Chapter 1
Introduction

This chapter gives a brief introduction to multivariate analysis, including some matrix optimization results, mixture distributions, and the special case of the location model. Section 1.2 gives an overview of the book along with a table of abbreviations. Truncated distributions, covered in Section 1.7, will be useful for large sample theory for the location model and for the regression model. See Chapter 14.

1.1 Introduction

Multivariate analysis is a set of statistical techniques used to analyze possibly correlated data containing observations on $p \geq 2$ random variables measured on a set of n cases. Let $\boldsymbol{x} = (x_1, ..., x_p)^T$ where $x_1, ..., x_p$ are p random variables. Usually context will be used to decide whether $\boldsymbol{x}$ is a random vector or the observed random vector. For multivariate location and dispersion, the ith case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T = (x_{i1}, ..., x_{ip})^T$.

Definition 1.1. A **case** or **observation** consists of p random variables measured for one person or thing. The ith case $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$.

Notation: Typically lowercase boldface letters such as $\boldsymbol{x}$ denote column vectors, while uppercase boldface letters such as $\boldsymbol{S}$ denote matrices with two or more columns. An exception may occur for random vectors which are usually denoted by $\boldsymbol{x}$, $\boldsymbol{y}$, or $\boldsymbol{z}$: if context is not enough to determine whether $\boldsymbol{x}$ is a random vector or an observed random vector, then $\boldsymbol{X} = (X_1, ..., X_p)^T$ and $\boldsymbol{Y}$ will be used for the random vectors, and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for the observed value of the random vector. This notation is used in Chapter 3 in order to study the conditional distribution of $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$. An uppercase letter such as Y will usually be a random variable. A lowercase letter such as x_1 will also often be a random variable. An exception to this notation is

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_1

the generic multivariate location and dispersion estimator $(T, \boldsymbol{C})$ where the location estimator T is a $p \times 1$ vector such as $T = \overline{\boldsymbol{x}}$. $\boldsymbol{C}$ is a $p \times p$ dispersion estimator and conforms to the above notation.

Assume that the data $\boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable x_j for $j = 1, ..., p$.

Often the n rows corresponding to the n cases are assumed to be independent and identically distributed (iid): a random sample from some multivariate distribution. The p columns correspond to n measurements on the p correlated random variables $x_1, ..., x_p$. The n cases are $p \times 1$ vectors, while the p columns are $n \times 1$ vectors.

Some techniques have a vector of response variables $(Y_1, ..., Y_m)^T$ that is predicted with a vector of predictor variables $(x_1, ..., x_p)^T$. See Chapters 10 and 12. Methods involving one response variable will not be covered in depth in this text. Such models include multiple linear regression, many experimental design models, and generalized linear models. Discrete multivariate analysis = categorical data analysis will also not be covered. Robust regression is briefly covered in Chapter 14.

Most of the multivariate techniques studied in this book will use estimators of multivariate location and dispersion. Typically the data will be assumed to come from a continuous distribution with a joint probability distribution function (pdf). Multivariate techniques that examine correlations among the p random variables $x_1, ..., x_p$ include principal component analysis, canonical correlation analysis, and factor analysis. Multivariate techniques that compare the n cases $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ include discriminant analysis and cluster analysis. *Data reduction* attempts to simplify the multivariate data without losing important information. Since the data matrix $\boldsymbol{W}$ has np terms, *data reduction* is an important technique. Prediction and hypothesis testing are also important techniques. Hypothesis testing is important for multivariate regression, Hotelling's T^2 test, and MANOVA. See Section 1.2 for a table of acronyms.

Robust multivariate analysis consists of i) techniques that are robust to non-normality or ii) techniques that are robust to outliers. Techniques that are robust to outliers tend to have some robustness to non-normality. The classical sample mean $\overline{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{S}$, defined in Section 2.2, are very robust to non-normality, but are not robust to outliers. Large sample theory is useful for both robust techniques. See Section 3.4.

Statistical Learning could be defined as the statistical analysis of multivariate data. Machine learning, data mining, Big Data, analytics, business analytics, data analytics, and predictive analytics are synonymous terms. The

techniques are useful for Data Science and Statistics, the science of extracting information from data. Often Statistical Learning methods are useful when n is large, but n/p is not large. Most of the methods in this text are for n/p large, but Section 4.7 shows how to detect outliers when n/p is not large. The outlier detection method gives a *covmb2 set* B of at least $n/2$ cases. If x is a matrix of predictor variables and Y is a vector of response variables, the following R commands produce cleaned data that can be used in Statistical Learning techniques, such as lasso, even if $p > n$. Section 8.9 uses a similar technique for discriminant analysis.

```
tem <- getB(x)
Yc <- Y[tem$indx]
xc <- x[tem$indx,]
```

Statistical Learning problems are supervised or unsupervised. For supervised learning, the goal is to predict a response variable given predictors. Discriminant analysis and regression are important examples. See Chapters 8 and 14. For unsupervised learning, the goal is to describe associations and patterns among the p variables. Clustering, described in Chapter 13, is an important example. Excellent texts for Statistical Learning are Efron and Hastie (2016), Hair et al. (2009), Hastie et al. (2015), and James et al. (2013). Also see Olive (2017c).

1.2 Overview and Acronyms

Chapters 1, 2, and 3 present some results useful for multivariate analysis, including matrix optimization results, the sample mean and covariance matrix, Mahalanobis distances, the multivariate normal distribution, and elliptically contoured distributions. This material is essential for any first course in multivariate analysis. Some of the sections of Chapter 1 are useful for robust regression which is covered in Chapter 14.

Chapters 4 and 5 are the most important chapters in the book and are needed for the following chapters except Chapter 13 on clustering. Chapter 4 discusses classical and outlier resistant methods of multivariate location and dispersion (MLD). Chapter 5 shows how to use the DD plot to detect outliers and gives a prediction region for multivariate data. Applying this prediction region on bootstrapped data gives a confidence region that can be used for hypothesis testing. This "prediction region method" will be used to perform inference on outlier resistant methods of multivariate analysis. There is a subset U that is used to compute the robust MLD estimator. Often applying a standard method, such as principal components, on the subset U results in a robust version of the standard method.

Most of the remaining chapters focus on standard methods of multivariate analysis such as principal component analysis, canonical correlation analysis,

Table 1.1 Acronyms

Acronym	Description
CCA	canonical correlation analysis
cdf	cumulative distribution function
cf	characteristic function
CI	confidence interval
CLT	central limit theorem
DA	discriminant analysis
Det-MCD	practical approximate MCD estimator not backed by theory
DGK	an MLD estimator (DGK are the initials of the paper's authors)
EC	elliptically contoured
ESP	estimated sufficient predictor
ESSP	estimated sufficient summary plot = response plot
Fast-MCD	a slow FMCD estimator
FCH	name of a fast, consistent, highly outlier resistant MLD estimator
FDA	Fisher's discriminant analysis
FLTS	practical approximate LTS estimator not backed by theory
FMCD	practical approximate MCD estimator not backed by theory
GAM	generalized additive model
GLM	generalized linear model
HB	high breakdown
hbreg	practical high breakdown regression estimator backed by theory
iid	independent and identically distributed
KNN	K-nearest neighbors discriminant analysis
LDA	linear discriminant analysis
LMS	least median of squares (robust regression)
LR	logistic regression
LTA	least trimmed sum of absolute deviations (robust regression)
LTS	least trimmed sum of squares (robust regression)
MAD	median absolute deviation
MANOVA	multivariate analysis of variance
MB	median ball estimator
MBA	an MLD estimator made obsolete by FCH
MBA	or the median ball algorithm is the mbareg estimator
mbareg	a resistant regression estimator backed by theory
MCD	the impractical minimum covariance determinant estimator
MCLT	multivariate central limit theorem
MED	the median
mgf	moment generating function
MLD	multivariate location and dispersion
MLR	multiple linear regression
MVE	the impractical minimum volume ellipsoid estimator
MVN	multivariate normal

(continued)

Table 1.1 (continued)

Acronym	Description
OGK	an MLD estimator not backed by theory
OLS	ordinary least squares
PCA	principal component analysis
pdf	probability density function
PI	prediction interval
pmf	probability mass function
QDA	quadratic discriminant analysis
RFCH	the reweighted FCH estimator
RMVN	a reweighted FCH estimator that works well for MVN data
SE	standard error
SSP	sufficient summary plot
SUR	seemingly unrelated regressions
TVREG	a resistant "trimmed views" regression estimator

discriminant analysis, MANOVA, factor analysis, and multivariate linear regression. Emphasis is on methods that are robust to normality: the methods have large sample theory that shows that the methods work on a large class of distributions. Of secondary importance is how to make outlier resistant methods that are backed by large sample theory. Chapter 14 considers other techniques, including robust regression.

Acronyms are widely used in robust statistics and multivariate analysis, and some of the more important acronyms are in Table 1.1 Also see the text's index. The letter "R" tends to stand for "robust" (RPCA) or "reweighted" (RFCH). The letter "F" before a brand-name robust estimator (FMCD) tends to mean a practical estimator that used a fixed number of trial fits, where the criterion of the brand-name estimator was used to select the trial fit used in the final estimator. The letter "C" before a brand-name estimator (CLTS) tends to mean a concentration algorithm that was used for the F-brand-name estimator. The letter "A," standing for "algorithm," was also used for concentration algorithms (ALTS). These acronyms (with A, C, F, or R) are often omitted from Table 1.1.

1.3 Some Things That Can Go Wrong with a Multivariate Analysis

In multivariate analysis, there is often a training data set used to predict or classify data in a future test data set. Many things can go wrong. For classification and prediction, it is usually assumed that the data in the training

set is from the same distribution as the data in the test set. Following Hand (2006), this crucial assumption is often not justified.

Population drift is a common reason why the above assumption, which assumes that the various distributions involved do not change over time, is violated. Population drift occurs when the population distribution does change over time. As an example, perhaps pot shards are classified after being sent to a laboratory for analysis. It is often the case that even if the shards are sent to the same laboratory twice, the two sets of laboratory measurements differ greatly. As another example, suppose there are several variables being used to produce greater yield of a crop or a chemical. If one journal paper out of 50 (the training set) finds a set of variables and variable levels that successfully increases yield, then the next 25 papers (the test set) are more likely to use variables and variable levels similar to the one successful paper than variables and variable levels of the 49 papers that did not succeed. Hand (2006) noted that classification rules used to predict whether applicants are likely to default on loans are updated every few months in the banking and credit scoring industries.

A second thing that can go wrong is that the training or test data set is distorted away from the population distribution. This could occur if outliers are present or if one of the data sets is not a random sample from the population. For example, the training data set could be drawn from three hospitals, and the test data set could be drawn from two more hospitals. These two data sets may not represent random samples from the same population of hospitals.

Often problems specific to the multivariate method can occur. Often simpler techniques can outperform sophisticated multivariate techniques because the user of the multivariate method does not have the expertise to get the most out of the sophisticated techniques. For supervised classification, Hand (2006) noted that there can be error in class labels, arbitrariness in class definitions, and data sets where different optimization criteria lead to very different classification rules. Hand (2006) suggested that simple rules, such as linear discriminant analysis, may perform almost as well or better than sophisticated classification rules because of all of the possible problems. See Chapter 8.

1.4 Some Matrix Optimization Results

The following results will be used throughout the text and are useful for principal component analysis, canonical correlation analysis, Fisher's discriminant analysis, and the Hotelling's T^2 test. Let $\boldsymbol{B} > 0$ denote that $\boldsymbol{B}$ is a positive definite matrix. The *generalized eigenvalue problem* finds eigenvalue eigenvector pairs $(\lambda, \boldsymbol{g})$ such that $\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{g} = \lambda\boldsymbol{g}$ which are also solutions to the equation $\boldsymbol{A}\boldsymbol{g} = \lambda\boldsymbol{C}\boldsymbol{g}$. Then the pairs are used to maximize or minimize

the *Rayleigh quotient* $\dfrac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$. Results from linear algebra show that if $\boldsymbol{C} > 0$ and $\boldsymbol{A}$ are both symmetric, then the p eigenvalues of $\boldsymbol{C}^{-1}\boldsymbol{A}$ are real, and the number of nonzero eigenvalues of $\boldsymbol{C}^{-1}\boldsymbol{A}$ is equal to $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$. Note that if $\boldsymbol{a}_1 = c_1\boldsymbol{g}_1$ is the maximizer and $\boldsymbol{a}_p = c_p\boldsymbol{g}_p$ is the minimizer of the Rayleigh quotient for any nonzero constants c_1 and c_p, then there is a vector $\boldsymbol{\beta}$ that is the maximizer or minimizer such that $\|\boldsymbol{\beta}\| = 1$.

Theorem 1.1. Let $\boldsymbol{B} > 0$ be a $p \times p$ symmetric matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$ and the orthonormal eigenvectors satisfy $\boldsymbol{e}_i^T\boldsymbol{e}_i = 1$ while $\boldsymbol{e}_i^T\boldsymbol{e}_j = 0$ for $i \neq j$. Let $\boldsymbol{d}$ be a given $p \times 1$ vector, and let $\boldsymbol{a}$ be an arbitrary nonzero $p \times 1$ vector. See Johnson and Wichern (1988, pp. 64–65, 184).

a) $\max\limits_{\boldsymbol{a}\neq\boldsymbol{0}} \dfrac{\boldsymbol{a}^T\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}} = \boldsymbol{d}^T\boldsymbol{B}^{-1}\boldsymbol{d}$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ for any constant $c \neq 0$. Note that the numerator $= (\boldsymbol{a}^T\boldsymbol{d})^2$.

b) $\max\limits_{\boldsymbol{a}\neq\boldsymbol{0}} \dfrac{\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{a}} = \max\limits_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} = \lambda_1$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_1$.

c) $\min\limits_{\boldsymbol{a}\neq\boldsymbol{0}} \dfrac{\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{a}} = \min\limits_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} = \lambda_p$ where the min is attained for $\boldsymbol{a} = \boldsymbol{e}_p$.

d) $\max\limits_{\boldsymbol{a}\perp\boldsymbol{e}_1,...,\boldsymbol{e}_k} \dfrac{\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{a}} = \max\limits_{\|\boldsymbol{a}\|=1,\boldsymbol{a}\perp\boldsymbol{e}_1,...,\boldsymbol{e}_k} \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} = \lambda_{k+1}$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_{k+1}$ for $k = 1, 2, ..., p-1$.

e) Let $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ be the observed sample mean and sample covariance matrix where $\boldsymbol{S} > 0$. Then $\max\limits_{\boldsymbol{a}\neq\boldsymbol{0}} \dfrac{n\boldsymbol{a}^T(\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{S}\boldsymbol{a}} = n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{\mu}) = T^2$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ for any constant $c \neq 0$.

f) Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix. Let $\boldsymbol{C} > 0$ be a $p \times p$ symmetric matrix. Then $\max\limits_{\boldsymbol{a}\neq\boldsymbol{0}} \dfrac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}} = \lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$, the largest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The value of $\boldsymbol{a}$ that achieves the max is the eigenvector $\boldsymbol{g}_1$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$. Similarly, $\min\limits_{\boldsymbol{a}\neq\boldsymbol{0}} \dfrac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}} = \lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$, the smallest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The value of $\boldsymbol{a}$ that achieves the min is the eigenvector $\boldsymbol{g}_p$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$.

Proof Sketch. For a), note that $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = 1$, where $\boldsymbol{C} = \boldsymbol{B}$ and $\boldsymbol{A} = \boldsymbol{d}\boldsymbol{d}^T$, since $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = \text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{d}) = 1$. Hence $\boldsymbol{C}^{-1}\boldsymbol{A}$ has one nonzero eigenvalue eigenvector pair $(\lambda_1, \boldsymbol{g}_1)$. Since

$$(\lambda_1 = \boldsymbol{d}^T\boldsymbol{B}^{-1}\boldsymbol{d}, \boldsymbol{g}_1 = \boldsymbol{B}^{-1}\boldsymbol{d})$$

is a nonzero eigenvalue eigenvector pair for $\boldsymbol{C}^{-1}\boldsymbol{A}$ and $\lambda_1 > 0$, the result follows by f).

Note that b) and c) are special cases of f) with $\boldsymbol{A} = \boldsymbol{B}$ and $\boldsymbol{C} = \boldsymbol{I}$.

Note that e) is a special case of a) with $\boldsymbol{d} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ and $\boldsymbol{B} = \boldsymbol{S}$.

(Also note that $(\lambda_1 = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}), \boldsymbol{g}_1 = \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}))$ is a nonzero eigenvalue eigenvector pair for the rank 1 matrix $\boldsymbol{C}^{-1}\boldsymbol{A}$ where $\boldsymbol{C} = \boldsymbol{S}$ and $\boldsymbol{A} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T$.)

For f), see Mardia et al. (1979, p. 480). □

Suppose $\boldsymbol{A} > 0$ and $\boldsymbol{C} > 0$ are $p \times p$ symmetric matrices, and let $\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{a} = \lambda \boldsymbol{a}$. Then $\boldsymbol{A}\boldsymbol{a} = \lambda \boldsymbol{C}\boldsymbol{a}$, or $\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{a} = \frac{1}{\lambda}\boldsymbol{a}$. Hence if $(\lambda_i(\boldsymbol{C}^{-1}\boldsymbol{A}), \boldsymbol{a})$ are eigenvalue eigenvector pairs of $\boldsymbol{C}^{-1}\boldsymbol{A}$, then $\left(\lambda_i(\boldsymbol{A}^{-1}\boldsymbol{C}) = \frac{1}{\lambda_i(\boldsymbol{C}^{-1}\boldsymbol{A})}, \boldsymbol{a}\right)$ are eigenvalue eigenvector pairs of $\boldsymbol{A}^{-1}\boldsymbol{C}$. Thus we can maximize $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ with the eigenvector $\boldsymbol{a}$ corresponding to the smallest eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$ and minimize $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ with the eigenvector $\boldsymbol{a}$ corresponding to the largest eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$.

Remark 1.1. Suppose $\boldsymbol{A}$ and $\boldsymbol{C}$ are symmetric $p \times p$ matrices, $\boldsymbol{A} > 0$, $\boldsymbol{C}$ is singular, and it is desired to make $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ large but finite. Hence $\frac{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}$ should be made small but nonzero. The above result suggests that the eigenvector $\boldsymbol{a}$ corresponding to the smallest nonzero eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$ may be useful. Similarly, suppose it is desired to make $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ small but nonzero. Hence $\frac{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}$ should be made large but finite. Then the eigenvector $\boldsymbol{a}$ corresponding to the largest eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$ may be useful.

1.5 The Location Model

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

The *location model* is

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{1.1}$$

where $e_1, ..., e_n$ are error random variables, often independent and identically distributed (iid) with zero mean. For example, if the Y_i are iid from a normal distribution with mean μ and variance σ^2, written $Y_i \sim N(\mu, \sigma^2)$, then the e_i are iid with $e_i \sim N(0, \sigma^2)$. The location model is often summarized by

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

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

Definition 1.2. The *sample mean*

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{1.2}$$

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

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

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

and

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

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

Definition 1.3. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,} \tag{1.3}$$

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

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

Definition 1.4. The *sample variance*

$$S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1}, \tag{1.4}$$

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

The sample median is a measure of location, while the sample standard deviation is a measure of scale. In terms of the "rod analogy," the median is a value m such that at least half of the rods are to the left of m and at least half of the rods are to the right of m. Hence the number of rods to the left and right of m rather than the lengths of the rods determines the sample median. The sample standard deviation is vulnerable to outliers and is a measure of the average value of the rod lengths $|r_i(\overline{Y})|$. The sample MAD, defined below, is a measure of the median value of the rod lengths $|r_i(\text{MED}(n))|$.

Definition 1.5. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{1.5}$$

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

Example 1.1. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and $\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

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

The following confidence interval provides considerable resistance to gross outliers while being very simple to compute. See Olive (2008, pp. 238, 261–262). The standard error $\text{SE}(\text{MED}(n))$ is due to Bloch and Gastwirth (1968), but the degrees of freedom p is motivated by the confidence interval for the trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (e.g., $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (e.g., $\lceil 7.7 \rceil = 8$).

Application 1.1: inference with the sample median. Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4} \rceil$ and use

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

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

$$\text{MED}(n) \pm t_{p,1-\alpha/2} SE(\text{MED}(n)). \tag{1.6}$$

Example 1.2. Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean μ, and find a 95% CI for μ.

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

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

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

Example 1.3. In the last example, what happens if the 6 becomes 66 and a 9 becomes 99?

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

Example 1.4. The Buxton (1920) data contains 87 heights of men, but five of the men were recorded to be about 0.75 inches tall! The mean height

is $\overline{Y} = 1598.862$, and the classical 95% CI is [1514.206, 1683.518]. MED$(n) = 1693.0$, and the resistant 95% CI based on the median is [1678.517, 1707.483].

The heights for the five men were recorded under their head lengths, so the outliers can be corrected. Then $\overline{Y} = 1692.356$, and the classical 95% CI is [1678.595, 1706.118]. Now MED$(n) = 1694.0$, and the 95% CI based on the median is [1678.403, 1709.597]. Notice that when the outliers are corrected, the two intervals are very similar although the classical interval length is slightly shorter. Also notice that the outliers roughly shifted the median confidence interval by about 1 mm, while the outliers greatly increased the length of the classical t-interval. See Problem 1.3 for *mpack* software.

1.6 Mixture Distributions

Mixture distributions are often used as outlier models, and certain mixtures of elliptically contoured distributions have an elliptically contoured distribution. See Problem 3.4. The following two definitions and proposition are useful for finding the mean and variance of a mixture distribution. Parts a) and b) of Proposition 1.2 below show that the definition of expectation given in Definition 1.7 is the same as the usual definition for expectation if Y is a discrete or continuous random variable. The two definitions and proposition can be extended to random vectors.

Definition 1.6. The distribution of a random variable Y is a *mixture distribution* if the cdf of Y has the form

$$F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y) \tag{1.7}$$

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

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y). \tag{1.8}$$

Proposition 1.2. a) If Y is a discrete random variable that has a probability mass function (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 probability distribution function (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 1.5. Proposition 1.2c 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 1.2. 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.

1.7 Truncated Distributions

Truncated and Winsorized random variables are important because they simplify the asymptotic theory of robust estimators. See Section 14.7. Let Y be a random variable with continuous cdf F, and let $\alpha = F(a) < F(b) = \beta$. Then α is the *left trimming proportion* and $1-\beta$ is the *right trimming proportion.* Let $F(a-) = P(Y < a)$. (Refer to Proposition 1.2 for the notation used below.)

Definition 1.8. The *truncated random variable* $Y_T \equiv Y_T(a,b)$ with *truncation points* a and b has cdf

$$F_{Y_T}(y|a,b) = G(y) = \frac{F(y) - F(a-)}{F(b) - F(a-)} \tag{1.9}$$

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

$$\mu_T = \mu_T(a,b) = \int_{-\infty}^{\infty} y dG(y) = \frac{\int_a^b y dF(y)}{\beta - \alpha} \tag{1.10}$$

and

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

See Cramér (1946, p. 247).

Definition 1.9. The *Winsorized random variable*

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

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

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

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

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

and

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

Truncated distributions can be used to simplify the asymptotic theory of robust estimators of location and regression. The following four subsections will be useful when the underlying distribution is exponential, double exponential, normal, or Cauchy. If Y has an exponential distribution, $Y \sim \text{EXP}(\lambda)$, then the pdf of Y is

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

where $\lambda > 0$ and the indicator $I(y \ge 0)$ is one if $y \ge 0$ and zero otherwise. If Y has a double exponential distribution (or Laplace distribution), $Y \sim \text{DE}(\theta, \lambda)$, then the pdf of Y is

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

where y is real and $\lambda > 0$. 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. 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 $\sigma > 0$.

Definitions 1.8 and 1.9 defined the truncated random variable $Y_T(a,b)$ and the Winsorized random variable $Y_W(a,b)$. Let Y have cdf F, and let the truncated random variable $Y_T(a,b)$ have the cdf $F_{T(a,b)}$. The following lemma illustrates the relationship between the means and variances of $Y_T(a,b)$ and $Y_W(a,b)$. Note that $Y_W(a,b)$ is a mixture of $Y_T(a,b)$ and two point masses at a and b. Let $c = \mu_T(a,b) - a$ and $d = b - \mu_T(a,b)$.

Lemma 1.3. 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

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

1.7.1 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 $k > 0$.

Lemma 1.4. 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

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

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

1.7.2 The Truncated Double Exponential Distribution

Suppose that X is a double exponential $DE(\mu, \lambda)$ random variable. Then $\text{MED}(X) = \mu$ 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

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

Lemma 1.5. 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 1.4 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 $\text{VAR}(Y) \approx 0.95(2\lambda^2)$.

1.7.3 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

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

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

Lemma 1.6. $E(Y) = \mu + \left[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\right]\sigma$, 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 $c = \frac{1}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$.

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

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

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

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

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

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

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 $E(Y^2)/c =$

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

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

Using

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

gives the result. □

Corollary 1.7. 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 the above lemma 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 1.2.

1.7.4 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

Table 1.2 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	$0.999\sigma^2$

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

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

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

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

Lemma 1.8. 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 et al. (2001).

1.8 Summary

1) Given a small data set, find $\overline{Y}$, S, MED(n), and MAD(n). Recall that $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$ and the *sample variance*

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

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

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

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

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

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

The *sample median absolute deviation*

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

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

1.9 Complements

Olive (2008, chapters 2–4) covered robust estimators for the location model, including alternatives to the Olive (2005b) confidence interval for the median given in Application 1.1.

Riani et al. (2009) found the population mean and covariance matrix of an elliptically trimmed (truncated) multivariate normal distribution, using Tallis (1963).

1.10 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

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

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

b) Find the standard deviation S

c) Find the sample median MED(n).

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

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

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

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

b) Find the sample standard deviation S.

c) Find the sample median MED(n).

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

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

R Problem

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, e.g., *medci*, will display the code for the function. Use the `args` command, e.g., *args(medci)*, to display the needed arguments for the function.

1.3. a) Use the commands

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

to simulate data similar to the Buxton heights.

Download the *mpack* functions, `cci` and `medci`, which produce a classical CI, and a CI using the median and the Bloch and Gastwirth (1968) SE. The default is a 95% CI.

b) Compute a 95% CI for the artificial height data set with the command *cci(height)*.

c) Compute a 95% CI for the artificial height data set with the command *medci(height)*.