

---

# Dependent data

---

Empirical likelihood was motivated by independent identically distributed data. As [Theorem 4.1](#) shows, the requirement for identically distributed data can be relaxed. When the observations are dependent, then this usually has to be accounted for in constructing confidence regions and tests. The ways of handling dependent data with empirical likelihood parallel the methods from parametric likelihood and the bootstrap. Failure to account for dependence among the data can destroy the coverage properties of confidence regions.

In order to construct nonparametric confidence regions for dependent data, we must assume something about the nature of the dependence. For some time series problems, we assume that the dependent data are driven by an unobserved set of independent random variables. For some other time series, we assume that there is possibly very strong dependence between relatively few pairs of observations. By contrast, some finite population sampling settings have very weak dependence between many or even all pairs of observations.

## 8.1 Time series

Chapter 8.10 gives some background references on time series. Here we provide some definitions. A time series is a sequence of observations  $Y_i \in \mathbb{R}^d$ ,  $i = 1, \dots, T$ , where  $Y_{i+1}$  is observed one time unit after  $Y_i$ . The time unit could be a fixed amount of real time, such as a day or year, or it could simply indicate the order in which values were observed.

Models with independent  $Y_i$  are seldom appropriate for time series data. There is generally some dependence among series values to account for. Time series are usually modeled as realizations of stochastic processes in which  $(Y_1, \dots, Y_T)$  is drawn from a joint distribution on  $\mathbb{R}^{dT}$ .

For some joint distributions of  $Y_1, \dots, Y_T$ , there is clearly no way to learn about the underlying stochastic process. As an extreme example, suppose that  $Y_i = Z + e_i$  for all  $i \geq 1$ , where  $e_i$  are independent of each other and of the random variable  $Z$ . Such data are less informative about the mean of  $Y_i$  than is one single data value from the distribution of  $Z$ . Another hard case has  $Y_i = \mu_i + e_i$  where  $\mu_i$  is an arbitrary unknown sequence of values in  $\mathbb{R}^d$ , and  $e_i$  are independent with mean zero. Some assumptions are needed in order that the amount of information in the  $Y_i$  about the underlying process should increase with  $T$ .

A widely used assumption is that  $Y_i$  for  $i = 1, \dots, T$  are  $T$  consecutive obser-

vations from an infinite series  $\dots, Y_{-1}, Y_0, Y_1, \dots$  having a stationary distribution. This means that the joint distribution of any finite set of observations is unaffected by a time shift of  $k$  units. Thus  $(Y_r, Y_s, \dots, Y_t)$  has the same distribution as  $(Y_{r+k}, Y_{s+k}, \dots, Y_{t+k})$ , for any  $k$  and any  $r, s, \dots, t$ . A weaker assumption has some low order stationary moments. For example, in a real-valued time series there may be some set of  $a, b, \dots, c$  values for which  $E(Y_r^a \times Y_s^b \times \dots \times Y_t^c)$  is unaffected by a time shift of  $k$  units.

A stationarity assumption addresses the problem in the second hard case. A stationary series would have to have a common value  $\mu_i = \mu$ . The first hard case,  $Y_i = Z + e_i$ , is stationary if the  $e_i$  are IID, so another condition is needed to rule this case out.

Under a mixing condition, the dependence between observations before and including time  $t$  and observations from  $t + k$  onward becomes negligible as  $k \rightarrow \infty$ . For a rigorous description of mixing, see the references in Chapter 8.10. Let  $A$  be a random variable that takes the value 0 or 1 depending on what the series does at times up to time  $t$ . Let  $B$  be a 0 or 1 random variable depending on what the series does from time  $t + k$  on. It is natural to write  $\Pr(A)$  for  $E(A)$ , identifying  $A$  with the event that  $A = 1$ . If the future is independent of the past then  $\Pr(A \cap B) = \Pr(A) \Pr(B)$ . Now measure the dependence through

$$\alpha(k) = \sup_t \sup_{A, B} |\Pr(A \cap B) - \Pr(A) \Pr(B)|. \quad (8.1)$$

The  $Y_i$  are  $\alpha$ -mixing if  $\alpha(k) \rightarrow 0$  as  $k \rightarrow \infty$ . If  $Y_i$  is stationary, then it is not necessary to maximize over  $t$  in (8.1).

A mixing condition rules out the first hard case. The series  $Z + e_i$  is not  $\alpha$ -mixing outside of trivial cases. Theorems that use mixing usually also stipulate that  $\alpha(k)$  goes to zero sufficiently fast as  $k \rightarrow \infty$ .

The discussion above emphasizes the time domain approach to time series. In the frequency domain approach, we study how much of the variance in the series comes from oscillations at different frequencies. Suppose that  $Y_t \in \mathbb{R}$  is a stationary time series, with mean  $E(Y_t) = \mu$  and autocovariances  $\gamma_k = E((Y_t - \mu)(Y_{t+k} - \mu))$ . The spectral density function of  $Y_t$  is defined as

$$f(\omega) = \frac{1}{\pi} \left[ \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \cos(\omega k) \right], \quad 0 \leq \omega \leq \pi,$$

when this exists. The variance of  $Y_t$  is  $\gamma_0 = \int_0^\pi f(\omega) d\omega$  and the interpretation of  $f(\omega)$  is that frequencies in the interval  $[\omega_1, \omega_2]$  contribute  $\int_{\omega_1}^{\omega_2} f(\omega) d\omega$  of this variance.

The periodogram is a sample version of the spectral density function

$$I(\omega_j) = \frac{1}{\pi} \left[ \hat{\gamma}_0 + 2 \sum_{k=1}^{\infty} \hat{\gamma}_k \cos(\omega k) \right],$$

using estimates  $\hat{\gamma}_k = T^{-1} \sum_{i=1}^{T-k} (Y_i - \bar{Y})(Y_{i+k} - \bar{Y})$ . Under mild conditions,

$E(I(\omega)) \rightarrow f(\omega)$  as  $T \rightarrow \infty$ , but because  $\text{Var}(I(\omega_j))$  does not converge to 0 as  $T \rightarrow \infty$ , the periodogram is usually smoothed somehow, when an estimate of  $f(\omega)$  is required.

## 8.2 Reducing to independence

Parametric likelihood methods often tackle dependent data by expressing the observations as functions of some other variables assumed to be statistically independent. One approach to empirical likelihood is to use the estimating equations from those models.

The autoregressive model is widely used in parametric modeling of time series data. As the name describes, the data series is generated by a regression on its own past. Suppose for example that  $e_i \sim N(0, \sigma^2)$  are independent, that

$$Y_i = \beta_0 + \beta_1 Y_{i-1} + e_i, \quad i = 1, \dots, T, \quad (8.2)$$

and that the series started off with an unobserved normally distributed  $Y_0$  independent of the  $e_i$ . If  $|\beta_1| < 1$ , the distribution of  $Y_i$  tends to an equilibrium distribution  $N(\mu, \sigma_y^2)$  as  $i \rightarrow \infty$ , where  $\mu = \beta_0/(1 - \beta_1)$  and  $\sigma_y^2 = \sigma^2/(1 - \beta_1^2)$ . If  $Y_0 \sim N(\mu, \sigma_y^2)$ , then the  $Y_i$  all have the same distribution.

We will suppose that  $|\beta_1| < 1$  and then reparameterize equation (8.2) as

$$Y_i - \mu = \beta_1(Y_{i-1} - \mu) + e_i, \quad i = 1, \dots, T. \quad (8.3)$$

The parametric likelihood for the autoregressive model (8.3) is

$$\begin{aligned} L &= \prod_{i=1}^T f(Y_i | Y_1, \dots, Y_{i-1}; \mu, \beta_1, \sigma) \\ &= \frac{e^{-\frac{1}{2\sigma_y^2}(Y_1 - \mu)^2}}{\sqrt{2\pi\sigma_y}} \prod_{i=2}^T \frac{e^{-\frac{1}{2\sigma^2}((Y_i - \mu) - \beta_1(Y_{i-1} - \mu))^2}}{\sqrt{2\pi}\sigma}. \end{aligned} \quad (8.4)$$

The special treatment of  $Y_1$  in (8.4) is awkward. The conditional (on  $Y_1$ ) likelihood

$$\begin{aligned} L_c &= \prod_{i=2}^T f(Y_i | Y_1, \dots, Y_{i-1}; \mu, \beta_1, \sigma) \\ &= \prod_{i=2}^T \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}((Y_i - \mu) - \beta_1(Y_{i-1} - \mu))^2\right) \end{aligned} \quad (8.5)$$

treats the data more symmetrically. Using  $L_c$  instead of  $L$  sacrifices some of the information available from  $Y_1$ . This information loss is small, especially when  $|\beta_1|$  is close to 1. Furthermore, there is the possibility that the series has not yet reached the equilibrium distribution. Then  $L$  is not the likelihood but  $L_c$  is still the conditional likelihood given  $Y_1$ .

We can use the conditional likelihood to generate estimating equations. For

$i \geq 2$ , let  $e_i = e_i(\mu, \beta_1) = Y_i - \mu - \beta_1(Y_{i-1} - \mu)$ ,  $\theta = (\mu, \beta_1, \sigma)^t$  and

$$Z_i = Z_i(\theta) = (e_i, (Y_i - \mu)e_i, e_i^2 - \sigma^2)^t.$$

Then the estimating equations are  $\sum_{i=2}^T Z_i = 0$ .

The empirical likelihood approach is then based on

$$\mathcal{R}(\theta) = \sup \left\{ \prod_{i=1}^n w_i \mid w_i > 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i Z_{1+i} = 0 \right\},$$

where  $n = T - 1$ . If the  $e_i$  are independent  $N(0, \sigma^2)$  then the limiting distribution of  $-2 \log \mathcal{R}$  with  $r$  parameters constrained is  $\chi_{(r)}^2$ . The empirical likelihood inferences go through under some weaker conditions, using results for the dual likelihood described below. The  $e_i$  do not have to be normal, nor identically distributed. They do have to be nearly independent, so that  $(1/n) \sum_i Z_i Z_i'$  estimates the variance matrix of  $(1/\sqrt{n}) \sum_i Z_i$ .

It is convenient that inferences may be based on the limiting distribution of  $\log \mathcal{R}(\theta)$ , though it is troubling that in time series models  $\mathcal{R}$  is no longer a likelihood ratio. If  $\theta_0$  is the true value of the parameter then  $Z_i(\theta_0)$  are independent, but for  $\theta \neq \theta_0$   $Z_i(\theta)$  are not independent, and so it is hard to consider  $\prod_i w_i$  to be the probability of the observations. The dual likelihood is one way to explain why  $\mathcal{R}$  has likelihood asymptotics. Write

$$\mathcal{D}_\theta(\lambda) = \prod_{i=1}^n (1 + \lambda' Z_i(\theta))^{-1}.$$

For the correct value of  $\theta$ , the  $Z_i$  are independent and the test for  $\theta = \theta_0$  using  $\mathcal{R}$  is the same as the test for  $\lambda = 0$  using  $\mathcal{D}_\theta$ .

The autoregressive model (8.3) is known as the AR(1) model because it uses a regression on one past data point. In an AR( $k$ ) model we write

$$Y_i - \mu = \sum_{j=1}^k \beta_j (Y_{i-j} - \mu) + e_i. \quad (8.6)$$

The estimating equations for the AR( $k$ ) are a natural extension of those for AR(1). If an AR( $k$ ) series has uncorrelated  $e_i$  with mean 0 and constant variance, then it will approach an equilibrium distribution, under conditions on  $\beta_j$ . Let  $u_1, \dots, u_k$  be the solutions to

$$1 - \sum_{j=1}^k \beta_j u^j = 0. \quad (8.7)$$

The  $u_i$  are complex numbers, not necessarily all distinct. The AR( $k$ ) series approaches an equilibrium if and only if all  $u_j$  lie outside the unit circle in the complex plane.

We can form estimating equations for very general regressions, linear or nonlinear, relating  $Y_i$  to past values  $Y_{i-j}$  as well as past and present values of covariates

## St. Lawrence River flow

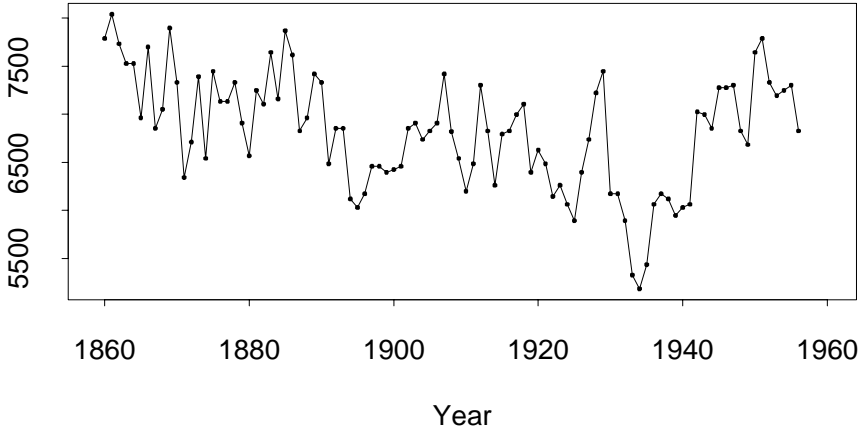


Figure 8.1 *Flow of the St. Lawrence river in cubic meters per second, at Ogdensburg, New York. The data are annual values from 1860 to 1957. Source: Yevjevich (1963).*

$X_i$ . An asymptotic  $\chi^2$  distribution for empirical likelihood inferences holds very generally for non-explosive series using dual likelihood.

For a model including predictors at lags of up to  $k$  time steps, we obtain  $n = T - k$  estimating function values to reweight. In software implementations, it can be a nuisance to have the number of estimating function values differ from the number of data values, or differ from model to model. A simple remedy is to define  $Z_i = 0$  for  $1 \leq i \leq k$ . It is easy to show that maximizing  $\prod_{i=1}^T \log(nw_i)$  subject to  $\sum_{i=1}^T w_i Z_i = 0$  and  $\sum_{i=1}^T w_i = 1$  places weight  $1/T$  on any  $Z_i$  that equals zero, and that the empirical likelihood ratio based on  $Z_{k+1}, \dots, Z_T$  does not change when  $Z_1 = \dots = Z_k = 0$  are adjoined to the sample.

Figure 8.1 shows annual flow of water in the St. Lawrence river. These values are clearly not independent. The correlation between one year's flow and the next is about 0.71.

We consider an AR(3) model for this data set. Let  $Y_1, \dots, Y_{97}$  be the raw values, and  $X_i = Y_i - \mu$  be centered values. We use estimating equations

$$Z_i = \begin{cases} (X_i, 0, 0, 0, 0)', & 1 \leq i \leq 3 \\ (X_i, e_i X_{i-1}, e_i X_{i-2}, e_i X_{i-3}, e_i^2 - \exp(2\tau))', & 4 \leq i \leq 97 \end{cases} \quad (8.8)$$

where  $e_i = X_i - \sum_{j=1}^3 \beta_j X_{i-j}$ . These describe an autoregression of  $Y_i$  on its past 3 lags, with an error standard deviation of  $\exp(\tau)$ . Instead of taking  $Z_1$  through  $Z_4$  equal to 0, the first component was modified slightly. It is customary in autore-

$j$	$\hat{\beta}_j$	$-2 \log \mathcal{R}_j(0)$
1	0.627	30.16
2	-0.093	0.48
3	0.214	4.05

Table 8.1 An AR(3) model was fit to the St. Lawrence River flow data. Shown are the coefficient estimates  $\hat{\beta}_j$ , and the empirical likelihood values for testing that  $\beta_j = 0$ .

gressive modeling to estimate  $\mu$  by  $\bar{Y} = (1/T) \sum_{i=1}^T Y_i$ . Neither the conditional nor the unconditional likelihood leads to  $\hat{\mu} = \bar{Y}$ , but equations (8.8) do.

The mean flow is estimated to be  $\hat{\mu} = 6818.6$  cubic meters per second. The standard deviation of  $e_i$  is estimated to be  $\exp(\hat{\tau}) = \exp(6.006) = 405.9$  cubic meters per second. This describes the uncertainty in a linear prediction of one year's river flow, based on the previous three years' data. Table 8.1 presents the estimated autoregressive coefficients as well as the empirical likelihood test statistics for each coefficient to be zero. The lag 1 coefficient  $\beta_1$  is clearly nonzero,  $\beta_2$  could reasonably be zero. A  $\chi^2_{(1)}$  test rejects  $\beta_3 = 0$  at just below the 5% level as does an  $F$  test.

It is interesting to consider whether  $\beta_2 = \beta_3 = 0$  is tenable. Imposing both constraints can only reduce the empirical likelihood compared to the test of  $\beta_3 = 0$  alone. This lower likelihood must, however, be compared to a distribution appropriate to a two-dimensional hypothesis. As is well known, omnibus tests that can detect multiple kinds of departure from a hypothesis often do so with reduced power compared to more specific tests. In this instance, a test of the hypothesis  $\beta_2 = \beta_3 = 0$  has  $p$ -value somewhat above 5%, while a test of  $\beta_3 = 0$  has a  $p$ -value below 5%.

Not being able to reject  $\beta_2 = \beta_3 = 0$  is not the same as proving that they are zero. We retain these coefficients in the model, judging that there is more to lose in dropping them should they matter than in retaining them if they do not. People can reasonably differ in these judgments. The model then gives a 95% confidence interval for  $\tau$  of (5.871, 6.134) using a  $\chi^2_{(1)}$  calibration. Exponentiating, we get a confidence interval of (354.6, 461.3) for  $\sigma$ .

### 8.3 Blockwise empirical likelihood

We do not always know a model in which the data are generated from a series of independent observations. A weaker assumption is that the data have a stationary distribution, or stationary moments, as described in Chapter 8.1. Stationarity alone cannot support a good asymptotic theory. An additional condition, such as one on the  $\alpha$ -mixing coefficients described in Chapter 8.1, is required.

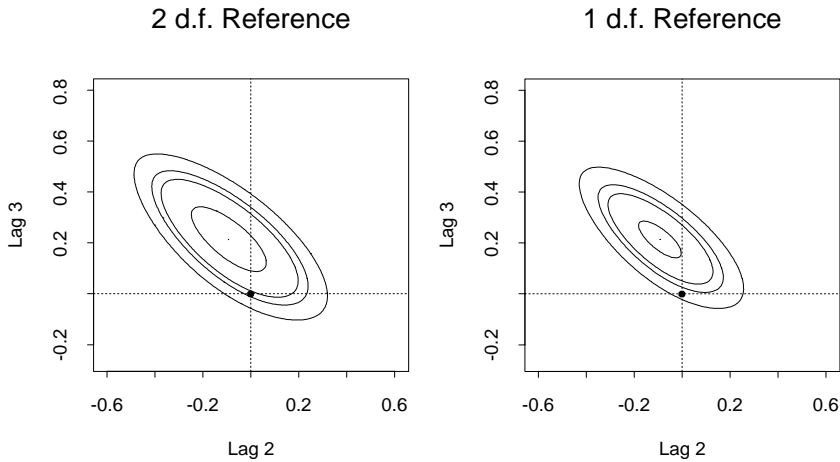


Figure 8.2 Both figures show empirical likelihood contours for two of the autoregressive parameters for the St. Lawrence river flows. The figure on the left uses two degrees of freedom, appropriate for two parameters. The contour levels correspond to 50, 90, 95, and 99 percent confidence. The solid circle shows that the origin has a statistical significance level between 5 and 10 percent. Here the hypothesis  $\beta_2 = \beta_3 = 0$  is not rejected at the 5% level, but the hypothesis  $\beta_3 = 0$  is rejected. The figure on the right illustrates how this can happen, by redrawing the contours using the  $\chi^2_{(1)}$  calibration. The hypothesis  $\beta_3 = 0$  was rejected because a horizontal line segment through  $\beta_3 = 0$  lies outside the 95% confidence contour based on 1 degree of freedom.

A bootstrap method for handling stationary mixing time series is to resample the data in blocks of length  $M > 1$ . By concatenating randomly sampled blocks of consecutive data points, the resampled series can capture some of the structure from the original series. For  $k$  small compared to  $M$ , resampled observations  $k$  apart are likely to be genuine observation pairs separated by  $k$  units in the original data. When  $k$  is large compared to  $M$  then resampled observation pairs  $k$  units apart are essentially independent in the resampled series, matching the weak dependence in the original series.

Suppose now that  $\theta$  is a parameter of the joint distribution of  $r \geq 1$  consecutive observations  $Y_{t-r+1}, \dots, Y_t$ , defined by

$$E(m(X_t, \theta)) = 0$$

where  $X_t = (Y'_{t-r+1}, \dots, Y'_t)'$  bundles  $r$  consecutive observations from the original series. Thus if  $Y_t \in \mathbb{R}^d$ , then  $X_t \in \mathbb{R}^{dr}$ .

The blocking idea can also be used in empirical likelihood. Starting with the series  $X_t$ , form blocks of length  $M$  with starting points separated by  $L$  time units.

That is,

$$B_i = (X_{(i-1)L+1}, \dots, X_{(i-1)L+M}), \quad i = 1, \dots, n$$

where

$$n = \left\lfloor \frac{T-M}{L} + 1 \right\rfloor.$$

Here  $\lfloor z \rfloor$  denotes the largest integer that is less than or equal to  $z$ . Values  $L$  between 1 and  $M$  inclusive are reasonable. With  $L = M$ , the blocks do not overlap. Taking  $L > M$  would leave some  $X_t$  values unused.

Now define the blockwise estimating function

$$b(B_i, \theta) = \frac{1}{M} \sum_{j=1}^M m(X_{(i-1)L+j}, \theta).$$

Of course, if  $E(m(X_t, \theta)) = 0$  then  $E(b(B_i, \theta)) = 0$  too. If  $L = M \rightarrow \infty$ , as  $T \rightarrow \infty$ , then with some assumptions, the dependencies among  $b(B_i, \theta)$  become negligible. Now blockwise empirical likelihood inferences for  $\theta$  are based on the empirical likelihood ratio

$$\mathcal{R}(\theta) = \sup \left\{ \prod_{i=1}^n w_i \mid w_i > 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i b(B_i, \theta) = 0 \right\}.$$

For  $L \doteq \alpha M$  with  $\alpha < 1$ , the dependencies do not become negligible, because there is a fixed fraction of overlap between consecutive blocks.

**Theorem 8.1** *Under conditions given in Kitamura (1997), including  $M \rightarrow \infty$  and  $MT^{-1/2} \rightarrow 0$*

$$-2 \left( \frac{T}{nM} \right) \log \mathcal{R}(\theta_0) \rightarrow \chi_{(q)}^2$$

as  $T \rightarrow \infty$ , where  $q$  is the dimension of  $\theta$ .

*Proof.* Kitamura (1997).  $\square$

The factor  $T/(nM)$  accounts for the overlap in the blocks. It would have to be there even if the time series were IID. If  $T$  is a multiple of the block size  $M$  and if  $L = M$ , so the blocks do not overlap, then  $T/(nM) = 1$ . For  $L = \alpha M$ , with  $\alpha < 1$ , the blocks overlap, and to compensate

$$\frac{T}{nM} = \frac{T}{\left\lfloor \frac{T-M}{\alpha M} + 1 \right\rfloor M} \doteq \alpha.$$

Thus when the blocks overlap, the asymptotic distribution of  $-2 \log \mathcal{R}(\theta_0)$  is approximately  $\alpha^{-1} \chi_{(q)}^2$ , ranging from  $\chi_{(q)}^2$  to  $M \chi_{(q)}^2$  as  $L$  ranges from  $M$  to 1.

Figure 8.3 shows the 5405 years of bristlecone pine tree ring widths from Campito Mountain in California. The last year in the data set is 1969. The units are 0.01 millimeters. The series values range from 0 to 99. There are several interesting features in this data, one of which is that downward spikes tend to be

## Campito tree ring data

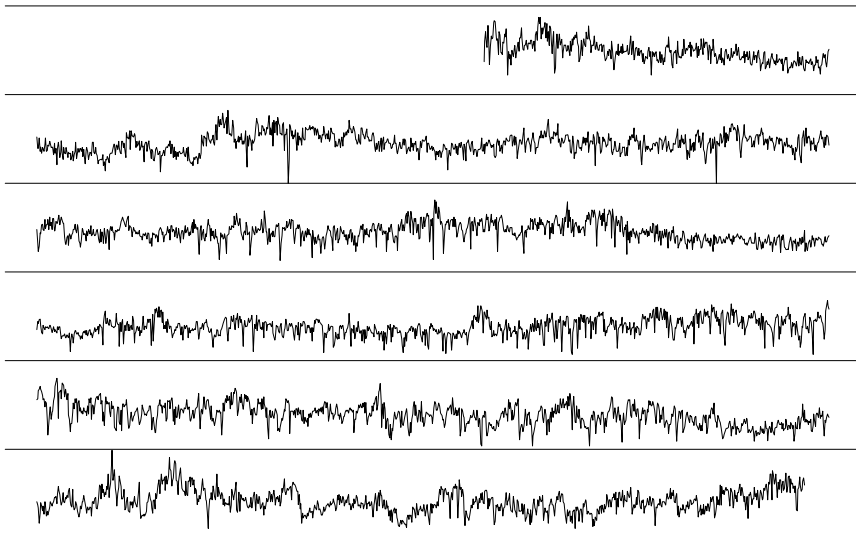


Figure 8.3 Shown are 5405 years of bristlecone pine tree ring data from Campito Mountain, California. Time increases from top to bottom of the figure, going from left to right within each of 6 ranges. The bottom range is for the years 1001 through 1969, where the series ends. Moving up one range corresponds to going back 1000 years. The data values are between 0 and 99. Within a range the data are plotted between lower and upper reference lines corresponding to values 0 and 100. The data are in units of 0.01 mm. The data are from Fritts et al. (1971) and are available on Statlib.

larger than upward ones. There were 39 years in which the tree ring width was more than 0.2mm larger than the average of the previous 10 years but 145 years in which the width was more than 0.2mm smaller than the average of the previous 10 years. We could not capture such asymmetry in an AR model with normally distributed errors.

The natural estimate of the probability of such a downward spike is  $145/(5405-10) = 0.027$ , because there are 145 successes in 5395 trials. A binomial confidence interval for this probability would not be appropriate because it would ignore the dependence in the data.

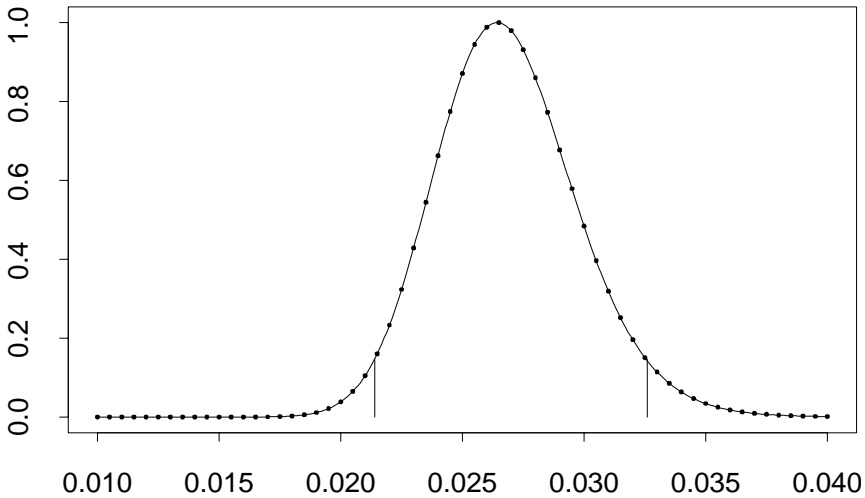


Figure 8.4 The horizontal axis shows the probability that the Campito tree ring width decreases by more than 0.2mm from its average over the previous 10 years. The vertical axis shows the block empirical likelihood ratio, including an adjustment for overlapping blocks. A 95% confidence interval extends from 0.0214 to 0.0326, as indicated by two vertical segments.

Define derived time series

$$W_i = Y_i - \frac{1}{10} \sum_{j=1}^{10} Y_{i-j}, \quad \text{and}$$

$$Z_i = \begin{cases} 1, & W_i < -20 \\ 0, & W_i \geq -20. \end{cases}$$

If  $Y_i$  is a stationary series then so are  $W_i$  and  $Z_i$ . We are interested in inferences on  $E(Z_i)$ . We take a block size of  $M = 50$  years, and starting points separated by  $L = 10$  years, and construct  $B_i$  as above.

For these data, the  $W$  and  $Z$  series have length 5395 so that is the appropriate value of  $T$ . The value of  $n$  is  $\lfloor (5395 - 50)/10 + 1 \rfloor = 535$ . Figure 8.4 shows the empirical likelihood curve for these blocked data. The empirical log likelihood is multiplied by  $T/(nM) = 5395/(535 \times 50) = 0.2017$  to adjust for block overlap. This is equivalent to multiplying the  $\chi^2_{(1)}$  threshold value by  $1/0.2017 = 4.96$ . The 95% confidence interval for the probability of a large downward spike ranges from 2.14% to 3.26%.

Choosing  $M$  and  $L$  can be difficult. Here are some guidelines, with the caveat that blocked empirical likelihood is a new method. Suppose at first that we take

$L = M$  and look for the right size of non-overlapping block. The asymptotic theory suggests that  $M$  should tend to infinity in order to control the dependence among blocks. If  $T/M = c$ , then for very large  $T$  the coverage of empirical likelihood should be like that for the mean of  $c$  independent blocked random vectors  $B_i$ . Sending  $c$  to infinity is necessary in order for the asymptotic coverage to set in. If we felt that ordinary empirical likelihood coverage properties were satisfactory for sample sizes  $c$  and above, then we might take  $M$  as the nearest integer to  $T/c$ . The value  $c$  that we would use would depend on the dimension of  $B_i$ .

Suppose that  $m(X_i, \theta)$  were really independent and normally distributed, and that we have grouped them. Then we have lost some efficiency. Instead of  $T$  observations with mean 0 and variance  $V_m$  say, we have only  $T/M$  observations  $b(B_i, \theta)$  having mean 0 and variance  $V_m/M$ . Our confidence regions for blocked data will not be as good as for the unblocked data, but it will primarily be the difference between using  $T/M$  degrees of freedom instead of  $T$  degrees of freedom. The sample size reduction by  $M$  is largely compensated by a variance reduction of  $M$ . Similar comments apply if the data are independent, but not normally distributed. This efficiency loss is explored in [Exercise 8.6](#).

In time series examples the data are not usually independent. The errors in treating small blocks as independent, when they are not, can be very large. Thus it seems that caution would dictate larger values of  $M$ .

The value of  $L$  would seem to be less crucial.  $L$  can range from 1 to  $M$ . Smaller values of  $L$  are usually more statistically efficient, though diminishing returns seem likely. The number of blocks grows as  $L$  decreases, increasing the computational effort.

## 8.4 Spectral method

The moving average model of order 1, denoted by MA(1) has

$$Y_i - \mu = e_i + \alpha_1 e_{i-1}$$

where  $|\alpha_1| < 1$ , and  $e_i$  are independent identically distributed random variables with mean 0 and variance  $\sigma^2$ . By substituting for  $e_{i-1}$  we find that the MA(1) model is an AR model of infinite order with  $\beta_k = \alpha_1^k$  for  $k \geq 1$ . We could base our inferences for  $\theta = (\mu, \sigma, \alpha_1)'$  on the expected values of  $Y_i - \mu$ ,  $(Y_i - \mu)(Y_{i-1} - \mu)$  and  $(Y_i - \mu)^2$ , written in terms of  $\theta$ . But such inferences are not efficient. They do not capture the information about  $\alpha_1$  in higher order lags than the first.

The moving average model of order  $\ell$ , MA( $\ell$ ) has  $Y_i - \mu = e_i + \sum_{j=1}^{\ell} \alpha_j e_{i-j}$ , and the ARMA( $k, \ell$ ) model has  $Y_i - \mu$  described as an AR( $k$ ) model with MA( $\ell$ ) errors. Estimating equations for MA and ARMA models are more complicated than those for AR models. See Chapter 8.10.

An alternative to reweighting the estimating equations of an ARMA model is to proceed through the spectrum, as outlined here. The spectrum may be written in terms of the parameters of the ARMA model, although  $\mu$  does not enter. The periodogram of a time series provides a noisy estimate of the spectrum. We can

take the estimate at  $\lfloor (T - 1)/2 \rfloor$  different frequencies. Asymptotically, these are independent exponential random variables with means equal to the true spectrum at the corresponding frequencies. The terms in the exponential log likelihood can be treated as a log likelihood and reweighted with a  $\chi^2$  limit. See Chapter 8.10 for references.

## 8.5 Finite populations

In many applications the sample is taken from a finite population. Theory and methods for sampling finite populations have historically had their impetus in survey sampling. The same problems now arise in some data mining applications. Computers that analyze a population of records now have to keep up with other computers that generate the data, making it attractive to work with a sample.

Suppose that the sample has  $n$  observations made on a population of  $N$  individuals. The statistical problem may be to estimate something about the whole finite population from the sample. In other settings, we seek inferences on an infinite superpopulation from which the  $N$  finite sample observations were drawn before we sampled  $n$  of them.

Several features make sampling finite populations different from the usual statistical problems. First, for inferences on the population, we get the answer without error if  $n = N$ . A related feature is that the problems of estimating a population total, or even of estimating  $N$  when it is unknown, may arise for finite populations, but not for infinite ones. Next, in finite populations it is especially common for there to be some variables with known population means or totals. These may be variables that were measured in a census, or they may be quantities that are constantly updated as records are added to a database. By taking account of the known population values, we can get sharper estimates for things that we do not know. Finally, there are a variety of strategies that can be employed in sampling to get better answers at lower cost. In stratification, we take separate samples within subpopulations perhaps overweighting an important rare group, such as records for fraudulent credit card transactions. In cluster sampling, we partition the population into groups of contiguous individuals, and take a sample of the groups.

In a simple random sample all  $N!/n!(N - n)!$  ways of selecting  $n$  of  $N$  observations are equally probable. When  $n \ll N$ , then the fact that the population is finite may often be ignored. But if  $n/N$  is not negligible, the finiteness of the population introduces a dependence described below that should not be ignored.

The customary notation for a simple random sample is that the population values are  $Y_i \in \mathbb{R}^q$ , measured on individuals  $i = 1, \dots, N$ . The simple random sample is then denoted by  $y_i$ , vectors measured on sampled individuals  $i = 1, \dots, n$ . There is understood to be no connection between the sample and population indices. In particular,  $y_1$  is not necessarily  $Y_1$ .

The population mean of  $Y$  is

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i,$$

which is estimated by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^N Y_i Z_i, \quad (8.9)$$

where  $Z_i$  is one if population element  $i$  is in the sample and zero otherwise. The population variance of  $Y$  is defined to be

$$S_{YY} = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})(Y_i - \bar{Y})',$$

which we suppose is invertible. By elementary calculations with equation (8.9), in which  $Z_i$  are random and  $Y_i$  are fixed, we find that

$$\text{Var}(\bar{y}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) S_{YY}. \quad (8.10)$$

The finite population correction factor  $1 - n/N$  arises from negative correlations among the  $Z_i$ . For two population members, if one is sampled then it is less likely that the other one is sampled. These negative correlations are usually very small, but there are  $O(N^2)$  of them, and together they cause an important variance reduction when  $n/N$  is not small.

The standard approach to inference for simple random samples is to obtain an unbiased estimate

$$s_{yy} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})'.$$

of  $S_{YY}$  and plug it into (8.10), getting

$$\widehat{\text{Var}}(\bar{y}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) s_{yy}. \quad (8.11)$$

Then, under a central limit theorem for  $\bar{y}$ , we have  $(\bar{y} - \bar{Y})'(\widehat{\text{Var}}(\bar{y}))^{-1}(\bar{y} - \bar{Y})$  is asymptotically  $\chi_{(q)}^2$ . When  $Y_i \in \mathbb{R}$ , the standard 95% confidence intervals for  $\bar{Y}$  are  $\bar{y} \pm 1.96(\widehat{\text{Var}}(\bar{y}))^{1/2}$ .

Central limit theorems for finite sampling are a bit more subtle than those for infinite populations. As we let  $n \rightarrow \infty$ , we must also have  $N \rightarrow \infty$  to keep  $n \leq N$ . Indeed, we assume that  $N - n \rightarrow \infty$ , for otherwise  $\bar{y}$  is determined by the average of a small number of excluded points. Finally there has to be a condition on the sequence of finite populations so that in the limit each  $Z_i Y_i$  is asymptotically negligible compared to  $\sum_{i=1}^n y_i$ . A commonly used condition is that  $(1/N) \sum_{i=1}^N |Y_i|^3 \leq B$  for all  $N$ .

## 8.6 MELE's using side information

In finite population settings, the effective use of side information (called auxiliary information in this context) is a very important issue. Maximum empirical likelihood estimates for finite populations have been more thoroughly studied than empirical likelihood ratios. MELE's allow us to incorporate side information while obeying range restrictions.

Suppose now the population consists of vectors  $U_i \in \mathbb{R}^u$  of underlying quantities with  $Y_i = \mathbb{Y}(U_i) \in \mathbb{R}^p$  for some function  $\mathbb{Y}$ . We take a simple random sample of values  $u_i$ , observing  $y_i = \mathbb{Y}(u_i)$ . Similarly, let  $X_i = \mathbb{X}(U_i) \in \mathbb{R}^q$  and  $x_i = \mathbb{X}(u_i)$ . We suppose that the population mean  $\bar{X}$  is known to us. Introducing the function  $\mathbb{X}$  allows us to encode a known mean or quantile of a component of  $U_i$  through  $\mathbb{X}(U_i) = U_{ij}$  or  $\mathbb{X}(U_i) = 1_{U_{ij} < Q} - \alpha$ . If one component  $U_{ij}$  represents a categorical variable taking a finite number  $c$  of values, and we know the population proportions in the  $c$  categories, we can encode this knowledge through  $c - 1$  components of  $\mathbb{X}(U_i)$  taking values 0 or 1. In general,  $\mathbb{X}$  encodes a finite number  $q$  of quantities whose population means are known.

The population variance of  $X$ , denoted  $S_{XX}$ , is defined analogously with  $S_{YY}$ . To avoid inessential complications, assume that  $S_{XX}$  has full rank. Define

$$S_{XY} = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})'$$

Suppose at first that  $p = 1$ . For any vector  $\beta$  of  $q$  components

$$\hat{Y}_\beta = \frac{1}{n} \sum_{i=1}^n y_i - (x_i - \bar{x})' \beta$$

is an unbiased estimate of  $\bar{Y}$ . The usual estimate (8.9) which ignores the  $x_i$  corresponds to  $\beta = 0$ . The variance of  $\hat{Y}_\beta$  is at a minimum for  $\beta_{LS} = S_{XX}^{-1} S_{XY}$ , which is usually unknown. The regression estimator of  $\bar{Y}$  is  $\hat{Y}_{REG} = \hat{Y}_\beta$  where

$$\hat{\beta} = \left( \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})' \right)^{-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

We may write the regression estimator as a weighted combination of data values

$$\hat{Y}_{REG} = \frac{1}{n} \sum_{i=1}^n \left( 1 - \frac{n}{n-1} (x_i - \bar{x})' s_{xx}^{-1} (\bar{x} - \bar{X}) \right) y_i, \quad (8.12)$$

where  $s_{xx}$  is the sample version of  $S_{XX}$ . For large  $n$  we expect the estimator  $\hat{\beta}$  to be close to  $\beta_{LS}$  and then  $\hat{Y}_\beta$  has variance near the optimal value.

If  $p > 1$ , then we may still estimate  $\bar{Y}$  by the weighted combination (8.12). The optimality results for a univariate  $\bar{Y}$  apply to any linear combination of components of the multivariate  $\bar{Y}$ .

A practical concern with equation (8.12) is that some of the  $y_i$  can receive negative weights. The result is that range restrictions are not obeyed. Estimated variances can be negative, and estimated cumulative distribution functions can be decreasing over some intervals.

An empirical likelihood approach enforces nonnegative weights. We find  $w_i$  satisfying

$$\sum_{i=1}^n w_i(x_i - \bar{X}) = 0, \quad \text{and} \quad \sum_{i=1}^n w_i = 1 \quad (8.13)$$

and maximizing  $\sum_{i=1}^n \log(nw_i)$  subject to (8.13). Then we estimate  $\bar{Y}$  by the maximum empirical likelihood estimator

$$\widehat{Y}_{\text{MELE}} = \sum_{i=1}^n w_i Y_i. \quad (8.14)$$

For small  $n$  or large  $q$  it may happen that  $\bar{X}$  is not in the convex hull of  $x_1, \dots, x_n$ . Then the weights required for  $\widehat{Y}_{\text{MELE}}$  do not exist, and so neither does the MELE. In that case we may have to use regression, and accept some negative weights, or decide not to impose some or all of the known population means.

The MELE has similar asymptotic properties to the regression estimator, but it respects range restrictions. Below is a theorem for the case  $p = q = 1$ .

**Theorem 8.2** *Suppose that  $n$ ,  $N$ , and  $N - n$  increase to  $\infty$  such that*

$$\frac{1}{N} \sum_{i=1}^N |Y_i|^3 < B, \quad \text{and} \quad \frac{1}{N} \sum_{i=1}^N |X_i|^3 < B$$

for some  $B < \infty$ . Then

$$\sqrt{n} \frac{\widehat{Y}_{\text{MELE}} - \bar{Y}}{\sigma_{Y|X}} \rightarrow N(0, 1)$$

where

$$\sigma_{Y|X}^2 = \left(1 - \frac{n}{N}\right) \left(S_{YY} - \frac{S_{XY}^2}{S_{XX}}\right).$$

*Proof.* Chen & Qin (1993).  $\square$

The asymptotic variance of  $\widehat{Y}_{\text{MELE}}$  in Theorem 8.2 is the same as that of  $\widehat{Y}_{\text{REG}}$ . Variance estimates for  $\widehat{Y}_{\text{REG}}$  can be used for  $\widehat{Y}_{\text{MELE}}$ .

## 8.7 Sampling designs

A simple random sample is not always the most efficient way to gather data. In stratified sampling one divides the population into a finite number  $H$  of strata.

Let the population have elements  $Y_{hi} = \mathbb{Y}(U_{hi})$  corresponding to individuals  $i = 1, \dots, N_h$  in strata  $h = 1, \dots, H$ . Stratified sampling takes a simple random sample of  $n_h$  observations from stratum  $h$ , yielding observations  $y_{hi} = \mathbb{Y}(u_{hi})$ . Each stratum is sampled independently.

In cluster sampling, the population is partitioned into groups as for stratified sampling. The groups comprise individuals that are conveniently sampled together, such as inhabitants of a block, or files created in a time window. Instead of taking a simple random sample from within each group, a simple random sample of the groups is taken. From each sampled group, we might take all individuals, or possibly a simple random sample, or even a cluster sample.

Clustering and stratification can be combined in complicated ways, as for example, clusters of clusters within strata, with auxiliary variables having known population and/or stratum means. The Horvitz-Thompson and generalized regression estimators provide a unified approach to estimating  $\bar{Y}$ . For each unit in the population let  $\Pi_i$  be the probability that it is included in the sample. Similarly, let  $\pi_i$  be these probabilities for the  $n$  observations actually included in the sample. The Horvitz-Thompson estimator of  $\bar{Y}$  is

$$\widehat{Y}_{\text{HT}} = \frac{1}{N} \sum_{i=1}^n \frac{y_i}{\pi_i} = \frac{1}{N} \sum_{i=1}^N \frac{Y_i Z_i}{\Pi_i}, \quad (8.15)$$

where  $Z_i$  is again an indicator variable for inclusion in the sample. The estimator (8.15) does not require that we know all the  $\Pi_i$ , only those for the individuals actually sampled. It weights items inversely to their inclusion probability, just as was done in Chapter 6.1.

Given known values for  $\bar{X}$ , the generalized regression estimator is

$$\widehat{Y}_{\text{GREG}} = \widehat{Y}_{\text{HT}} - (\widehat{X}_{\text{HT}} - \bar{X})' \widehat{\beta}_{\text{HT}} \quad (8.16)$$

where, for scalar  $Y_i$ ,  $\widehat{\beta}_{\text{HT}}$  minimizes the weighted sum of squares

$$\sum_{i=1}^n \pi_i^{-1} \left( (y_i - \widehat{Y}_{\text{HT}}) - (x_i - \widehat{X}_{\text{HT}})' \beta \right)^2.$$

Known stratum sizes can be incorporated into  $X$  as known population means of stratum indicator variables.

The estimator (8.16) may be written as a weighted sum of  $y_i$  values, and the same weights are employed for multivariate  $Y_i$ . The generalized regression estimator does not necessarily respect range restrictions. Once again, a solution is available using an MELE, subject to a convex hull condition. Non-existence of the MELE does provide a diagnostic that the GREG estimator is using some negative weights, constituting a form of extrapolation.

To construct an MELE that takes account of  $\pi_i$ , we maximize

$$L(w) = \sum_{i \in s} d_i \log w_i \quad (8.17)$$

where  $d_i = 1/\pi_i$  is called a design weight, subject to constraints  $\sum_{i \in s} w_i = 1$  and  $\sum_{i \in s} w_i (X_i - \bar{X}) = 0$ . Here  $s$  is the sample, and writing the summation limits as  $i \in s$  reminds us that the sample size may be random. The motivation for  $L$  is that it is an unbiased estimate of the log likelihood  $\sum_{i=1}^N \log W_i$  that we would use for inferences on a superpopulation, if we had observed the entire finite population.

Using some foresight, we construct the Lagrangian

$$G = \sum_{i \in s} d_i \log(w_i) - D\lambda' \sum_{i \in s} w_i (X_i - \bar{X}) + \gamma \left(1 - \sum_{i \in s} w_i\right),$$

where  $D = \sum_{i \in s} d_i$ . Setting  $\sum_{i \in s} w_i \partial G / \partial w_i = 0$  gives  $\gamma = D$ , and  $\partial G / \partial w_i = 0$  gives

$$w_i = \frac{d_i}{D} \frac{1}{1 + \lambda'(X_i - \bar{X})}$$

where  $\lambda$  satisfies

$$0 = \sum_{i \in s} \frac{d_i (X_i - \bar{X})}{1 + \lambda'(X_i - \bar{X})}.$$

The MELE

$$\hat{Y}_{\text{MELE}} = \sum_{i \in s} w_i Y_i,$$

respects range restrictions and is close to the generalized regression estimator:

**Theorem 8.3** *If as  $N$  and  $n$  increase to  $\infty$ ,  $\max_{i \in s} \|X_i - \bar{X}\| = o_p(n^{-1/2})$ ,*

$$\left( \sum_{i \in s} d_i (X_i - \bar{X})(X_i - \bar{X})' \right)^{-1} \left( \sum_{i \in s} d_i (X_i - \bar{X}) \right) = O_p(n^{-1/2}),$$

*and  $\max_{1 \leq i \leq N} \|Y_i\|$  is bounded, then  $\hat{Y}_{\text{MELE}} = \hat{Y}_{\text{GREG}} + o_p(n^{-1/2})$ .*

*Proof.* Chen & Sitter (1999) show that the MELE and GREG weights differ by  $o_p(n^{-1/2})$  and then the bound on  $Y_i$  completes the proof.  $\square$

**Theorem 8.3** applies to various forms of cluster sampling. For probability sampling within  $L$  strata, we introduce  $L(w) = \sum_{h=1}^L \sum_{i \in s_h} d_{hi} \log(w_{hi})$ . See Chapter 8.10.

## 8.8 Empirical likelihood ratios for finite populations

Now we consider empirical likelihood ratios. For simple random sampling, let

$$\mathcal{R}(\mu) = \max \left\{ \prod_{i=1}^n n w_i | w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i y_i = \mu \right\}.$$

From **Theorem 8.2**, it is reasonable to expect that  $-2(1 - n/N)^{-1} \log \mathcal{R}(\bar{Y})$  will have an asymptotic  $\chi_{(p)}^2$  distribution. A more general result, including stratified

sampling is cited in Chapter 8.10. The factor of  $1 - n/N$  plays a similar role to the factor used in blockwise empirical likelihood in Chapter 8.3. It is not surprising that a correction must be employed, because  $\mathcal{R}$  was derived for  $N = \infty$  and is not a finite population likelihood ratio.

One way to build the finite population assumption into the likelihood is to suppose that there are  $K$  distinct values of  $Y_i$  in the population. Let them be  $\mathcal{Y}_i$ , for  $i = 1, \dots, K$  and suppose that  $\mathcal{Y}_i$  appears  $N_i$  times in the population and  $n_i$  times in a simple random sample. If the  $\mathcal{Y}_i$  are known then the population is described by the parameter  $(N_1, \dots, N_K)$ , which has a hypergeometric likelihood

$$L(N_1, \dots, N_K) = \binom{N}{n}^{-1} \prod_{i=1}^K \binom{N_i}{n_i}. \quad (8.18)$$

This likelihood is difficult to work with, because it is only defined over integer  $N_i$ . However, in the limit with  $N_i/N \rightarrow w_i$  and  $n/N \rightarrow 0$ , a likelihood proportional to  $\prod_{i=1}^n w_i$  emerges.

## 8.9 Other dependent data

This section describes some other settings with dependent data, where an empirical likelihood analysis might add value.

Longitudinal data arise as repeated measures, usually over time, on a series of subjects. Such data are commonly found in biomedical applications. They can be arranged into multiple time series, one per subject. The statistical issues may be to describe the typical time trend of a subject, the subject-to-subject variation in the trends, or the effects of covariates such as treatments.

Random fields are generalizations of time series to higher dimensional index spaces. The observations may be taken on a grid in  $\mathbb{R}^g$ , or at scattered sites, or continuously, or on some hybrid such as along line transects, as was done for the shrub width data in Chapter 6.1. Spatial point processes are scattered observations, such as the locations of trees or galaxies, where the random locations themselves are under study.

## 8.10 Bibliographic notes

### *Time series*

Box, Jenkins & Reinsel (1994), Cryer (1986), and Anderson (1994) provide background material on time series. Politis, Romano & Wolf (1999) provide an appendix with results on mixing on which equation (8.1) is based. There are many different nomenclatures for describing spectral densities and their estimates. The account in Chapter 8.1 follows Chatfield (1989).

The idea of describing a dependent series through a series of independent “shocks” is a powerful one that Box et al. (1994) attribute to Yule (1927). This approach underlies most parametric work on time series in the time domain. Efron

& Tibshirani (1986) propose a bootstrap based on resampling residuals from an autoregression.

The dual likelihood, due to Mykland (1995), takes the Lagrange multiplier  $\lambda$  to be the parameter. For each fixed parameter value  $\theta$  the test of  $\theta$  corresponds to a test of  $\lambda = 0$  for the  $Z_i = m(X_i, \theta)$ . For IID data the dual and empirical likelihoods coincide. The dual likelihood applies also to time series and survival analysis settings with martingale estimating equations. Very generally the dual likelihood statistic is close to a quadratic statistic (like the Euclidean likelihood), and has a  $\chi^2$  limit. Mykland (1995) also presents a notion of Bartlett correctability for this martingale setting.

Hipel & McLeod (1994) analyze the St. Lawrence river flow data. They identified an AR(3) model for it, and recommend constraining the lag 2 coefficient to be zero. The data are from Yevjevich (1963). They are repeated in Hipel & McLeod (1994) and are also available from Statlib.

Chuang & Chan (2001) study unstable autoregressions in which (8.7) has at least one root on the unit circle, but no roots inside the unit circle. They show that both the empirical log likelihood ratio statistic and the usual quadratic test statistic have the same (non  $\chi^2$ ) limiting distribution.

Politis et al. (1999) trace the development of blockwise approaches for the bootstrap. Carlstein (1986) proposed non-overlapping blocks for variance estimation. Künsch (1989) and Liu & Singh (1992) developed versions for confidence regions. Politis & Romano (1994) propose a method of sampling with random block lengths, so that the resampled series are stationary, conditionally on the observed one. Politis et al. (1999) remark that more overlap among blocks (smaller  $L$ ) gives more efficiency.

Blockwise empirical likelihood was developed by Kitamura (1997), for estimating equations and for smooth functions of means. Kitamura (1997) also extends the results from Qin & Lawless (1994) to stationary time series, and establishes Bartlett correctability for some time series versions of empirical likelihood. The Bartlett correction supposes that  $M$  is of exact order  $T^{1/3}$ . Then Bartlett correction improves the order of coverage error from  $T^{-2/3}$  to  $T^{-5/6}$ . The blocking used in Chapter 8.3 takes simple averages over blocks. Kitamura (1997) raises the possibility of taking weighted averages within blocks and relates this idea to kernel methods of smoothing the spectrum.

Kitamura (1999) applies a pre-whitening filter to the time series before applying blockwise empirical likelihood. The filter subtracts a linear combination of past series values  $Y_{t-k}$  from  $Y_t$ . The linear combination is chosen to make the filtered series more nearly, even if not exactly, uncorrelated, allowing a smaller block size.

Hipel & McLeod (1994) give an estimate of the spectrum for the Campito tree ring data. Those data are available on Statlib, with an attribution to Fritts et al. (1971).

Some properties of a time series, such as its spectrum, are not functions of a finite dimensional margin, but depend instead on the whole infinite dimensional

joint distribution of the data. For these Kitamura (1997) describes an approach based on blocks of blocks, paralleling the blocks of blocks bootstrap of Politis & Romano (1992).

Estimation in MA and ARMA models is described in Box et al. (1994) and Hipel & McLeod (1994). Maximum likelihood algorithms usually require back forecasting of error terms from before the start of the data and the formation of a sum of squares of estimated errors. As a result, the estimating equations being solved are not explicit.

Forming a likelihood from the distribution of the periodogram is known as Whittle's method after Whittle (1953). Bootstrap-resampled periodograms have been used by Ramos (1989) to generate new estimators by Rao-Blackwellization. Franke & Härdle (1992), Janas (1994), and Dahlhaus & Janas (1996) propose inferences based on resampled periodograms.

The spectral approach to empirical likelihood is due to Monti (1997), who also proposes a Bartlett correction. Monti (1997) presents a confidence region for the parameters of an ARMA(1, 1) model fit to a series of 197 chemical process concentration readings from Box et al. (1994). The region is asymmetric, extending farther toward the origin where the series would be independent than away from it where the series would be explosive. The parametric region is elliptical.

Monti (1997) simulates some MA(1) processes with parameter values in  $(0, 1)$ . Two error distributions are considered:  $N(0, 1)$  and  $\chi_{(5)}^2 - 5$ . For simulations with Gaussian errors, methods based on the Gaussian likelihood did best, but for non-Gaussian errors, empirical likelihood inferences had better coverage than the customary asymptotic ones, particularly for parameter values close to the boundary of the invertibility region.

### *Finite populations*

Hartley & Rao (1968) provide one of the earliest NPMLE arguments, using the discrete likelihood (8.18). They also show how to optimize that likelihood over integer values to find the MLE. Hartley & Rao (1968) also provide what may be the very first MELE, maximizing a continuous version of the likelihood subject to a constraint on the mean. They do not consider nonparametric likelihood ratios, but show instead that the MELE closely approximates the regression estimator for which there are well-known variance estimates. They also consider a Bayesian formulation using a Dirichlet prior.

Chen & Qin (1993) present empirical likelihood for samples from a finite population. In addition to Theorem 8.2, they also present a consistent estimate of the variance of the MELE based on the jackknife. Under a superpopulation model with a continuous CDF for  $Y$ , Chen & Qin (1993) characterize the asymptotic behavior of the reweighted CDF of  $Y$ , using empirical likelihood weights based on known  $\bar{X}$ . Chen & Qin (1993) show that the MELE reproduces several well known estimates from survey sampling. A categorical  $X$  with known category frequencies gives rise to the post-stratified estimator of  $\bar{Y}$ . The MELE of the me-

dian of one variable, using as auxiliary information the known median of another variable, gives rise to a raking estimator.

Chen & Sitter (1999) formulate an empirical likelihood that respects design weights, and they use it to construct MELE's. Their statement of [Theorem 8.3](#) does not put any conditions on  $Y_i$ . In a personal communication, Jiahua Chen indicates that the conditions on the  $Y_i$  should be the same as those on the  $X_i$ . Chen & Sitter (1999) show that the conditions in [Theorem 8.3](#) are satisfied for sampling proportional to population size (pps) with replacement, for the Rao-Hartley-Cochran method (of pps without replacement), and for cluster sampling. Chen & Sitter (1999) also show how to define an MELE for sampling designs within strata, using side information. Zhong & Rao (2000) provide a central limit theorem for the MELE based on independent simple random samples within strata. They also show that the empirical likelihood ratio can be used to form confidence regions, if a correction generalizing  $1 - f$  to the stratified case is applied.

Sitter & Wu (2000) consider estimating quadratic population quantities defined as  $\sum_{i=1}^N \sum_{j=1}^N \phi(Y_i, Y_j)$  for some function  $\phi$ . Variances and covariances are the motivating statistics. They modify the design effect likelihood ([8.17](#)) to take account of pairwise inclusion probabilities  $\pi_{ij} = \Pr(Z_i Z_j = 1)$ , and obtain range respecting estimators that incorporate side information.

Wu & Sitter (2001) consider a setting where the entire population  $X_1, \dots, X_N$  is known but only the sampled  $y_1, \dots, y_n$  are available. Then using a working model to link the mean and variance of  $Y_i$  to  $X_i$ , they develop estimators of  $\bar{Y}$  that are consistent generally, and efficient if the working model holds.

Zhong, Chen & Rao (2001) consider combining multiple finite samples to estimate a common feature, such as a mean or CDF, when some of the samples have distorted observations.

MELE's for survey sampling have a lot in common with variance reduction methods in Monte Carlo. Hesterberg (1995*b*) presents the usual Monte Carlo variance reduction methods in terms of reweighted sample points.

### *Other dependencies*

Longitudinal data are the subject of Diggle, Liang & Zeger (1994). They are usually analyzed by techniques in the companion papers Liang & Zeger (1986) and Zeger & Liang (1986). Bootstrap methods and references for spatial processes are discussed by Davison & Hinkley (1997, Chapter 8). A parametric model may be used, or the data can be resampled in spatial blocks. Loh (1996) considers empirical likelihood confidence regions for the mean based on Latin hypercube samples.

## **8.11 Exercises**

**Exercise 8.1** The GARCH(1,1) model is widely used for financial time series. Let  $e_i$  be a sequence of independent random variables with mean 0 and variance

1. The series values are  $Y_i = e_i \sigma_i$  where the variance  $\sigma_i^2$  evolves in time as

$$\sigma_i^2 = \alpha_0 + \alpha_1 Y_{i-1}^2 + \beta_1 \sigma_{i-1}^2.$$

This model captures the volatility clustering phenomenon often seen in financial data, where increases in variance have been seen to persist. It can also give rise to fatter than normal tails for the  $Y_i$  distribution, even if the  $e_i$  are normally distributed.

Let  $\theta = (\alpha_0, \alpha_1, \beta)$  and suppose that  $e_i$  are independent  $N(0, 1)$ . Obtain conditional likelihood estimating equations for  $\theta$ .

**Exercise 8.2** Suppose that we suspect that the random variables  $e_i$  in [Exercise 8.1](#) are not normally distributed but instead have some skewness. If  $Y_i$  are returns to holding an asset, negative skewness corresponds to downward price movements having a fatter tail than upward ones. Formulate estimating equations for the GARCH(1,1) model that support the construction of confidence intervals for the skewness.

**Exercise 8.3** For the block approach to empirical likelihood, observation  $X_1$  is always contained in block  $B_1$ . If  $T$  leaves a remainder of  $M$  when divided by  $L$ , then  $X_T$  appears in block  $B_n$ , but otherwise  $X_T$  does not appear in any block. Redefine the blocks, so that they have length  $M$ , have starting points separated by  $L$  units, and so that  $X_T$  is always used (in the last block) while  $X_1$  may or may not be used at all.

**Exercise 8.4** For the Campito tree ring data, the event of interest happened 145 times in 5395 yearly trials. The confidence interval for the event probability, taking account of dependence in the data, extends from 0.0214 to 0.0326. How much narrower (or wider) would the 95% confidence interval be for a problem with 145 occurrences in 5395 independent trials?

**Exercise 8.5** What is the first year in the Campito tree series? (Hint: the year before 1 A.D. was 1 B.C., there being no year 0.)

**Exercise 8.6** Suppose that  $Z_1, \dots, Z_{200} \sim N(\mu, \sigma^2)$ , independently, with both  $\mu$  and  $\sigma$  unknown. Statistician A has all the data and constructs an exact 95% confidence interval, of random length  $L_A$ , for  $\mu$  using the  $t_{(199)}$  distribution as usual. These same observations are then averaged in blocks of 10 and sent to Statistician B. This statistician gets  $Y_j = (Z_{10(j-1)+1} + \dots + Z_{10(j-1)+10})/10 \sim N(\mu, \sigma^2/10)$  for  $j = 1, \dots, 20$ , and constructs an exact 95% confidence interval for  $\mu$  of random length  $L_B$ , using the  $t_{(19)}$  distribution. A is clearly better off than B. Find the mean, variance, and histogram of  $L_B/L_A$ , by simulation.