CHAPTER 15

Robust Statistical Procedures

While "nonparametric statistics" deals with developing statistical procedures that make fewer and weaker assumptions (see Chapter 13), "robust statistics" deals with developing statistical procedures which are insensitive to violation of the assumptions under which they were developed. A major activity of statisticians in the last two decades has been the development of robust statistical procedures for a wide variety of problems, where such procedures have now often become feasible in practice because of the availability of the modern digital computer (lack of which prevented such procedures from being of use earlier, as robust procedures often—though not always—require substantial computations that until recently were not feasible). In this chapter we introduce this important area by looking at problems that carry the flavor of the area, especially the interplay between theory, data, and computation. These are the **influence curve** used to find observations that (if incorrect) would greatly influence the analysis (studied in Sections 15.1 through 15.3), **M-estimation** (a robust generalization of maximum likelihood estimation, studied in Section 15.4), the **jackknife estimator** (a technique for reducing bias in estimation, studied in Section 15.5), and **bootstrap methods** (which amount to simulation and Monte Carlo methods, given a new name in their statistical context, studied in Section 15.6), which are a special case of the general **Statistical Simulation Procedure (SSP)** given in Section 15.6.

15.1. THE INFLUENCE CURVE

Suppose that $X_1, X_2, \ldots, X_n$ are independent r.v.'s, each with d.f. $F(x)$, and that of interest to us is some function of $F(\cdot)$, which we will call $\theta(F)$. For example, $\theta(F)$ could be the mean of the d.f.:

$$\theta(F) = \int_{-\infty}^{\infty} xf(x)\,dx \tag{15.1.1}$$

where $f(x)$ is the p.d.f. corresponding to the d.f. $F(x)$. Often $F(x)$ will be **unknown**; then one might use $\theta(F_n)$ to estimate $\theta(F)$, where $F_n(x)$ is the empiric

d.f. defined in Definition 4.8.1. For example, one might estimate the mean by

$$\theta(F_n) = \int_{-\infty}^{\infty} x\, dF_n(x) = (X_1 + \cdots + X_n)/n = \bar{X}. \tag{15.1.2}$$

[When F in equation (15.1.1) assigns probabilities to points, the integral is replaced by a sum over those points, with weightings of the probabilities of the points; technically, this is then called a Stieltjes integral, although that technicality need not concern us here.]

The basic item of concern in robust estimation is the effect of throwing in a small amount of "contamination" at some point x. **The effect of this contamination is often measured by the "influence curve" at x:**

Definition 15.1.3. The **influence curve at x** is given by

$$w(x) = \lim_{\varepsilon \to 0} \frac{\theta((1-\varepsilon)F + \varepsilon I_x) - \theta(F)}{\varepsilon} \tag{15.1.4}$$

where I_x is the d.f. that puts all probability at the point x.

Example 15.1.5. If $\theta(F)$ is the mean of the d.f. as in equation (15.1.1), then we find that the influence curve at x is (denoting the mean of F by μ)

$$\begin{aligned} w(x) &= \lim_{\varepsilon \to 0} \frac{\theta((1-\varepsilon)F + \varepsilon I_x) - \theta(F)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{(1-\varepsilon)\mu + \varepsilon x - \mu}{\varepsilon} = x - \mu. \end{aligned}$$

Thus, in this problem contamination at the mean will have influence 0, whereas in general contamination at a point x has an influence that is proportional to how far x is from the mean μ.

PROBLEMS FOR SECTION 15.1

15.1.1 Let $X_1, X_2, \ldots, X_{10}$ be independent exponential r.v.'s with mean λ. Let $\theta(F) = \mathrm{Var}(F)$. *Find* an estimator $\theta(F_n)$ of $\theta(F)$ based on the empiric d.f. F_n. *Derive* $w(x)$, the influence curve at x. *Plot* $w(x)$ as a function of $\mathrm{Var}(F)$.

15.2. THE INFLUENCE CURVE IN TWO DIMENSIONS

If our observations are bivariate, say n independent r.v.'s (X_{11}, X_{12}), $(X_{21}, X_{22}), \ldots, (X_{n1}, X_{n2})$ each with d.f. $F(x_{11}, x_{12})$, then some function of $F(\cdot, \cdot)$ may be of interest to us. For example,

$$\theta(F) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_{11} x_{12} f(x_{11}, x_{12})\, dx_{11}\, dx_{12} \tag{15.2.1}$$

where $f(x_{11}, x_{12})$ is the p.d.f. corresponding to the d.f. $F(x_{11}, x_{12})$. When $F(\cdot, \cdot)$ is *unknown*, one may use $\theta(F_n)$ to estimate $\theta(F)$, where $F_n(x_{11}, x_{12})$ is the

bivariate empiric d.f.:

Definition 15.2.2. Suppose that r.v.'s $(X_{11}, X_{12}), (X_{21}, X_{22}), \ldots, (X_{n1}, X_{n2})$ are independent with d.f. $F(x_{11}, x_{12})$. The **(bivariate) empiric d.f.** (based on this sample) is defined by

$$F_n(x_{11}, x_{12} | (X_{11}, X_{12}), (X_{21}, X_{22}), \ldots, (X_{n1}, X_{n2})) = \frac{(\text{Number of the pairs } (X_{i1}, X_{i2}) \text{ for which } X_{i1} \leqslant x_{11} \text{ and } X_{i2} \leqslant x_{12})}{n}.$$

Example 15.2.3. **If $\theta(F)$ is the correlation coefficient** of the bivariate d.f. F, that is, if

$$\theta(F) = \operatorname{Cov}(X_{11}, X_{12}) / (\operatorname{Var}(X_{11}) \operatorname{Var}(X_{12}))^{0.5},$$

then it can be shown that **the influence curve turns out to be**

$$w((x_1, x_2)) = \lim_{\epsilon \to 0} \frac{\theta\left((1-\epsilon)F + \epsilon I_{(x_1, x_2)}\right) - \theta(F)}{\epsilon} \tag{15.2.4}$$

$$= \frac{x_1 - \mu_1}{\sigma_1} \frac{x_2 - \mu_2}{\sigma_2} - \frac{\rho}{2}\left(\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right)$$

where $\rho, \mu_1, \mu_2, \sigma_1, \sigma_2$ are the respective correlation coefficient, means, and standard deviations of the bivariate d.f. F.

Contours of constant influence are given by the equation $w(x_1, x_2) = c$; algebraic manipulation shows that these contours can be expressed as

$$(1 - \rho^2) v_1 v_2 = c \tag{15.2.5}$$

where $y_1 = (x_1 - \mu_1)/\sigma_1$, $y_2 = (x_2 - \mu_2)/\sigma_2$,

$$v_1 = 2^{-1}\left(\frac{y_1 + y_2}{(1+\rho)^{0.5}} + \frac{y_1 - y_2}{(1-\rho)^{0.5}}\right)$$

and

$$v_2 = 2^{-1}\left(\frac{y_1 + y_2}{(1+\rho)^{0.5}} - \frac{y_1 - y_2}{(1-\rho)^{0.5}}\right),$$

that is, **hyperbolas**.

PROBLEMS FOR SECTION 15.2

15.2.1 In Example 15.2.3, suppose $\theta(F) = \rho^2$. Find the influence curve. Derive and plot the contours of constant influence $w(x_1, x_2) = c$.

15.3. THE EMPIRIC INFLUENCE CURVE IN TWO DIMENSIONS

In practical applications of the material of Section 15.2, we have n bivariate data points, and we calculate their sample correlation, sample means, and sample variances r, m_1, m_2, s_1^2, s_2^2. These sample estimates are then used (respectively) to replace ρ, μ_1, μ_2, σ_1^2, σ_2^2 in equation (15.2.5) to yield **contours of equal**

TABLE 15.3-1
Electric utility data as submitted to the Federal Power Commission

Observation Number	*Consumption*	*Generation*
1	71	12074
2	4	681
3	3	1072
4	5	1204
5	4	1372
6	3	1090
7	48	22040
8	11	5130
9	11	4656
10	5	1910
11	4	2590
12	16	7380
13	3	143
14	4	2090
15	6	2290
16	15	7100
17	5	2190
18	3	1190
19	4	1910
20	13	6410
21	6	2880
22	5	2030
23	93	330
24	4	1830
25	8	3700
26	4	2200
27	5	1990
28	5	2320
29	4	1920
30	6	2670
31	19	9540
32	4	2117
33	3	960
34	3	1210
35	5	1760
36	4	1830

estimated influence. In a paper that won the 1983 Jacob Wolfowitz Prize, Chernick (1982) proposed to **superimpose estimated influence contours over a scatter plot of bivariate data in order to detect very influential observations—which then should be subject to extensive checking** (as they will have affected the correlation estimate markedly, and if they are in error then the correlation estimate will likewise be badly in error).

Example 15.3.1. Table 15.3-1 gives 36 bivariate observations from an electric utility company, as submitted to the Federal Power Commission on FPC Form 4.

```
      PROGRAM INFL(INFDAT,OUTPUT,TAPE5 = INFDAT,TAPE6 = OUTPUT)
C FINF IS THE INFLUENCE FUNCTION
C X AND Y ARE THE BIVARIATE OBSERVATIONS
      DIMENSION FINF(1000),X(1000),Y(1000),XS(1000),YS(1000)
      REWIND 5
C NCASE IS THE NUMBER OF CASES IN THE RUN
      READ(5,10) NCASE
 10 FORMAT(I2)
      DO 20 M = 1,NCASE
C N IS THE SAMPLE SIZE
C IM IS A FLAG,IF IM = 1, THEN THE MEANS OF X AND Y ARE INPUT
C IF IM = 0, THE MEANS ARE COMPUTED
C IS IS A FLAG, IF IS = 1, THEN THE STANDARD DEVIATIONS ARE INPUT
C IF IS = 0, THE STANDARD DEVIATIONS ARE COMPUTED
      WRITE(6,333) M
333 FORMAT(" THE CASE NUMBER IS ",I2)
      READ(5,100)N,IM,IS
100 FORMAT(2X,I4,2(2X,I2)
      NM1 = N - 1
      READ(5,200) (X(I), Y(I), I = 1,N)
200 FORMAT(6(2X,F10.5))
      IF(IM.EQ.1) GO TO 1
C MEANS ARE CALCULATED HERE
      XB = 0
      YB = 0
      DO 11 I = 1,N
      XB = XB + X(I)
      YB = YB + Y(I)
 11 CONTINUE
      RN = N
      XB = XB / RN
      YB = YB / RN
      GO TO 3
  1 READ(5,300)XB,YB
300 FORMAT(2(2X,F10.5))
  3 IF(IS.EQ.1) GO TO 4
      SX = 0.
      SY = 0.
C STANDARD DEVIATIONS ARE COMPUTED HERE
      DO 21 I = 1,N
      SX = (X(I) - XB)*(X(I) - XB) + SX
      SY = (Y(I) - YB)*(Y(I) - YB) + SY
```

Figure 15.3-1. Computer program for influence curve calculations.

```
 21 CONTINUE
    RN = N
    RN = SQRT(RN)
    SX = SQRT(SX) / RN
    SY = SQRT(SY) / RN
    GO TO 5
  4 READ(5,300)SX, SY
  5 Z = SX*SY
C THE CORRELATION IS CALCULATED HERE
    COV = 0.
    DO 22 I = 1,N
    COV = (X(I) - XB)*(Y(I) - YB) + COV
 22 CONTINUE
    RN = N
    COV = COV / RN
    COR = COV / Z
C COMPUTATION OF THE INFLUENCE FUNCTION
    WRITE(6,999)
999 FORMAT(2X,"THE NORMALIZED OBSERVATIONS")
    DO 25 I = 1,N
    XS(I) = (X(I) - XB) / SX
    YS(I) = (Y(I) - YB) / SY
    FINF(I) = XS(I)*YS(I) - COR*(XS(I)*XS(I) + YS(I)*YS(I)) / 2.
C THE NORMALIZED OBSERVATIONS ARE PRINTED HERE
    WRITE(6,990)I,XS(I),YS(I)
990 FORMAT(I5,5X,F10.5,5X,F10.5)
 25 CONTINUE
C THE CORRELATION COEFFICIENT IS PRINTED
    WRITE(6,700) COR
700 FORMAT(2X,"THE CORRELATION IS",F10.5)
C THE MEANS ARE PRINTED
    WRITE(6,800) XB,YB
800 FORMAT(2X,"XBAR = ",F15.5,5X,"YBAR = ",F15.5)
C THE STANDARD DEVIATIONS ARE PRINTED
    WRITE(6,900) SX,SY
900 FORMAT(2X,"SX = ",F15.5,2X,"SY = ",F15.5)
C THE INFLUENCE FUNCTION ESTIMATES ARE PRINTED
    WRITE(6,600)
600 FORMAT(2X,"THE INFLUENCE FUNCTION ESTIMATES")
    DO 26 I = 1,N
    WRITE(6,500)I,FINF(I)
500 FORMAT(2X,I3,5X,F10.5)
 26 CONTINUE
 20 CONTINUE
    END
```

Figure 15.3-1. continued

The first component of each observation is fuel consumption; the second component is electricity generation. Perform an influence analysis of this data set, with special attention to the correlation.

In Figure 15.3-1, we have a computer program for performing the calculations for the influence curve estimates, with the output given in Figure 15.3-2. [This program and example were first given by Chernick (1982).] A plot of these results

```
THE CASE NUMBER IS 2
THE NORMALIZED OBSERVATIONS

     1      3.12807      2.12473
     2      -.39759      -.67868
     3      -.45021      -.58247
     4      -.34496      -.54999
     5      -.39759      -.50865
     6      -.45021      -.57804
     7      1.91777      4.57701
     8      -.02923       .41606
     9      -.02923       .29943
    10      -.34496      -.37627
    11      -.39759      -.20894
    12       .23387       .96971
    13      -.45021      -.81106
    14      -.39759      -.33197
    15      -.29234      -.28276
    16       .18125       .90081
    17      -.34496      -.30737
    18      -.45021      -.55343
    19      -.39759      -.37627
    20       .07601       .73102
    21      -.29234      -.13758
    22      -.34496      -.34674
    23      4.28574      -.76505
    24      -.39759      -.39595
    25      -.18710       .06419
    26      -.39759      -.30491
    27      -.34496      -.35658
    28      -.34496      -.27538
    29      -.39759      -.37381
    30      -.29234      -.18926
    31       .39174      1.50120
    32      -.39759      -.32533
    33      -.45021      -.61003
    34      -.45021      -.54851
    35      -.34496      -.41318
    36      -.39759      -.39595

THE CORRELATION IS .48442
XBAR  =  11.55556     YBAR  =  3439.13889
  SX  =  19.00357       SY  =  4063.97810
```

Figure 15.3-2. Output of influence program (Figure 15.3-1) for data of Table 15.3-1.

is given in Figure 15.3-3, where the contours are those for influence $c = \pm 1$. From the influence figures in Figure 15.3-2, we know that three points will fall outside these contours, and that is, of course, confirmed in Figure 15.3-3.

In interpretation, note that high correlation is expected between consumption and generation, while in this case $r = .48442$ has been found (see Figure 15.3-2). The approximate effect of an observation on the correlation is the influence

```
THIS IS CASE NUMBER 2
THE INFLUENCE FUNCTION ESTIMATES
                    1        3.18290
                    2         .11998        19         .07702
                    3         .13097        20        -.07527
                    4         .08764        21         .01494
                    5         .10128        22         .06167
                    6         .13022        23       -7.86934
                    7        2.81283        24         .08117
                    8        -.05430        25        -.02149
                    9        -.03068        26         .06042
                   10         .06668        27         .06339
                   11         .03421        28         .04781
                   12        -.01421        29         .07649
                   13         .15672        30         .02595
                   14         .06701        31         .00507
                   15         .04260        32         .06542
                   16        -.04122        33         .13541
                   17         .05433        34         .12498
                   18         .12588        35         .07236
                                            36         .08117
```

Figure 15.3-2. continued

function divided by the sample size, hence observation number 23 (with influence function estimate of -7.86934) has decreased the r value by $-7.86934/36 = -.219$. Indeed, we find that if we remove observation number 23, then the correlation increase is even more dramatic, as r now changes to .8457.

In conclusion, we note that **robustness is a key topic in modern statistics, and that the methods studied in this chapter will become of increasing use in applications as they allow assessment of influence on the statistic of interest in the**

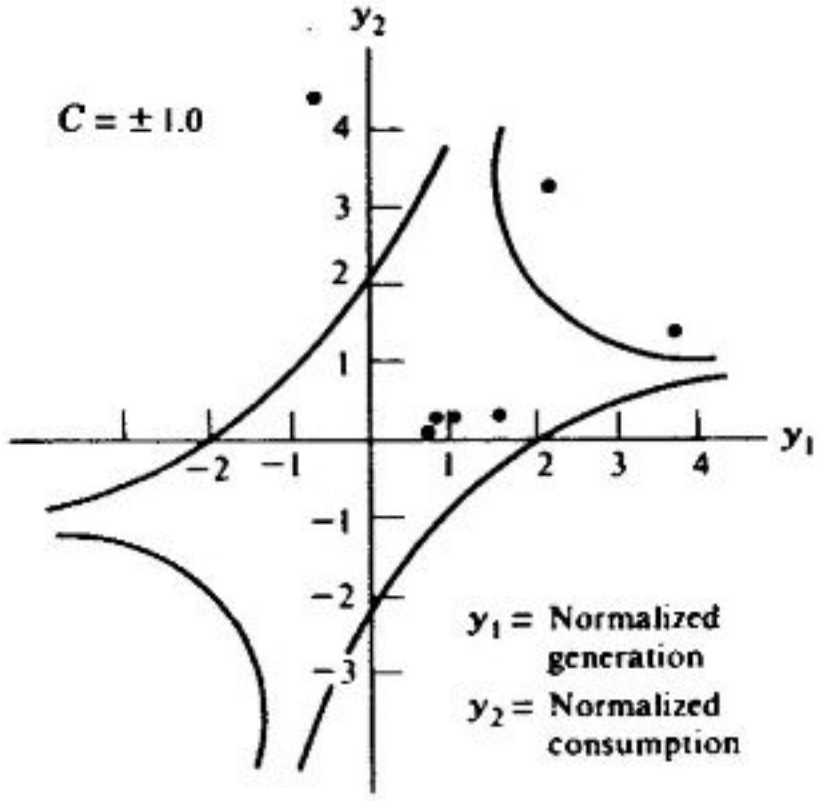

Figure 15.3-3. Scatterplot of normalized observations (data set of Table 15.3-1), with estimated influence function contours.

application (and here we have concentrated on the correlation). Extensions to multiple regression are studied by Chernick (1982).

PROBLEMS FOR SECTION 15.3

15.3.1 For the data (X_{1i}, X_{2i}), $i = 1, 2, \ldots, 7$, of Example 14.5.9, perform an analysis as in Example 15.3.1, seeking to detect very influential observations. Include a plot of estimated influence contours over a scatter plot of the bivariate data. What is r? Do you recommend this estimate be revised? (If so, why, and to what value?)

15.4. M-ESTIMATION

For many parametric families of distributions the maximum likelihood estimator (MLE) (studied in detail in Chapter 7) is a consistent and asymptotically minimum variance estimator of the parameters (see Section 7.4). However, there are cases of interest where the MLE is not consistent, and (in certain cases) the MLE either may not exist, or may not exist in a simple closed form (such as $\hat{\theta} = \bar{X}$). Also, if we do not know the exact distribution of the data (up to within some unknown parameters), then an MLE cannot be obtained. Even if we do know the exact distribution (up to within some unknown parameters), one wonders: "If the true distribution deviates from the assumed underlying distribution, how good is the MLE estimator?"

The idea of **M-estimation, generalizing the MLE**, was introduced by P. J. Huber in 1964. It seeks to have "good" properties of estimation under distributions "closely resembling" the underlying assumed distribution, that is, to be a **distributionally robust estimator**. Huber's original work concerned estimating the "center" of a distribution, and it is this aspect on which we will concentrate in this section.

As a motivating example, suppose we desire to estimate the mean θ of a normal population with known variance σ^2, based on a random sample $X_1, X_2, \ldots, X_n$ of size n. The maximum likelihood method seeks a θ that maximizes

$$(\sqrt{2\pi}\sigma)^{-n} \exp\left(-\sum(X_i - \theta)^2 \Big/ (2\sigma^2)\right),$$

or equivalently minimizes $\sum(X_i - \theta)^2$. **Another method of estimating** θ is to use the θ value that is closest to the sample values $X_1, X_2, \ldots, X_n$ in the sense of the Euclidean distance $\sum(X_i - E(X_i))^2$, that is, to find the θ value that minimizes $\sum(X_i - \theta)^2$ with respect to θ; this is known as **least squares estimation** [see equation (14.1.21)]. (Note that the MLE and LS estimators here both turn out to be $\hat{\theta} = \bar{X}$, but that is not the case in general.) Huber's **M-estimation** generalizes the Euclidean distance between the sample values and the expectation (which involves the parameter value) to a **general distance function** $d(\cdot)$ **with the**

properties:

$$d(x) \geq 0 \text{ for all } x,$$

$$d(x) = d(-x) \text{ for all } x,$$

$$d(0) = 0,$$

$$d(x) \text{ has a derivative } d'(x) \text{ at all } x,$$

and

$$d''(x) \geq 0$$

(i.e., $d(x)$ is a nondecreasing function of x). (Although it is assumed that $d(x)$ is differentiable for all x, later we will examine a case where the derivative fails to exist at a finite number of points x.)

Definition 15.4.1. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from d.f. $F(x|\theta)$, where θ is a location parameter, that is, $F(x|\theta) = F(x - \theta)$. The **M-estimator** for distance function $d(\cdot)$ is the value $T_n(X_1, X_2, \ldots, X_n)$ of θ that minimizes the **objective function** $\sum_{i=1}^{n} d(X_i - \theta)$, that is, the T_n that solves

$$\sum_{i=1}^{n} d'(X_i - T_n) = 0. \tag{15.4.2}$$

Example 15.4.3. **(a)** If $d(x) = x^2$, or if $d'(x) = x$, and if the p.d.f. is symmetric about θ, the M-estimator equals the least squares estimator of θ. **(b)** If $d'(x) = f'(x)/f(x)$, then the M-estimator equals the MLE of θ, since then $d(x) = \ln(f(x))$.

Note that **equation (15.4.2) can be written as**

$$\sum_{i=1}^{n} w_i(X_i - T_n) = 0 \tag{15.4.4}$$

where

$$w_i = \frac{d'(X_i - T_n)}{X_i - T_n}. \tag{15.4.5}$$

Thus,

$$T_n = \frac{\sum w_i X_i}{\sum w_i}, \tag{15.4.6}$$

so that **the M-estimator T_n of θ can be thought of as a weighted average of the sample values, with the weights depending on the data.**

In Example 15.4.3 we noted how different choices of $d(x)$ (or $d'(x)$) lead to different estimators. We next examine some additional choices of $d(x)$ and their resulting estimators.

Example 15.4.7. Consider minimizing $\sum_{i=1}^{n} d(X_i - \theta)$ where

$$d(x) = \begin{cases} x^2 & \text{if } |x| \le k \\ k^2 & \text{if } |x| > k \end{cases}$$

for some $k > 0$. Here

$$d'(x) = \begin{cases} 2x & \text{if } |x| \le k \\ 0 & \text{if } |x| > k, \end{cases}$$

hence solving

$$\sum_{i=1}^{n} d'(X_i - T_n) = 0$$

yields

$$\begin{aligned} T_n &= (\text{Average of those of } X_1, X_2, \ldots, X_n \text{ for which } |X_i| < k) \\ &= \overline{X}_{n,k} \text{ (say)}. \end{aligned}$$

Note that this estimator discards all sample values that are less than $-k$ or greater than k, and then averages those that are left. Such an average is called a **trimmed mean**. In practice one might choose k so that the upper $100p\%$ and lower $100p\%$ of the observations would be trimmed from the data. By such trimming one hopes to remove any extreme observations so that the estimator may be less sensitive to outliers.

Note that, **in Example 15.4.7, the function $d(x)$ used was not differentiable at $x = k$**, thus technically violating the properties the function $d(\cdot)$ was assumed to have. However, if we are dealing with a continuous case (so that $P(X_i = k) = 0$), or a discrete case where k is not a possible value of X_i (so that again $P(X_i = k) = 0$), then clearly the value of $d'(\cdot)$ at k will have no effect on the estimator (since the probability is zero that an $X_i = k$ will be encountered). Thus, in this case (and many similar cases) it is reasonable to allow $d'(\cdot)$ to fail to exist at a finite number of points.

Example 15.4.8. Consider minimizing $\sum_{i=1}^{n} d(X_i - \theta)$ where

$$d(x) = \begin{cases} x^2/2, & \text{if } |x| \le k \\ k|x| - k^2/2, & \text{if } |x| > k \end{cases}$$

where $k > 0$ is a constant. Since $d(x)$ grows linearly as $x \to \pm\infty$, the M-estimator will be less sensitive to outliers. Here we have

$$d'(x) = \begin{cases} -k, & \text{if } x < -k \\ x, & \text{if } |x| \le k \\ k, & \text{if } x > k, \end{cases}$$

and solving

$$\sum_{i=1}^{n} d'(X_i - T_n) = 0$$

yields (see (15.4.6))

$$T_n = \frac{\sum w_i X_i}{\sum w_i} \tag{15.4.9}$$

with

$$w_i = \frac{d'(X_i - T_n)}{X_i - T_n} = \begin{cases} 1, & \text{if } |X_i - T_n| \le k \\ k/|X_i - T_n|, & \text{if } |X_i - T_n| > k \end{cases}$$

which decreases as X_i increases.

In Example 15.4.8 the weights assigned to outliers decrease (but do not abruptly drop to zero as in the case of Example 15.4.7), yielding what is called a **Winsorized mean**. To find a p-Winsorized mean based on a random sample of size n, we let $k = [np]$ with $0 < p < 0.5$, let $Y_1 \le Y_2 \le \cdots \le Y_n$ denote the order statistics, and then (with $[y]$ denoting the largest integer $\le y$) set

$$T_n = \frac{(k+1)Y_{k+1} + Y_{k+2} + \cdots + Y_{n-k-1} + (k+1)Y_{n-k}}{n}. \tag{15.4.10}$$

As a final example, we consider the distance function $d(x) = |x|$.

Example 15.4.11. Consider minimizing $\sum_{i=1}^{n} d(X_i - \theta)$ where

$$d(x) = |x|. \tag{15.4.12}$$

Here we have

$$d'(x) = \operatorname{sgn}(x)$$

where

$$\operatorname{sgn}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0. \end{cases}$$

(Actually, $d(x)$ is not differentiable at $x = 0$, but, as in the discussion following Example 15.4.7, this does not matter as long as the probability $X_i = 0$ is zero.) To find the M-estimator we solve

$$\sum d'(X_i - T_n) = \sum \operatorname{sgn}(X_i - T_n) = 0.$$

Since this is a sum of $\operatorname{sgn}(X_i - T_n)$, the sum counts $+1$ for each observation above T_n and -1 for each observation below T_n, hence $T_n = \text{Median}(X_1, X_2, \ldots, X_n)$ yields a sum of zero. If n is even, T_n is any point between the two middle ordered observations; while, if n is odd, T_n is the middle observation. Thus, T_n **is the median**, which we have previously seen is in some senses not as sensitive to extreme observations as is the mean.

Additional sources for further reading on M-estimation include Huber (1981); Hoaglin, Mosteller, and Tukey (1983); and Serfling (1980). Some additional key results are:

Influence Curve for M-estimators. Huber (1981), p. 45, showed that for an M-estimator the influence curve at x (see Definition 15.1.3) is

$$w(x) = \frac{d'(x-\theta)}{Ed'(X_i - \theta)} = \frac{d'(x-\theta)}{-\dfrac{d}{d\theta} Ed'(X_i - \theta)}. \tag{15.4.13}$$

Thus, the influence curve of an M-estimator is proportional to $d'(\cdot)$ (since the denominator is a constant). Therefore the desired properties of an influence curve may be achieved by choosing a d'-function with those desired properties.

Limiting Distribution of M-estimators. Under conditions, $T_n \to \theta$ w.p. 1, and

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \tau^2) \tag{15.4.14}$$

where

$$\tau^2 = \frac{E(d'(X_i - \theta))^2}{(Ed''(X_i - \theta))^2}.$$

PROBLEMS FOR SECTION 15.4

15.4.1 Show that in equation (15.4.13), $E(w(X_i)) = 0$.

15.4.2 Draw the influence curve of the M-estimator in Example 15.4.7. Find the limiting distribution of T_n if the parent distribution is

(a) Uniform on $[0, \theta]$, so that $f(x|\theta) = (1/\theta) \cdot I_{[0,\theta]}(x)$.

(b) A symmetric triangular distribution with median at θ.

15.4.3 Suppose a random sample of size 10 from a $N(\theta, 1)$ distribution yields (in some order) the data {1.2, 2.3, 3.4, 3.6, 3.8, 4.1, 4.8, 4.9, 5.0, 5.8}. Compute the estimate of θ discussed in

(a) Example 15.4.7 with $p = .1$.

(b) Example 15.4.8 with $p = .2$.

(c) Example 15.4.11.

(d) Compute the MLE of θ.

15.5. THE JACKKNIFE ESTIMATOR

The **jackknife estimator** was introduced by Quenouille in 1949 and named by Tukey in 1958. This technique's purpose is **to decrease the bias of an estimator and provide an approximate confidence interval** for the parameter of interest.

As discussed in Chapters 7 and 8, for many d.f.'s the MLE and a UMVUE exist. If a parameter has a UMVUE associated with it, then clearly there is no chance of improving such an estimator's bias (since it is already zero). However, MLE's are often biased, and hence improvement may be possible in the sense of an estimator with lower bias. Jackknifing is an important technique for accomplishing such bias reduction. The name "jackknife" was coined due to the fact that, like a Scout's trusty jackknife, it is rough and ready in all situations. (To some this terminology seems tacky, and even to obscure matters; however, the terminology is well-established and will therefore be used here.)

The procedure operates as follows. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with real-valued parameter θ. Let $\hat{\theta}$ be an estimator of θ. Divide the random sample into N groups of equal size $m = n/N$ observations each (where, of course, N is one of the factors of n). Delete one group at a time, and estimate θ based on the remaining $(N-1)m$ observations, using the same estimation procedure previously used with a sample of size n. Denote the estimator of θ obtained with the ith group deleted by $\hat{\theta}_i$, called a **jackknife statistic** ($i = 1, 2, \ldots, N$). For $i = 1, 2, \ldots, N$, form **pseudovalues**

$$J_i = N\hat{\theta} - (N-1)\hat{\theta}_i, \tag{15.5.1}$$

and consider

$$J(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left(N\hat{\theta} - (N-1)\hat{\theta}_i\right) = N\hat{\theta} - (N-1)\bar{\hat{\theta}}_i \tag{15.5.2}$$

where

$$\bar{\hat{\theta}}_i = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i.$$

Definition 15.5.3. $J(\hat{\theta})$ is called the **jackknife estimator of** θ.

Remark 15.5.4. Note that the jackknife estimator can be written as

$$J(\hat{\theta}) = \hat{\theta} + (N-1)\left(\hat{\theta} - \bar{\hat{\theta}}_i\right),$$

which shows the estimator $J(\hat{\theta})$ as an adjustment to $\hat{\theta}$, with the amount of adjustment depending on the difference between $\hat{\theta}$ and $\hat{\theta}_i$. The special case $m = 1$ **is the most commonly used jackknife**, in which case

$$J(\hat{\theta}) = n\hat{\theta} - (n-1)\bar{\hat{\theta}}_i. \tag{15.5.5}$$

Example 15.5.6. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a $N(\theta, 1)$ distribution. Then $\hat{\theta} = \bar{X}$. With $m = 1$ as in Remark 15.5.4 and equation (15.5.5),

$$\hat{\theta}_i = \frac{1}{n-1}\sum_{j\neq i} X_j, \qquad \bar{\hat{\theta}}_i = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i = \bar{X},$$

and hence $J(\hat{\theta}) = n\bar{X} - (n-1)\bar{X} = \bar{X}$. Thus, the jackknife estimator is still $\bar{X}$ (which is desirable since $\bar{X}$ is a UMVUE of θ).

Example 15.5.7. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ distribution with μ and σ^2 both unknown. Consider estimating σ^2. The MLE of σ^2 is $\hat{\theta} = \hat{\sigma}^2 = \Sigma(X_i - \bar{X})^2/n = \Sigma X_i^2/n - \bar{X}^2$, which is biased. Taking $m = 1$ (see Remark 15.5.4), we find

$$\hat{\sigma}_i^2 = \frac{1}{n-1}\sum_{j\neq i} X_j^2 - \left(\frac{1}{n-1}\sum_{j\neq i} X_j\right)^2 = \frac{1}{n-1}\sum_{j\neq i} X_j^2 - \left(\frac{n\bar{X} - X_i}{n-1}\right)^2$$

$$= \frac{1}{n-1}\sum_{j\neq i} X_j^2 - \frac{n^2\bar{X}^2 - 2nX_i\bar{X} + X_i^2}{(n-1)^2},$$

hence

$$\bar{\hat{\sigma}}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}_i^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} X_j^2 - \frac{n^2\bar{X}^2 - 2n\bar{X}^2 + \sum_{i=1}^{n} X_i^2}{(n-1)^2}$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{(n^2-2n)\bar{X}^2}{(n-1)^2} - \frac{1}{n(n-1)^2}\sum_{i=1}^{n} X_i^2$$

$$= \frac{(n-1)^2 - 1}{n(n-1)^2}\sum_{i=1}^{n} X_i^2 - \frac{(n^2-2n)\bar{X}^2}{(n-1)^2} = \frac{n(n-2)}{(n-1)^2}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)$$

$$= \frac{n(n-2)}{(n-1)^2}\hat{\sigma}^2$$

and **the jackknife variance estimator is**

$$J(\hat{\sigma}^2) = n\hat{\sigma}^2 - (n-1)\bar{\hat{\sigma}}_i^2 = n\hat{\sigma}^2 - \frac{n(n-2)}{n-1}\hat{\sigma}^2$$

$$= \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which (as seen at (7.1.13)) is unbiased for σ^2. Thus, the bias of $\hat{\sigma}^2$ has not only been reduced, but an unbiased estimator has been found.

In the next example, the range of possible values of the X_i's depends on θ.

Example 15.5.8. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the displaced exponential p.d.f. $e^{-(x-\theta)} \cdot I_{[\theta,\infty)}(x)$. Let $Y_1 \le Y_2 \le \cdots \le Y_n$ be the order statistics. Note that

$$E(Y_1) = \theta + \frac{1}{n}, \qquad E(Y_2) = \theta + \frac{2n-1}{n(n-1)}.$$

The MLE of θ is Y_1, which is a biased estimator. With $m = 1$ as in Remark 15.5.4,

$$\hat{\theta}_i = \begin{cases} Y_1 & \text{if } X_i \neq Y_1 \\ Y_2 & \text{if } X_i = Y_1, \end{cases}$$

and

$$\bar{\hat{\theta}}_i = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i = \frac{1}{n}((n-1)Y_1 + Y_2),$$

hence the jackknife estimator is

$$J(\hat{\theta}) = n\hat{\theta} - (n-1)\tilde{\hat{\theta}}_i = nY_1 - (n-1)\left(\frac{n-1}{n}Y_1 + \frac{1}{n}Y_2\right)$$

$$= nY_1 - \frac{(n-1)^2}{n}Y_1 - \frac{n-1}{n}Y_2 = Y_1 - \frac{n-1}{n}(Y_2 - Y_1). \tag{15.5.9}$$

Since

$$E(J(\hat{\theta})) = E(Y_1) + \frac{n-1}{n}(E(Y_1) - E(Y_2))$$

$$= \theta + \frac{1}{n} + \frac{n-1}{n}\left(\theta + \frac{1}{n} - \theta - \frac{2n-1}{n(n-1)}\right) = \theta,$$

$J(\hat{\theta})$ is unbiased for θ.

Our final example is for a discrete case.

Example 15.5.10. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n with $P(X_1 = 0) = 1 - p$, $P(X_1 = 1) = p$. Let $X = X_1 + X_2 + \cdots + X_n$. Here $\hat{p} = X/n$ = (Number of successes)/(Number of trials) is the MLE of p. Hence (see Theorem 7.2.15), the MLE of $\theta = pq$ is

$$\hat{\theta} = \hat{p}\hat{q} = \frac{X}{n}\left(1 - \frac{X}{n}\right)$$

and

$$E(\hat{\theta}) = pq - \frac{pq}{n} = \frac{n-1}{n}pq,$$

so $\hat{p}\hat{q}$ is biased for pq.

To find the jackknife estimator, let $m = 1$ as in Remark 15.5.4. Then

$$\hat{p}_i = \begin{cases} \dfrac{X-1}{n-1}, & \text{if } X_i = 1 \text{ (i.e. if the } i\text{th trial is a success)} \\ \dfrac{X}{n-1}, & \text{if } X_i = 0 \text{ (i.e., if the } i\text{th trial is a failure),} \end{cases}$$

so

$$\hat{\theta}_i = \begin{cases} \left(\dfrac{X-1}{n-1}\right)\left(1 - \dfrac{X-1}{n-1}\right) = \left(\dfrac{X-1}{n-1}\right)\left(\dfrac{n-X}{n-1}\right) & \text{if } X_i = 1 \\ \dfrac{X}{n-1}\left(1 - \dfrac{X}{n-1}\right) = \dfrac{X}{n-1}\left(\dfrac{n-X-1}{n-1}\right) & \text{if } X_i = 0, \end{cases}$$

and (since there are X successes and $n - X$ failures to be removed)

$$\bar{\hat{\theta}}_i = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i = \frac{1}{n}\left(X\left(\frac{X-1}{n-1}\right)\left(\frac{n-X}{n-1}\right) + (n-X)\left(\frac{X}{n-1}\right)\left(\frac{n-X-1}{n-1}\right)\right)$$

$$= \frac{n(n-2)}{(n-1)^2}\hat{p}\hat{q}.$$

Therefore the jackknife estimator of pq is

$$J(\hat{\theta}) = n\hat{\theta} - (n-1)\bar{\hat{\theta}}_i = n\hat{p}\hat{q} - \frac{n(n-2)}{n-1}\hat{p}\hat{q} \qquad (15.5.11)$$

$$= \frac{n}{n-1}\hat{p}\hat{q}.$$

Note that $E(J(\hat{\theta})) = pq$, and $J(\hat{\theta})$ is unbiased and UMVUE for $\theta = pq$.

The **jackknife estimator of θ of Definition 15.5.3 can be generalized as follows.**

Definition 15.5.12. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of θ, and let $R \neq 1$ be a constant. The **generalized jackknife estimator** $G(\hat{\theta}_1, \hat{\theta}_2)$ is defined as

$$G(\hat{\theta}_1, \hat{\theta}_2) = \frac{\hat{\theta}_1 - R\hat{\theta}_2}{1 - R}. \tag{15.5.13}$$

Remark 15.5.14. If we choose $R = (N-1)/N$ in equation (15.5.13), and also $\hat{\theta}_1 = \hat{\theta}$, $\hat{\theta}_2 = \hat{\theta}_i$, then we obtain the jackknife estimator of equation (15.5.2). Therefore the name "generalized jackknife estimator" is valid (since the new estimator does generalize the previous jackknife estimation procedure). We will not consider the generalized jackknife further in this chapter, but will touch on *confidence intervals for* θ. Also, note that although all the preceding examples are in parametric settings the jackknife can be extended to bias reduction in nonparametric settings as well.

Use of the jackknife estimator to obtain approximate confidence intervals (and tests) for θ is the final jackknife technique we will consider. Tukey (one of the first to deal with this topic) suggested that the pseudovalues J_i of equation (15.5.1) be treated as independent and identically distributed r.v.'s. Then (with $m = 1$ as in Remark 15.5.4)

$$t = \frac{\sqrt{n}\left(J(\hat{\theta}) - \theta\right)}{\left(\sum_{i=1}^{n} \frac{\left(J_i - J(\hat{\theta})\right)^2}{n-1}\right)^{0.5}} = \frac{\sqrt{n}\left(J(\hat{\theta}) - \theta\right)}{\sqrt{\sigma^2_{J(\theta)}}} \tag{15.5.15}$$

is approximately a Student's t r.v., which (for large n) converges to a standard normal distribution. [For proofs under various conditions, see Schucany and Gray (1972).] Therefore an approximate $100(1-\alpha)\%$ confidence interval for θ is

$$\left(J(\hat{\theta}) - \Phi^{-1}(1-\alpha/2)\sqrt{\sigma^2_{J(\theta)}/n}\,, \qquad J(\hat{\theta}) + \Phi^{-1}(1-\alpha/2)\sqrt{\sigma^2_{J(\theta)}/n}\right). \tag{15.5.16}$$

Example 15.5.17. A survey was conducted in Mobile, Alabama, to estimate the percentage of households that have at least one videocassette recorder (VCR). A sample of 200 households from West Mobile yielded 75 homes with at least one VCR. Construct a 95.44% confidence interval for pq.

Referring to Example 15.5.10 for $J(\widehat{pq})$, we use this in conjunction with (15.5.16), and a large-sample approximation uses $\Phi^{-1}(.9772) = 2.0$. Here $n = 200$,

$X = 75$, and

$$J_i = \begin{cases} n\hat{p}\hat{q} - (n-1)\left(\dfrac{X-1}{n-1}\right)\left(\dfrac{n-X}{n-1}\right) \\ \qquad = 200\left(\dfrac{75}{200}\right)\left(\dfrac{125}{200}\right) - 74\left(\dfrac{125}{199}\right) = .3925879397 & \text{if } X_i = 1 \\ n\hat{p}\hat{q} - (n-1)\left(\dfrac{X}{n-1}\right)\left(\dfrac{n-X-1}{n-1}\right) \\ \qquad = 200\left(\dfrac{75}{200}\right)\left(\dfrac{125}{200}\right) - 75\left(\dfrac{124}{199}\right) = .1413316583 & \text{if } X_i = 0. \end{cases}$$

Therefore

$$J(\hat{\theta}) = \frac{n}{n-1}\hat{p}\hat{q} = \frac{200}{199}\left(\frac{75}{200}\right)\left(\frac{125}{200}\right) = .2355527638,$$

and

$$\sigma^2_{J(\theta)} = \sum_{i=1}^{n} \left(J_i - J(\hat{\theta})\right)^2 \Big/ (n-1)$$

$$= \left(X(.3926 - J(\hat{\theta}))^2 + (n-X)(.1413 - J(\hat{\theta}))^2\right)\Big/(n-1)$$

$$= \left(75(.3926 - .2356)^2 + 125(.1413 - .2356)^2\right)\Big/199 = .014066552.$$

Hence the approximate 95.44% confidence interval for pq is

$$\left(.2356 - 2.0\sqrt{.014/200}\,,\quad .2356 + 2.0\sqrt{.014/200}\,\right) = (.218, .253).$$

PROBLEMS FOR SECTION 15.5

15.5.1 Let $X_1, X_2, \ldots, X_n$ be independent uniform r.v.'s on $(0, \theta)$. Find $J(\hat{\theta})$ for estimation of θ. Compare the biases of the MLE of θ and of $J(\hat{\theta})$.

15.5.2 Let $X_1, X_2, \ldots, X_n$ be independent r.v.'s with $P(X_i = 0) = 1 - p$, $P(X_i = 1) = p$. Find $J(p^2)$ for estimation of $\theta = p^2$. Compare the biases of the MLE of θ and of $J(p^2)$.

15.5.3 Let $X_1, X_2, \ldots, X_n$ be independent r.v.'s with p.d.f. $f(x|\alpha) = x^{\alpha-1}e^{-x}/\Gamma(\alpha)$ if $x > 0$ (and $= 0$ otherwise). Find $J(\hat{\alpha})$ for estimation of $\theta = \alpha$. Compare the biases of the MME of α and of $J(\hat{\alpha})$. [The MLE of α is difficult to find.]

15.5.4 Construct an approximate 95% confidence interval for p^2 based on $n = 200$, $X = 75$, and the estimator developed in the setting of Problem 15.5.2.

15.6. THE GENERAL STATISTICAL SIMULATION PROCEDURE (SSP) AND BOOTSTRAP METHODS

If we know the distribution function $F(\cdot)$ of a random variable X and wish to evaluate some function of it, say $\theta(F)$, we can proceed in two ways: (1) evaluate $\theta(F)$ exactly; and (2) simulate to estimate $\theta(F)$.

Example 15.6.1. Suppose that X is a r.v. which is $N(0,1)$ so that $F(x) = \Phi(x)$, and that we wish to know the 8th moment of X, so that $\theta(F) = \int_{-\infty}^{\infty} x^8 \phi(x)\, dx$. Then **one way we may proceed is to (1) evaluate the integral exactly.**

If the integral involved in the preceding method is such that no simple way to evaluate it exists, an alternate method may be sought. **A second way we may proceed is (2) simulate to estimate $\theta(F)$.** From Example 4.7.11, we know how to generate (on a computer) r.v.'s $X_1, X_2, \ldots, X_n$ that are $N(0,1)$ and independent. From those we can estimate $\theta(F)$ using the principles of estimation in Chapters 7 and 8. Thus, we might estimate $\theta(F)$ by $(X_1^8 + X_2^8 + \cdots + X_n^8)/n$, which is a consistent estimator (as shown in Theorem 7.1.12, where it was called m_8').

The preceding discussion assumed that we knew the d.f. $F(\cdot)$, and that is sometimes the case. However, **often we have** not a known d.f. $F(\cdot)$ for which we wish to know $\theta(F)$, but rather **a random sample $X_1, X_2, \ldots, X_n$ drawn from d.f. $F(\cdot)$ with $F(\cdot)$ unknown, and wish to estimate $\theta(F)$.** Since in settings like that of Example 15.6.1 we used the known $F(\cdot)$ only to draw the random sample, and we now have a random sample from $F(\cdot)$ given to use, we can proceed as in that example's part (2)—and with even greater simplicity since we do not need to generate $X_1, X_2, \ldots, X_n$, but they are instead given to us.

Now suppose we are taking approach (2), and wish to specify the variance of our estimator. (For example, making the variance of the estimator suitably small is one rational way to choose the sample size n.) With $F(\cdot)$ known, we can proceed as follows to solve this problem:

Step 1. Generate $X_1, X_2, \ldots, X_n$ and estimate $\theta(F)$; call the estimate $\hat{\theta}_1$.

Step 2. Generate $X_{n+1}, X_{n+2}, \ldots, X_{2n}$ and estimate $\theta(F)$ from these new r.v.'s (which are to be independent of all r.v.'s previously generated); call the estimate $\hat{\theta}_2$.

Step 3. Generate $X_{2n+1}, X_{2n+2}, \ldots, X_{3n}$ and estimate $\theta(F)$ from these new n r.v.'s (which are to be independent of all previously generated r.v.'s); call the estimate $\hat{\theta}_3$.

⋮

Step N. Generate $X_{(N-1)n+1}, X_{(N-1)n+2}, \ldots, X_{Nn}$ and estimate $\theta(F)$ from these new n r.v.'s (which are to be independent of all previously generated r.v.'s); call the estimate $\hat{\theta}_N$.

Then $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_N$ are N independent and identically distributed r.v.'s, each estimating $\theta(F)$. Their variance may be estimated by

$$\hat{\sigma}^2 = \sum_{i=1}^{N} (\hat{\theta}_i - \bar{\theta})^2 / (N-1), \qquad \bar{\theta} = (\hat{\theta}_1 + \hat{\theta}_2 + \cdots + \hat{\theta}_N)/N,$$

and the variance of the estimator $\bar{\theta}$ is estimated by $\hat{\sigma}^2/N$. We call this procedure the **Statistical Simulation Procedure SSP(θ, F, n, N).**

If $F(\cdot)$ is unknown, the SSP(θ, F, n, N) cannot be used. However, then one will have a random sample $X_1, X_2, \ldots, X_n$ taken from the unknown d.f. $F(\cdot)$. Using the random sample, the d.f. may be estimated, say by some estimator $\hat{F}$, and then Step 1, Step 2, ..., Step N followed (with all r.v.'s being generated from the estimated d.f. $\hat{F}$). We call this the General **Statistical Simulation Procedure SSP($\theta, \hat{F}, X_1, \ldots, X_n, N$).**

There are many **ways to choose the estimate $\hat{F}$ of F** in the SSP (θ, $\hat{F}$, $X_1, \ldots, X_n$, N). **One of the very simplest is to take $\hat{F}$ to be the empiric d.f.** based on the sample $X_1, \ldots, X_n$. If that is done, then the procedure is called the **bootstrap procedure**, and any similar methods are called **bootstrap methods**. Bootstrap methods essentially rely on sampling the empiric d.f., which is (unless n is large) a rough step-function estimate of the true $F(x)$, which is often known to be continuous. For this reason, **a smoothed version, such as the empiric p.d.f. (see Section 4.8), which is also easy to sample from, should yield more reliable estimates.** (State-of-the-art practice in simulation is to sample from a smoothed d.f. or p.d.f. estimate that incorporates all knowledge the experimenter has about the true d.f.)

Example 15.6.2. Estimating the Standard Error of $\bar{X}$. Let $\theta = E(X)$ and $\sigma^2 = \mathrm{Var}(X)$. Then from a random sample $X_1, X_2, \ldots, X_n$ with the same d.f. as X, we find $\bar{X}$ (the sample mean) and it has mean θ and $\mathrm{Var}(\bar{X}) = \sigma^2/n$.

The bootstrap method of estimating $\mathrm{Var}(\bar{X})$ is the SSP(θ, $\hat{F}$, $X_1, \ldots, X_n$, N) with $\hat{F}$ taken to be the empiric d.f., and **proceeds as follows:**

Step 1. Take a sample of size n (with replacement) from $\{X_1, X_2, \ldots, X_n\}$, say

$$\{X_{11}, X_{12}, X_{13}, \ldots, X_{1n}\}, \text{ and calculate its sample mean } \bar{X}_1.$$

Step 2. Repeat Step 1 independently $N-1$ additional times, finding $\bar{X}_2, \ldots, \bar{X}_N$. The bootstrap estimate of $\mathrm{Var}(\bar{X})$ is

$$\hat{\sigma}^2 = \sum_{i=1}^{N} (\bar{X}_i - \bar{X}_\cdot)^2 / (N-1), \qquad \bar{X}_\cdot = (\bar{X}_1 + \bar{X}_2 + \cdots + \bar{X}_N)/N. \quad (15.6.3)$$

Example 15.6.4. Estimating Bias. Suppose that, based on a random sample $X_1, X_2, \ldots, X_n$, some quantity θ of interest is estimated by $\hat{\theta}$. The estimator $\hat{\theta}$ has some bias $b = E(\hat{\theta} - \theta)$. To estimate the bias, consider use of SSP

$(b, \hat{F}, X_1, \ldots, X_n, N)$ with $\hat{F}$ taken to be the empiric d.f. (so we have a bootstrap estimate). Based on N bootstrap samples of size n each, one finds the estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N$, with $\bar{\theta} = (\hat{\theta}_1 + \hat{\theta}_2 + \cdots + \hat{\theta}_N)/N$, and estimates the bias of $\hat{\theta}$ by

$$b = \bar{\theta} - \hat{\theta}. \tag{15,6.5}$$

Example 15.6.6. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a Poisson distribution with unknown mean λ. If the parameter of interest is $\theta = P(X \leq 1) = e^{-\lambda}(1 + \lambda)$, the MLE is $e^{-\bar{X}}(1 + \bar{X})$, which is biased. To reduce the bias, let us investigate the bootstrap method.

Let X_{ij} $(i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, n)$ be the N bootstrap samples, that is, samples taken at random with replacement from $\{X_1, X_2, \ldots, X_n\}$, and (for $i = 1, 2, \ldots, N$)

$$\hat{\theta}_i = e^{-\bar{X}_i}(1 + \bar{X}_i) - (\text{Number of } X_{i1}, X_{i2}, \ldots, X_{in} \text{ that are } \leq 1)/n.$$

Then the bootstrap estimate of the bias of θ is

$$\bar{\theta} = \left(\hat{\theta}_1 + \hat{\theta}_2 + \cdots + \hat{\theta}_N\right)/N, \tag{15.6.7}$$

i.e. the associated estimate of bias is simply

$$\hat{b} = \bar{\theta}. \tag{15.6.8}$$

Then, one might use $e^{-\bar{X}}(1 + \bar{X}) - \hat{b}$ to estimate θ.

Remark 15.6.9. Note that an approximate $100(1 - \alpha)\%$ confidence interval for θ can be constructed using bootstrap methods, as follows. If $\bar{\theta}$ is the bootstrap estimate of θ and $\hat{\sigma}^2$ its sample variance based on $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N$, for N large we take the interval

$$\left(\hat{\theta} - \Phi^{-1}(1 - \alpha/2)\hat{\sigma},\quad \hat{\theta} + \Phi^{-1}(1 - \alpha/2)\hat{\sigma}\right). \tag{15.6.10}$$

Note that we use the original estimate of θ (not the bootstrap estimate), and the bootstrap procedure has been used to provide us with an estimate of variability for $\hat{\theta}$.

Note that the exact same details apply to the more general Statistical Simulation Procedure SSP$(\theta, \hat{F}, X_1, \ldots, X_n, N)$, in which the only difference is what estimate of F is being sampled from.

PROBLEMS FOR SECTION 15.6

15.6.1 Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ d.f. with μ and σ^2 both unknown. Detail how to proceed in a bootstrap method for obtaining an estimate of $\theta = \Phi((2 - \mu)/\sigma)$.

15.6.2 The MLE for the θ of Problem 15.6.1 is $\hat{\theta} = \Phi((2 - \bar{X})/s_n)$, where $s_n^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$. Find a bootstrap estimator of the variance of $\hat{\theta}$, and use it to construct a $100(1 - \alpha)\%$ confidence interval for θ.

15.6.3 Let the sample in Problem 15.6.1 be {3.1, 2.9, 1.6, 3.4, 4.2, 1.08, 2.4, 3.0, 2.9, 2.5}. Take $N = 4$ random samples of size 10 each (bootstrap samples) and find the bootstrap estimate of θ described in Problem 15.6.1. Compare this with the MLE of Problem 15.6.2. Estimate the bias of each estimator.

15.6.4 Based on the bootstrap samples obtained in Problem 15.6.3, compute a numerical estimate of the variance of the MLE, and then construct a 95% (approximate) confidence interval for θ using
(a) The bootstrap estimate of θ.
(b) The MLE of θ.
[*Note*: Use the Student's t-distribution with $N - 1$ degrees of freedom in (15.6.10) since here N is not very large.]

15.6.5 As in Problem 15.6.3, with $N = 7$.

15.6.6 As in Problem 15.6.3, with $N = 10$.

15.6.7 As in Problem 15.6.3, but use the Statistical Simulation Procedure instead of the bootstrap and draw your samples from the empiric p.d.f. for the data set given.

PROBLEMS FOR CHAPTER 15

15.1 As in Problem 15.4.2, but for Example 15.4.8.

15.2 As in Problem 15.4.2, but for Example 15.4.11.

15.3 Let $X_1, X_2, \ldots, X_n$ be independent Poisson r.v.'s, each with mean λ. Find $J(\hat{\theta})$ for estimation of $\theta = \lambda^2$. Compare the biases of the MLE of λ^2 and of $J(\hat{\theta})$.

15.4 Develop approximate $100(1 - \alpha)\%$ confidence limits for λ^2 based on the jackknife estimator developed in Problem 15.3.

15.5 Let $X_1, X_2, \ldots, X_n$ be a random sample of size n with $P(X_i = 0) = 1 - p$, $P(X_i = 1) = p$. A jackknife technique for constructing a confidence interval for pq was given in Section 15.5. Use bootstrap techniques to develop a $100(1 - \alpha)\%$ confidence interval for pq.

15.6 Take a random sample of size 10 from a uniform density on $(0, \theta)$ with $\theta = 8$. Then generate 50 bootstrap samples of size 10 each to estimate the variance of $\hat{\theta}$, the MLE of θ. Construct a 95% confidence interval for θ.