Chapter 4

Truncated Distributions

This chapter presents a simulation study of several of the confidence intervals first presented in Chapter 2. Theorem 2.2 on p. 50 shows that the (α, β) trimmed mean T_n is estimating a parameter μ_T with an asymptotic variance equal to $\sigma_W^2/(\beta - \alpha)^2$. The first five sections of this chapter provide the theory needed to compare the different confidence intervals. Many of these results will also be useful for comparing multiple linear regression estimators.

Mixture distributions are often used as outlier models. The following two definitions and proposition are useful for finding the mean and variance of a mixture distribution. Parts a) and b) of Proposition 4.1 below show that the definition of expectation given in Definition 4.2 is the same as the usual definition for expectation if Y is a discrete or continuous random variable.

Definition 4.1. The distribution of a random variable Y is a *mixture distribution* if the cdf of Y has the form

$$F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y) \tag{4.1}$$

where $0 < \alpha_i < 1$, $\sum_{i=1}^k \alpha_i = 1$, $k \geq 2$, and $F_{W_i}(y)$ is the cdf of a continuous or discrete random variable W_i, $i = 1, ..., k$.

Definition 4.2. Let Y be a random variable with cdf $F(y)$. Let h be a function such that the expected value $Eh(Y) = E[h(Y)]$ exists. Then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y). \tag{4.2}$$

Proposition 4.1. a) If Y is a discrete random variable that has a pmf $f(y)$ with support $\mathcal{Y}$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y)dF(y) = \sum_{y \in \mathcal{Y}} h(y)f(y).$$

b) If Y is a continuous random variable that has a pdf $f(y)$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y)dF(y) = \int_{-\infty}^{\infty} h(y)f(y)dy.$$

c) If Y is a random variable that has a mixture distribution with cdf $F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y)$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y)dF(y) = \sum_{i=1}^{k} \alpha_i E_{W_i}[h(W_i)]$$

where $E_{W_i}[h(W_i)] = \int_{-\infty}^{\infty} h(y)dF_{W_i}(y)$.

Example 4.1. Proposition 4.1c implies that the pmf or pdf of W_i is used to compute $E_{W_i}[h(W_i)]$. As an example, suppose the cdf of Y is $F(y) = (1-\epsilon)\Phi(y) + \epsilon\Phi(y/k)$ where $0 < \epsilon < 1$ and $\Phi(y)$ is the cdf of $W_1 \sim N(0,1)$. Then $\Phi(y/k)$ is the cdf of $W_2 \sim N(0,k^2)$. To find EY, use $h(y) = y$. Then

$$EY = (1-\epsilon)EW_1 + \epsilon EW_2 = (1-\epsilon)0 + \epsilon 0 = 0.$$

To find EY^2, use $h(y) = y^2$. Then

$$EY^2 = (1-\epsilon)EW_1^2 + \epsilon EW_2^2 = (1-\epsilon)1 + \epsilon k^2 = 1 - \epsilon + \epsilon k^2.$$

Thus $\text{VAR}(Y) = E[Y^2] - (E[Y])^2 = 1 - \epsilon + \epsilon k^2$. If $\epsilon = 0.1$ and $k = 10$, then $EY = 0$, and $\text{VAR}(Y) = 10.9$.

Remark 4.1. Warning: Mixture distributions and linear combinations of random variables are very different quantities. As an example, let

$$W = (1-\epsilon)W_1 + \epsilon W_2$$

where ϵ, W_1 and W_2 are as in the previous example and suppose that W_1 and W_2 are independent. Then W, a linear combination of W_1 and W_2, has a normal distribution with mean

$$EW = (1-\epsilon)EW_1 + \epsilon EW_2 = 0$$

and variance

$$\text{VAR}(W) = (1-\epsilon)^2\text{VAR}(W_1) + \epsilon^2\text{VAR}(W_2) = (1-\epsilon)^2 + \epsilon^2 k^2 < \text{VAR}(Y)$$

where Y is given in the example above. Moreover, W has a unimodal normal distribution while Y does not follow a normal distribution. In fact, if $X_1 \sim N(0,1)$, $X_2 \sim N(10,1)$, and X_1 and X_2 are independent, then $(X_1+X_2)/2 \sim N(5, 0.5)$; however, if Y has a mixture distribution with cdf

$$F_Y(y) = 0.5F_{X_1}(y) + 0.5F_{X_2}(y) = 0.5\Phi(y) + 0.5\Phi(y-10),$$

then the pdf of Y is bimodal.

Truncated distributions can be used to simplify the asymptotic theory of robust estimators of location and regression. Sections 4.1, 4.2, 4.3, and 4.4 will be useful when the underlying distribution is exponential, double exponential, normal, or Cauchy (see Chapter 3). Sections 4.5 and 4.6 examine how the sample median, trimmed means and two stage trimmed means behave at these distributions.

Definitions 2.17 and 2.18 defined the truncated random variable $Y_T(a,b)$ and the Winsorized random variable $Y_W(a,b)$. Let Y have cdf F and let the truncated random variable $Y_T(a,b)$ have the cdf $F_{T(a,b)}$. The following lemma illustrates the relationship between the means and variances of $Y_T(a,b)$ and $Y_W(a,b)$. Note that $Y_W(a,b)$ is a mixture of $Y_T(a,b)$ and two point masses at a and b. Let $c = \mu_T(a,b) - a$ and $d = b - \mu_T(a,b)$.

Lemma 4.2. Let $a = \mu_T(a,b) - c$ and $b = \mu_T(a,b) + d$. Then
a)

$$\mu_W(a,b) = \mu_T(a,b) - \alpha c + (1-\beta)d, \text{ and}$$

b)

$$\sigma^2_W(a,b) = (\beta-\alpha)\sigma^2_T(a,b) + (\alpha-\alpha^2)c^2$$
$$+[(1-\beta)-(1-\beta)^2]d^2 + 2\alpha(1-\beta)cd.$$

c) If $\alpha = 1-\beta$ then

$$\sigma^2_W(a,b) = (1-2\alpha)\sigma^2_T(a,b) + (\alpha-\alpha^2)(c^2+d^2) + 2\alpha^2 cd.$$

d) If $c = d$ then

$$\sigma^2_W(a,b) = (\beta-\alpha)\sigma^2_T(a,b) + [\alpha-\alpha^2+1-\beta-(1-\beta)^2+2\alpha(1-\beta)]d^2.$$

e) If $\alpha = 1 - \beta$ and $c = d$, then $\mu_W(a,b) = \mu_T(a,b)$ and

$$\sigma_W^2(a,b) = (1-2\alpha)\sigma_T^2(a,b) + 2\alpha d^2.$$

Proof. We will prove b) since its proof contains the most algebra. Now

$$\sigma_W^2 = \alpha(\mu_T - c)^2 + (\beta - \alpha)(\sigma_T^2 + \mu_T^2) + (1-\beta)(\mu_T + d)^2 - \mu_W^2.$$

Collecting terms shows that

$$\sigma_W^2 = (\beta-\alpha)\sigma_T^2 + (\beta - \alpha + \alpha + 1 - \beta)\mu_T^2 + 2[(1-\beta)d - \alpha c]\mu_T$$
$$+\alpha c^2 + (1-\beta)d^2 - \mu_W^2.$$

From a),

$$\mu_W^2 = \mu_T^2 + 2[(1-\beta)d - \alpha c]\mu_T + \alpha^2 c^2 + (1-\beta)^2 d^2 - 2\alpha(1-\beta)cd,$$

and we find that

$$\sigma_W^2 = (\beta-\alpha)\sigma_T^2 + (\alpha - \alpha^2)c^2 + [(1-\beta) - (1-\beta)^2]d^2 + 2\alpha(1-\beta)cd. \quad QED$$

4.1 The Truncated Exponential Distribution

Let Y be a (one sided) truncated exponential $TEXP(\lambda, b)$ random variable. Then the pdf of Y is

$$f_Y(y|\lambda, b) = \frac{\frac{1}{\lambda}e^{-y/\lambda}}{1 - \exp(-\frac{b}{\lambda})}$$

for $0 < y \le b$ where $\lambda > 0$. Let $b = k\lambda$, and let

$$c_k = \int_0^{k\lambda} \frac{1}{\lambda} e^{-y/\lambda} dx = 1 - e^{-k}.$$

Next we will find the first two moments of $Y \sim TEXP(\lambda, b = k\lambda)$ for $k > 0$.

Lemma 4.3. If Y is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, then

$$a)\ E(Y) = \lambda \left[\frac{1-(k+1)e^{-k}}{1-e^{-k}}\right],$$

and

$$b)\ E(Y^2) = 2\lambda^2 \left[\frac{1 - \frac{1}{2}(k^2 + 2k + 2)e^{-k}}{1 - e^{-k}} \right].$$

See Problem 4.9 for a related result.

Proof. a) Note that

$$c_k E(Y) = \int_0^{k\lambda} \frac{y}{\lambda} e^{-y/\lambda} dy$$

$$= -ye^{-y/\lambda}|_0^{k\lambda} + \int_0^{k\lambda} e^{-y/\lambda} dy$$

(use integration by parts). So $c_k E(Y) =$

$$-k\lambda e^{-k} + (-\lambda e^{-y/\lambda})|_0^{k\lambda}$$

$$= -k\lambda e^{-k} + \lambda(1 - e^{-k}).$$

Hence

$$E(Y) = \lambda \left[\frac{1 - (k+1)e^{-k}}{1 - e^{-k}} \right].$$

b) Note that

$$c_k E(Y^2) = \int_0^{k\lambda} \frac{y^2}{\lambda} e^{-y/\lambda} dy.$$

Since

$$\frac{d}{dy}[-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}]$$

$$= \frac{1}{\lambda} e^{-y/\lambda}(y^2 + 2\lambda y + 2\lambda^2) - e^{-y/\lambda}(2y + 2\lambda)$$

$$= y^2 \frac{1}{\lambda} e^{-y/\lambda},$$

we have $c_k E(Y^2) =$

$$[-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}]_0^{k\lambda}$$

$$= -(k^2\lambda^2 + 2\lambda^2 k + 2\lambda^2)e^{-k} + 2\lambda^2.$$

So the result follows. QED

Since as $k \to \infty$, $E(Y) \to \lambda$, and $E(Y^2) \to 2\lambda^2$, we have $\text{VAR}(Y) \to \lambda^2$. If $k = 9\log(2) \approx 6.24$, then $E(Y) \approx .998\lambda$, and $E(Y^2) \approx 0.95(2\lambda^2)$.

4.2 The Truncated Double Exponential Distribution

Suppose that X is a double exponential $DE(\mu, \lambda)$ random variable. Chapter 3 states that MED$(X) = \mu$ and MAD$(X) = \log(2)\lambda$. Let $c = k\log(2)$, and let the truncation points $a = \mu - k\text{MAD}(X) = \mu - c\lambda$ and $b = \mu + kMAD(X) = \mu + c\lambda$. Let $X_T(a, b) \equiv Y$ be the truncated double exponential $TDE(\mu, \lambda, a, b)$ random variable. Then the pdf of Y is

$$f_Y(y|\mu, \lambda, a, b) = \frac{1}{2\lambda(1 - \exp(-c))} \exp(-|y - \mu|/\lambda)$$

for $a \le y \le b$.

Lemma 4.4. a) $E(Y) = \mu$.

$$b)\ \text{VAR}(Y) = 2\lambda^2 \left[\frac{1 - \frac{1}{2}(c^2 + 2c + 2)e^{-c}}{1 - e^{-c}}\right].$$

Proof. a) follows by symmetry and b) follows from Lemma 4.3 b) since VAR$(Y) = E[(Y - \mu)^2] = E(W_T^2)$ where W_T is $TEXP(\lambda, b = c\lambda)$. QED

As $c \to \infty$, VAR$(Y) \to 2\lambda^2$. If $k = 9$, then $c = 9\log(2) \approx 6.24$ and VAR$(Y) \approx 0.95(2\lambda^2)$.

4.3 The Truncated Normal Distribution

Now if X is $N(\mu, \sigma^2)$ then let Y be a truncated normal $TN(\mu, \sigma^2, a, b)$ random variable. Then $f_Y(y) =$

$$\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} I_{[a,b]}(y)$$

where Φ is the standard normal cdf. The indicator function

$$I_{[a,b]}(y) = 1 \ \text{ if } \ \text{a} \le \text{y} \le \text{b}$$

and is zero otherwise. Let ϕ be the standard normal pdf.

Lemma 4.5.

$$E(Y) = \mu + \left[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right] \sigma,$$

and VAR$(Y) =$

$$\sigma^2 \left[1 + \frac{(\frac{a-\mu}{\sigma})\phi(\frac{a-\mu}{\sigma}) - (\frac{b-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right] \; - \; \sigma^2 \left[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right]^2 .$$

(See Johnson and Kotz 1970a, p. 83.)

Proof. Let $c =$

$$\frac{1}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}.$$

Then

$$E(Y) = \int_a^b y f_Y(y) dy.$$

Hence

$$\frac{1}{c} E(Y) = \int_a^b \frac{y}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy \; +$$

$$\frac{\mu}{\sigma} \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$+\mu \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy.$$

Note that the integrand of the last integral is the pdf of a $N(\mu, \sigma^2)$ distribution. Let $z = (y - \mu)/\sigma$. Thus $dz = dy/\sigma$, and $E(Y)/c =$

$$\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z}{\sqrt{2\pi}} e^{-z^2/2} dz + \frac{\mu}{c}$$

$$= \frac{\sigma}{\sqrt{2\pi}} (-e^{-z^2/2}) \Big|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \frac{\mu}{c}.$$

Multiplying both sides by c gives the expectation result.

$$E(Y^2) = \int_a^b y^2 f_Y(y) dy.$$

Hence

$$\frac{1}{c} E(Y^2) = \int_a^b \frac{y^2}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \sigma \int_a^b \left(\frac{y^2}{\sigma^2} - \frac{2\mu y}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$+ \sigma \int_a^b \frac{2y\mu - \mu^2}{\sigma^2} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \sigma \int_a^b \left(\frac{y-\mu}{\sigma}\right)^2 \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy + 2\frac{\mu}{c} E(Y) - \frac{\mu^2}{c}.$$

Let $z = (y-\mu)/\sigma$. Then $dz = dy/\sigma$, $dy = \sigma dz$, and $y = \sigma z + \mu$. Hence $E(Y^2)/c =$

$$2\frac{\mu}{c} E(Y) - \frac{\mu^2}{c} + \sigma \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Next integrate by parts with $w = z$ and $dv = ze^{-z^2/2}dz$. Then $E(Y^2)/c =$

$$2\frac{\mu}{c} E(Y) - \frac{\mu^2}{c} +$$

$$\frac{\sigma^2}{\sqrt{2\pi}} [(-ze^{-z^2/2})|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-z^2/2} dz]$$

$$= 2\frac{\mu}{c} E(Y) - \frac{\mu^2}{c} + \sigma^2 \left[\left(\frac{a-\mu}{\sigma}\right)\phi\left(\frac{a-\mu}{\sigma}\right) - \left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right) + \frac{1}{c} \right].$$

Using

$$\text{VAR}(Y) = c\frac{1}{c} E(Y^2) - (E(Y))^2$$

gives the result. QED

Corollary 4.6. Let Y be $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$. Then $E(Y) = \mu$ and VAR$(Y) =$

$$\sigma^2 \left[1 - \frac{2k\phi(k)}{2\Phi(k) - 1} \right].$$

Table 4.1: Variances for Several Truncated Normal Distributions

k	VAR(Y)
2.0	$0.774\sigma^2$
2.5	$0.911\sigma^2$
3.0	$0.973\sigma^2$
3.5	$0.994\sigma^2$
4.0	$0.999\sigma^2$

Proof. Use the symmetry of ϕ, the fact that $\Phi(-x) = 1 - \Phi(x)$, and the above lemma to get the result. QED

Examining VAR(Y) for several values of k shows that the $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$ distribution does not change much for $k > 3.0$. See Table 4.1.

4.4 The Truncated Cauchy Distribution

If X is a Cauchy $C(\mu, \sigma)$ random variable, then MED(X) $= \mu$ and MAD(X) $= \sigma$. If Y is a truncated Cauchy $TC(\mu, \sigma, \mu - a\sigma, \mu + b\sigma)$ random variable, then

$$f_Y(y) = \frac{1}{\tan^{-1}(b) + \tan^{-1}(a)} \; \frac{1}{\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

for $\mu - a\sigma < y < \mu + b\sigma$. Moreover,

$$E(Y) = \mu + \sigma \left(\frac{\log(1 + b^2) - \log(1 + a^2)}{2[\tan^{-1}(b) + \tan^{-1}(a)]} \right),$$

and VAR(Y) =

$$\sigma^2 \left[\frac{b + a - \tan^{-1}(b) - \tan^{-1}(a)}{\tan^{-1}(b) + \tan^{-1}(a)} - \left(\frac{\log(1 + b^2) - \log(1 + a^2)}{\tan^{-1}(b) + \tan^{-1}(a)} \right)^2 \right].$$

Lemma 4.7. If $a = b$, then $E(Y) = \mu$, and

$$\text{VAR}(Y) = \sigma^2 \left[\frac{b - \tan^{-1}(b)}{\tan^{-1}(b)} \right].$$

See Johnson and Kotz (1970a, p. 162) and Dahiya, Staneski and Chaganty (2001).

4.5 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,

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

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

Definition 4.3. 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 4.2. 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 4.2. 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 4.3. 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

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

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

Definition 4.4. 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 4.4. 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

$$\mu_D(F_X) \equiv \mu_D[F_Y(\frac{y-\alpha}{\beta})] = \alpha + \beta \mu_D[F_Y(y)]. \tag{4.3}$$

Let $D_n(\boldsymbol{Y}) \equiv D_n(Y_1, ..., Y_n)$. If

$$\sqrt{n}[D_n(\boldsymbol{Y}) - \mu_D(F_Y(y))] \xrightarrow{D} N(0, \sigma^2_D),$$

then

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

This result is especially useful when F is a cdf from a location–scale family with parameters μ and σ. In this case, Equation (4.3) 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).

Recall the following facts about two stage trimmed means from Chapter 2. Let $a = \text{MED}(Y) - k_1\text{MAD}(Y)$ and $b = \text{MED}(Y) + k_2\text{MAD}(Y)$ where MED(Y) and MAD(Y) are the population median and median absolute deviation respectively. Usually we will take $k_1 = k_2 = k$. Assume that the underlying cdf F is continuous. Let $\alpha = F(a)$ 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 $\beta_o = F(b_o)$. Let $L(M_n)$ count the number of $Y_i < \hat{a} = \text{MED}(n) - k_1\text{MAD}(n)$ and let $n - U(M_n)$ count the number of $Y_i > \hat{b} = \text{MED}(n) + k_2\text{MAD}(n)$. 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 asymmetrically trimmed $T_{A,n}$ is asymptotically equivalent to the $(\alpha_o, 1 - \beta_o)$ trimmed mean and the symmetrically trimmed mean $T_{S,n}$ is asymptotically equivalent to the $\max(\alpha_o, 1 - \beta_o)$ 100% trimmed mean. Then from Corollary 2.5,

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

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 $b = \text{MED}(Y) + k\text{MAD}(Y)$.

Lemma 4.8. 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,

$$n\ AV = \left(\frac{1 - \frac{2z\phi(z)}{2\Phi(z)-1}}{1 - 2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \right) \sigma^2 \qquad (4.4)$$

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 (4.4).

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 Corollary 4.6,

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

Use Remark 4.3 with $z = k\Phi^{-1}(0.75)$, and $\alpha = \Phi(-z)$ to get Equation (4.4).

Lemma 4.9. Suppose that Y comes from a double exponential DE(0,1) distribution. Then for the α trimmed mean,

$$n\ AV = \frac{\frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \tag{4.5}$$

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 4.3 with

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

Lemma 4.10. Suppose that Y comes from a Cauchy (0,1) distribution. Then for the α trimmed mean,

$$n\ AV = \frac{z - \tan^{-1}(z)}{(1-2\alpha)\tan^{-1}(z)} + \frac{2\alpha(\tan[\pi(\alpha - \frac{1}{2})])^2}{(1-2\alpha)^2} \tag{4.6}$$

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, MAD$(Y) = 1$. If the C(0,1) distribution is truncated at $-z$ and z, then use Remark 4.3 with

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

4.6 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 MAD(Y)) that are otherwise hard to compute and to check whether software is running correctly.

Example 4.3. The *R/Splus* 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 4.4. 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 4.5. 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/Splus* 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 4.5. 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

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

Table 4.2: 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

so

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

Example 4.6. 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 (4.4). 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.

Example 4.7. 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 (4.5). 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 4.2. 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.

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. Six different CIs

$$D_n \pm t_{d,.975} SE(D_n)$$

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 2.16 on p. 45.
(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 2.15.
(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 2.14.
(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 Bloch and Gastwirth (1968), $SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)})$. See Application 2.2.
(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 $U_n = n - L_n$.

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

Table 4.3: 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

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 =

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

Table 4.4: 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

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.

Tables 4.3 and 4.4 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 4.3, 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 4.4 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.

4.7 Complements

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, plausible 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 plausible 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 (4.4) and (4.5). Gross (1976), Guenther (1969) and Lax (1985) are useful references for confidence intervals. Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972) is a well known simulation study for robust location estimators.

In Section 4.6, 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 2.2 and the column "MED" in Tables 4.3 and 4.4) 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.

The estimators $T_{S,n}$ and $T_{A,n}$ depend on a parameter k. Smaller values of k should have smaller CI lengths if the data has heavy tails while larger values of k should perform better for light tailed distributions. In simulations, $T_{S,n}$ performed well for $k > 1$, but the variability of $T_{A,n}$ was too large for $n \leq 100$ for Gaussian data if $1 < k < 5$. These estimators also depend on the grid C of trimming proportions. Using $C = \{0, 0.01, 0.02, ..., 0.49, 0.5\}$ makes the estimators easy to compute, but $T_{S,n}$ will perform better if the much coarser grid $C_c = \{0, 0.01, 0.10, 0.25, 0.40, 0.49, 0.5\}$ is used. The performance does not change much for symmetric data, but can improve considerably if the data is skewed. The estimator can still perform rather poorly if the data is asymmetric and the trimming proportion of the metrically trimmed mean is near one of these allowed trimming proportions. For example if $k = 3.5$ and the data is exponential(1), the metrically trimmed mean trims approximately 9.3% of the cases. Hence the $T_{S,n}$ is often the 25% and the 10% trimmed mean for small n. When $k = 4.5$, $T_{S,n}$ with grid C_c is usually the 10% trimmed mean and hence performs well on exponential(1) data.

$T_{A,n}$ is the estimator most like high breakdown M–estimators proposed in the literature. These estimators basically use a random amount of trimming and work well on symmetric data. Estimators that give zero weight to distant outliers ("hard rejection") can work well on "contaminated normal" populations such as $(1 - \epsilon)N(0, 1) + \epsilon N(\mu_s, 1)$. Of course $\epsilon \in (0, 0.5)$ and μ_s can always be chosen so that these estimators perform poorly. Stigler (1977) argues that complicated robust estimators are not needed.

4.8 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

4.1*. Suppose the random variable X has cdf $F_X(x) = 0.9\ \Phi(x - 10) + 0.1\ F_W(x)$ where $\Phi(x - 10)$ is the cdf of a normal $N(10, 1)$ random variable with mean 10 and variance 1 and $F_W(x)$ is the cdf of the random variable W that satisfies $P(W = 200) = 1$.
a) Find $E(W)$.
b) Find $E(X)$.

4.2. Suppose the random variable X has cdf $F_X(x) = 0.9\ F_Z(x) + 0.1\ F_W(x)$ where F_Z is the cdf of a gamma($\nu = 10, \lambda = 1$) random variable with mean 10 and variance 10 and $F_W(x)$ is the cdf of the random variable W that satisfies $P(W = 400) = 1$.
a) Find $E(W)$.
b) Find $E(X)$.

4.3. a) Prove Lemma 4.2 a).

b) Prove Lemma 4.2 c).
c) Prove Lemma 4.2 d).
d) Prove Lemma 4.2 e).

4.4. Suppose that F is the cdf from a distribution that is symmetric about 0. Suppose $a = -b$ and $\alpha = F(a) = 1 - \beta = 1 - F(b)$. Show that

$$\frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2} = \frac{\sigma^2_T(a,b)}{1-2\alpha} + \frac{2\alpha(F^{-1}(\alpha))^2}{(1-2\alpha)^2}.$$

4.5. Recall that $L(M_n) = \sum_{i=1}^n I[Y_i < \text{MED}(n) - k \ \text{MAD}(n)]$ and $n - U(M_n) = \sum_{i=1}^n I[Y_i > \text{MED}(n) + k \ \text{MAD}(n)]$ where the *indicator variable* $I(A) = 1$ if event A occurs and is zero otherwise. Show that $T_{S,n}$ is a randomly trimmed mean. (Hint: round

$$100 \max[L(M_n), n - U(M_n)]/n$$

up to the nearest integer, say J_n. Then $T_{S,n}$ is the J_n% trimmed mean with $L_n = \lfloor (J_n/100) \ n \rfloor$ and $U_n = n - L_n$.)

4.6. Show that $T_{A,n}$ is a randomly trimmed mean. (Hint: To get L_n, round $100L(M_n)/n$ up to the nearest integer J_n. Then $L_n = \lfloor (J_n/100) \ n \rfloor$. Round $100[n - U(M_n)]/n$ up to the nearest integer K_n. Then $U_n = \lfloor (100 - K_n)n/100 \rfloor$.)

4.7*. Let F be the $N(0,1)$ cdf. Show that the ARE of the sample median MED(n) with respect to the sample mean $\overline{Y}_n$ is $ARE \approx 0.64$.

4.8*. Let F be the $DE(0,1)$ cdf. Show that the ARE of the sample median MED(n) with respect to the sample mean $\overline{Y}_n$ is $ARE \approx 2.0$.

4.9. If Y is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, show that a)

$$E(Y) = \lambda \left[1 - \frac{k}{e^k - 1} \right].$$

b)

$$E(Y^2) = 2\lambda^2 \left[1 - \frac{(0.5k^2 + k)}{e^k - 1} \right].$$

R/Splus problems

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

4.10. a) Download the *R/Splus* function `nav` that computes Equation (4.4) from Lemma 4.8.

b) Find the asymptotic variance of the α trimmed mean for $\alpha = 0.01, 0.1$, 0.25 and 0.49.

c) Find the asymptotic variance of $T_{A,n}$ for $k = 2, 3, 4, 5$ and 6.

4.11. a) Download the *R/Splus* function `deav` that computes Equation (4.5) from Lemma 4.9.

b) Find the asymptotic variance of the α trimmed mean for $\alpha = 0.01, 0.1$, 0.25 and 0.49.

c) Find the asymptotic variance of $T_{A,n}$ for $k = 2, 3, 4, 5$ and 6.

4.12. a) Download the *R/Splus* function `cav` that finds n AV for the Cauchy(0,1) distribution.

b) Find the asymptotic variance of the α trimmed mean for $\alpha = 0.01, 0.1$, 0.25 and 0.49.

c) Find the asymptotic variance of $T_{A,n}$ for $k = 2, 3, 4, 5$ and 6.

4.13. a) Download the *R/Splus* function `rcisim` to reproduce Tables 4.3 and 4.4. 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 2.21. The functions `medci`, `tmci`, `atmci`, `stmci`, `med2ci`, `cgci` and `bg2ci` given in Problems 2.22 – 2.28 are also interesting.

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

4.14. a) Download the *R/Splus* 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 4.3 and 4.4 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).*