Chapter 11
Appendix

11.1 Tips for Doing Research

As a student or new researcher, you will probably encounter researchers who think that their method of doing research is the only correct way of doing research, but there are dozens of methods that have proven effective.

Familiarity with the literature is important since your research should be original. This text and Olive (2017ab,2020) present much of the author's applied research in the fields of regression and high breakdown robust statistics from 1990–2020. Several other important contributions follow. Gnanadesikan and Kettenring (1972) suggested an algorithm similar to concentration. Hampel (1975) introduced the least median of squares estimator. The LTA estimator was an interesting extension. Devlin, Gnanadesikan, and Kettenring (1975, 1981) introduced the concentration technique. Siegel (1982) suggested using elemental sets to find robust regression estimators. Rousseeuw (1984) popularized LMS and extended the LTS/MCD location estimator to the LTS regression estimator and the MCD estimator of multivariate location and dispersion. Ruppert (1992) used concentration for resistant regression. Cook and Nachtsheim (1994) showed that robust Mahalanobis distances could be used to reduce the bias of 1D regression estimators. Rousseeuw and Van Driessen (1999) introduced the DD plot.

Beginners can have a hard time determining whether a robust algorithm estimator is consistent or not. As a rule of thumb, assume that the approximations (including those for depth, LTA, LMS, LTS, MCD, MVE, S, projection estimators and two stage estimators) are inconsistent unless the authors show that they understand this text, Hawkins and Olive (2002), and Olive (2008, 2017b). In particular, the elemental or basic resampling algorithms, concentration algorithms, and algorithms based on random projections should be considered inconsistent until you can prove otherwise.

After finding a research topic, **paper trailing** is an important technique for finding related literature. To use this technique, find a paper on the topic,

go to the bibliography of the paper, find one or more related papers and repeat. Often your university's library will have useful internet resources for finding literature. Often a research university will subscribe to either *The Web of Knowledge* with a link to ISI Web of Science or to the *Current Index to Statistics*. Both of these resources allow you to search for literature by author, e.g. Olive, or by topic, e.g. robust statistics. Both of these methods search for recent papers. With Web of Knowledge, find an article with *Search*, click on the article and then click on the *view related reference* icon to get a list of related articles. The Google search engine and "Google Scholar" are also useful. When searching, enter a topic and the word *robust* or *outliers*. For example, enter the keywords *robust factor analysis* or *factor analysis and outliers*. Statistical journals often have websites that make abstracts and preprints available.

Finally, a Ph.D. student needs an advisor or **mentor** and most researchers will find collaboration valuable. Attending conferences and making your research available over the internet can lead to contacts.

Some references on research, including technical writing and presentations, include American Society of Civil Engineers (1950), Becker and Keller-McNulty (1996), Ehrenberg (1982), Freeman, Gonzalez, Hoaglin and Kilss (1983), Hamada and Sitter (2004), Rubin (2004), and Smith (1997).

11.2 R

R is available from the **CRAN** website (https://cran.r-project.org/). As of August 2020, the author's personal computer has Version 3.3.1 (June 21, 2016) of *R*. The *R* software is similar to *Splus*, but is free. *R* is very versatile since many people have contributed useful code, often as packages. A useful *R* link is (www.r-project.org/#doc).

Many of the homework problems use *R* functions contained in the book's website (http://parker.ad.siu.edu/Olive/robbook.htm) under the file name *rpack.txt*. The following two *R* commands can be copied and pasted into *R* from near the top of the file (http://parker.ad.siu.edu/Olive/robRhw.txt).

Downloading the book's R functions *rpack.txt* and *R* data sets *robdata.txt* into *R*: The commands

```
source("http://parker.ad.siu.edu/Olive/rpack.txt")
source("http://parker.ad.siu.edu/Olive/robdata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Nearly 110 *R* functions from *rpack* should appear. In *R*, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* to remove the functions from the computer (clicking on *Yes* saves the functions on *R*, but the functions and data are easily obtained with the source commands).

For Windows, the functions can be saved on a flash drive G, say. Then use the following command.

```
source("G:/rpack.txt")
```

This section gives tips on using *R*, but is no replacement for books such as Becker et al. (1988), Crawley (2005, 2013), Fox and Weisberg (2011), or Venables and Ripley (2010). Also see Mathsoft (1999ab) and use the website (www.google.com) to search for useful websites. For example enter the search words *R documentation*.

The command *q()* gets you out of *R*.

Least squares regression is done with the function *lsfit* or *lm*.

The commands *help(fn)* and *args(fn)* give information about the function fn, e.g. if fn = lsfit.

Type the following commands.

```
x <- matrix(rnorm(300),nrow=100,ncol=3)
y <- x%*%1:3 + rnorm(100)
out<- lsfit(x,y)
out$coef
ls.print(out)
```

The first line makes a 100 by 3 matrix x with N(0,1) entries. The second line makes $y[i] = 0 + 1 * x[i,1] + 2 * x[i,2] + 3 * x[i,2] + e$ where e is N(0,1). The term 1:3 creates the vector $(1,2,3)^T$ and the matrix multiplication operator is $\%*\%$. The function `lsfit` will automatically add the constant to the model. Typing "out" will give you a lot of irrelevant information, but *out$coef* and *out$resid* give the OLS coefficients and residuals respectively.

To make a residual plot, type the following commands.

```
fit <- y - out$resid
plot(fit,out$resid)
title("residual plot")
```

The first term in the plot command is always the horizontal axis while the second is on the vertical axis.

To put a graph in *Word,* hold down the *Ctrl* and *c* buttons simultaneously. Then select "paste" from the *Word* Edit menu, or hit *Ctrl* and *v* at the same time.

To enter data, open a data set in *Notepad* or *Word*. You need to know the number of rows and the number of columns. Assume that each case is entered in a row. For example, assuming that the file *cyp.lsp* has been saved on your flash drive from the webpage for this book, open *cyp.lsp* in *Word.* It has 76 rows and 8 columns. In *R* , write the following command.

```
cyp <- matrix(scan(),nrow=76,ncol=8,byrow=T)
```

A data frame is a two-dimensional array in which the values of different variables are stored in different named columns.

Then copy the data lines from *Word* and paste them in *R*. If a cursor does not appear, hit *enter*. The command *dim(cyp)* will show if you have entered the data correctly.

Enter the following commands

```
cypy <- cyp[,2]
cypx<- cyp[,-c(1,2)]
lsfit(cypx,cypy)$coef
```

to produce the output below.

```
   Intercept            X1            X2            X3
205.40825985   0.94653718   0.17514405   0.23415181
         X4            X5            X6
0.75927197   -0.05318671  -0.30944144
```

Making functions in R is easy.

For example, type the following commands.

```
mysquare <- function(x){
# this function squares x
r <- x^2
r }
```

The second line in the function shows how to put comments into functions.

Modifying your function is easy.

Store a function as text file, modify the function in *Notepad*, and copy and paste the function into *R*.

To save data or a function in *R*, when you exit, click on *Yes* when the "*Save worksheet image?*" window appears. When you reenter *R*, type *ls()*. This will show you what is saved. You should rarely need to save anything for this book. To remove unwanted items from the worksheet, e.g. *x*, type *rm(x)*, *pairs(x)* makes a scatterplot matrix of the columns of *x*, *hist(y)* makes a histogram of *y*, *boxplot(y)* makes a boxplot of *y*, *stem(y)* makes a stem and leaf plot of y, *scan()*, *source()*, and *sink()* can be useful. To type a simple list, use *y <− c(1,2,3.5)*. The commands *mean(y)*, *median(y)*, *var(y)* are self explanatory.

The following commands are useful for a scatterplot created by the command *plot(x,y)*.
lines(x,y), *lines(lowess(x,y,f=.2))*,
identify(x,y),
abline(out$coef), *abline(0,1)*

The usual arithmetic operators are $2+4$, $3-7$, $8*4$, $8/4$, and

```
2^10, 2^(10) or 2^{10}.
```

The ith element of vector y is $y[i]$ while the ij element of matrix x is $x[i, j]$. The second row of x is $x[2,]$ while the 4th column of x is $x[, 4]$. The transpose of x is *t(x)*.

The command *apply(x,1,fn)* will compute the row means if fn = mean. The command *apply(x,2,fn)* will compute the column variances if fn = var. The commands *cbind* and *rbind* combine column vectors or row vectors with an existing matrix or vector of the appropriate dimension.

Citing packages

We will use R packages often in this book. The following R command is useful for citing the Venables and Ripley (2010) `MASS` package.

```
citation("MASS")
```

Other packages cited in this book include `glmnet`: Friedman et al. (2015), `leaps`: Lumley (2009), and `robustbase`: Rousseeuw et al. (2016).

Getting information about a library in R

In R, a *library* is a built in package or add–on package of R code. The command `library()` shows the available packages and libraries, and information about a specific library, such as `MASS` for robust estimators like `cov.mcd` or `ts` for time series estimation, can be found, e.g., with the command *library(help=MASS)*.

Downloading a library into R

Many researchers have contributed a *library* or *package* of R code that can be downloaded for use. To see what is available, go to the website (http://cran.us.r-project.org/) and click on the Packages icon.

Following Crawley (2013, p. 8), you may need to "Run as administrator" before you can install packages (right click on the R icon to find this). Then use the following command to install the *glmnet* package.

```
install.packages("glmnet")
```

Open R and type the following command.

library(glmnet)

Next type *help(glmnet)* to make sure that the library is available for use.

Warning: R is free but not fool proof. If you have an old version of R and want to download a library, you may need to update your version of R. The libraries for robust statistics may be useful for outlier detection, but the methods have not been shown to be consistent or high breakdown. All software has some bugs. For example, Version 1.1.1 (August 15, 2000) of R had a random generator for the Poisson distribution that produced variates with too small of a mean θ for $\theta \geq 10$. Hence simulated 95% confidence intervals might contain θ 0% of the time. This bug seems to have been fixed in Versions 2.4.1 and later. Also, some functions in *rpack* may no longer work in new versions of R.

11.3 Projects

Straightforward Projects

1) Run a *rpack* simulation function for a range of values of n, p, error distributions, estimators, et cetera. Functions problem pairs include (*rcisim*, 2.37), (*cisim*, 2.38), (*pisim*, 5.21), (*rcovsim*, 10.14), (*ddsim*, 11.2) and (*corrsim*, 11.3). Also see the *rpack* functions *concsim*, *corrsim2*, *covesim*,*covsim2*, *ddsim*, *ddsim3*, *drsim5*, *drsim6*, *drsim7*, *fysim*, *hbregsim*, *locsim*, *lpisim*, *mbsim*, *mldsim*, *mldsim6*, *pisim3*, *pisim4*, *pisim5*, *predsim* and *prsim*. For example, `lpisim` can be used to simulate the asymptotically optimal PI for the location model, while Remark 3.3 estimates the percentage of outliers that the FMCD algorithm can tolerate. Near the beginning of Section 3.8, data is generated such that the FMCD estimator works well for $p = 4$ but fails for $p = 8$. Generate similar data sets for $p = 8, 9, 10, 12, 15, 20, 25, 30, 35, 40,$ 45, and 50. For each value of p find the smallest integer valued percentage of outliers needed to cause the FMCD and FCH estimators to fail. Use the `rpack` function `concsim`. If `concsim` is too slow for large p, use `covsim2` which will only give counts for the fast FCH estimator. As a criterion, a count ≥ 16 is good. Compare these observed FMCD percentages with Remark 3.3 (use the `gamper2` function). Do not forget the *library(MASS)* command if you use *R*.

2) Run a *mpack* simulation function described in Olive (2017b).

3) Are robust estimators needed for multiple linear regression? Examine whether using the OLS response plot is as effective as robust methods for detecting outliers. See Park, Kim, and Kim (2012).

4) Find some benchmark multiple linear regression outlier data sets such as those used by Park, Kim, and Kim (2012). Fit OLS, L_1 and M-estimators from *R*. Are any of the M-estimators as good as L_1?

5) Find some large data sets or data sets with $p > n$ and try to detect outliers using $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|$, the Euclidean distance of $\boldsymbol{x}_i$ from the coordinatewise median MED($\boldsymbol{W}$).

6) DD plots: compare, for example, classical–RFCH vs classical–cov.mcd DD plots on real and simulated data. Do problems 10.15, 11.2 and 11.3 but with a wider variety of data sets, n, p and gamma.

7) Resistant regression: use `tvreg` to compare the OLS–covfch combination with the OLS–cov.mcd combination. (L_1–cov.mcd and L_1–covfch are also interesting.) The `tvreg` and `covfch` functions are in *rpack.txt*.

8) *Using ESP to Search for the Missing Link*: Compare `trimmed views` which uses OLS and `FCH` with another regression–MLD combo. There are several possible projects: i) OLS–RFCH, ii) OLS-RMVN, iii) OLS-`cov.mcd`, iv) OLS–Classical (use `ctrviews`), v) SIR–cov.mcd (`sirviews`), vi) SIR–FCH, vii) SIR–classical, viii) lmsreg–cov.mcd (`lmsviews`), ix) lmsreg–FCH, x) lmsreg–RFCH, xi) lmsreg–RMVN ,and xii) lmsreg–classical. Do Problem 12.7ac (but just copy and paste the best view instead of using the essp(nx,ncuby,M=40) command) with both your estimator and the OLS-

FCH `trimmed views`. Try to see what types of functions work for both estimators, when OLS-FCH `trimmed views` is better and when the procedure i)–xii) is better. If you can invent interesting 1D functions, do so. See Problem 12.8.

9) Many 1D regression models where Y_i is independent of $\boldsymbol{x}_i$ given the sufficient predictor $\boldsymbol{x}_i^T\boldsymbol{\beta}$ can be made resistant by making response plots of the estimated sufficient predictor $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus Y_i for the 10 trimming proportions. Since 1D regression is the study of the conditional distribution of Y_i given $\boldsymbol{x}_i^T\boldsymbol{\beta}$, the response plot is used to visualize this distribution and needs to be made anyway. See how well trimmed views work when outliers are present.

11.4 Some Useful Distributions

The distributions in this section are discussed in much greater detail in Olive (2014, ch. 10). Also see Olive (1998). The two stage trimmed means of Chapter 2 are asymptotically equivalent to a classical trimmed mean provided that $A_n = \text{MED}(n) - k_1\text{MAD}(n) \xrightarrow{D} a$, $B_n = \text{MED}(n) + k_2\text{MAD}(n) \xrightarrow{D} b$ and if $100F(a-)$ and $100F(b)$ are not integers. This result will also hold if k_1 and k_2 depend on n. For example take $k_1 = k_2 = c_1 + c_2/n$. Then $\text{MED}(n) \pm k_1\text{MAD}(n) \xrightarrow{D} \text{MED}(Y) \pm c_1\text{MAD}(Y)$. A *trimming rule* suggests values for c_1 and c_2 and depends on the distribution of Y. Sometimes the rule is obtained by transforming the random variable Y into another random variable W (e.g. transform a lognormal into a normal) and then using the rule for W. These rules may not be as resistant to outliers as rules that do not use a transformation. For example, an observation which does not seem to be an outlier on the log scale may appear as an outlier on the original scale.

Several of the trimming rules in this section have been tailored so that the probability is high that none of the observations are trimmed when the sample size is moderate. Robust (but perhaps ad hoc) analogs of classical procedures can be obtained by applying the classical procedure to the data that remains after trimming.

Relationships between the distribution's parameters and $\text{MED}(Y)$ and $\text{MAD}(Y)$ are emphasized. Note that for location–scale families, highly outlier resistant estimates for the two parameters can be obtained by replacing $\text{MED}(Y)$ by $\text{MED}(n)$ and $\text{MAD}(Y)$ by $\text{MAD}(n)$.

Definition 11.1. The *indicator function* $I_A(x) \equiv I(x \in A) = 1$ if $x \in A$ and 0, otherwise. Sometimes an indicator function such as $I_{(0,\infty)}(y)$ will be denoted by $I(y > 0)$.

11.4.1 The Binomial Distribution

If Y has a binomial distribution, $Y \sim \text{BIN}(\text{k}, \rho)$, then the probability mass function (pmf) of Y is

$$P(Y = y) = \binom{k}{y} \rho^y (1-\rho)^{k-y}$$

for $0 < \rho < 1$ and $y = 0, 1, \ldots, k$.

The following normal approximation is often used.

$$Y \approx N(k\rho, k\rho(1-\rho))$$

when $k\rho(1-\rho) > 9$. Hence

$$P(Y \leq y) \approx \Phi\left(\frac{y + 0.5 - k\rho}{\sqrt{k\rho(1-\rho)}}\right).$$

This normal approximation suggests that $\text{MED}(Y) \approx k\rho$, and $\text{MAD}(Y) \approx 0.6745\sqrt{k\rho(1-\rho)}$. Hamza (1995) states that $|E(Y) - \text{MED}(Y)| \leq \max(\rho, 1-\rho)$ and shows that

$$|E(Y) - \text{MED}(Y)| \leq \log(2).$$

11.4.2 The Burr Type XII Distribution

If Y has a Burr Type XII distribution, $Y \sim \text{BTXII}(\phi, \lambda)$, then the probability density function (pdf) of Y is

$$f(y) = \frac{1}{\lambda} \frac{\phi y^{\phi-1}}{(1+y^\phi)^{\frac{1}{\lambda}+1}}$$

where y, ϕ, and λ are all positive. The cumulative distribution function (cdf) of Y is

$$F(y) = 1 - \exp\left[\frac{-\log(1+y^\phi)}{\lambda}\right] = 1 - (1+y^\phi)^{-1/\lambda} \quad \text{for} \quad \text{y} > 0.$$

$\text{MED}(Y) = [e^{\lambda \log(2)} - 1]^{1/\phi}$. See Patel, Kapadia, and Owen (1976, p. 195).

Assume that ϕ is known. Since $W = \log(1 + Y^\phi)$ is $EXP(\lambda)$,

$$\hat{\lambda} = \frac{\text{MED}(W_1, ..., W_n)}{\log(2)}$$

is a robust estimator. If all the $y_i \geq 0$ then a trimming rule is keep y_i if

$$0.0 \leq w_i \leq 9.0(1 + \frac{2}{n})\text{med}(n)$$

where med(n) is applied to $w_1, \ldots, w_n$ with $w_i = \log(1 + y_i^{\phi})$.

11.4.3 The Cauchy Distribution

If Y has a Cauchy distribution, $Y \sim C(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{\sigma}{\pi} \frac{1}{\sigma^2 + (y-\mu)^2} = \frac{1}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

where y and μ are real numbers and $\sigma > 0$.
The cdf of Y is $F(y) = \frac{1}{\pi}[\arctan(\frac{y-\mu}{\sigma}) + \pi/2]$. See Ferguson (1967, p. 102).
This family is a location–scale family that is symmetric about μ. MED$(Y) = \mu$, the upper quartile $= \mu + \sigma$, and the lower quartile $= \mu - \sigma$.
MAD$(Y) = F^{-1}(3/4) - \text{MED}(Y) = \sigma$. For a standard normal random variable, 99% of the mass is between -2.58 and 2.58 while for a standard Cauchy $C(0, 1)$ random variable 99% of the mass is between -63.66 and 63.66. Hence a rule which gives weight one to almost all of the observations of a Cauchy sample will be more susceptible to outliers than rules which do a large amount of trimming.

11.4.4 The Chi Distribution

If Y has a chi distribution, $Y \sim \chi_p$, then the pdf of Y is

$$f(y) = \frac{y^{p-1}e^{-y^2/2}}{2^{\frac{p}{2}-1}\Gamma(p/2)}$$

where $y \geq 0$ and p is a positive integer.
MED$(Y) \approx \sqrt{p - 2/3}$.
See Patel, Kapadia, and Owen (1976, p. 38). Since $W = Y^2$ is χ^2_p, a trimming rule is keep y_i if $w_i = y_i^2$ would be kept by the trimming rule for χ^2_p.

11.4.5 The Chi–square Distribution

If Y has a chi–square distribution, $Y \sim \chi^2_p$, then the pdf of Y is

$$f(y) = \frac{y^{\frac{p}{2}-1}e^{-\frac{y}{2}}}{2^{\frac{p}{2}}\Gamma(\frac{p}{2})}$$

where $y \geq 0$ and p is a positive integer.
$E(Y) = p$.
$\mathrm{VAR}(Y) = 2p$.

$\mathrm{MED}(Y) \approx p - 2/3$. See Pratt (1968, p. 1470) for more terms in the expansion of $\mathrm{MED}(Y)$. Empirically,

$$\mathrm{MAD}(Y) \approx \frac{\sqrt{2p}}{1.483}(1 - \frac{2}{9p})^2 \approx 0.9536\sqrt{p}.$$

Note that $p \approx \mathrm{MED}(Y) + 2/3$, and $\mathrm{VAR}(Y) \approx 2\mathrm{MED}(Y) + 4/3$. Let i be an integer such that $i \leq w < i + 1$. Then define $rnd(w) = i$ if $i \leq w \leq i + 0.5$ and $rnd(w) = i + 1$ if $i + 0.5 < w < i + 1$. Then $p \approx rnd(\mathrm{MED}(Y) + 2/3)$, and the approximation can be replaced by equality for $p = 1, \ldots, 100$.

Assume all $y_i > 0$. Let $\hat{p} = rnd(\mathrm{med}(n) + 2/3)$. Then a trimming rule is keep y_i if

$$\frac{1}{2}(-3.5 + \sqrt{2\hat{p}})^2 I(\hat{p} \geq 15) \leq y_i \leq \hat{p}[(3.5 + 2.0/n)\sqrt{\frac{2}{9\hat{p}}} + 1 - \frac{2}{9\hat{p}}]^3.$$

Another trimming rule would be to let

$$w_i = \left(\frac{y_i}{\hat{p}}\right)^{1/3}.$$

Then keep y_i if the trimming rule for the normal distribution keeps the w_i.

11.4.6 The Double Exponential Distribution

If Y has a double exponential distribution (or Laplace distribution), $Y \sim \mathrm{DE}(\theta, \lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{2\lambda}\exp\left(\frac{-|y - \theta|}{\lambda}\right)$$

where y is real and $\lambda > 0$. The cdf of Y is

$$F(y) = 0.5\exp\left(\frac{y - \theta}{\lambda}\right) \quad \text{if} \quad y \leq \theta,$$

and

$$F(y) = 1 - 0.5\exp\left(\frac{-(y - \theta)}{\lambda}\right) \quad \text{if} \quad y \geq \theta.$$

This family is a location–scale family which is symmetric about θ.
MAD$(Y) = \log(2)\lambda \approx 0.693\lambda$.
Hence $\lambda =$ MAD$(Y)/\log(2) \approx 1.443MAD(Y)$.
To see that MAD$(Y) = \lambda \log(2)$, note that $F(\theta + \lambda \log(2)) = 1 - 0.25 = 0.75$.

A trimming rule is keep y_i if

$$y_i \in [\text{med}(n) \pm 10.0(1 + \frac{2.0}{n})\text{mad}(n)].$$

Note that $F(\theta + \lambda \log(1000)) = 0.9995 \approx F(\text{MED}(Y) + 10.0\text{MAD}(Y))$.

11.4.7 The Exponential Distribution

If Y has an exponential distribution, $Y \sim \text{EXP}(\lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{\lambda} \exp\left(\frac{-y}{\lambda}\right) I(y \geq 0)$$

where $\lambda > 0$ and the indicator $I(y \geq 0)$ is one if $y \geq 0$ and zero otherwise. The cdf of Y is

$$F(y) = 1 - \exp(-y/\lambda), \ y \geq 0.$$

$E(Y) = \lambda$,
and VAR$(Y) = \lambda^2$.

MED$(Y) = \log(2)\lambda$ and
MAD$(Y) \approx \lambda/2.0781$ since it can be shown that

$$\exp(\text{MAD}(Y)/\lambda) = 1 + \exp(-\text{MAD}(Y)/\lambda).$$

Hence 2.0781 MAD$(Y) \approx \lambda$.
A robust estimator is $\hat{\lambda} =$ MED$(n)/\log(2)$.

If all the $y_i \geq 0$, then the trimming rule is keep y_i if

$$0.0 \leq y_i \leq 9.0(1 + \frac{c_2}{n})\text{med}(n)$$

where $c_2 = 2.0$ seems to work well. Note that $P(Y \leq 9.0\text{MED}(Y)) \approx 0.998$.

11.4.8 The Two Parameter Exponential Distribution

If Y has a two parameter exponential distribution, $Y \sim \text{EXP}(\theta, \lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{\lambda} \exp\left(\frac{-(y-\theta)}{\lambda}\right) I(y \geq \theta)$$

where $\lambda > 0$ and θ is real. The cdf of Y is

$$F(y) = 1 - \exp[-(y-\theta)/\lambda)], \ y \geq \theta.$$

This family is an asymmetric location-scale family.

$$\text{MED}(Y) = \theta + \lambda \log(2)$$

and

$$\text{MAD}(Y) \approx \lambda/2.0781.$$

Hence $\theta \approx \text{MED}(Y) - 2.0781 \log(2)\text{MAD}(Y)$. See Rousseeuw and Croux (1993) for similar results. Note that $2.0781 \log(2) \approx 1.44$.

A trimming rule is keep y_i if

$$\text{med}(n) - 1.44(1.0 + \frac{c_4}{n})\text{mad}(n) \leq y_i \leq$$

$$\text{med}(n) - 1.44\text{mad}(n) + 9.0(1 + \frac{c_2}{n})\text{med}(n)$$

where $c_2 = 2.0$ and $c_4 = 2.0$ may be good choices.

To see that $2.0781 \text{ MAD}(Y) \approx \lambda$, note that

$$0.5 = \int_{\theta+\lambda \log(2)-\text{MAD}}^{\theta+\lambda \log(2)+\text{MAD}} \frac{1}{\lambda} \exp(-(y-\theta)/\lambda) dy$$

$$= 0.5[-e^{-\text{MAD}/\lambda} + e^{\text{MAD}/\lambda}]$$

assuming $\lambda \log(2) > \text{MAD}$. Plug in $\text{MAD} = \lambda/2.0781$ to get the result.

11.4.9 The Gamma Distribution

If Y has a gamma distribution, $Y \sim G(\nu, \lambda)$, then the pdf of Y is

$$f(y) = \frac{y^{\nu-1} e^{-y/\lambda}}{\lambda^\nu \Gamma(\nu)}$$

where ν, λ, and y are positive. $E(Y) = \nu\lambda$.
$\text{VAR}(Y) = \nu\lambda^2$.
Chen and Rubin (1986) show that $\lambda(\nu - 1/3) < \text{MED}(Y) < \lambda\nu = E(Y)$.
Empirically, for $\nu > 3/2$,

$$\text{MED}(Y) \approx \lambda(\nu - 1/3),$$

and

$$\text{MAD}(Y) \approx \frac{\lambda\sqrt{\nu}}{1.483}.$$

This family is a scale family for fixed ν, so if Y is $G(\nu, \lambda)$ then cY is $G(\nu, c\lambda)$ for $c > 0$. If W is EXP(λ) then W is $G(1, \lambda)$. If W is χ^2_p, then W is $G(p/2, 2)$. For some M–estimators, see Marazzi and Ruffieux (1996).

Next we give some trimming rules. Assume each $y_i > 0$. Assume $\nu \geq 0.5$. Rule 1. Assume λ is known. Let $\hat{\nu} = (\text{med}(n)/\lambda) + (1/3)$. Keep y_i if $y_i \in [lo, hi]$ where

$$lo = \max(0, \hat{\nu}\lambda\ [-(3.5 + 2/n)\sqrt{\frac{1}{9\hat{\nu}}} + 1 - \frac{1}{9\hat{\nu}}]^3),$$

and

$$hi = \hat{\nu}\lambda\ [(3.5 + 2/n)\sqrt{\frac{1}{9\hat{\nu}}} + 1 - \frac{1}{9\hat{\nu}}]^3.$$

Rule 2. Assume ν is known. Let $\hat{\lambda} = \text{med}(n)/(\nu - (1/3))$. Keep y_i if $y_i \in [lo, hi]$ where

$$lo = \max(0, \nu\hat{\lambda}\ [-(3.5 + 2/n)\sqrt{\frac{1}{9\nu}} + 1 - \frac{1}{9\nu}]^3),$$

and

$$hi = \nu\hat{\lambda}\ \left[(3.5 + 2/n)\sqrt{\frac{1}{9\nu}} + 1 - \frac{1}{9\nu}\right]^3.$$

Rule 3. Let $d = \text{med}(n) - c\ \text{mad}(n)$. Keep y_i if

$$dI[d \geq 0] \leq y_i \leq \text{med}(n) + c\ \text{mad}(n)$$

where

$$c \in [9, 15].$$

11.4.10 The Half Cauchy Distribution

If Y has a half Cauchy distribution, $Y \sim \text{HC}(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{2}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

where $y \geq \mu$, μ is a real number and $\sigma > 0$. The cdf of Y is

$$F(y) = \frac{2}{\pi} \arctan(\frac{y - \mu}{\sigma})$$

for $y \geq \mu$ and is 0, otherwise. This distribution is a right skewed location-scale family.

MED$(Y) = \mu + \sigma$.

MAD$(Y) = 0.73205\sigma$.

11.4.11 The Half Logistic Distribution

If Y has a half logistic distribution, $Y \sim \text{HL}(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{2 \exp\left(-(y-\mu)/\sigma\right)}{\sigma[1 + \exp\left(-(y-\mu)/\sigma\right)]^2}$$

where $\sigma > 0$, $y \geq \mu$ and μ are real. The cdf of Y is

$$F(y) = \frac{\exp[(y-\mu)/\sigma] \ - \ 1}{1 + \exp[(y-\mu)/\sigma)]}$$

for $y \geq \mu$ and 0 otherwise. This family is a right skewed location–scale family.

MED$(Y) = \mu + \log(3)\sigma$.

MAD$(Y) = 0.67346\sigma$.

11.4.12 The Half Normal Distribution

If Y has a half normal distribution, $Y \sim \text{HN}(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{2}{\sqrt{2\pi}\ \sigma} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $y \geq \mu$ and μ is real. Let $\Phi(y)$ denote the standard normal cdf. Then the cdf of Y is

$$F(y) = 2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1$$

for $y > \mu$ and $F(y) = 0$, otherwise. This is an asymmetric location–scale family that has the same distribution as $\mu + \sigma|Z|$ where $Z \sim N(0,1)$. Note that $Z^2 \sim \chi^2_1$. MED$(Y) = \mu + 0.6745\sigma$.

MAD$(Y) = 0.3990916\sigma$.

Thus $\hat{\mu} \approx \text{MED}(n) - 1.6901\text{MAD}(n)$ and $\hat{\sigma} \approx 2.5057\text{MAD}(n)$.

11.4.13 The Inverse Exponential Distribution

If Y has an inverse exponential distribution, $Y \sim \text{IEXP}(\theta)$, then the pdf of Y is

$$f(y) = \frac{\theta}{y^2} \exp\left(\frac{-\theta}{y}\right)$$

where $y > 0$ and $\theta > 0$. The cdf $F(y) = \exp(-\theta/y)$ for $y > 0$. $E(Y)$ and $V(Y)$ do not exist. $\text{MED}(Y) = \theta/\log(2)$. This distribution is a scale family with scale parameter θ. $W = 1/Y \sim EXP(1/\theta)$.

11.4.14 The Largest Extreme Value Distribution

If Y has a largest extreme value distribution (or extreme value distribution for the max, or Gumbel distribution), $Y \sim \text{LEV}(\theta, \sigma)$, then the pdf of Y is

$$f(y) = \frac{1}{\sigma} \exp(-(\frac{y-\theta}{\sigma})) \exp[-\exp(-(\frac{y-\theta}{\sigma}))]$$

where y and θ are real and $\sigma > 0$. (Then $-Y$ has the smallest extreme value distribution or the log–Weibull distribution, see Section 11.4.26.) The cdf of Y is

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

This family is an asymmetric location–scale family with a mode at θ.

$$\text{MED}(Y) = \theta - \sigma \log(\log(2)) \approx \theta + 0.36651\sigma$$

and

$$\text{MAD}(Y) \approx 0.767049\sigma.$$

$W = \exp(-(Y-\theta)/\sigma) \sim \text{EXP}(1)$.

A trimming rule is keep y_i if

$$\text{med}(n) - 2.5\text{mad}(n) \leq y_i \leq \text{med}(n) + 7\text{mad}(n).$$

11.4.15 The Logistic Distribution

If Y has a logistic distribution, $Y \sim L(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{\exp\left(-(y-\mu)/\sigma\right)}{\sigma[1 + \exp\left(-(y-\mu)/\sigma\right)]^2}$$

where $\sigma > 0$ and y and μ are real. The cdf of Y is

$$F(y) = \frac{1}{1 + \exp\left(-(y-\mu)/\sigma\right)} = \frac{\exp\left((y-\mu)/\sigma\right)}{1 + \exp\left((y-\mu)/\sigma\right)}.$$

$\text{MED}(Y) = \mu$.
$\text{MAD}(Y) = \log(3)\sigma \approx 1.0986\ \sigma$.
Hence $\sigma = \text{MAD}(Y)/\log(3)$.

A trimming rule is keep y_i if

$$\text{med}(n) - 7.6(1 + \frac{c_2}{n})\text{mad}(n) \leq y_i \leq \text{med}(n) + 7.6(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 7.0. Note that if

$$q = F_{L(0,1)}(c) = \frac{e^c}{1 + e^c} \quad \text{then} \quad c = \log(\frac{q}{1-q}).$$

Taking $q = .9995$ gives $c = \log(1999) \approx 7.6$. To see that $\text{MAD}(Y) = \log(3)\sigma$, note that $F(\mu + \log(3)\sigma) = 0.75$, while $F(\mu - \log(3)\sigma) = 0.25$ and $0.75 = \exp\left(\log(3)\right)/(1 + \exp(\log(3)))$.

11.4.16 The Log-Cauchy Distribution

If Y has a log–Cauchy distribution, $Y \sim LC(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{1}{\pi\sigma y[1 + (\frac{\log(y)-\mu}{\sigma})^2]}$$

where $y > 0$, $\sigma > 0$ and μ is a real number. This family is a scale family with scale parameter $\tau = e^\mu$ if σ is known.

$W = \log(Y)$ has a Cauchy(μ, σ) distribution.

Robust estimators are $\hat{\mu} = \text{MED}(W_1, ..., W_n)$ and $\hat{\sigma} = \text{MAD}(W_1, ..., W_n)$.

11.4.17 The Log-Logistic Distribution

If Y has a log–logistic distribution, $Y \sim LL(\phi, \tau)$, then the pdf of Y is

$$f(y) = \frac{\phi\tau(\phi y)^{\tau-1}}{[1 + (\phi y)^\tau]^2}$$

where $y > 0$, $\phi > 0$ and $\tau > 0$. The cdf of Y is

$$F(y) = 1 - \frac{1}{1 + (\phi y)^\tau}$$

for $y > 0$. This family is a scale family with scale parameter ϕ^{-1} if τ is known.

MED$(Y) = 1/\phi$.

$W = \log(Y)$ has a logistic$(\mu = -\log(\phi), \sigma = 1/\tau)$ distribution. Hence $\phi = e^{-\mu}$ and $\tau = 1/\sigma$.

Robust estimators are $\hat{\tau} = \log(3)/\text{MAD}(W_1, ..., W_n)$ and $\hat{\phi} = 1/\text{MED}(Y_1, ..., Y_n)$ since MED$(Y) = 1/\phi$.

11.4.18 The Lognormal Distribution

If Y has a lognormal distribution, $Y \sim \text{LN}(\mu, \sigma^2)$, then the pdf of Y is

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

where $y > 0$ and $\sigma > 0$ and μ is real. The cdf of Y is

$$F(y) = \Phi\left(\frac{\log(y) - \mu}{\sigma}\right) \quad \text{for} \quad y > 0$$

where $\Phi(y)$ is the standard normal N(0,1) cdf. This family is a scale family with scale parameter $\tau = e^{\mu}$ if σ^2 is known.

MED$(Y) = \exp(\mu)$ and

$\exp(\mu)[1 - \exp(-0.6744\sigma)] \leq \text{MAD}(Y) \leq \exp(\mu)[1 + \exp(0.6744\sigma)]$.

Since $W = \log(Y) \sim N(\mu, \sigma^2)$, robust estimators are

$$\hat{\mu} = \text{MED}(W_1, ..., W_n) \quad \text{and} \quad \hat{\sigma} = 1.483\text{MAD}(W_1, ..., W_n).$$

Assume all $y_i \geq 0$. Then a trimming rule is keep y_i if

$$\text{med}(n) - 5.2(1 + \frac{c_2}{n})\text{mad}(n) \leq w_i \leq \text{med}(n) + 5.2(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 7.0. Here med(n) and mad(n) are applied to $w_1, \ldots, w_n$ where $w_i = \log(y_i)$.

11.4.19 The Maxwell-Boltzmann Distribution

If Y has a Maxwell–Boltzmann distribution, $Y \sim MB(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{\sqrt{2}(y - \mu)^2 e^{\frac{-1}{2\sigma^2}(y-\mu)^2}}{\sigma^3\sqrt{\pi}}$$

where μ is real, $y \geq \mu$ and $\sigma > 0$. This is a location–scale family.

MED$(Y) = \mu + 1.5381722\sigma$ and MAD$(Y) = 0.460244\sigma$.
Note that $W = (Y - \mu)^2 \sim G(3/2, 2\sigma^2)$.

11.4.20 The Normal Distribution

If Y has a normal distribution (or Gaussian distribution), $Y \sim N(\mu, \sigma^2)$, then the pdf of Y is

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

where $\sigma > 0$ and μ and y are real. Let $\Phi(y)$ denote the standard normal cdf. Then $\Phi(y) = 1 - \Phi(-y)$. MED$(Y) = \mu$ and

$$\text{MAD}(Y) = \Phi^{-1}(0.75)\sigma \approx 0.6745\sigma.$$

Hence $\sigma = [\Phi^{-1}(0.75)]^{-1}\text{MAD}(Y) \approx 1.483\text{MAD}(Y)$.
This family is a location–scale family which is symmetric about μ.

A trimming rule is keep y_i if

$$\text{med}(n) - 5.2(1 + \frac{c_2}{n})\text{mad}(n) \leq y_i \leq \text{med}(n) + 5.2(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 7.0. Using $c_2 = 4.0$ seems to be a good choice.

Note that

$$P(\mu - 3.5\sigma \leq Y \leq \mu + 3.5\sigma) = 0.9996.$$

To see that MAD$(Y) = \Phi^{-1}(0.75)\sigma$, note that $3/4 = F(\mu + \text{MAD})$ since Y is symmetric about μ. However,

$$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$$

and

$$\frac{3}{4} = \Phi\left(\frac{\mu + \Phi^{-1}(3/4)\sigma - \mu}{\sigma}\right).$$

So $\mu + \text{MAD} = \mu + \Phi^{-1}(3/4)\sigma$. Cancel μ from both sides to get the result.

11.4.21 The One Sided Stable Distribution

If Y has a one sided stable distribution (with index 1/2, also called a Lévy distribution), $Y \sim OSS(\sigma)$, then the pdf of Y is

$$f(y) = \frac{1}{\sqrt{2\pi y^3}} \sqrt{\sigma} \exp\left(\frac{-\sigma}{2} \frac{1}{y}\right)$$

for $y > 0$ and $\sigma > 0$. The cdf

$$F(y) = 2\left[1 - \Phi\left(\sqrt{\frac{\sigma}{y}}\right)\right]$$

for $y > 0$ where $\Phi(x)$ is the cdf of a $N(0, 1)$ random variable.

$$\text{MED}(Y) = \frac{\sigma}{[\Phi^{-1}(3/4)]^2}.$$

This distribution is a scale family with scale parameter σ. It can be shown that $W = 1/Y \sim G(1/2, 2/\sigma)$. This distribution is even more outlier prone than the Cauchy distribution. See Feller (1971, p. 52) and Lehmann (1999, p. 76). For applications see Besbeas and Morgan (2004).

11.4.22 The Pareto Distribution

If Y has a Pareto distribution, $Y \sim \text{PAR}(\sigma, \lambda)$, then the pdf of Y is

$$f(y) = \frac{\frac{1}{\lambda}\sigma^{1/\lambda}}{y^{1+1/\lambda}}$$

where $y \geq \sigma$, $\sigma > 0$, and $\lambda > 0$. The cdf of Y is $F(y) = 1 - (\sigma/y)^{1/\lambda}$ for $y > \sigma$. This family is a scale family when λ is fixed. $\text{MED}(Y) = \sigma 2^{\lambda}$.

$X = \log(Y/\sigma)$ is $\text{EXP}(\lambda)$ and $W = \log(Y)$ is $\text{EXP}(\theta = \log(\sigma), \lambda)$. Let $\hat{\theta} = \text{MED}(W_1, ..., W_n) - 1.440\text{MAD}(W_1, ..., W_n)$. Then robust estimators are

$$\hat{\sigma} = e^{\hat{\theta}} \quad \text{and} \quad \hat{\lambda} = 2.0781\text{MAD}(W_1, ..., W_n).$$

A trimming rule is keep y_i if

$$\text{med}(n) - 1.44\text{mad}(n) \leq w_i \leq 10\text{med}(n) - 1.44\text{mad}(n)$$

where med(n) and mad(n) are applied to $w_1, \ldots, w_n$ with $w_i = \log(y_i)$.

11.4.23 The Poisson Distribution

If Y has a Poisson distribution, $Y \sim \text{POIS}(\theta)$, then the pmf of Y is

$$P(Y = y) = \frac{e^{-\theta}\theta^y}{y!}$$

for $y = 0, 1, \ldots$, where $\theta > 0$.
$E(Y) = \theta$, and Chen and Rubin (1986) and Adell and Jodrá (2005) show that $-1 < \text{MED}(Y) - E(Y) < 1/3$.
$\text{VAR}(Y) = \theta$.

11.4.24 The Power Distribution

If Y has a power distribution, $Y \sim \text{POW}(\lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{\lambda} y^{\frac{1}{\lambda}-1},$$

where $\lambda > 0$ and $0 < y \leq 1$. The cdf of Y is $F(y) = y^{1/\lambda}$ for $0 < y \leq 1$.
$\text{MED}(Y) = (1/2)^{\lambda}$. $W = -\log(Y)$ is $\text{EXP}(\lambda)$.

If all the $y_i \in [0, 1]$, then a cleaning rule is keep y_i if

$$0.0 \leq w_i \leq 9.0(1 + \frac{2}{n})\text{med}(n)$$

where $\text{med}(n)$ is applied to $w_1, \ldots, w_n$ with $w_i = -\log(y_i)$. See Problem 11.5 for robust estimators.

11.4.25 The Rayleigh Distribution

If Y has a Rayleigh distribution, $Y \sim R(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{y - \mu}{\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{y - \mu}{\sigma}\right)^2\right]$$

where $\sigma > 0$, μ is real, and $y \geq \mu$. See Cohen and Whitten (1988, Ch. 10). This is an asymmetric location–scale family. The cdf of Y is

$$F(y) = 1 - \exp\left[-\frac{1}{2}\left(\frac{y - \mu}{\sigma}\right)^2\right]$$

for $y \geq \mu$, and $F(y) = 0$, otherwise. $\text{MED}(Y) = \mu + \sigma\sqrt{\log(4)} \approx \mu + 1.17741\sigma$.
Hence $\mu \approx \text{MED}(Y) - 2.6255\text{MAD}(Y)$ and $\sigma \approx 2.230\text{MAD}(Y)$.
Let $\sigma D = \text{MAD}(Y)$. If $\mu = 0$, and $\sigma = 1$, then

$$0.5 = \exp[-0.5(\sqrt{\log(4)} - D)^2] - \exp[-0.5(\sqrt{\log(4)} + D)^2].$$

Hence $D \approx 0.448453$ and $\text{MAD}(Y) \approx 0.448453\sigma$.
It can be shown that $W = (Y - \mu)^2 \sim \text{EXP}(2\sigma^2)$.

Other parameterizations for the Rayleigh distribution are possible. See Problem 11.7.

11.4.26 The Smallest Extreme Value Distribution

If Y has a smallest extreme value distribution (or log-Weibull distribution), $Y \sim SEV(\theta, \sigma)$, then the pdf of Y is

$$f(y) = \frac{1}{\sigma} \exp(\frac{y - \theta}{\sigma}) \exp[-\exp(\frac{y - \theta}{\sigma})]$$

where y and θ are real and $\sigma > 0$. The cdf of Y is

$$F(y) = 1 - \exp[-\exp(\frac{y - \theta}{\sigma})].$$

This family is an asymmetric location-scale family with a longer left tail than right.

$\text{MED}(Y) = \theta - \sigma \log(\log(2))$.

$\text{MAD}(Y) \approx 0.767049\sigma$.

If Y has a $\text{SEV}(\theta, \sigma)$ distribution, then $W = -Y$ has an $\text{LEV}(-\theta, \sigma)$ distribution.

11.4.27 The Student's t Distribution

If Y has a Student's t distribution, $Y \sim t_p$, then the pdf of Y is

$$f(y) = \frac{\Gamma(\frac{p+1}{2})}{(p\pi)^{1/2}\Gamma(p/2)}(1 + \frac{y^2}{p})^{-(\frac{p+1}{2})}$$

where p is a positive integer and y is real. This family is symmetric about 0. The t_1 distribution is the Cauchy$(0, 1)$ distribution. If Z is $N(0, 1)$ and is independent of $W \sim \chi^2_p$, then

$$\frac{Z}{(\frac{W}{p})^{1/2}}$$

is t_p.
$E(Y) = 0$ for $p \geq 2$.
$\text{MED}(Y) = 0$.

VAR$(Y) = p/(p-2)$ for $p \geq 3$, and
MAD$(Y) = t_{p,0.75}$ where $P(t_p \leq t_{p,0.75}) = 0.75$.

A trimming rule for $p \geq 3$ is keep y_i if $y_i \in [\pm 5.2(1 + 10/n)\text{mad}(n)]$.

11.4.28 The Topp-Leone Distribution

If Y has a Topp–Leone distribution, $Y \sim TL(\nu)$, then pdf of Y is

$$f(y) = \nu(2-2y)(2y-y^2)^{\nu-1}$$

for $\nu > 0$ and $0 < y < 1$. The cdf of Y is $F(y) = (2y - y^2)^\nu$ for $0 < y < 1$.
MED$(Y) = 1 - \sqrt{1-(1/2)^{1/\nu}}$, and $W = -\log(2Y - Y^2) \sim EXP(1/\nu)$.

11.4.29 The Truncated Extreme Value Distribution

If Y has a truncated extreme value distribution, $Y \sim \text{TEV}(\lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{\lambda} \exp\left(y - \frac{e^y - 1}{\lambda}\right)$$

where $y > 0$ and $\lambda > 0$. The cdf of Y is

$$F(y) = 1 - \exp\left[\frac{-(e^y - 1)}{\lambda}\right]$$

for $y > 0$.
MED$(Y) = \log(1 + \lambda \log(2))$.

$W = e^Y - 1$ is EXP(λ).

If all the $y_i > 0$, then a trimming rule is keep y_i if

$$0.0 \leq w_i \leq 9.0(1 + \frac{2}{n})\text{med}(n)$$

where med(n) is applied to $w_1, \ldots, w_n$ with $w_i = e^{y_i} - 1$. See Problem 11.6 for robust estimators.

11.4.30 The Uniform Distribution

If Y has a uniform distribution, $Y \sim U(\theta_1, \theta_2)$, then the pdf of Y is

$$f(y) = \frac{1}{\theta_2 - \theta_1} I(\theta_1 \leq y \leq \theta_2).$$

The cdf of Y is $F(y) = (y - \theta_1)/(\theta_2 - \theta_1)$ for $\theta_1 \le y \le \theta_2$.
This family is a location-scale family which is symmetric about $(\theta_1 + \theta_2)/2$.
$\text{MED}(Y) = (\theta_1 + \theta_2)/2$.
$\text{MAD}(Y) = (\theta_2 - \theta_1)/4$.
Note that $\theta_1 = \text{MED}(Y) - 2\text{MAD}(Y)$ and $\theta_2 = \text{MED}(Y) + 2\text{MAD}(Y)$.

A trimming rule is keep y_i if

$$\text{med}(n) - 2.0(1 + \frac{c_2}{n})\text{mad}(n) \le y_i \le \text{med}(n) + 2.0(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 5.0. Replacing 2.0 by 2.00001 yields a rule for which the cleaned data will equal the actual data for large enough n (with probability increasing to one).

11.4.31 The Weibull Distribution

If Y has a Weibull distribution, $Y \sim W(\phi, \lambda)$, then the pdf of Y is

$$f(y) = \frac{\phi}{\lambda} y^{\phi - 1} e^{-\frac{y^\phi}{\lambda}}$$

where λ, y, and ϕ are all positive. For fixed ϕ, this is a scale family in $\sigma = \lambda^{1/\phi}$.
The cdf of Y is $F(y) = 1 - \exp(-y^\phi/\lambda)$ for $y > 0$. $\text{MED}(Y) = (\lambda \log(2))^{1/\phi}$.
Note that

$$\lambda = \frac{(\text{MED}(Y))^\phi}{\log(2)}.$$

Since $W = Y^\phi$ is EXP(λ), if all the $y_i > 0$ and if ϕ is known, then a cleaning rule is keep y_i if

$$0.0 \le w_i \le 9.0(1 + \frac{2}{n})\text{med}(n)$$

where med(n) is applied to $w_1, \ldots, w_n$ with $w_i = y_i^\phi$.

$W = \log(Y)$ has a smallest extreme value SEV($\theta = \log(\lambda^{1/\phi}), \sigma = 1/\phi$) distribution.

See Olive (2006) and Problem 11.8c for robust estimators of ϕ and λ.

11.5 Truncated Distributions

Truncated distributions are useful for the location model and for comparing multiple linear regression estimators. This section follow Olive (1998, 2017b: ∮ 1.7) closely. Theorem 2.2 shows that the (α, β) trimmed mean T_n is esti-

mating a parameter μ_T with an asymptotic variance equal to $\sigma_W^2/(\beta-\alpha)^2$.

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 Theorem 11.1 below show that the definition of expectation given in Definition 11.3 is the same as the usual definition for expectation if Y is a discrete or continuous random variable. Section 11.7 has more on mixture distributions.

Definition 11.2. The distribution of a random variable Y is a *mixture distribution* if the cdf of Y has the form

$$F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y) \tag{11.1}$$

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 11.3. 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

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y). \tag{11.2}$$

Theorem 11.1. 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 11.1. Theorem 11.1c 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 $\text{VAR}(Y) = 10.9$.

To generate a random variable Y with the above mixture distribution, generate a uniform (0,1) random variable U which is independent of the W_i. If $U \leq 1-\epsilon$, then generate W_1 and take $Y = W_1$. If $U > 1-\epsilon$, then generate W_2 and take $Y = W_2$. Note that the cdf of Y is $F_Y(y) = (1-\epsilon)F_{W_1}(y) + \epsilon F_{W_2}(y)$.

Remark 11.1. 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 W_1 and W_2 are independent random variables and $0 < \epsilon < 1$. Then the random variable W is a linear combination of W_1 and W_2, and W can be generated by generating two independent random variables W_1 and W_2. Then take $W = (1-\epsilon)W_1 + \epsilon W_2$.

If W_1 and W_2 are as in the previous example then the random variable W is a linear combination that 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 11.5.1, 11.5.2, 11.5.3, and 11.5.4 will be useful when the underlying distribution is exponential, double exponential, normal, or Cauchy (see Section 11.4). Sections 2.13 and 2.14 examine how the sample median, trimmed means and two stage trimmed means behave at these distributions.

Definitions 2.27 and 2.28 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 theorem 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 $d = b - \mu_T(a,b)$.

Theorem 11.2. Let $a = \mu_T(a,b) - c$ and $b = \mu_T(a,b) + d$. Then
a) $\mu_W(a,b) = \mu_T(a,b) - \alpha c + (1-\beta)d$, 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

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

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

$$\sigma^2_W = \alpha(\mu_T - c)^2 + (\beta-\alpha)(\sigma^2_T + \mu^2_T) + (1-\beta)(\mu_T + d)^2 - \mu^2_W.$$

Collecting terms shows that

$$\sigma^2_W = (\beta-\alpha)\sigma^2_T + (\beta - \alpha + \alpha + 1 - \beta)\mu^2_T + 2[(1-\beta)d - \alpha c]\mu_T$$
$$+\alpha c^2 + (1-\beta)d^2 - \mu^2_W.$$

From a),

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

and we find that

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

11.5.1 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 $0 < y \leq b$ where $\lambda > 0$. Let $b = k\lambda$, and let

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

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

Theorem 11.3. 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 11.32 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. □

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 $E(Y^2) \approx 0.95(2\lambda^2)$.

11.5.2 The Truncated Double Exponential Distribution

Suppose that X is a double exponential $DE(\mu, \lambda)$ random variable. Chapter 3 states that $\text{MED}(X) = \mu$ and $\text{MAD}(X) = \log(2)\lambda$. 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 for $a \le y \le b$, the pdf of Y is

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

Theorem 11.4. 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 $\text{VAR}(Y) = E[(Y - \mu)^2] = E(W_T^2)$ where W_T is $TEXP(\lambda, b = c\lambda)$. □

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

11.5.3 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.

Theorem 11.5. $E(Y) = \mu + \left[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right] \sigma$, and

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

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

Proof. Let $c = \dfrac{1}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$.

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

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

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

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

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

Using

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

gives the result. □

Theorem 11.6. Let Y be $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$. Then $E(Y) = \mu$ and $V(Y) = \sigma^2\left[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}\right]$.

Proof. Use the symmetry of ϕ, the fact that $\Phi(-x) = 1 - \Phi(x)$, and Theorem 11.5 to get the result. □

Examining $V(Y)$ 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 11.1.

Table 11.1 Variances for Several Truncated Normal Distributions

k	$V(Y)$
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	$0.999\sigma^2$

11.5.4 The Truncated Cauchy Distribution

If X is a Cauchy $C(\mu, \sigma)$ random variable, then $\text{MED}(X) = \mu$ and $\text{MAD}(X) = \sigma$. 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,

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

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

Theorem 11.7. If $a = b$, then $E(Y) = \mu$, and $V(Y) = \sigma^2\left[\frac{b - \tan^{-1}(b)}{\tan^{-1}(b)}\right]$.

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

11.6 Large Sample Theory

This section follows Olive (2014: ch. 8, 2017b: § 3.4) closely. The first three subsections will review large sample theory for the univariate case, then multivariate theory will be given.

11.6.1 The CLT and the Delta Method

Large sample theory, also called asymptotic theory, is used to approximate the distribution of an estimator when the sample size n is large. This theory is extremely useful if the exact sampling distribution of the estimator is complicated or unknown. To use this theory, one must determine what the estimator is estimating, the rate of convergence, the asymptotic distribution, and how large n must be for the approximation to be useful. Moreover, the (asymptotic) standard error (SE), an estimator of the asymptotic standard deviation, must be computable if the estimator is to be useful for inference. Often the bootstrap can be used to compute the SE.

Theorem 11.8: the Central Limit Theorem (CLT). Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

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

Hence

$$\sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \sqrt{n}\left(\frac{\sum_{i=1}^n Y_i - n\mu}{n\sigma}\right) \xrightarrow{D} N(0, 1).$$

Note that the sample mean is estimating the *population mean* μ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the SE $= S/\sqrt{n}$ where S is the *sample standard deviation.* For distributions "close" to the

normal distribution, the central limit theorem provides a good approximation if the sample size $n \geq 30$. Hesterberg (2014, pp. 41, 66) suggests $n \geq 5000$ is needed for moderately skewed distributions. A special case of the CLT is proven after Theorem 11.21.

Notation. The notation $X \sim Y$ and $X \stackrel{D}{=} Y$ both mean that the random variables X and Y have the same distribution. Hence $F_X(x) = F_Y(y)$ for all real y. The notation $Y_n \stackrel{D}{\rightarrow} X$ means that for large n we can approximate the cdf of Y_n by the cdf of X. The distribution of X is the limiting distribution or asymptotic distribution of Y_n. For the CLT, notice that

$$Z_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \left(\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}\right)$$

is the z–score of $\overline{Y}$. If $Z_n \stackrel{D}{\rightarrow} N(0,1)$, then the notation $Z_n \approx N(0,1)$, also written as $Z_n \sim AN(0,1)$, means approximate the cdf of Z_n by the standard normal cdf. See Definition 11.4. Similarly, the notation

$$\overline{Y}_n \approx N(\mu, \sigma^2/n),$$

also written as $\overline{Y}_n \sim AN(\mu, \sigma^2/n)$, means approximate the cdf of $\overline{Y}_n$ as if $\overline{Y}_n \sim N(\mu, \sigma^2/n)$. The distribution of X does not depend on n, but the approximate distribution $\overline{Y}_n \approx N(\mu, \sigma^2/n)$ does depend on n.

The two main applications of the CLT are to give the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ and the limiting distribution of $\sqrt{n}(Y_n/n - \mu_X)$ for a random variable Y_n such that $Y_n = \sum_{i=1}^{n} X_i$ where the X_i are iid with $E(X) = \mu_X$ and $\text{VAR}(X) = \sigma_X^2$.

Example 11.2. a) Let $Y_1, ..., Y_n$ be iid Ber(ρ). Then $E(Y) = \rho$ and $\text{VAR}(Y) = \rho(1-\rho)$. (The Bernoulli ($\rho$) distribution is the binomial (1,ρ) distribution.) Hence

$$\sqrt{n}(\overline{Y}_n - \rho) \stackrel{D}{\rightarrow} N(0, \rho(1-\rho))$$

by the CLT.

b) Now suppose that $Y_n \sim BIN(n, \rho)$. Then $Y_n \stackrel{D}{=} \sum_{i=1}^{n} X_i$ where $X_1, ..., X_n$ are iid Ber(ρ). Hence

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \stackrel{D}{\rightarrow} N(0, \rho(1-\rho))$$

since

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \stackrel{D}{=} \sqrt{n}(\overline{X}_n - \rho) \stackrel{D}{\rightarrow} N(0, \rho(1-\rho))$$

by a).

c) Now suppose that $Y_n \sim BIN(k_n, \rho)$ where $k_n \to \infty$ as $n \to \infty$. Then

$$\sqrt{k_n}\left(\frac{Y_n}{k_n} - \rho\right) \approx N(0, \rho(1-\rho))$$

or

$$\frac{Y_n}{k_n} \approx N\left(\rho, \frac{\rho(1-\rho)}{k_n}\right) \quad \text{or} \quad \mathrm{Y_n} \approx \mathrm{N}\left(\mathrm{k_n}\rho, \mathrm{k_n}\rho(1-\rho)\right).$$

Theorem 11.9: the Delta Method. If g does not depend on n, $g'(\theta) \neq 0$, and

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2),$$

then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2).$$

Example 11.3. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\mathrm{VAR}(Y) = \sigma^2$. Then by the CLT,

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

Let $g(\mu) = \mu^2$. Then $g'(\mu) = 2\mu \neq 0$ for $\mu \neq 0$. Hence

$$\sqrt{n}((\overline{Y}_n)^2 - \mu^2) \xrightarrow{D} N(0, 4\sigma^2\mu^2)$$

for $\mu \neq 0$ by the delta method.

Example 11.4. Let $X \sim \mathrm{Binomial}(n, p)$ where the positive integer n is large and $0 < p < 1$. Find the limiting distribution of $\sqrt{n}\left[\left(\frac{X}{n}\right)^2 - p^2\right]$.

Solution. Example 11.2b gives the limiting distribution of $\sqrt{n}(\frac{X}{n} - p)$. Let $g(p) = p^2$. Then $g'(p) = 2p$ and by the delta method,

$$\sqrt{n}\left[\left(\frac{X}{n}\right)^2 - p^2\right] = \sqrt{n}\left(g\left(\frac{X}{n}\right) - g(p)\right) \xrightarrow{D}$$

$$N(0, p(1-p)(g'(p))^2) = N(0, p(1-p)4p^2) = N(0, 4p^3(1-p)).$$

Example 11.5. Let $X_n \sim \mathrm{Poisson}(n\lambda)$ where the positive integer n is large and $\lambda > 0$.

a) Find the limiting distribution of $\sqrt{n}\left(\frac{X_n}{n} - \lambda\right)$.

b) Find the limiting distribution of $\sqrt{n}\left[\sqrt{\frac{X_n}{n}} - \sqrt{\lambda}\right]$.

Solution. a) $X_n \stackrel{D}{=} \sum_{i=1}^{n} Y_i$ where the Y_i are iid Poisson(λ). Hence $E(Y) = \lambda = Var(Y)$. Thus by the CLT,

$$\sqrt{n}\left(\frac{X_n}{n} - \lambda\right) \stackrel{D}{=} \sqrt{n}\left(\frac{\sum_{i=1}^{n} Y_i}{n} - \lambda\right) \stackrel{D}{\to} N(0, \lambda).$$

b) Let $g(\lambda) = \sqrt{\lambda}$. Then $g'(\lambda) = \frac{1}{2\sqrt{\lambda}}$ and by the delta method,

$$\sqrt{n}\left[\sqrt{\frac{X_n}{n}} - \sqrt{\lambda}\right] = \sqrt{n}\left(g\left(\frac{X_n}{n}\right) - g(\lambda)\right) \stackrel{D}{\to}$$

$$N(0, \lambda\,(g'(\lambda))^2) = N\left(0, \lambda\frac{1}{4\lambda}\right) = N\left(0, \frac{1}{4}\right).$$

Example 11.6. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma(α, β) distribution.

a) Find the limiting distribution of $\sqrt{n}\,(\,\overline{Y} - \,\alpha\beta\,)$.

b) Find the limiting distribution of $\sqrt{n}\,(\,(\overline{Y})^2 - \,c\,)$ for appropriate constant c.

Solution: a) Since $E(Y) = \alpha\beta$ and $V(Y) = \alpha\beta^2$, by the CLT $\sqrt{n}\,(\,\overline{Y} - \,\alpha\beta\,) \stackrel{D}{\to} N(0, \alpha\beta^2)$.

b) Let $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$. Let $g(\mu) = \mu^2$ so $g'(\mu) = 2\mu$ and $[g'(\mu)]^2 = 4\mu^2 = 4\alpha^2\beta^2$. Then by the delta method, $\sqrt{n}\,(\,(\overline{Y})^2 - \,c\,) \stackrel{D}{\to} N(0, \sigma^2[g'(\mu)]^2) = N(0, 4\alpha^3\beta^4)$ where $c = \mu^2 = \alpha^2\beta^2$.

11.6.2 Modes of Convergence and Consistency

Definition 11.4. Let $\{Z_n, n = 1, 2, ...\}$ be a sequence of random variables with cdfs F_n, and let X be a random variable with cdf F. Then Z_n **converges in distribution to** X, written

$$Z_n \stackrel{D}{\to} X,$$

or Z_n *converges in law to* X, written $Z_n \stackrel{L}{\to} X$, if

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

at each continuity point t of F. The distribution of X is called the **limiting distribution** or the **asymptotic distribution** of Z_n.

An important fact is that **the limiting distribution does not depend on the sample size** n. Notice that the CLT and delta method give the limiting distributions of $Z_n = \sqrt{n}(\overline{Y}_n - \mu)$ and $Z_n = \sqrt{n}(g(T_n) - g(\theta))$, respectively.

Convergence in distribution is useful if the distribution of X_n is unknown or complicated and the distribution of X is easy to use. Then for large n we can approximate the probability that X_n is in an interval by the probability that X is in the interval. To see this, notice that if $X_n \xrightarrow{D} X$, then $P(a < X_n \leq b) = F_n(b) - F_n(a) \rightarrow F(b) - F(a) = P(a < X \leq b)$ if F is continuous at a and b.

Warning: Convergence in distribution says that the cdf $F_n(t)$ of X_n gets close to the cdf of $F(t)$ of X as $n \rightarrow \infty$ provided that t is a continuity point of F. Hence for any $\epsilon > 0$ there exists N_t such that if $n > N_t$, then $|F_n(t) - F(t)| < \epsilon$. Notice that N_t depends on the value of t. Convergence in distribution does not imply that the random variables $X_n \equiv X_n(\omega)$ converge to the random variable $X \equiv X(\omega)$ for all ω.

Example 11.7. Suppose that $X_n \sim U(-1/n, 1/n)$. Then the cdf $F_n(x)$ of X_n is

$$F_n(x) = \begin{cases} 0, & x \leq \frac{-1}{n} \\ \frac{nx}{2} + \frac{1}{2}, & \frac{-1}{n} \leq x \leq \frac{1}{n} \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Sketching $F_n(x)$ shows that it has a line segment rising from 0 at $x = -1/n$ to 1 at $x = 1/n$ and that $F_n(0) = 0.5$ for all $n \geq 1$. Examining the cases $x < 0$, $x = 0$, and $x > 0$ shows that as $n \rightarrow \infty$,

$$F_n(x) \rightarrow \begin{cases} 0, & x < 0 \\ \frac{1}{2} & x = 0 \\ 1, & x > 0. \end{cases}$$

Notice that the right hand side is not a cdf since right continuity does not hold at $x = 0$. Notice that if X is a random variable such that $P(X = 0) = 1$, then X has cdf

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0. \end{cases}$$

Since $x = 0$ is the only discontinuity point of $F_X(x)$ and since $F_n(x) \rightarrow F_X(x)$ for all continuity points of $F_X(x)$ (i.e. for $x \neq 0$),

$$X_n \xrightarrow{D} X.$$

Example 11.8. Suppose $Y_n \sim U(0, n)$. Then $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\rightarrow\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and $n > t$, then $F_n(t) = t/n \rightarrow 0$ as $n \rightarrow \infty$. Thus $\lim_{n\rightarrow\infty} F_n(t) = 0$ for all t, and Y_n does not converge in distribution to any random variable Y since $H(t) \equiv 0$ is not a cdf.

Definition 11.5. A sequence of random variables X_n *converges in distribution to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{D} \tau(\theta), \quad \text{if} \quad \mathrm{X_n} \xrightarrow{\mathrm{D}} \mathrm{X}$$

where $P(X = \tau(\theta)) = 1$. The distribution of the random variable X is said to be *degenerate at* $\tau(\theta)$ or to be a *point mass at* $\tau(\theta)$.

Definition 11.6. A sequence of random variables X_n *converges in probability to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{P} \tau(\theta),$$

if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \quad \text{or, equivalently,} \quad \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \tau(\theta)| \geq \epsilon) = 0.$$

The sequence X_n **converges in probability to** X, written

$$X_n \xrightarrow{P} X,$$

if $X_n - X \xrightarrow{P} 0$.

Notice that $X_n \xrightarrow{P} X$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \quad \text{or, equivalently,} \quad \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \mathrm{X}| \geq \epsilon) = 0.$$

Definition 11.7. Let the *parameter space* Θ be the set of possible values of θ. A sequence of estimators T_n of $\tau(\theta)$ is **consistent** for $\tau(\theta)$ if

$$T_n \xrightarrow{P} \tau(\theta)$$

for every $\theta \in \Theta$. If T_n is consistent for $\tau(\theta)$, then T_n is a **consistent estimator** of $\tau(\theta)$.

Consistency is a weak property that is usually satisfied by good estimators. T_n is a consistent estimator for $\tau(\theta)$ if the probability that T_n falls in any neighborhood of $\tau(\theta)$ goes to one, regardless of the value of $\theta \in \Theta$.

Definition 11.8. For a real number $r > 0$, Y_n *converges in* r*th mean* to a random variable Y, written

$$Y_n \xrightarrow{r} Y,$$

if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, Y_n **converges in quadratic mean** to Y, written

$$Y_n \xrightarrow{2} Y \quad \text{or} \quad \mathrm{Y_n} \xrightarrow{\mathrm{qm}} \mathrm{Y},$$

if

$$E[(Y_n - Y)^2] \to 0$$

as $n \to \infty$.

Theorem 11.10: Generalized Chebyshev's Inequality. Let $u : \mathbb{R} \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

$$P[u(Y) \geq c] \leq \frac{E[u(Y)]}{c}.$$

If $\mu = E(Y)$ exists, then taking $u(y) = |y - \mu|^r$ and $\tilde{c} = c^r$ gives
Markov's Inequality: for $r > 0$ and any $c > 0$,

$$P[|Y - \mu| \geq c] = P[|Y - \mu|^r \geq c^r] \leq \frac{E[|Y - \mu|^r]}{c^r}.$$

If $r = 2$ and $\sigma^2 = \mathrm{VAR}(Y)$ exists, then we obtain
Chebyshev's Inequality:

$$P[|Y - \mu| \geq c] \leq \frac{\mathrm{VAR}(Y)}{c^2}.$$

Proof. The proof is given for pdfs. For pmfs, replace the integrals by sums. Now

$$E[u(Y)] = \int_{\mathbb{R}} u(y)f(y)dy = \int_{\{y:u(y)\geq c\}} u(y)f(y)dy + \int_{\{y:u(y)<c\}} u(y)f(y)dy$$

$$\geq \int_{\{y:u(y)\geq c\}} u(y)f(y)dy$$

since the integrand $u(y)f(y) \geq 0$. Hence

$$E[u(Y)] \geq c \int_{\{y:u(y)\geq c\}} f(y)dy = cP[u(Y) \geq c]. \ \square$$

The following theorem gives sufficient conditions for T_n to be a consistent estimator of $\tau(\theta)$. Notice that $E_\theta[(T_n - \tau(\theta))^2] = MSE_{\tau(\theta)}(T_n) \to 0$ for all $\theta \in \Theta$ is equivalent to $T_n \xrightarrow{qm} \tau(\theta)$ for all $\theta \in \Theta$.

Theorem 11.11. a) If

$$\lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

b) If

$$\lim_{n\to\infty} \mathrm{VAR}_\theta(T_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Proof. a) Using Theorem 11.10 with $Y = T_n$, $u(T_n) = (T_n - \tau(\theta))^2$ and $c = \epsilon^2$ shows that for any $\epsilon > 0$,

$$P_\theta(|T_n - \tau(\theta)| \geq \epsilon) = P_\theta[(T_n - \tau(\theta))^2 \geq \epsilon^2] \leq \frac{E_\theta[(T_n - \tau(\theta))^2]}{\epsilon^2}.$$

Hence

$$\lim_{n\to\infty} E_\theta[(T_n - \tau(\theta))^2] = \lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) \to 0$$

is a sufficient condition for T_n to be a consistent estimator of $\tau(\theta)$.

b) Recall that

$$MSE_{\tau(\theta)}(T_n) = \mathrm{VAR}_\theta(T_n) + [\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n})]^2$$

where $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta)$. Since $MSE_{\tau(\theta)}(T_n) \to 0$ if both $\mathrm{VAR}_\theta(T_n) \to 0$ and $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta) \to 0$, the result follows from a). □

The following result shows estimators that converge at a $\sqrt{n}$ rate are consistent. Use this result and the delta method to show that $g(T_n)$ is a consistent estimator of $g(\theta)$. Note that b) follows from a) with $X_\theta \sim N(0, v(\theta))$. The WLLN shows that $\overline{Y}$ is a consistent estimator of $E(Y) = \mu$ if $E(Y)$ exists.

Theorem 11.12. a) Let X_θ be a random variable with distribution depending on θ, and $0 < \delta \leq 1$. If

$$n^\delta(T_n - \tau(\theta)) \xrightarrow{D} X_\theta$$

then $T_n \xrightarrow{P} \tau(\theta)$.

b) If

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N(0, v(\theta))$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Definition 11.9. A sequence of random variables X_n *converges almost everywhere* (or *almost surely*, or *with probability 1*) to X if

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

This type of convergence will be denoted by

$$X_n \xrightarrow{ae} X.$$

Notation such as "X_n converges to X ae" will also be used. Sometimes "ae" will be replaced with "as" or "wp1." We say that X_n *converges almost everywhere* to $\tau(\theta)$, written

$$X_n \xrightarrow{ae} \tau(\theta),$$

if $P(\lim_{n\to\infty} X_n = \tau(\theta)) = 1$.

Theorem 11.13. Let Y_n be a sequence of iid random variables with $E(Y_i) = \mu$. Then

a) **Strong Law of Large Numbers (SLLN):** $\overline{Y}_n \xrightarrow{ae} \mu$, and

b) **Weak Law of Large Numbers (WLLN):** $\overline{Y}_n \xrightarrow{P} \mu$.

Proof of WLLN when $V(Y_i) = \sigma^2$: By Chebyshev's inequality, for every $\epsilon > 0$,

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{V(\overline{Y}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$. □

In proving consistency results, there is an infinite sequence of estimators that depend on the sample size n. Hence the subscript n will be added to the estimators.

Definition 11.10. Lehmann (1999, pp. 53-54): a) A sequence of random variables W_n is *tight* or *bounded in probability,* written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$.

b) The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

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

c) W_n has the *same order as X_n in probability*, written $W_n \asymp_P X_n$, if for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P\left(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon\right) = P\left(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}\right) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$.

d) Similar notation is used for a $k \times r$ matrix $\boldsymbol{A}_n = \boldsymbol{A} = [a_{i,j}(n)]$ if each element $a_{i,j}(n)$ has the desired property. For example, $\boldsymbol{A} = O_P(n^{-1/2})$ if each $a_{i,j}(n) = O_P(n^{-1/2})$.

Definition 11.12. Let $W_n = \|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|$.

a) If $W_n \asymp_P n^{-\delta}$ for some $\delta > 0$, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have (tightness) **rate** n^δ.

b) If there exists a constant κ such that

$$n^{\delta}(W_n - \kappa) \xrightarrow{D} X$$

for some nondegenerate random variable X, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have *convergence rate* n^{δ}.

Theorem 11.14. Suppose there exists a constant κ such that

$$n^{\delta}(W_n - \kappa) \xrightarrow{D} X.$$

a) Then $W_n = O_P(n^{-\delta})$.
b) If X is not degenerate, then $W_n \asymp_P n^{-\delta}$.

The above result implies that if W_n has convergence rate n^{δ}, then W_n has tightness rate n^{δ}, and the term "tightness" will often be omitted. Part a) is proved, for example, in Lehmann (1999, p. 67).

The following result shows that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$, and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^{δ} is a lower bound on the rate of W_n. As an example, if the CLT holds then $\overline{Y}_n = O_P(n^{-1/3})$, but $\overline{Y}_n \asymp_P n^{-1/2}$.

Theorem 11.15. a) If $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$.
b) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$.
c) If $W_n \asymp_P X_n$, then $X_n = O_P(W_n)$.
d) $W_n \asymp_P X_n$ iff $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

Proof. a) Since $W_n \asymp_P X_n$,

$$P\left(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon\right) = P\left(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}\right) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Hence $X_n \asymp_P W_n$.

b) Since $W_n \asymp_P X_n$,

$$P(|W_n| \leq |X_n D_\epsilon|) \geq P\left(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon\right) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Hence $W_n = O_P(X_n)$.

c) Follows by a) and b).

d) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$ by b) and c). Now suppose $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Then

$$P(|W_n| \leq |X_n| D_{\epsilon/2}) \geq 1 - \epsilon/2$$

for all $n \geq N_1$, and

$$P(|X_n| \leq |W_n| 1/d_{\epsilon/2}) \geq 1 - \epsilon/2$$

for all $n \geq N_2$. Hence

$$P(A) \equiv P\left(\left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}\right) \geq 1 - \epsilon/2$$

and

$$P(B) \equiv P\left(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right|\right) \geq 1 - \epsilon/2$$

for all $n \geq N = \max(N_1, N_2)$. Since $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$,

$$P(A \cap B) = P(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}) \geq 1 - \epsilon/2 + 1 - \epsilon/2 - 1 = 1 - \epsilon$$

for all $n \geq N$. Hence $W_n \asymp_P X_n$. $\square$

The following result is used to prove the following Theorem 11.17 which says that if there are K estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$, and if T_n^* picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Theorem 11.16: Pratt (1959). Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{11.3}$$

Proof.

$$P(\max\{X_{1,n}, ..., X_{K,n}\} \leq x) = P(X_{1,n} \leq x, ..., X_{K,n} \leq x) \leq$$

$$F_{W_n}(x) \leq P(\min\{X_{1,n}, ..., X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, ..., X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1 - \epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1, ..., K$. Bonferroni's inequality states that $P(\cap_{i=1}^K A_i) \geq \sum_{i=1}^K P(A_i) \ - \ (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B, ..., X_{K,n} \leq B) \geq$$

$$K(1 - \epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, ..., X_{K,n} > -B) \geq$$

$$-1 + K(1 - \epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \ \text{ for } \ \text{n} > \text{N}. \ \square$$

Theorem 11.17. Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{11.4}$$

Proof. Let $X_{j,n} = n^{\delta}\|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by Theorem 11.16, $n^{\delta}\|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. □

11.6.3 Slutsky's Theorem and Related Results

Theorem 11.18: Slutsky's Theorem. Suppose $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{P} w$ for some constant w. Then

a) $Y_n + W_n \xrightarrow{D} Y + w$,

b) $Y_n W_n \xrightarrow{D} wY$, and

c) $Y_n/W_n \xrightarrow{D} Y/w$ if $w \neq 0$.

Theorem 11.19. a) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

b) If $X_n \xrightarrow{ae} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

c) If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

d) $X_n \xrightarrow{P} \tau(\theta)$ iff $\mathrm{X_n} \xrightarrow{\mathrm{D}} \tau(\theta)$.

e) If $X_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{P} \tau(\theta)$.

f) If $X_n \xrightarrow{D} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{D} \tau(\theta)$.

Suppose that for all $\theta \in \Theta$, $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$, or $T_n \xrightarrow{ae} \tau(\theta)$. Then T_n is a consistent estimator of $\tau(\theta)$ by Theorem 11.19. We are assuming that the function τ does not depend on n.

Example 11.9. Let $Y_1, ..., Y_n$ be iid with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Then the sample mean $\overline{Y}_n$ is a consistent estimator of μ since i) the SLLN holds (use Theorems 11.13 and 11.19), ii) the WLLN holds, and iii) the CLT holds (use Theorem 11.12). Since

$$\lim_{n\to\infty} \mathrm{VAR}_{\mu}(\overline{Y}_n) = \lim_{n\to\infty} \sigma^2/n = 0 \quad \text{and} \quad \lim_{n\to\infty} E_{\mu}(\overline{Y}_n) = \mu,$$

$\overline{Y}_n$ is also a consistent estimator of μ by Theorem 11.11b. By the delta method and Theorem 11.12b, $T_n = g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 1.19e, $g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if g is continuous at μ for all $\mu \in \Theta$.

Theorem 1.20. Assume that the function g does not depend on n.

a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

b) **Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Remark 11.2. For Theorem 11.19, a) follows from Slutsky's Theorem by taking $Y_n \equiv X = Y$ and $W_n = X_n - X$. Then $Y_n \xrightarrow{D} Y = X$ and $W_n \xrightarrow{P} 0$. Hence $X_n = Y_n + W_n \xrightarrow{D} Y + 0 = X$. The convergence in distribution parts of b) and c) follow from a). Part f) follows from d) and e). Part e) implies that if T_n is a consistent estimator of θ and τ is a continuous function, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$. Theorem 11.20 says that convergence in distribution is preserved by continuous functions, and even some discontinuities are allowed as long as the set of continuity points is assigned probability 1 by the asymptotic distribution. Equivalently, the set of discontinuity points is assigned probability 0.

Example 11.10. (Ferguson 1996, p. 40): If $X_n \xrightarrow{D} X$, then $1/X_n \xrightarrow{D} 1/X$ if X is a continuous random variable since $P(X = 0) = 0$ and $x = 0$ is the only discontinuity point of $g(x) = 1/x$.

Example 11.11. Show that if $Y_n \sim t_n$, a t distribution with n degrees of freedom, then $Y_n \xrightarrow{D} Z$ where $Z \sim N(0, 1)$.

Solution: $Y_n \xrightarrow{D} Z/\sqrt{V_n/n}$ where $Z \perp\!\!\!\perp V_n \sim \chi^2_n$. If $W_n = \sqrt{V_n/n} \xrightarrow{P} 1$, then the result follows by Slutsky's Theorem. But $V_n \xrightarrow{D} \sum_{i=1}^{n} X_i$ where the iid $X_i \sim \chi^2_1$. Hence $V_n/n \xrightarrow{P} 1$ by the WLLN and $\sqrt{V_n/n} \xrightarrow{P} 1$ by Theorem 11.19e.

Theorem 1.21: Continuity Theorem. Let Y_n be sequence of random variables with characteristic functions $\phi_n(t)$. Let Y be a random variable with characteristic function (cf) $\phi(t)$.

a)

$$Y_n \xrightarrow{D} Y \text{ iff } \phi_n(t) \to \phi(t) \ \forall t \in \mathbb{R}.$$

b) Also assume that Y_n has moment generating function (mgf) m_n and Y has mgf m. Assume that all of the mgfs m_n and m are defined on $|t| \leq d$ for some $d > 0$. Then if $m_n(t) \to m(t)$ as $n \to \infty$ for all $|t| < c$ where $0 < c < d$, then $Y_n \xrightarrow{D} Y$.

Application: Proof of a Special Case of the CLT. Following Rohatgi (1984, pp. 569-9), let $Y_1, ..., Y_n$ be iid with mean μ, variance σ^2, and mgf $m_Y(t)$ for $|t| < t_o$. Then

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

has mean 0, variance 1, and mgf $m_Z(t) = \exp(-t\mu/\sigma) m_Y(t/\sigma)$ for $|t| < \sigma t_o$. We want to show that

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

Notice that $W_n =$

$$n^{-1/2}\sum_{i=1}^{n} Z_i = n^{-1/2}\sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right) = n^{-1/2}\frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma} = \frac{n^{-1/2}}{\frac{1}{n}}\,\frac{\overline{Y}_n - \mu}{\sigma}.$$

Thus

$$m_{W_n}(t) = E(e^{tW_n}) = E[\exp(tn^{-1/2}\sum_{i=1}^{n} Z_i)] = E[\exp(\sum_{i=1}^{n} tZ_i/\sqrt{n})]$$

$$= \prod_{i=1}^{n} E[e^{tZ_i/\sqrt{n}}] = \prod_{i=1}^{n} m_Z(t/\sqrt{n}) = [m_Z(t/\sqrt{n})]^n.$$

Set $\psi(x) = \log(m_Z(x))$. Then

$$\log[m_{W_n}(t)] = n\log[m_Z(t/\sqrt{n})] = n\psi(t/\sqrt{n}) = \frac{\psi(t/\sqrt{n})}{\frac{1}{n}}.$$

Now $\psi(0) = \log[m_Z(0)] = \log(1) = 0$. Thus by L'Hôpital's rule (where the derivative is with respect to n), $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\lim_{n\to\infty}\frac{\psi(t/\sqrt{n}\,)}{\frac{1}{n}} = \lim_{n\to\infty}\frac{\psi'(t/\sqrt{n}\,)[\frac{-t/2}{n^{3/2}}]}{(\frac{-1}{n^2})} = \frac{t}{2}\lim_{n\to\infty}\frac{\psi'(t/\sqrt{n}\,)}{\frac{1}{\sqrt{n}}}.$$

Now

$$\psi'(0) = \frac{m_Z'(0)}{m_Z(0)} = E(Z_i)/1 = 0,$$

so L'Hôpital's rule can be applied again, giving $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\frac{t}{2}\lim_{n\to\infty}\frac{\psi''(t/\sqrt{n}\,)[\frac{-t}{2n^{3/2}}]}{(\frac{-1}{2n^{3/2}})} = \frac{t^2}{2}\lim_{n\to\infty}\psi''(t/\sqrt{n}\,) = \frac{t^2}{2}\psi''(0).$$

Now

$$\psi''(t) = \frac{d}{dt}\frac{m_Z'(t)}{m_Z(t)} = \frac{m_Z''(t)m_Z(t) - (m_Z'(t))^2}{[m_Z(t)]^2}.$$

So

$$\psi''(0) = m_Z''(0) - [m_Z'(0)]^2 = E(Z_i^2) - [E(Z_i)]^2 = 1.$$

Hence $\lim_{n\to\infty}\log[m_{W_n}(t)] = t^2/2$ and

$$\lim_{n\to\infty} m_{W_n}(t) = \exp(t^2/2)$$

which is the N(0,1) mgf. Thus by the continuity theorem,

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1). \ \square$$

11.6.4 Multivariate Limit Theorems

Many of the univariate results of the previous 3 subsections can be extended to random vectors. For the limit theorems, the vector $\boldsymbol{X}$ is typically a $k \times 1$ column vector and $\boldsymbol{X}^T$ is a row vector. Let $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_k^2}$ be the Euclidean norm of $\boldsymbol{x}$.

Definition 11.13. Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

c) Let $r > 0$ be a real number. Then $\boldsymbol{X}_n$ **converges in rth mean** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{X}$, if $E(\|\boldsymbol{X}_n - \boldsymbol{X}\|^r) \to 0$ as $n \to \infty$.

d) $\boldsymbol{X}_n$ **converges almost everywhere** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{ae} \boldsymbol{X}$, if $P(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.

Theorems 11.22 and 11.23 below are the multivariate extensions of the limit theorems in subsection 11.6.1. When the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta}))$ is multivariate normal $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, approximate the joint cdf of $\boldsymbol{Z}_n$ with the joint cdf of the $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. Thus to find probabilities, manipulate $\boldsymbol{Z}_n$ as if $\boldsymbol{Z}_n \approx N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$, and $V(X) = \boldsymbol{\Sigma}_{\boldsymbol{X}} = \sigma^2$.

Theorem 11.22: the Multivariate Central Limit Theorem (MCLT). If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

To see that the delta method is a special case of the multivariate delta method, note that if T_n and parameter θ are real valued, then $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = g'(\theta)$.

Theorem 11.23: the Multivariate Delta Method. If $\boldsymbol{g}$ does not depend on n and

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

then

$$\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} N_d(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{\Sigma} \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})})$$

where the $d \times k$ Jacobian matrix of partial derivatives

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_1(\boldsymbol{\theta}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial \theta_1} g_d(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_d(\boldsymbol{\theta}) \end{bmatrix}.$$

Here the mapping $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^d$ needs to be differentiable in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^k$.

Definition 11.14. If the estimator $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{g}(\boldsymbol{T}_n)$ is a **consistent estimator** of $\boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 11.24. If $0 < \delta \leq 1$, $\boldsymbol{X}$ is a random vector, and

$$n^{\delta}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} \boldsymbol{X},$$

then $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 11.25. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$, and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then

a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$ and

b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{ae} \boldsymbol{\mu}$.

Theorem 11.26: Continuity Theorem. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic functions $\phi_n(\boldsymbol{t})$, and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $\phi(\boldsymbol{t})$. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \phi_{\mathrm{n}}(\boldsymbol{t}) \to \phi(\boldsymbol{t})$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Theorem 11.27: Cramér Wold Device. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Application: Proof of the MCLT Theorem 11.22. Note that for fixed $\boldsymbol{t}$, the $\boldsymbol{t}^T \boldsymbol{X}_i$ are iid random variables with mean $\boldsymbol{t}^T \boldsymbol{\mu}$ and variance

$\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}$. Hence by the CLT, $\boldsymbol{t}^T\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. The right hand side has distribution $\boldsymbol{t}^T\boldsymbol{X}$ where $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence by the Cramér Wold Device, $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. $\square$

Theorem 11.28. a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.
b)

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta}) \text{ iff } \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{g}(\boldsymbol{\theta}).$$

Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, e.g. $g(n) = \sqrt{n}$. See White (1984, p. 15). If a $k \times 1$ random vector $\boldsymbol{T}_n - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $\boldsymbol{T}_n$ has (tightness) rate $\sqrt{n}$.

Definition 11.15. Let $\boldsymbol{A}_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A}_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A}_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A}_n \asymp_P (1/(g(n))$ if $a_{i,j}(n) \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_{1,n} = \boldsymbol{T}_n - \boldsymbol{\mu}$ and $\boldsymbol{A}_{2,n} = \boldsymbol{C}_n - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_{1,n} \asymp_P (1/(g(n))$ and $\boldsymbol{A}_{2,n} \asymp_P (1/(g(n))$, then $(\boldsymbol{T}_n, \boldsymbol{C}_n)$ has (tightness) rate $g(n)$.

Theorem 11.29: Continuous Mapping Theorem. Let $\boldsymbol{X}_n \in \mathbb{R}^k$. If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^j$ is continuous, then $\boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{X})$.

The following two theorems are taken from Severini (2005, pp. 345-349, 354).

Theorem 11.30. Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

a) $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$ iff $X_{in} \xrightarrow{P} X_i$ for $i = 1, ..., k$.

b) **Slutsky's Theorem**: If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{c}$ for some constant $k \times 1$ vector $\boldsymbol{c}$, then i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{X} + \boldsymbol{c}$ and

ii) $\boldsymbol{Y}_n^T\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{c}^T\boldsymbol{X}$.

c) If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, then $\boldsymbol{W}_n\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}\boldsymbol{X}$, $\boldsymbol{X}_n^T\boldsymbol{W}_n \xrightarrow{D} \boldsymbol{X}^T\boldsymbol{C}$, $\boldsymbol{W}_n^{-1}\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}^{-1}\boldsymbol{X}$, and $\boldsymbol{X}_n^T\boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{X}^T\boldsymbol{C}^{-1}$.

Theorem 11.31. Let W_n, X_n, Y_n, and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, e.g. $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_nX_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1)O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1)o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n)O_P(Z_n) = O_P(Y_n Z_n)$.

Theorem 11.32. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) Let $\boldsymbol{\Sigma} > 0$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iii) Let $\boldsymbol{\Sigma} > 0$. If $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular, $n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) =$
$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$
$= (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{x} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - T)$
$+(\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) + (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu})$
$+(\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(1)$.

(Note that $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a continuous function of $(T, \boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) =$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) +$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since $\sqrt{n}(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}] \sqrt{n}(T - \boldsymbol{\mu}) = O_P(1)o_P(1)O_P(1) = o_P(1)$. $\square$

Example 11.12. Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Then

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$$

by Theorem 11.26. To see this, let $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, $\boldsymbol{z}_n = (\boldsymbol{x}_n^T, \boldsymbol{y}_n^T)^T$, and $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Since $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ and $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, the characteristic function

$$\phi_{\boldsymbol{z}_n}(\boldsymbol{t}) = \phi_{\boldsymbol{x}_n}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}_n}(\boldsymbol{t}_2) \to \phi_{\boldsymbol{x}}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}}(\boldsymbol{t}_2) = \phi_{\boldsymbol{z}}(\boldsymbol{t}).$$

Hence $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{z})$ by Theorem 11.29.

11.7 Mixture Distributions

Mixture distributions are useful for model and variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{I_j,0}$, and the lasso estimator $\hat{\boldsymbol{\beta}}_L$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{L,\lambda_i}$ for $i = 1, ..., M$. See Sections 2.3, 3.2, and 3.6. A random vector $\boldsymbol{u}$ has a mixture distribution if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$. See Definition 3.8 for the population mean and population covariance matrix of a random vector. Definitions 11.2 and 11.3 and Theorem 11.1 were for a mixture distribution of random variables.

Definition 11.16. The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{11.5}$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j.

Theorem 11.30. Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{11.6}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{11.7}$$

and $Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$
$\sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \tag{11.8}$$

If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and

$$Cov(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j).$$

This theorem is easy to prove if the $\boldsymbol{u}_j$ are continuous random vectors with (joint) probability density functions (pdfs) $f_{\boldsymbol{u}_j}(\boldsymbol{t})$. Then $\boldsymbol{u}$ is a continuous random vector with pdf

$$f_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j f_{\boldsymbol{u}_j}(\boldsymbol{t}), \quad \text{and} \quad \mathrm{E(h(}\boldsymbol{u}\mathrm{))} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{h}(\boldsymbol{t})\mathrm{f}_{\boldsymbol{u}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}$$

$$= \sum_{j=1}^{J} \pi_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{u}_j}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]$$

where $E[h(\boldsymbol{u}_j)]$ is the expectation with respect to the random vector $\boldsymbol{u}_j$. Note that

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \tag{11.9}$$

Alternatively, with respect to a Riemann Stieltjes integral, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t})$ provided the expected value exists, and the integral is a linear operator with respect to both h and F. Hence for a mixture distribution, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t}) =$

$$\int h(\boldsymbol{t}) \; d \left[\sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \right] = \sum_{j=1}^{J} \pi_j \int h(\boldsymbol{t}) dF_{\boldsymbol{u}_j}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)].$$

Remark 11.3. Suppose the random vector $\boldsymbol{u}$ is equal to random vectors $\boldsymbol{u}_j$ with probabilities π_j. Let $\boldsymbol{u} = (u_1, ..., u_g)^T$ and $P(\boldsymbol{u} \leq \boldsymbol{t}) = P(u_1 \leq t_1, ..., u_g \leq t_g) = F_{\boldsymbol{u}}(\boldsymbol{t})$. Let $P(A|B) = 0$ if $P(B) = 0$. Then

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j P(\boldsymbol{u} \leq \boldsymbol{t} | \boldsymbol{u} = \boldsymbol{u}_j)\pi_j = \sum_j P(\boldsymbol{w}_j \leq \boldsymbol{t})\pi_j = \sum_j F_{\boldsymbol{w}_j}(\boldsymbol{t})\pi_j$$

where $\boldsymbol{w}_j$ is a random vector with a distribution equal to the conditional distribution of $\boldsymbol{u}|\boldsymbol{u} = \boldsymbol{u}_j$. Hence $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_k. The $\boldsymbol{w}_j = \boldsymbol{u}_j$ if there is no selection bias, e.g. if the $\boldsymbol{u}_j$ are randomly selected with probabilities π_j. Random selection can be done by generating a uniform (0,1) random variable W where W is independent of the $\boldsymbol{u}_j$. If $0 \leq W \leq \pi_1$, let $\boldsymbol{u} = \boldsymbol{u}_1$. If $\pi_1 < W \leq \pi_1 + \pi_2$, let $\boldsymbol{u} = \boldsymbol{u}_2$, etc. Often selection bias is present which changes the distribution of $\boldsymbol{u}_j$ to $\boldsymbol{w}_j$. This happened for the variable selection estimator $\boldsymbol{\beta}_{VS}$. The estimator $\hat{\boldsymbol{\beta}}_{MIX}$ used random selection.

As an analogy, consider generating $X_{11}, ..., X_{1n}$ iid $N(\mu, \sigma^2)$, but you see randomly selected $X_{1,j_1} = Y_1$. Another sample is generated, and you see $Y_2 = X_{2,j_2}$, and the process is continued to generate $Y_1, ..., Y_B$. If B is large, the sample will look like it is from a $N(\overline{X}, S^2) \approx N(\mu, \sigma^2)$ distribution. If random selection is replaced by using $W_j = \min(X_{j1}, ..., X_{jn})$, the selection bias is such that $W_1, ..., W_B$ no longer come from a normal distribution.

11.8 Complements

Many of the trimming rules and robust point estimators in this chapter are due to Olive (1998, 2006). These robust estimators are usually inefficient, but can be used as starting values for iterative procedures such as maximum likelihood and as a quick check for outliers. These estimators can also be used to create a robust fully efficient cross checking estimator.

If no outliers are present and the sample size is large, then the robust and classical methods should give similar estimates. If the estimates differ, then outliers may be present or the assumed distribution may be incorrect. Although a plot is the best way to check for univariate outliers, many users of statistics plug in data and then take the result from the computer without checking assumptions. If the software would print the robust estimates besides the classical estimates and warn that the assumptions might be invalid if the robust and classical estimates disagree, more users of statistics would use plots and other diagnostics to check model assumptions.

11.9 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

11.1. Verify the formula for the cdf F for the following distributions.
a) Cauchy (μ, σ).
b) Double exponential (θ, λ).
c) Exponential (λ).
d) Logistic (μ, σ).
e) Pareto (σ, λ).
f) Power (λ).
g) Uniform (θ_1, θ_2).
h) Weibull $W(\phi, \lambda)$.

11.2*. Verify the formula for MED(Y) for the following distributions.
a) Exponential (λ).
b) Lognormal (μ, σ^2). (Hint: $\Phi(0) = 0.5$.)
c) Pareto (σ, λ).
d) Power (λ).
e) Uniform (θ_1, θ_2).
f) Weibull (ϕ, λ).

11.3*. Verify the formula for MAD(Y) for the following distributions. (Hint: Some of the formulas may need to be verified numerically. Find the cdf in the appropriate section of Chapter 3. Then find the population median

MED$(Y) = M$. The following trick can be used except for part c). If the distribution is symmetric, find $U = y_{0.75}$. Then $D = \text{MAD}(Y) = U - M$.)
a) Cauchy (μ, σ).
b) Double exponential (θ, λ).
c) Exponential (λ).
d) Logistic (μ, σ).
e) Normal (μ, σ^2).
f) Uniform (θ_1, θ_2).

11.4. Assume that Y is gamma (ν, λ). Let

$$\alpha = P[Y \leq G_\alpha].$$

Using

$$Y^{1/3} \approx N((\nu\lambda)^{1/3}(1 - \frac{1}{9\nu}), (\nu\lambda)^{2/3}\frac{1}{9\nu}),$$

show that

$$G_\alpha \approx \nu\lambda[z_\alpha\sqrt{\frac{1}{9\nu}} + 1 - \frac{1}{9\nu}]^3$$

where z_α is the standard normal percentile, $\alpha = \Phi(z_\alpha)$.

11.5. Suppose that $Y_1, ..., Y_n$ are iid from a power (λ) distribution. Suggest a robust estimator for λ

a) based on Y_i and

b) based on $W_i = -\log(Y_i)$.

11.6. Suppose that $Y_1, ..., Y_n$ are iid from a truncated extreme value TEV(λ) distribution. Find a robust estimator for λ

a) based on Y_i and

b) based on $W_i = e^{Y_i} - 1$.

11.7. Other parameterizations for the Rayleigh distribution are possible. For example, take $\mu = 0$ and $\lambda = 2\sigma^2$. Then W is Rayleigh RAY(λ), if the pdf of W is

$$f(w) = \frac{2w}{\lambda}\exp(-w^2/\lambda)$$

where λ and w are both positive.
The cdf of W is $F(w) = 1 - \exp(-w^2/\lambda)$ for $w > 0$.
$E(W) = \lambda^{1/2}\ \Gamma(1 + 1/2)$.
VAR$(W) = \lambda\Gamma(2)\ -\ (E(W))^2$.

$$E(W^r) = \lambda^{r/2}\ \Gamma(1 + \frac{r}{2}) \quad \text{for} \quad r > -2.$$

MED$(W) = \sqrt{\lambda \log(2)}$.
W is RAY(λ) if W is Weibull $W(\lambda, 2)$. Thus $W^2 \sim$ EXP(λ). If all $w_i > 0$, then a trimming rule is keep w_i if $0 \leq w_i \leq 3.0(1 + 2/n)MED(n)$.

a) Find the median MED(W).

b) Suggest a robust estimator for λ.

11.8. Suppose Y has a smallest extreme value distribution, $Y \sim SEV(\theta, \sigma)$. See Section 11.4.26.

a) Find MED(Y).

b) Find MAD(Y).

c) If X has a Weibull distribution, $X \sim W(\phi, \lambda)$, then $Y = \log(X)$ is SEV(θ, σ) with parameters

$$\theta = \log(\lambda^{\frac{1}{\phi}}) \text{ and } \sigma = 1/\phi.$$

Use the results of a) and b) to suggest estimators for ϕ and λ.

11.9. Suppose that Y has a half normal distribution, $Y \sim HN(\mu, \sigma)$.

a) Show that MED$(Y) = \mu + 0.6745\sigma$.

b) Show that MAD$(Y) = 0.3990916\sigma$ numerically.

11.10. Suppose that Y has a half Cauchy distribution, $Y \sim$ HC(μ, σ). See Section 11.4.10 for $F(y)$.

a) Find MED(Y).

b) Find MAD(Y) numerically.

11.11. If Y has a log–Cauchy distribution, $Y \sim LC(\mu, \sigma)$, then $W = \log(Y)$ has a Cauchy(μ, σ) distribution. Suggest robust estimators for μ and σ based on an iid sample $Y_1, ..., Y_n$.

11.12. Suppose Y has a half logistic distribution, $Y \sim$ HL(μ, σ). See Section 11.4.11 for $F(y)$. Find MED(Y).

11.13. Suppose Y has a log–logistic distribution, $Y \sim LL(\phi, \tau)$, then $W = \log(Y)$ has a logistic$(\mu = -\log(\phi), \sigma = 1/\tau)$ distribution. Hence $\phi = e^{-\mu}$ and $\tau = 1/\sigma$.

a) Using $F(y) = 1 - \dfrac{1}{1 + (\phi y)^\tau}$ for $y > 0$, find MED(Y).

b) Suggest robust estimators for τ and ϕ.

11.14. If Y has a geometric distribution, $Y \sim geom(p)$, then the pmf of Y is $P(Y = y) = p(1 - p)^y$ for $y = 0, 1, 2, ...$ and $0 \leq p \leq 1$. The cdf for Y

is $F(y) = 1 - (1-p)^{\lfloor y+1 \rfloor}$ for $y \geq 0$ and $F(y) = 0$ for $y < 0$. Use the cdf to find an approximation for MED(Y).

11.15. Suppose Y has a Maxwell–Boltzmann distribution, $Y \sim MB(\mu, \sigma)$. Show that MED$(Y) = \mu + 1.5381722\sigma$ and MAD$(Y) = 0.460244\sigma$.

11.16 If Y is Fréchet (μ, σ, ϕ), then the cdf of Y is

$$F(y) = \exp\left[-\left(\frac{y-\mu}{\sigma}\right)^{-\phi}\right]$$

for $y \geq \mu$ and 0 otherwise where $\sigma, \phi > 0$. Find MED(Y).

11.17. If Y has an F distribution with degrees of freedom p and $n-p$, then

$$Y \stackrel{D}{=} \frac{\chi^2_p/p}{\chi^2_{n-p}/(n-p)} \approx \chi^2_p/p$$

if n is much larger than p ($n >> p$). Find an approximation for MED(Y) if $n >> p$.

11.18. If Y has a Topp–Leone distribution, $Y \sim TL(\phi)$, then the cdf of Y is $F(y) = (2y - y^2)^\phi$ for $\phi > 0$ and $0 < y < 1$. Find MED(Y).

11.19. If Y has a one sided stable distribution (with index 1/2), then the cdf

$$F(y) = 2\left[1 - \Phi\left(\sqrt{\frac{\sigma}{y}}\right)\right]$$

for $y > 0$ where $\Phi(x)$ is the cdf of a $N(0,1)$ random variable. Find MED(Y).

11.20. If Y has a two parameter power distribution, then the pdf

$$f(y) = \frac{1}{\tau\lambda}\left(\frac{y}{\tau}\right)^{\frac{1}{\lambda}-1}$$

for $0 < y \leq \tau$ where $\lambda > 0$ and $\tau > 0$. Suggest robust estimators for τ and λ using $W = -\log(Y) \sim EXP(-\log(\tau), \lambda)$.

11.21. If Y has an inverse exponential distribution, then the cdf

$$F(y) = \exp\left(\frac{-\theta}{y}\right)$$

for $y > 0$ and $\theta > 0$. Find MED(Y).

11.22. If Y has a Birnbaum Saunders distribution, $Y \sim BS(\nu, \theta)$, then the cdf of Y is

$$F(y) = \Phi\left[\frac{1}{\nu}\left(\sqrt{\frac{y}{\theta}} - \sqrt{\frac{\theta}{y}}\right)\right]$$

where $\Phi(x)$ is the N(0,1) cdf and $y > 0$. Find MED(Y).

11.23. If Y has a Burr Type X distribution, $Y \sim$ BTX(τ), then the pdf of Y is

$$f(y) = I(y > 0)\ 2\ \tau\ y\ e^{-y^2}\ (1 - e^{-y^2})^{\tau - 1} =$$

$$I(y > 0)\ 2y\ e^{-y^2}\ \tau\ \exp[(1 - \tau)(-\log(1 - e^{-y^2}))]$$

where $\tau > 0$. Then $W = -\log(1 - e^{-Y^2}) \sim EXP(1/\tau)$ and MED(W) $= \log(2)/\tau$. Find a robust estimator of τ.

11.24*. 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 $E(X)$.

11.25. 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 $E(X)$.

11.26. a) Prove Theorem 11.2 a).
b) Prove Theorem 11.2 c).
c) Prove Theorem 11.2 d).
d) Prove Theorem 11.2 e).

11.27. 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

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

11.28. 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 $U_n = n - L_n$.)

11.29. 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 $U_n = \lfloor (100 - K_n)n/100 \rfloor$.)

11.30*. 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 $ARE \approx 0.64$.

11.31*. 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 $ARE \approx 2.0$.

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

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

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

11.33. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors from a multivariate t-distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with d degrees of freedom. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}) = \frac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. Assuming $d > 2$, find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

11.34. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where

$$\boldsymbol{x}_i \sim (1 - \gamma)N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

with $0 < \gamma < 1$ and $c > 0$. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}_i) = [1 + \gamma(c - 1)]\boldsymbol{\Sigma}$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{d})$ for appropriate vector $\boldsymbol{d}$.

11.35. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where $E(\boldsymbol{x}_i) = e^{0.5}\mathbf{1}$ and $\mathrm{Cov}(\boldsymbol{x}_i) = (e^2 - e)\boldsymbol{I}_p$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

11.36. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid 2×1 random vectors from a multivariate lognormal LN$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Let $\boldsymbol{x}_i = (X_{i1}, X_{i2})^T$. Following Press (2005, pp. 149-150), $E(X_{ij}) = \exp(\mu_j + \sigma_j^2/2)$,
$V(X_{ij}) = \exp(\sigma_j^2)[\exp(\sigma_j^2) - 1]\exp(2\mu_j)$ for $j = 1, 2$, and
$\mathrm{Cov}(X_{i1}, X_{i2}) = \exp[\mu_1 + \mu_2 + 0.5(\sigma_1^2 + \sigma_2^2) + \sigma_{12}][\exp(\sigma_{12}) - 1]$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

R problems

Warning: 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. *rcisim*, will display the code for the function. Use the

`args` command, e.g. *args(rcisim)*, to display the needed arguments for the function.

11.33. a) Download the R function `nav` that computes Equation (4.4) from Theorem 2.14.

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.

11.34. a) Download the R function `deav` that computes Equation (2.44) from Theorem 2.15.

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.

11.35. a) Download the R 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.

11.10 Hints for Selected Problems

Chapter 1

1.1 $\|r_{i,1} - r_{i,2}\| = \|Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_1 - (Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_2)\| = \|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_2 - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_1\| = \|\hat{Y}_{2,i} - \hat{Y}_{1,i}\| = \|\hat{Y}_{1,i} - \hat{Y}_{2,i}\|$.

1.2 The plot should be similar to Figure 1.5, but since the data is simulated, may not be as smooth.

1.3 c) The histograms should become more like a normal distribution as n increases from 1 to 200. In particular, when $n = 1$ the histogram should be right skewed while for $n = 200$ the histogram should be nearly symmetric. Also the scale on the horizontal axis should decrease as n increases.

d) Now $\overline{Y} \sim N(0, 1/n)$. Hence the histograms should all be roughly symmetric, but the scale on the horizontal axis should be from about $-3/\sqrt{n}$ to $3/\sqrt{n}$.

1.4 e) The plot should be strongly nonlinear, having a "V" shape.

1.5 You could save the data set from the text's website on a flash drive, and then open the data in *Arc* from the flash drive.

c) Most students should delete cases 5, 47, 75, 95, 168, 181, and 199.

f) The response plot looks like a line while the residual plot looks like a curve. A residual plot emphasizes lack of fit while the response plot emphasizes goodness of fit.

h) The quadratic model looks good.

Chapter 2

2.2. $F_W(w) = P(W \leq w) = P(Y \leq w - \mu) = F_Y(w - \mu)$. So $f_W(w) = \frac{d}{dw} F_Y(w - \mu) = f_Y(w - \mu)$.

2.3. $F_W(w) = P(W \leq w) = P(Y \leq w/\sigma) = F_Y(w/\sigma)$. So $f_W(w) = \frac{d}{dw} F_Y(w/\sigma) = f_Y(w/\sigma)\frac{1}{\sigma}$.

2.4. $F_W(w) = P(W \leq w) = P(\sigma Y \leq w - \mu) = F_Y(\frac{w-\mu}{\sigma})$. So $f_W(w) = \frac{d}{dw} F_Y(\frac{w-\mu}{\sigma}) = f_Y(\frac{w-\mu}{\sigma})\frac{1}{\sigma}$.

2.5 $N(0, \sigma_M^2)$

2.9 a) $8.25 \pm 0.7007 = (6.020, 10.480)$

b) 8.75 ± 1.1645 = (7.586, 9.914).

2.10 a) $\overline{Y} = 24/5 = 4.8$.

b)

$$S^2 = \frac{138 - 5(4.8)^2}{4} = 5.7$$

so $S = \sqrt{5.7} = 2.3875$.

c) The ordered data are 2,3,5,6,8 and MED$(n) = 5$.

d) The ordered $|Y_i - \text{MED}(n)|$ are 0,1,2,3,3 and MAD$(n) = 2$.

2.11 a) $\overline{Y} = 15.8/10 = 1.58$.

b)

$$S^2 = \frac{38.58 - 10(1.58)^2}{9} = 1.5129$$

so $S = \sqrt{1.5129} = 1.230$.

c) The ordered data set is 0.0,0.8,1.0,1.2,1.3,1.3,1.4,1.8,2.4,4.6 and MED$(n) = 1.3$.

d) The ordered $|Y_i - \text{MED}(n)|$ are 0,0,0.1,0.1,0.3,0.5,0.5,1.1,1.3,3.3 and MAD$(n) = 0.4$.

e) 4.6 is unusually large.

2.12 a) $S/\sqrt{n} = 3.2150$.

b) $n - 1 = 9$.

c) 94.0

d) $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\ \rceil = \lfloor 10/2 \rfloor - \lceil \sqrt{10/4}\ \rceil = 5 - 2 = 3.$

e) $U_n = n - L_n = 10 - 3 = 7.$

f) $p = U_n - L_n - 1 = 7 - 3 - 1 = 3.$

g) $\text{SE}(\text{MED}(n)) = (Y_{(U_n)} - Y_{(L_n+1)})/2 = (95 - 90.0)/2 = 2.5.$

2.13 a) $L_n = \lfloor n/4 \rfloor = \lfloor 2.5 \rfloor = 2.$

b) $U_n = n - L_n = 10 - 2 = 8.$

c) $p = U_n - L_n - 1 = 8 - 2 - 1 = 5.$

d) $(89.7 + 90.0 + \cdots + 95.3)/6 = 558/6 = 93.0.$

e) 89.7 89.7 89.7 90.0 94.0 94.0 95.0 95.3 95.3 95.3

f) $(\sum d_i)/n = 928/10 = 92.8.$

g) $(\sum d_i^2 - n(\overline{d})^2)/(n-1) = (86181.54 - 10(92.8)^2)/9 = 63.14/9 = 7.0156.$

h)

$$V_{SW} = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2} = \frac{7.0156}{(\frac{8-2}{10})^2} = 19.4877,$$

so

$$SE(T_n) = \sqrt{V_{SW}/n} = \sqrt{19.4877/10} = 1.3960.$$

2.14 a) $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\ \rceil = \lfloor 5/2 \rfloor - \lceil \sqrt{5/4}\ \rceil = 2 - 2 = 0.$

$U_n = n - L_n = 5 - 0 = 5.$

$p = U_n - L_n - 1 = 5 - 0 - 1 = 4.$

$\text{SE}(\text{MED}(n)) = (Y_{(U_n)} - Y_{(L_n+1)})/2 = (8 - 2)/2 = 3.$

b) $L_n = \lfloor n/4 \rfloor = \lfloor 1 \rfloor = 1.$

$U_n = n - L_n = 5 - 1 = 4.$

$p = U_n - L_n - 1 = 4 - 1 - 1 = 2.$

$T_n = (3 + 5 + 6)/3 = 4.6667.$

The $d's$ are 3 3 5 6 6.

$(\sum d_i)/n = 4.6$

$(\sum d_i^2 - n(\overline{d})^2)/(n-1) = (115 - 5(4.6)^2)/4 = 9.2/4 = 2.3.$

$$V_{SW} = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2} = \frac{2.3}{(\frac{4-1}{5})^2} = 6.3889,$$

so

$$SE(T_n) = \sqrt{V_{SW}/n} = \sqrt{6.3889/5} = 1.1304.$$

The *R* functions for Problems 2.26–2.35 are available from the text's website file *rpack* and should have been entered into the computer using a command like *source("G:/rpack.txt")*, as described in the preface or Section 11.2.

2.23 Simulated data: a) about 0.669 b) about 0.486.

2.24 Simulated data: a) about 0.0 b) $\overline{Y} \approx 1.00$ and $T_n \approx 0.74$.

2.28 Simulated data gives about (1514,1684).

2.29 Simulated data gives about (1676,1715).

2.30 Simulated data gives about (1679,1712).

2.39b i) Coverages should be near 0.95. The lengths should be about 4.3 for $n = 10$, 4.0 for $n = 50$ and 3.96 for $n = 100$.

ii) Coverage should be near 0.78 for $n = 10$ and 0 for $n = 50, 100$. The lengths should be about 187 for $n = 10$, 173 for $n = 50$ and 171 for $n = 100$. (It can be shown that the expected length for large n is 169.786.)

Chapter 3

3.1 a) $X_2 \sim N(100, 6)$.

b)

$$\begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 49 \\ 17 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \right).$$

c) $X_1 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.

d)

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_3)}{\sqrt{\text{VAR}(X_1)\text{VAR}(X_3)}} = \frac{-1}{\sqrt{3}\sqrt{4}} = -0.2887.$$

3.2 a) $Y|X \sim N(49, 16)$ since $Y \perp\!\!\!\perp X$. (Or use $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 0(1/25)(X - 100) = 49$ and $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 0(1/25)0 = 16$.)

b) $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 10(1/25)(X - 100) = 9 + 0.4X$.

c) $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 10(1/25)10 = 16 - 4 = 12$.

3.4 The proof is identical to that given in Example 3.2.

3.6 a) Sort each column, then find the median of each column. Then $\text{MED}(\boldsymbol{W}) = (1430, 180, 120)^T$.

b) The sample mean of $(X_1, X_2, X_3)^T$ is found by finding the sample mean of each column. Hence $\overline{\boldsymbol{x}} = (1232.8571, 168.00, 112.00)^T$.

3.11 $\boldsymbol{\Sigma B} = E[E(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{B})] = E(\boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{\Sigma B}$. Hence $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$.

3.20 a)

$$N_2 \left(\begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \right).$$

b) $X_2 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.

c) $\frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{33}}} = \frac{1}{\sqrt{2}\sqrt{3}} = 1/\sqrt{6} = 0.4082.$

3.29 a) The 4 plots should look nearly identical with the five cases 61–65 appearing as outliers.

3.30 Not only should none of the outliers be highlighted, but the highlighted cases should be ellipsoidal.

3.31 Answers will vary since this is simulated data, but should get gam near 0.4, 0.3, 0.2 and 0.1 as p increases from 2 to 20.

3.32 b) Ideally the answer to this problem and Problem 11.3b would be nearly the same, but students seem to want correlations to be very high and use n too high. Values of n around 20, 40 and 50 for $p = 2, 3$ and 4 should be enough.

3.33 b) Values of n should be near 20, 40 and 50 for $p = 2, 3$ and 4.

3.34 This is simulated data, but for most plots the slope is near 2 to 2.5.

Chapter 5

5.3 c) $F_o = 265.96$, pvalue = 0.0, reject Ho, there is a MLR relationship between the response variable height and the predictors sternal height and finger to ground.

5.4 No, the relationship should be linear.

5.5 No, since 0 is in the CI. X_2 could be a very useful predictor for Y, e.g. if $Y = X_2^2$.

5.6 c) The plot should have $\log(Z)$ on the vertical axis.

e) Since randomly generated data is used, answers vary slightly, but $\widehat{\log(Y)} \approx 4 + X_1 + X_2 + X_3$.

5.8 b) Masking since 3 outliers are good cases with respect to Cook's distances.

c) and d) usually the MBA residuals will be large in magnitude, but for some students MBA, ALMS and ALTS will be highly correlated.

Chapter 6

6.3 Adding $\mathbf{1}$ to $\boldsymbol{Y}$ is equivalent to using $\boldsymbol{u} = (1, 0, ..., 0)^T$ in Equation (7.7), and the result follows.

Chapter 7

7.4 b) The line should go through the left and right cluster but not through the middle cluster of outliers.

c) The identity line should NOT PASS through the cluster of outliers with Y near 0 and the residuals corresponding to these outliers should be large in magnitude.

8.5 e) Usually the MBA esitmator based on the median squared residual will pass through the outliers, while the MBA LATA estimator gives zero weight to the outliers (so that the outliers are large in magnitude).

Chapter 8

8.1 Approximately 2 $n^{\delta}\ f(0)$ cases have small errors.

8.35 b) The identity line should NOT PASS through the cluster of outliers with Y near 0. The amount of trimming seems to vary some with the computer (which should not happen unless there is a bug in the `tvreg2` function or if the computers are using different versions of `cov.mcd`), but most students liked 70% or 80% trimming.

Chapter 9

9.1
a) $\hat{e}_i = Y_i - T(Y)$.
b) $\hat{e}_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$.
c)

$$\hat{e}_i = \frac{Y_i}{\hat{\beta}_1 \exp[\hat{\beta}_2(x_i - \bar{x})]}.$$

d) $\hat{e}_i = \sqrt{w_i}(Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$.

9.2
a) Since Y is a (random) scalar and $E(\boldsymbol{w}) = \boldsymbol{0}$, $\Sigma_{\boldsymbol{x},Y} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))^T] = E[\boldsymbol{w}(Y - E(Y))] = E(\boldsymbol{w}Y) - E(\boldsymbol{w})E(Y) = E(\boldsymbol{w}Y)$.

b) Using the definition of z and $\boldsymbol{r}$, note that $Y = m(z) + e$ and $\boldsymbol{w} = \boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$. Hence $E(\boldsymbol{w}Y) = E[(\boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})(m(z) + e)] = E[(\boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})m(z)] + E[\boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}]E(e)$ since e is independent of $\boldsymbol{x}$. Since $E(e) = 0$, the latter term drops out. Since $m(z)$ and $\boldsymbol{\beta}^T\boldsymbol{w}m(z)$ are (random) scalars, $E(\boldsymbol{w}Y) = E[m(z)\boldsymbol{r}] + E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\Sigma_{\boldsymbol{x}}\boldsymbol{\beta}$.

c) Using result b), $\Sigma_{\boldsymbol{x}}^{-1}\Sigma_{\boldsymbol{x},Y} = \Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] + \Sigma_{\boldsymbol{x}}^{-1}E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\Sigma_{\boldsymbol{x}}\boldsymbol{\beta}$ =
$E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\Sigma_{\boldsymbol{x}}^{-1}\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} + \Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] = E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\beta} + \Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}]$ and the result follows.

d) $E(\boldsymbol{w}z) = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T\boldsymbol{\beta}] = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T) + E(\boldsymbol{x}^T))\boldsymbol{\beta}]$
$= E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T))]\boldsymbol{\beta} + E[\boldsymbol{x} - E(\boldsymbol{x})]E(\boldsymbol{x}^T)\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$.

e) If $m(z) = z$, then $c(\boldsymbol{x}) = E(\boldsymbol{\beta}^T \boldsymbol{w} z) = \boldsymbol{\beta}^T E(\boldsymbol{w} z) = \boldsymbol{\beta}^T \Sigma_{\boldsymbol{x}} \boldsymbol{\beta} = 1$ by result d).

f) Since z is a (random) scalar, $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = E[(\boldsymbol{w} - (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})z]$ $= E(\boldsymbol{w}z) - (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T E(\boldsymbol{w}z)$. Using result d), $E(\boldsymbol{r}z) = \Sigma_{\boldsymbol{x}}\boldsymbol{\beta} - \Sigma_{\boldsymbol{x}}\boldsymbol{\beta}\boldsymbol{\beta}^T\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} =$ $\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} - \Sigma_{\boldsymbol{x}}\boldsymbol{\beta} = \boldsymbol{0}$.

g) Since z and $\boldsymbol{r}$ are linear combinations of $\boldsymbol{x}$, the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Since $E(\boldsymbol{r}) = \boldsymbol{0}$, z and $\boldsymbol{r}$ are uncorrelated and thus independent. Hence $m(z)$ and $\boldsymbol{r}$ are independent and $\boldsymbol{u}(\boldsymbol{x}) = \Sigma_{\boldsymbol{x}}^{-1} E[m(z)\boldsymbol{r}] =$ $\Sigma_{\boldsymbol{x}}^{-1} E[m(z)]E(\boldsymbol{r}) = \boldsymbol{0}$.

9.4 The submodel I that uses a constant and A, C, E, F, H looks best since it is the minimum $C_p(I)$ model and I has the smallest value of k such that $C_p(I) \leq 2k$.

9.6 a) No strong nonlinearities for MVN data but there should be some nonlinearities present for the non–EC data.

b) The plot should look like a cubic function.

c) The plot should use 0% trimming and resemble the plot in b), but may not be as smooth.

d) The plot should be linear and for many students some of the trimmed views should be better than the OLS view.

e) The response plot should look like a cubic with trimming greater than 0%.

f) The plot should be linear.

9.7 b) and c) It is possible that none of the trimmed views look much like the sinc(ESP) = sin(ESP)/ESP function.

d) Now at least one of the trimmed views should be good.

e) More lmsreg trimmed views should be good than the views from the other 2 methods, but since simulated data is used, one of the plots from b) or c) could be as good or even better than the plot in d).

Chapter 10

10.2 a) ESP = 1.11108, $\exp(ESP) = 3.0376$ and $\hat{\rho} = \exp(ESP)/(1 + \exp(ESP)) = 3.0376/(1 + 3.0376) = 0.7523$.

10.3 $G^2(O|F) = 62.7188 - 13.5325 = 49.1863$, df = 3, p–value = 0.00, reject Ho, there is a LR relationship between ape and the predictors lower jaw, upper jaw and face length.

10.4 $G^2(R|F) = 17.1855 - 13.5325 = 3.653$, df = 1, $0.05 <$ p–value < 0.1, fail to reject Ho, the reduced model is good.

10.5a ESP = 0.2812465 and $\hat{\mu} = \exp(ESP) = 1.3248$.

10.6 $G^2(O|F) = 187.490 - 138.685 = 48.805$, df = 2, p–value = 0.00, reject Ho, there is a PR relationship between possums and the predictors habitat and stags.

10.8 a) B4

b) EE plot

c) B3 is best. B3 has 12 fewer predictors than B2 but the AIC increased by less than 3. B1 has too many predictors with large Wald p–values, B2 = I_I still has too many predictors (want $\leq 300/10 = 30$ predictors) while B4 has too small of a p–value for the change in deviance test.

10.12 a) A good submodel uses a constant, Bark, Habitat and Stags as predictors.

d) The response and EE plots are good as are the Wald p–values. Also AIC(full) = 141.506 while AIC(sub) = 139.644.

10.14 b) Use the log rule: (max age)/(min age) = 1400 > 10.

e) The slice means track the logistic curve very well if 8 slices are used.

i) The EE plot is linear.

j) The slice means track the logistic curve very well if 8 slices are used.

10.15 c) Should have 200 cases, df = 178 and deviance = 112.168.

d) The response plot with 12 slices suggests that the full model is good.

e) The submodel I_1 that uses a constant, AGE, CAN, SYS, TYP and FLOC and the submodel I_2 that is the same as I_1 but also uses FRACE seem to be competitive. If the factor FRACE is not used, then the response plot follows 3 lines, one for each race. The Wald p–values suggest that FRACE is not needed, but FRACE is needed since the EE plot is inadequate for model I_I.

10.16 b) The response plot (e.g. with 4 slices) is bad, so the LR model is bad.

d) Now the response plot (e.g. with 12 slices) is good in that slice smooth and the logistic curve are close where there is data (also the LR model is good at classifying 0's and 1's).

f) For this problem, $G^2(O|F) = 62.7188 - 0.00419862 = 62.7146$, df = 1, p–value = 0.00, so reject Ho and conclude that there is an LR relationship between ape and the predictor x_3.

g) The MLE does not exist since there is perfect classification (and the logistic curve can get close to but never equal a discontinuous step function). Hence Wald p–values tend to have little meaning; however, the change in

deviance test tends to correctly suggest that there is an LR relationship when there is perfect classification.

10.18 k) The deleted point is certainly influential. Without this case, there does not seem to be a PR relationship between the predictors and the response.

m) The weighted residual plot suggests that something is wrong with the model since the plotted points scatter about a line with positive slope rather than a line with 0 slope. The deviance residual plot does not suggest that anything is wrong with the model.

10.19 The response plot should look ok, but the function uses a default number of slices rather than allowing the user to select the number of slices using a "slider bar" (a useful feature of *Arc*).

10.20 a) Since this is simulated PR data, the response plot should look ok, but the function uses a default lowess smoothing parameter rather than allowing the user to select smoothing parameter using a "slider bar" (a useful feature of *Arc*).

b) The data should the identity line in the weighted fit response plots. In about 1 in 20 plots there will be a very large count that looks like an outlier. The weighted residual plot based on the MLE usually looks better than the plot based on the minimum chi-square estimator (the MLE plot tends to have less of a "left opening megaphone shape").

10.22 b) Model I_1 is better since it has fewer predictors and lower AIC than model I_2.

10.23 a)

```
Number in Model  Rsquare C(p)   Variables in model
      6          0.2316  7.0947 X3 X4 X6 X7 X9 X10
```

c) The slice means follow the logistic curve fairly well with 8 slices.

e) The EE plot is linear.

f) The slice means follow the logistic curve fairly well with 8 slices.

Chapter 11

11.2 a) $F(y) = 1 - \exp(-y/\lambda)$ for $y \geq 0$. Let $M = \text{MED}(Y) = \log(2)\lambda$. Then $F(M) = 1 - \exp(-\log(2)\lambda/\lambda) = 1 - \exp(-\log(2)) = 1 - \exp(\log(1/2)) = 1 - 1/2 = 1/2$.

b) $F(y) = \Phi([\log(y) - \mu]/\sigma)$ for $y > 0$. Let $M = \text{MED}(Y) = \exp(\mu)$. Then $F(M) = \Phi([\log(\exp(\mu)) - \mu]/\sigma) = \Phi(0) = 1/2$.

11.3 a) $M = \mu$ by symmetry. Since $F(U) = 3/4$ and $F(y) = 1/2 + (1/\pi)\text{arctan}([\text{y} - \mu]/\sigma)$, want $\text{arctan}([\text{U} - \mu]/\sigma) = \pi/4$ or $(U - \mu)/\sigma = 1$. Hence $U = \mu + \sigma$ and $\text{MAD}(Y) = D = U - M = \mu + \sigma - \mu = \sigma$.

b) $M = \theta$ by symmetry. Since $F(U) = 3/4$ and $F(y) = 1 - 0.5\exp(-[y - \theta]/\lambda)$ for $y \geq \theta$, want $0.5\exp(-[U-\theta]/\lambda) = 0.25$ or $\exp(-[U-\theta]/\lambda) = 1/2$. So $-(U-\theta)/\lambda = \log(1/2)$ or $U = \theta - \lambda\log(1/2) = \theta - \lambda(-\log(2)) = \theta + \lambda\log(2)$. Hence $\text{MAD}(Y) = D = U - M = U - \theta = \lambda\log(2)$.

11.7 a) $\text{MED}(W) = \sqrt{\lambda\log(2)}$.

11.8 a) $\text{MED}(W) = \theta - \sigma\log(\log(2))$.

b) $\text{MAD}(W) \approx 0.767049\sigma$.

c) Let $W_i = \log(X_i)$ for $i = 1, ..., n$. Then
$\hat{\sigma} = \text{MAD}(W_1, ..., W_n)/0.767049$ and $\hat{\theta} = \text{MED}(W_1, ..., W_n) - \hat{\sigma}\log(\log(2))$. So take $\hat{\phi} = 1/\hat{\sigma}$ and $\hat{\lambda} = \exp(\hat{\theta}/\hat{\sigma})$.

11.10 a) $\text{MED}(Y) = \mu + \sigma$.

b) $\text{MAD}(Y) = 0.73205\sigma$.

11.11 Let $\hat{\mu} = \text{MED}(W_1, ..., W_n)$ and $\hat{\sigma} = \text{MAD}(W_1, ..., W_n)$.

11.12 $\mu + \log(3)\sigma$

11.13 a) $\text{MED}(Y) = 1/\phi$
b) $\hat{\tau} = \log(3)/\text{MAD}(W_1, ..., W_n)$ and $\hat{\phi} = 1/\text{MED}(Y_1, ..., Y_n)$.

11.17 $\text{MED}(Y) \approx (p - 2/3)/p \approx 1$ if p is large.

11.19.

$$\text{MED}(Y) = \frac{\sigma}{[\Phi^{-1}(3/4)]^2}.$$

11.20. Let $\text{MED}(n)$ and $\text{MAD}(n)$ be computed using $W_1, ..., W_n$. Use $-\log(\hat{\tau}) = \text{MED}(n) - 1.440\text{MAD}(n) \equiv A$, so $\hat{\tau} = e^{-A}$. Also $\hat{\lambda} = 2.0781\text{MAD}(n)$.

11.21. $\text{MED}(Y) = \theta/\log(2)$.
11.22. θ

11.23. Given data $Y_1, ..., Y_n$, a robust estimator of τ is $\hat{\tau} = \log(2)/\text{MED}(n)$ where $\text{MED}(n)$ is the sample median of $W_1, ..., W_n$ and $W_i = -\log(1 - e^{-Y_i^2})$.

11.24 a) 200

b) $0.9(10) + 0.1(200) = 29$

11.25 a) $400(1) = 400$

b) $0.9(10) + 0.1(400) = 49$

11.11 Tables

Tabled values are F(0.95,k,d) where $P(F < F(0.95, k, d)) = 0.95$.
00 stands for ∞. Entries produced with the `qf(.95,k,d)` command in *R*.
The numerator degrees of freedom are k while the denominator degrees of freedom are d.

k	1	2	3	4	5	6	7	8	9	00
d										
1	161	200	216	225	230	234	237	239	241	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	1.62
00	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.00

Tabled values are $t_{\alpha,d}$ where $P(t < t_{\alpha,d}) = \alpha$ where t has a t distribution with d degrees of freedom. If $d > 29$ use the $N(0,1)$ cutoffs $d = Z = \infty$.

					alpha					pvalue
d	0.005	0.01	0.025	0.05	0.5	0.95	0.975	0.99	0.995	left tail
1	-63.66	-31.82	-12.71	-6.314	0	6.314	12.71	31.82	63.66	
2	-9.925	-6.965	-4.303	-2.920	0	2.920	4.303	6.965	9.925	
3	-5.841	-4.541	-3.182	-2.353	0	2.353	3.182	4.541	5.841	
4	-4.604	-3.747	-2.776	-2.132	0	2.132	2.776	3.747	4.604	
5	-4.032	-3.365	-2.571	-2.015	0	2.015	2.571	3.365	4.032	
6	-3.707	-3.143	-2.447	-1.943	0	1.943	2.447	3.143	3.707	
7	-3.499	-2.998	-2.365	-1.895	0	1.895	2.365	2.998	3.499	
8	-3.355	-2.896	-2.306	-1.860	0	1.860	2.306	2.896	3.355	
9	-3.250	-2.821	-2.262	-1.833	0	1.833	2.262	2.821	3.250	
10	-3.169	-2.764	-2.228	-1.812	0	1.812	2.228	2.764	3.169	
11	-3.106	-2.718	-2.201	-1.796	0	1.796	2.201	2.718	3.106	
12	-3.055	-2.681	-2.179	-1.782	0	1.782	2.179	2.681	3.055	
13	-3.012	-2.650	-2.160	-1.771	0	1.771	2.160	2.650	3.012	
14	-2.977	-2.624	-2.145	-1.761	0	1.761	2.145	2.624	2.977	
15	-2.947	-2.602	-2.131	-1.753	0	1.753	2.131	2.602	2.947	
16	-2.921	-2.583	-2.120	-1.746	0	1.746	2.120	2.583	2.921	
17	-2.898	-2.567	-2.110	-1.740	0	1.740	2.110	2.567	2.898	
18	-2.878	-2.552	-2.101	-1.734	0	1.734	2.101	2.552	2.878	
19	-2.861	-2.539	-2.093	-1.729	0	1.729	2.093	2.539	2.861	
20	-2.845	-2.528	-2.086	-1.725	0	1.725	2.086	2.528	2.845	
21	-2.831	-2.518	-2.080	-1.721	0	1.721	2.080	2.518	2.831	
22	-2.819	-2.508	-2.074	-1.717	0	1.717	2.074	2.508	2.819	
23	-2.807	-2.500	-2.069	-1.714	0	1.714	2.069	2.500	2.807	
24	-2.797	-2.492	-2.064	-1.711	0	1.711	2.064	2.492	2.797	
25	-2.787	-2.485	-2.060	-1.708	0	1.708	2.060	2.485	2.787	
26	-2.779	-2.479	-2.056	-1.706	0	1.706	2.056	2.479	2.779	
27	-2.771	-2.473	-2.052	-1.703	0	1.703	2.052	2.473	2.771	
28	-2.763	-2.467	-2.048	-1.701	0	1.701	2.048	2.467	2.763	
29	-2.756	-2.462	-2.045	-1.699	0	1.699	2.045	2.462	2.756	
Z	-2.576	-2.326	-1.960	-1.645	0	1.645	1.960	2.326	2.576	
CI						90%	95%		99%	
	0.995	0.99	0.975	0.95	0.5	0.05	0.025	0.01	0.005	right tail
	0.01	0.02	0.05	0.10	1	0.10	0.05	0.02	0.01	two tail