\overline{T}^*$ just long enough to contain $U_B \approx \lceil B(1-\delta) \rceil$ of the T_j^*. The modified Bickel and Ren CI is $[T_n - b, T_n + b]$, which is the interval centered at T_n just long enough to contain U_B of the T_j^*. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. The percentile CI is $[T_{(k_1)}^*, T_{(k_2)}^*]$, which deletes the $k_1 - 1$ smallest and $B - k_2$ largest T_j^*.

12) For a large sample level δ test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, reject H_0 if θ_0 is not in the large sample $100(1-\delta)\%$ confidence interval (CI) for θ. A bootstrap test corresponds to a bootstrap CI.

2.17 Complements

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

Huber and Roncheti (2009, p. 72-73) shows that the sample median minimizes the asymptotic bias for estimating MED(Y) for the family of symmetric contaminated distributions, and concludes that since the asymptotic variance is going to zero for reasonable estimators, MED(n) is the estimator of choice for large n. Also see Chen (1998). Hampel et al. (1986, p. 133-134, 142-143) contains some other optimality properties of MED(n) and MAD(n). See Olive (1998) and Serfling and Mazumder (2009) for large sample theory for MAD(n).

The prediction region method CI (2.16) is due to Olive (2017b: pp. 168-169). CIs (2.17) and (2.18) are due to Pelawa Watagoda and Olive (2019).

CI (2.19) from Application 2.4 is due to Olive (2005b, 2017b: p. 11). Several other approximations for the standard error of the sample median $SE(\text{MED}(n))$ could be used. Also see Baszczyńska and Pekasiewicz (2010), Larocque and Randles (2008), and Woodruff (1952).

a) McKean and Schrader (1984) proposed

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

where $c = (n+1)/2 - z_{1-\delta/2}\sqrt{n/4}$ is rounded up to the nearest integer. This estimator was based on the half length of a distribution free 100 $(1-\delta)$% 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 2.26. Use the t_p approximation with $p = U_n - L_n - 1$.

In a small simulation study (see Section 2.14), the proposal in Application 2.4 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 and Ronchetti (2009), Hampel et al. (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. 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 et al. (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). For variance estimation of L-estimators, see Wang et al. (2012).

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 et al. (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 references.

Scale estimators are essential for testing and are discussed in Falk (1997), Hall and Welsh (1985), Lax (1985), Rousseeuw and Croux (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. Hettmansperger and McKean (2010) consider rank procedures.

Wilcox (2012) 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 (2012) 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. Ullah et al. (2006) list some of the better M-estimators. Jureckova and Sen (1996, p. 208) show that under symmetry a large class of M-estimators is asymptotically nor-

mal, 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 2.20 do not appear in the paper) are adequate for finding outliers.

The median can be computed with $O(n \log(n))$ complexity by sorting the data, but faster $O(n)$ complexity algorithms exist. Google *quickselect* or see Blum et al. (1973) for references.

Several points about resistant location estimators need to be made. First, **by far the most important step in analyzing location data is to check whether outliers are present with a plot of the data**. Secondly, no single procedure will dominate all other procedures. In particular, it is unlikely that the sample mean will be replaced by a robust estimator. The sample mean often works well for distributions with second moments. In particular, the sample mean works well for many skewed and discrete distributions. Thirdly, the mean and the median should usually both be computed. If a CI is needed and the data is thought to be symmetric, several resistant CIs should be computed and compared with the classical interval. Fourthly, in order to perform hypothesis testing, reasonable values for the unknown parameter must be given. The mean and median of the population are fairly simple parameters even if the population is skewed while the truncated population mean is considerably more complex.

With some robust estimators, it very difficult to determine what the estimator is estimating if the population is not symmetric. In particular, the difficulty in finding reasonable values of the population quantities estimated by M, L, and R estimators may be one reason why these estimators are not widely used. For testing hypotheses, the following population quantities are listed in order of increasing complexity.
1) The population median MED(Y).
2) The population mean $E(Y)$.
3) The truncated mean μ_T as estimated by the α trimmed mean.
4) The truncated mean μ_T as estimated by the (α, β) trimmed mean.
5) The truncated mean μ_T as estimated by the $T_{S,n}$.
6) The truncated mean μ_T as estimated by the $T_{A,n}$.

Bickel (1965), Prescott (1978), and Olive (2001) give formulas similar to Equations (2.43) and (2.4). Gross (1976), Guenther (1969) and Lax (1985) are useful references for confidence intervals. Andrews et al. (1972) is a well known simulation study for robust location estimators.

In Section 2.14, only intervals that are simple to compute by hand for sample sizes of ten or so were considered. The interval based on MED(n) (see Application 2.4 and the column "MED" in Tables 2.6 and 2.7) is even easier to compute than the classical interval, kept its coverage pretty well, and was frequently shorter than the classical interval.

Stigler (1973a) showed that the trimmed mean has a limiting normal distribution even if the population is discrete provided that the asymptotic

truncation points a and b have zero probability; however, in finite samples the trimmed mean can perform poorly if there are gaps in the distribution near the trimming proportions. Stigler (1977) argues that complicated robust estimators are not needed.

Warning: Simulations for confidence intervals and prediction intervals should include both length and coverage while simulations for tests of hypothesis should include both coverage and power.

The Shorth: Useful papers for the shorth include Chen and Shao (1999), Einmahl and Mason (1992), Frey (2013), Grübel (1988) and Pelawa Watagoda and Olive (2019).

The Bootstrap:

Buckland (1984) shows that the expected coverage of the nominal $100(1-\delta)\%$ percentile confidence interval is approximately correct, but the standard deviation of the coverage is proportional to $1/\sqrt{B}$. Hence the percentile CI is a large sample confidence interval, in that the true coverage converges in probability to the nominal coverage, only if $B \to \infty$ as $n \to \infty$. These results are good reasons for using $B = \max(1000, \lfloor n \log(n) \rfloor)$ samples for the location model. Also see Olive (2014, pp. 279-283) and Robinson (1988). Efron (1982) and Efron and Tibshirani (1993) are good books for the bootstrap.

2.18 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 2.2 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 2.26 reduces to the SE for the sample mean if there is no trimming.

b) Prove Remark 2.8:

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

2.9. Find a 95% CI for μ_T based on the 25% trimmed mean for the following data sets. Follow Examples 2.16 and 2.17 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 $S^2(d_1, ..., d_{10})$.

h) Find SE(T_n).

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

a) Referring to Application 2.4, find L_n, U_n, p and SE(MED(n)).

b) Referring to Application 2.5, let T_n be the 25% trimmed mean. Find L_n, U_n, p, T_n and SE(T_n).

2.15. Consider the Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) listed below. Find shorth(7). Show work.

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

2.16. Find shorth(5) for the following data set. Show work.

```
6   76   90   90   94   94   95   97   97   1008
```

2.17. Find shorth(5) for the following data set. Show work.

```
66  76   90   90   94   94   95   95   97   98
```

2.18. Suppose you are estimating the mean θ of losses with the maximum likelihood estimator (MLE) $\overline{X}$ assuming an exponential (θ) distribution. Compute the sample mean of the fourth bootstrap sample.

actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$
bootstrap samples:
2, 10, 1, 2, 2: $\overline{X} = 3.4$
50, 10, 50, 2, 2: $\overline{X} = 22.8$
10, 50, 2, 1, 1: $\overline{X} = 12.8$
5, 2, 5, 1, 50: $\overline{X} =$?

2.19. The data below are a sorted residuals from a least squares regression where $n = 100$ and $p = 4$. Find shorth(97) of the residuals.

```
number      1      2      3      4    ...   97    98    99    100
residual -2.39 -2.34 -2.03 -1.77 ... 1.76 1.81 1.83  2.16
```

2.20. To find the sample median of a list of n numbers where n is odd, order the numbers from smallest to largest and the median is the middle ordered number. The sample median estimates the population median. Suppose the sample is $\{14, 3, 5, 12, 20, 10, 9\}$. Find the sample median for each of the three samples listed below.
Sample 1: 9, 10, 9, 12, 5, 14, 3
Sample 2: 3, 9, 20, 10, 9, 5, 14
Sample 3: 14, 12, 10, 20, 3, 3, 5

2.21. Suppose you are estimating the mean μ of losses with $T = \overline{X}$.
actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$,
a) Compute $T_1^*, ..., T_4^*$, where T_i^* is the sample mean of the ith sample. samples:

2, 10, 1, 2, 2:

50, 10, 50, 2, 2:

10, 50, 2, 1, 1:

5, 2, 5, 1, 50:

b) Now compute the bagging estimator which is the sample mean of the T_i^*: the bagging estimator $\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^*$ where $B = 4$ is the number of samples.

R problems Some R code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

2.22*. Use the commands

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

to simulate data similar to the Buxton heights. Paste the commands for this problem into R to make a plot similar to Figure 2.1.

2.23*. 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)
```

2.24***.** 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 by pasting the commands for this problem into *R*.

b) Compute the mean and 25% trimmed mean of 10000 simulated EXP(1) random variables by pasting the commands for this problem into *R*.

2.25. The following *R* 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 by pasting the commands for this problem into *R*.

Warning: For the following problems, use a command like *source("G:/rpack.txt")* **to download the programs. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *ratmn*, will display the code for the function. Use the `args` command, e.g. *args(ratmn)*, to display the needed arguments for the function.

2.26. Download the *R* 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 by pasting the commands for this problem into *R*.

2.27. Download the *R* 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 by pasting the commands for this problem into *R*.

2.28***.** 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 2.22. Use the command *cci(height)*.

2.29***.** a) Download the *R* 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 2.22. Use the command *medci(height)*.

2.30***.** a) Download the R 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 2.22. Use the command *tmci(height)*.

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

b) Compute a 95% CI for the artificial height data set created in Problem 2.22. Use the command *atmci(height)*.

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

b) Compute a 95% CI for the artificial height data set created in Problem 2.22. Use the command *stmci(height)*.

2.33. a) Download the R 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 2.22. Use the command *med2ci(height)*.

2.34. a) Download the R 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 2.22. Use the command *cgci(height)*.

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

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

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 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 2.22. Use the command *bg2ci(height)*.

2.36. 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 2.22 with the command *qplot(height)*.

c) Make a Q plot for $N(0, 1)$ data by pasting the commands for this problem into R.

2.37. a) Download the *R* function `rcisim` to reproduce Tables 2.6 and 2.7. Two lines need to be changed with each CI. One line is the output line that calls the CI and the other line is the parameter estimated for exponential(1) data. The default is for the classical interval. Thus the program calls the function *cci* used in Problem 2.28. The functions `medci`, `tmci`, `atmci`, `stmci`, `med2ci`, `cgci` and `bg2ci` given in Problems 2.29 – 2.35 are also interesting. The program gives the proportion of times 0 is in the classical CI. For type ii) data which has 25% outliers, this proportion will be low.

b) Enter the following commands, obtain the output and explain what the output shows.
i) rcisim(n,type=1) for n = 10, 50, 100
ii) rcisim(n,type=2) for n = 10, 50, 100
iii) rcisim(n,type=3) for n = 10, 50, 100
iv) rcisim(n,type=4) for n = 10, 50, 100
v) rcisim(n,type=5) for n = 10, 50, 100

2.38. a) Download the *R* functions `cisim` and `robci`. Download the data set `cushny`. That is, use the source command twice to download `rpack.txt` and `robdata.txt`.

b) An easier way to reproduce Tables 2.6 and 2.7 is to evaluate the six CIs on the same data. Type the command *cisim(100)* and interpret the results.

c) To compare the six CIs on the Cushny Peebles data described in Problem 2.11, type the command *robci(cushny).*