Chapter 9
Other Estimation Methods

9.1 Estimation Using Empirical Distributions

9.1.1 Empirical Distribution Functions

Suppose that we have sample data $x_1, x_2, \ldots, x_n$ assumed to be observed values of independent random variables each having the same distribution function F where

$$F(x) = \mathbb{P}(X \leq x)$$

Define new random variables Z_i as the indicator functions of the interval $(-\infty, x]$, i.e.,

$$Z_i(x) = \begin{cases} 1 \ X_i \leq x \\ 0 \text{ otherwise} \end{cases}$$

Note that the $Z_i(x)$ are independent and are Bernoulli random variables with parameter $F(x)$, i.e.,

$$\mathbb{P}(Z_i(x) = 1) = \mathbb{P}(X_i \leq x) = F(x)$$

It follows that, for any fixed x, we have that

$$S_n(x) = \sum_{i=1}^{n} Z_i(x)$$

has a binomial distribution with parameters $F(x)$ and n.

C.A. Rohde, *Introductory Statistical Inference with the Likelihood Function*, DOI 10.1007/978-3-319-10461-4_9

Definition 9.1.1. The **empirical distribution function**, $\widehat{F}_n(x)$ is defined as

$$\widehat{F}_n(x) = \frac{S_n(x)}{n}$$

The empirical distribution function is the natural estimator of F, the population distribution function, for the following reasons:

1. $\widehat{F}_n(x)$ is **unbiased**, i.e.,

$$\mathbb{E}[\widehat{F}_n(x)] = F(x) \text{ for any } x$$

2. The variance of $\widehat{F}_n(x)$ is given by

$$\mathbb{V}\left[\widehat{F}_n(x)\right] = \frac{F(x)[1 - F(x)]}{n}$$

3. $\widehat{F}_n(x)$ is **consistent**, i.e.,

$$\widehat{F}_n(x) \xrightarrow{p} F(x) \text{ for any } x$$

The above results follow from the fact that $n\widehat{F}_n(x) = S_n$ is binomial with parameters n and $F(x)$.

There are two important additional properties of $\widehat{F}_n(x)$:

1. **Glivenko–Cantelli Theorem**
   Under the assumption of iid X_i's we have

$$\sup_x \left|\widehat{F}_n(x) - F(x)\right| \xrightarrow{p} 0$$

   i.e., the maximum difference between $\widehat{F}_n(x)$ and $F(x)$ is small for large n.
2. **Dvoretzky–Kiefer–Wolfowitz (DKW) Inequality**
   Under the assumption that the X_i's are iid

$$\mathbb{P}\left(\sup_x |\widehat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \text{ for any } \epsilon > 0$$

The implication of the last result is that if we define

$$L(x) = \max\{\widehat{F}_n(x) - \epsilon_n,\, 0\} \text{ and } U(x) = \min\{\widehat{F}_n(x) + \epsilon_n,\, 1\}$$

where

$$\epsilon_n = \sqrt{\frac{\ln(2/\alpha)}{2n}}$$

then we have that

$$\mathbb{P}\{L(x) \leq F(x) \leq U(x) \text{ for all } x\} \geq 1 - \alpha$$

i.e., we have a $100(1-\alpha)\%$ confidence interval for $F(x)$.

1. The previous two results, particularly the first, have been called the fundamental theorems of mathematical statistics because they show that we can, with high probability, learn about F using a random sample from a population assumed to have distribution F.
2. In most statistical applications we can do better (use smaller n) since we assume that F is specified by a small number of parameters.
3. In fact, in many cases, we are not interested in F itself but some other function such as the mean or variance of the population.

9.1.2 Statistical Functionals

In mathematics a **functional** is a function whose domain is a set of functions.

Definition 9.1.2. A **statistical functional**, $\theta = T(F)$, is any function of the distribution function F.

Almost any parameter of interest is a statistical functional, e.g., the mean, median, and quantiles. Since the sample distribution function is the natural estimate of the distribution function the following gives the natural estimates of statistical functionals.

Definition 9.1.3. The **plug-in estimator** of the statistical functional $\theta = T(F)$ is

$$\widehat{\theta}_n = T(\widehat{F}_n)$$

i.e., to estimate $T(F)$ plug in (substitute) $\widehat{F}_n$ for F.

9.1.3 Linear Statistical Functionals

One important class of statistical functionals are the linear statistical functionals.

Definition 9.1.4. A **linear statistical functional** is a statistical functional of the form

$$T(F) = \int r(x)dF(x) \text{ for some function } r$$

where by $\int r(x)dF(x)$ we mean

$$\sum_{x\in\mathcal{X}} r(x)f(x) \text{ or } \int_{x\in\mathcal{X}} r(x)f(x)dx$$

depending on whether F is discrete or continuous.

For linear functionals we have the following two important results:

(i) The plug-in estimator for a linear functional is

$$T(\widehat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} r(X_i)$$

(ii) Assuming that we can find an estimate, $\widehat{s.e.}$, of the standard error of $T(\widehat{F}_n)$, an approximate $100(1-\alpha)$ confidence interval for $T(F)$ is given by

$$T(\widehat{F}_n) \pm z_{1-\alpha/2}\widehat{s.e.}$$

The reason for the second statement is that it is often true that

$$\frac{T(\widehat{F}_n - T(F))}{\widehat{\text{se}}} \xrightarrow{d} \text{N}(0,1)$$

We then use the standard pivotal argument for the normal distribution to obtain the approximate confidence interval for $T(F)$.

9.1.4 Quantiles

One other class of statistical functionals is of major importance, the quantiles of a distribution.

Definition 9.1.5. If F has a density function f then the pth **quantile** of F is defined by

$$T(F) = F^{-1}(p)$$

The plug-in estimate of the p quantile is

$$T(\widehat{F}_n) = \inf_x\{x : \widehat{F}_n(x) \geq p\}$$

and is called the pth **sample quantile**.

The following are important quantiles:

p	Name	Estimate
$\frac{1}{10}, \ldots, \frac{9}{10}$	Deciles	Sample deciles
$\frac{1}{4}, \frac{3}{4}$	Quartiles	Sample quartiles
$\frac{1}{2}$	Median	Sample median

9.1.5 Confidence Intervals for Quantiles

Let $X_1, X_2, \ldots, X_n$ be independent with distribution function F. Suppose that we want a confidence interval for η_p, the pth quantile of F, i.e.,

$$p = F(\eta_p) = \mathbb{P}(X_i \leq \eta_p)$$

Define $Z_1, Z_2, \ldots, Z_n$ by

$$Z_i = \begin{cases} 1 \text{ if } X_i < \eta_p \\ 0 \text{ otherwise} \end{cases}$$

The Z_i are independent Bernoulli with

$$\mathrm{P}(Z_i = 1) = \mathbb{P}(X_i < \eta_p) = p$$

It follows that

$$S_n = \sum_{i=1}^{n} Z_i \text{ is binomial } (n, p)$$

Now define the **order statistics**, $X_{n1}, X_{n2}, \ldots, X_{nn}$, as the ordered values of $X_1, X_2, \ldots, X_n$ from smallest to largest.

Note that

$$S_n \geq j \iff X_{nj} < \eta_p$$

and

$$S_n \leq k-1 \iff X_{nk} \geq \eta_p$$

These last two facts allow us to determine confidence limits for η_p since

$$\mathbb{P}(X_{nj} < \eta_p \leq X_{nk}) = \mathbb{P}(j \leq S_n \leq k-1)$$

The last probability can be obtained from the binomial distribution with parameter p, i.e.,

$$\mathbb{P}(j \leq S_n \leq k-1) = \mathbb{P}(S_n \leq k-1) - \mathbb{P}(S_n \leq j-1)$$

Thus all we need to do is find j and k such that

$$\mathbb{P}(S_n \leq k-1) - \mathbb{P}(S_n \leq j-1) \ \geq 1-\alpha$$

and we will have a $100(1-\alpha)\%$ confidence interval for η_p.

This interval is **nonparametric** since we do not need to assume the specific form of F. In cases where we are willing to assume a specific form for F we can do better, i.e., have a shorter confidence interval.

Where to start for j and k? Note that

$$\frac{S_n - np}{\sqrt{np(1-p)}} \ \approx \ \mathrm{N}(0,1)$$

so that

$$\mathbb{P}(S_n \leq k-1) \ \approx \ \mathbb{P}\left(Z \leq \frac{k-1-np}{\sqrt{np(1-p)}}\right)$$

i.e.,

$$k-1 \ \approx \ np + z_{1-\alpha/2}\sqrt{np(1-p)}$$

Similarly

$$j-1 \ \approx \ np - z_{1-\alpha/2}\sqrt{np(1-p)}$$

Start with this j and k and iterate.

9.2 Method of Moments

The **method of moments** is related to the plug-in method. If

$$\alpha_j = \mathbb{E}(X^j)$$

the plug-in method of estimation equates α_j to the sample moment

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

where x_i denotes the ith observation from a random sample. This defines the estimate of α_j

The method of moments uses the fact that the population moments are functions of the parameters θ and solves the equations

$$\alpha_j(\theta) = \widehat{\alpha}_j \quad \text{for } j = 1, 2, \ldots, k$$

assuming that there are k parameters

$$\theta_1, \theta_2, \ldots, \theta_k$$

The method of moments enjoys some reasonable properties in the frequentist paradigm:

1. Consistency, i.e., $\widehat{\theta}_n \xrightarrow{p} \theta$
2. Asymptotic normality, i.e.,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathrm{N}(0, \Sigma)$$

where Σ is determined by the solution to the equations defining the estimates.

9.2.1 Technical Details of the Method of Moments

Consider n iid random variables $X_1, X_2, \ldots, X_n$ and define the sample moments by

$$\overline{X}^1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \ldots, \quad \overline{X}^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

and let

$$\alpha_r = E(X^r) \text{ and } \mu_r = E(X - \mu)^r \text{ where } \mu = E(X)$$

be the corresponding population moments (moments of the distribution of X).

Provided that the expected value of X^{2k} exists the central limit theorem guarantees that

$$\mathbf{Y} = (\overline{X}^1, \overline{X}^2, \ldots, \overline{X}^k)$$

are jointly asymptotically normal. More precisely,

$$\sqrt{n}[\mathbf{Y} - \mathbb{E}(\mathbf{Y})] \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$$

where

$$\mathbb{E}(\mathbf{Y}) = \begin{bmatrix} E(X) \\ E(X^2) \\ \vdots \\ E(X^k) \end{bmatrix}$$

and $\mathbf{\Sigma}$ is given by

$$\begin{bmatrix} \text{var}(X) & \text{cov}(X, X^2) & \cdots & \text{cov}(X, X^k) \\ \text{cov}(X^2, X) & \text{var}(X^2) & \cdots & \text{cov}(X^2, X^k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X^k, X) & \text{cov}(X^k, X^2) & \cdots & \text{var}(X^k) \end{bmatrix}$$

To obtain approximations to the sampling distributions of method of moment estimators we use the Delta method. Let g be a continuous and differentiable function and define

$$\nabla_g(\boldsymbol{\alpha}) = \begin{bmatrix} \frac{\partial g(y_1, y_2, \ldots, y_k)}{\partial y_1} \\ \frac{\partial g(y_1, y_2, \ldots, y_k)}{\partial y_2} \\ \vdots \\ \frac{\partial g(y_1, y_2, \ldots, y_k)}{\partial y_k} \end{bmatrix}_{y_1=\alpha_1, y_2=\alpha_2, \ldots, y_k=\alpha_k}$$

then the Delta method applies and we have that

$$\sqrt{n}[g(\overline{X}^1, \overline{X}^2, \ldots, \overline{X}^k) - g(\alpha_1, \alpha_2, \ldots, \alpha_k)]$$

converges in distribution to a

$$\text{N}\left(0, \theta^2\right)$$

distribution where

$$\theta^2 = \nabla_g^\top(\boldsymbol{\alpha}) \mathbf{\Sigma} \nabla_g(\boldsymbol{\alpha})$$

More generally if $g_1, g_2, \ldots, g_r$ are continuous and differentiable functions let $\mathbf{g} = (g_1, g_2, \ldots, g_r)$ and let

$$\nabla_g(\mathbf{y}) = \frac{\partial \mathbf{g}(\mathbf{y})}{\partial \mathbf{y}} = \begin{bmatrix} \frac{\partial g_1(\mathbf{y})}{\partial y_1} & \frac{\partial g_2(\mathbf{y})}{\partial y_1} & \ldots & \frac{\partial g_r(\mathbf{y})}{\partial y_1} \\ \frac{\partial g_1(\mathbf{y})}{\partial y_2} & \frac{\partial g_2(\mathbf{y})}{\partial y_2} & \ldots & \frac{\partial g_r(\mathbf{y})}{\partial y_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\mathbf{y})}{\partial y_k} & \frac{\partial g_2(\mathbf{y})}{\partial y_k} & \ldots & \frac{\partial g_r(\mathbf{y})}{\partial y_k} \end{bmatrix}$$

be evaluated at

$$y_1 = \alpha_1, y_2 = \alpha_2, \ldots, y_k = \alpha_k$$

to obtain $\nabla_g(\boldsymbol{\alpha})$, then

$$\sqrt{n}\begin{bmatrix} g_1(\overline{X}_n^1, \overline{X}_n^2, \ldots, \overline{X}_n^k) - g_1(\alpha_1, \alpha_2, \ldots, \alpha_k) \\ g_2(\overline{X}_n^1, \overline{X}_n^2, \ldots, \overline{X}_n^k) - g_2(\alpha_1, \alpha_2, \ldots, \alpha_k) \\ \vdots \\ g_r(\overline{X}_n^1, \overline{X}_n^2, \ldots, \overline{X}_n^k) - g_r(\alpha_1, \alpha_2, \ldots, \alpha_k) \end{bmatrix}$$

converges in distribution to a

$$\mathrm{N}\ (\mathbf{0}, \mathbf{V})$$

distribution where

$$\mathbf{V} = \nabla_g^\top(\boldsymbol{\alpha})\boldsymbol{\Sigma}\nabla_g(\boldsymbol{\alpha})$$

9.2.2 Application to the Normal Distribution

The following are some general relationships between the central moments (the μ's) and the moments (the α's) which are valid for any distribution.

$$\begin{aligned} \mu_1 &= 0 \\ \alpha_1 &= \mu \\ \mu_2 &= \alpha_2 - \mu^2 \\ \alpha_2 &= \mu_2 + \mu^2 \\ \mu_3 &= \alpha_3 - 3\alpha_2 + \mu^3 \\ \alpha_3 &= \mu_3 + 3\alpha^2\mu - \mu^3 \\ \mu_4 &= \alpha_4 - 4\alpha_3\mu + 6\alpha_2\mu^2 - 3\mu^4 \\ \alpha_4 &= \mu_4 + 4\alpha_3\mu - 6\alpha_2\mu^2 + 3\mu^4 \end{aligned}$$

Suppose now that X is normal with mean μ and variance σ^2. Then we have

$$\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= \sigma^2 \\ \mu_3 &= 0 \\ \mu_4 &= 3\sigma^4 \end{aligned}$$

and hence for the normal distribution

$$\begin{aligned}\alpha_1 &= \mu \\ \alpha_2 &= \sigma^2 + \mu^2 \\ \alpha_3 &= 3\sigma^2\mu + \mu^3 \\ \alpha_4 &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4\end{aligned}$$

It follows that

$$\begin{aligned}\text{var}(X) &= \sigma^2 \\ \text{cov}(X, X^2) &= E(X^3) - E(X^2)E(X) \\ &= 3\sigma^2\mu + \mu^3 - (\sigma^2 + \mu^2)\mu \\ &= 2\mu\sigma^2 \\ \text{var}(X^2) &= E(X^4) - [E(X^2)]^2 \\ &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 - (\sigma^2 + \mu^2)^2 \\ &= 2\sigma^4 + 4\mu^2\sigma^2\end{aligned}$$

Thus

$$\sqrt{n}\begin{bmatrix}\overline{X}^1 - E(X) \\ \overline{X}^2 - E(X^2)\end{bmatrix}$$

converges in distribution to

$$\text{N}\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 2\sigma^4 + 4\mu^2\sigma^2\end{bmatrix}\right)$$

Example 1. Asymptotic distribution of s^2. If we let

$$g(\overline{x}^1, \overline{x}^2) = \overline{x}^2 - [\overline{x}^1]^2$$

Then

$$g(\overline{x}^1, \overline{x}^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

and hence

$$\frac{\partial g(\overline{x}^1, \overline{x}^2)}{\partial \overline{x}^1} = -2\overline{x}^1$$

$$\frac{\partial g(\overline{x}^1, \overline{x}^2)}{\partial \overline{x}^2} = 1$$

Evaluating at $\overline{x}^1 = \mu$ and $\overline{x}^2 = \sigma^2 + \mu^2$ yields

$$\frac{\partial g(\mu, \sigma^2)}{\partial \mu} = -2\mu$$

and

$$\frac{\partial g(\mu, \sigma^2)}{\partial \sigma^2} = 1$$

It follows that the asymptotic distribution of S^2 satisfies

$$\sqrt{n}(S^2 - \sigma^2) \xrightarrow{d} \mathrm{N}(0, v^2)$$

where

$$\begin{aligned} v^2 &= [-2\mu, 1] \begin{bmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 2\sigma^4 + 4\mu^2\sigma^2 \end{bmatrix} \begin{bmatrix} -2\mu \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 0, 2\sigma^4 \end{bmatrix} \begin{bmatrix} -2\mu \\ 1 \end{bmatrix} \\ &= 2\sigma^4 \end{aligned}$$

Since $\overline{X}^1 = \overline{X}$ and S^2 are independent it follows that their joint distribution satisfies

$$\sqrt{n} \begin{bmatrix} \overline{X} - \mu \\ S^2 - \sigma^2 \end{bmatrix} \xrightarrow{d} \mathrm{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \right)$$

Example 2 (Effect size). The **effect size** is defined as

$$\frac{\mu}{\sigma}$$

It is a widely used measure of the importance of a variable. If we let

$$g(\overline{x}, s^2) = \overline{x}_1 (s^2)^{-1/2}$$

then we have a natural estimate of the effect size based on the method of moments.

Note that

$$\frac{\partial g(\overline{x}, s^2)}{\partial \overline{x}} = (s^2)^{-1/2}$$

$$\frac{\partial g(\overline{x}, s^2)}{\partial s^2} = -\overline{x}(s^2)^{-3/2}/2$$

Evaluating these at $\overline{x} = \mu$ and $s^2 = \sigma^2$ we have

$$\frac{\partial g(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma}$$

and

$$\frac{\partial g(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{\mu}{2\sigma^3}$$

It follows that

$$\sqrt{n}\left(\frac{\overline{x}}{s} - \frac{\mu}{\sigma}\right) \xrightarrow{d} \mathrm{N}(0, v^2)$$

where

$$v^2 = \left[\begin{array}{cc} \frac{1}{\sigma} & -\frac{\mu}{2\sigma^3} \end{array}\right] \left[\begin{array}{cc} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{array}\right] \left[\begin{array}{c} \frac{1}{\sigma} \\ -\frac{\mu}{2\sigma^3} \end{array}\right]$$

which reduces to

$$v^2 = 1 + \frac{\mu^2}{2\sigma^2}$$

Example 3 (Coefficient of variation). The **coefficient of variation** is defined as

$$\frac{\sigma}{\mu}$$

and is a widely used measure of variability.

If we let $cv = \frac{1}{\overline{x}/s}$ then we have a natural estimate of the coefficient of variation.

It follows that

$$\sqrt{n}\left(\frac{s}{\overline{x}} - \frac{\sigma}{\mu}\right) \xrightarrow{d} \mathrm{N}(0, v_1^2)$$

where

$$v_1^2 = \left(-\frac{\sigma^2}{\mu^2}\right)\left(1 + \frac{\mu^2}{2\sigma^2}\right)\left(-\frac{\sigma^2}{\mu^2}\right)$$

which reduces to

$$v_1^2 = \frac{\sigma^2}{\mu^2}\left(\frac{1}{2} + \frac{\sigma^2}{\mu^2}\right)$$

9.3 Estimating Functions

The method of moments and maximum likelihood are examples of obtaining estimates using estimating functions.

Definition 9.3.1. A function $\mathbf{g}$ such that the equation

$$\mathbf{g}(\mathbf{y}; \widehat{\boldsymbol{\theta}}) = \mathbf{0}$$

defines $\widehat{\boldsymbol{\theta}}$ as an estimate of $\boldsymbol{\theta}$ is called an **estimating function**. The equation itself is called an **estimating equation**.

Definition 9.3.2. The estimating function $\mathbf{g}$ is an unbiased estimating function if

$$E[\mathbf{g}(\mathbf{Y}; \boldsymbol{\theta})] = \mathbf{0} \text{ for all } \boldsymbol{\theta}$$

9.3.1 General Linear Model

Example 1. In a general linear model, i.e.,

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \;\; ; \;\; \text{var}\,(\mathbf{Y}) = \sigma^2\mathbf{I}$$

where $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}$ is $n \times (p+1)$, the estimating function

$$\mathbf{g}(\mathbf{y}; \boldsymbol{\beta}) = \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

defines the **least squares estimate** of $\boldsymbol{\beta}$.

9.3.2 Maximum Likelihood

Example 2. If $\mathbf{Y}$ has density $f(\mathbf{y}\,;\,\boldsymbol{\theta})$ the estimating function

$$\mathbf{g}(\mathbf{y}\,;\,\boldsymbol{\theta}) = \frac{\partial \ln[f(\mathbf{y}\,;\,\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}$$

defines the **maximum likelihood estimate** of $\boldsymbol{\theta}$.

9.3.3 Method of Moments

Example 3. Let $Y_1, Y_2, \ldots, Y_n$ be iid $f(y\,;\,\boldsymbol{\theta})$ and define

$$\bar{y^r} = \frac{\sum_{i=1}^{n} y_i}{n} \;\;;\;\; \mu_r(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(Y^r)$$

Then the estimating function $\mathbf{g}(\mathbf{y};\boldsymbol{\theta})$ with rth component equal to

$$\bar{y^r} - \mu_r(\boldsymbol{\theta})$$

defines the **moment estimator** of $\boldsymbol{\theta}$.

9.3.4 Generalized Linear Models

Example 4. Let $Y_1, Y_2, \ldots, Y_n$ be independent where f_i is of the exponential type, i.e.,

$$f_i(y_i\,;\,\boldsymbol{\theta}) = \exp\left\{\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)}\right] + c_i(y_i\,;\,\phi)\right\}$$

Then it is easy to show that

$$\mu_i = \mathbb{E}(Y_i) = b^{(1)}(\theta_i)$$

A function h such that

$$h(\mu_i) = h[b^{(1)}(\theta_i)] = \mathbf{x}_i^{\top}\boldsymbol{\beta}$$

is called a **link function** and $\eta_i = h(\mu_i)$ is called a **linear predictor**.

The link is called canonical if

$$\eta_i = \mathbf{x}_i^T\boldsymbol{\beta} = \theta_i$$

and in this case

$$\mu_i(\boldsymbol{\beta}) = b^{(1)}(\theta_i)$$

For canonical links the maximum likelihood estimating equations are given by

$$\sum_{i=1}^{n}\left[\frac{y_i - \mu_i(\boldsymbol{\beta})}{a_i(\phi)}\right]\frac{\partial\theta_i}{\partial\beta_j} = 0$$

for $j = 1, 2, \ldots, p$

Note that

$$\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \beta_j} = b^{(2)}(\theta_i)\frac{\partial \theta_i}{\partial \beta_j}$$

so that

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \beta_j}}{b^{(2)}(\theta_i)}$$

Thus the maximum likelihood equations are

$$\sum_{i=1}^{n}\left[\frac{y_i - \mu_i(\boldsymbol{\beta})}{v_i}\right]\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \beta_j} \text{ for } j = 1, 2, \ldots, p$$

where v_i is the variance of Y_i. In matrix form the maximum likelihood equations are

$$\sum_{i=1}^{n}\left[\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{\top}\left[\frac{y_i - \mu_i(\boldsymbol{\beta})}{v_i}\right] = \mathbf{0}$$

These equations specialize to the general linear model, the logistic regression model, the log linear model, and many other commonly used models.

Each of the above examples yields an unbiased estimating function. In R we have the packages LM and GLM.

9.3.5 Quasi-Likelihood

Example 5. The estimating function

$$\sum_{i=1}^{n}\left[\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{\top} v_i^{-1}(y_i - \mu_i(\boldsymbol{\beta}))$$

where v_i is the variance of Y_i defines the **quasi-likelihood estimator** and it can be used regardless of whether the family is of the exponential type since it depends only on the mean and variance of Y_i.

9.3.6 Generalized Estimating Equations

Example 6. Consider clustered data (either defined as repeated measures over time on the same individual or as clusters defined by family or environmental facts). Specifically let the observations (responses) from the ith cluster be

$$(y_{i1}, y_{i2}, \ldots, y_{in_i}) \text{ for } i = 1, 2, \ldots, m$$

Let

$$\mathbb{E}(Y_{ij}) = \mu_{ij} \text{ where } h(\mu_{ij}) = \mathbf{x}_{ij}^\top \theta$$

where h is a link function and let

$$\mu_i(\theta)^\top = (\mu_{i1}(\theta), \mu_{i2}(\theta), \ldots, \mu_{in_i}(\theta))$$

for $i = 1, 2, \ldots, m$.

The **GEE** estimating equations are defined by

$$\sum_{i=1}^{m} \left[\frac{\partial \mu_i(\theta)}{\partial \theta}\right]^\top [\mathbb{V}(\mathbf{Y}_i)]^{-1} [\mathbf{y}_i - \mu_i(\theta)] = \mathbf{0}$$

In R there is a package GEEpack (and others). These methods were introduced by Liang and Zeger. See Diggle et al. [11] for details.

9.4 Generalized Method of Moments

Suppose there exists a function $g\ \mathcal{X} \times \Theta \mapsto \mathbb{R}^p$ such that

$$\mu_g(\theta_0) = \mathbb{E}\{g(\mathbf{X}, \theta_0)\} = \mathbf{0}$$

where $\mu_g(\theta_0) \neq \mathbf{0}$ for $\theta \neq \theta_0$.

The **generalized method of moments** replaces $\mathbb{E}$ by $\widehat{E}$, the sample average, to obtain

$$\widehat{\mu}_g(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta)$$

Then $\widehat{\theta}$ is chosen to minimize

$$\widehat{\mu}_g(\theta)^\top \mathbf{W} \widehat{\mu}_g(\theta)$$

where W is a weighting matrix (assumed positive definite). The optimum choice of $\mathbf{W}$ is $\mathbf{\Sigma}$ where

$$\mathbf{\Sigma} = \text{Var}_{\theta_0}\left\{\frac{\partial g(Y, \theta)}{\partial \theta}\right\}$$

Under weak conditions, such a $\widehat{\theta}$ satisfies:

- Consistency
- Asymptotic normality, i.e.,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \text{MVN}(\mathbf{0}, \mathbf{G}^\top \mathbf{\Sigma} \mathbf{G})$$

where

$$\mathbf{G} = \frac{\partial g(Y, \theta)}{\partial \theta}$$

All of these facts arise from routine Taylor's expansions. In R there is a package GMM.

9.5 The Bootstrap

Most estimation methods have the property that they produce estimators which have the property that

$$\frac{\widehat{\theta}_n - \theta}{\text{s.e.}(\widehat{\theta}_n)} \xrightarrow{d} \text{N}(0, 1)$$

so that

$$\widehat{\theta}_n \pm z_{1-\alpha/2} \text{s.e.}(\widehat{\theta}_n)$$

is an approximate $100(1-\alpha)\%$ confidence interval for θ.

9.5.1 Basic Ideas

The bootstrap, developed by Bradley Efron, is a method which can be used, with few assumptions, to estimate the standard error of a statistic and to calculate approximate confidence intervals for the parameter the statistic estimates.

Assume that T_n is a statistic, that is, T_n is some function of the observed data which is a random sample from F. The variance of T_n and distribution of T_n is of interest.

We write

$$\mathbb{V}_F(T_n)$$

to denote this variance and note that it depends on the unknown F.

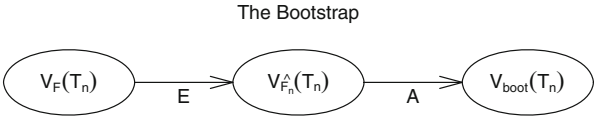

Fig. 9.1 The bootstrap

The following two steps constitute the basis of the bootstrap (Fig. 9.1):

1. **Estimate** $\mathbb{V}_F(T_n)$ by $\mathbb{V}_{\widehat{F}_n}(T_n)$
2. **Approximate** $\mathbb{V}_{\widehat{F}_n}(T_n)$ by **simulation**

1. The approximation error of F by the sample distribution function is the most likely of the approximations to be large since it requires that the sample distribution function be close, in some sense, to the true distribution function.
2. Thus it will work well if the sample is "representative" and if n is not too small.
3. The approximation error of the sampling distribution of T_n, assuming that $\widehat{F}_n$ is the true distribution function, by simulation is expected to be small.

9.5.2 Simulation Background

1. If $Y_1, Y_2, \ldots, Y_B$ is a random sample from a population with distribution G then the law of large numbers implies that

$$\frac{1}{B}\sum_{i=1}^{B} Y_j \quad \xrightarrow{p} \quad \mathbb{E}(Y)$$

   i.e., if we draw a (large) sample from population G we can approximate $\mathbb{E}(Y)$ by the sample mean.
2. This result is easily generalizable to any function of Y, say $h(Y)$, which has finite mean, i.e.,

$$\frac{1}{B}\sum_{i=1}^{B} h(Y_i) \quad \xrightarrow{p} \quad \mathbb{E}[h(Y)]$$

3. Assuming that variances exist it follows that

$$\frac{1}{B}\sum_{i=1}^{B}(Y_i - \overline{Y}_B)^2 = \frac{1}{B}\sum_{i=1}^{B} Y_i^2 - (\overline{Y}_B)^2$$

   converges in probability to $\mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2$, i.e., to $\mathbb{V}(Y)$.

4. Thus the sample variance of the Y_i's can be used to approximate the variance of G. It follows that if we can simulate random samples from a population with distribution G, then we can get a good approximation to the expected value and variance of G.

R and other computer packages provide functions which allow selection of random samples from a variety of distributions, e.g., rnorm, rgamma, rbinom, etc. For other distributions and to understand how random samples are generated recall the following basic result from probability theory.

If X is a random variable with a continuous distribution function F then the random variable $U = F(X)$ has a uniform distribution on the interval $[0, 1]$.

Proof.

$$\begin{aligned} F_U(u) &= \mathbb{P}(U \leq u) \\ &= \mathbb{P}(\{x \ : \ F(x) \leq u\}) \\ &= \mathbb{P}(\{x \ : \ x \leq F^{-1}(u)\}) \\ &= F[F^{-1}(u)] \\ &= u \end{aligned}$$

This result is called the **probability integral transformation** and provides, among other things, a method of obtaining a random observation from any continuous distribution. Simply generate a random uniform, then, $F^{-1}(U)$has distribution F. More generally, generate $u_1, u_2, \ldots, u_n$, independent with each observation on a uniform on $[0, 1]$. Then

$$x_1 = F^{-1}(u_1), x_2 = F^{-1}(u_2), \ldots, x_n = F^{-1}(u_n)$$

is a random sample from F.

Computer scientists have discovered much more efficient ways to generate such samples, but the above result is important because it shows that we can always simulate from any distribution function.

9.5.3 Variance Estimation Using the Bootstrap

It is clear from the previous section that we can use simulation to approximate $\mathbb{V}_{\widehat{F}_n}(T_n)$. This requires the simulation of the distribution of T_n when the data are assumed to have population distribution $\widehat{F}_n$.

Note that $\widehat{F}_n$ puts probability mass $1/n$ on each sample point. Thus, drawing an observation from $\widehat{F}_n$ is equivalent to drawing one point at random from the original data set, i.e., to simulate

$$X_1^*, X_2^*, \ldots, X_n^*$$

from $\widehat{F}_n$ it is sufficient to draw n observations from the original data set $x_1, x_2, \ldots, x_n$ **with replacement**.

Assuming that $\widehat{F}_n$ adequately estimates F we thus have one sample from the original distribution function. Hence we can, by simulation, approximate the sampling variance of the statistic T_n.

Here is the bootstrap method for variance estimation.

1. Draw n observations $x_1^*, x_2^*, \ldots, x_n^*$ at random, with replacement from the original data set.
2. Compute the statistic $T_n^* = g(x_1^*, x_2^*, \ldots, x_n^*)$.
3. Repeat steps 1 and 2 a large number, B, of times to obtain

$$T_{n1}^*, T_{n2}^*, \ldots, T_{nB}^*$$

called the **bootstrap replicates** and their sample mean.

$$\overline{T}_n^* = \frac{1}{B}\sum_{i=1}^{n} T_{ni}^*$$

4. The bootstrap estimate of the variance of T_n is then given by

$$\text{var}_{bs} = \frac{1}{B}\sum_{i=1}^{B}(T_{ni}^* - \overline{T}_n^*)^2$$

Note that $T_{n1}^*, T_{n2}^*, \ldots, T_{nB}^*$ can be used to estimate the distribution function of T_n (Fig. 9.2).

9.6 Confidence Intervals Using the Bootstrap

9.6.1 Normal Interval

There are many ways to find confidence intervals using the bootstrap.

If the distribution of $T_n = \widehat{\theta}_n$ is approximately normal, then use

$$\widehat{\theta}_n \pm z_{1-\alpha/2}\widehat{s.e.}_{bs}$$

In the R function **boot.ci** this method is called "norm."

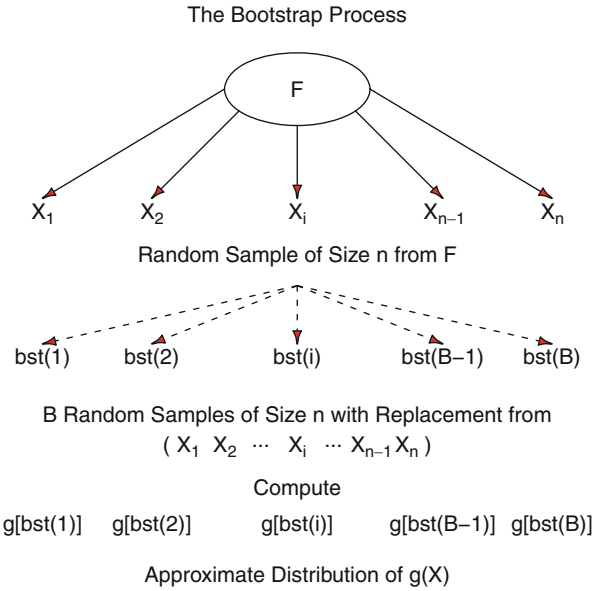

Fig. 9.2 The bootstrap process

9.6.2 Pivotal Interval

Recall that a **pivot**, $p(Y, \theta)$, is any function of a random variable Y and a parameter θ such that the distribution of $p(Y, \theta)$ does not depend on θ.

Example. The best known example of a pivot is

$$Z_n = \frac{\sqrt{n}(\overline{Y}_n - \theta)}{\sigma}$$

where

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and the Y_i's are iid each normal with mean μ and known variance σ^2. The distribution of Z_n is normal with mean 0 and variance 1 and does not depend on μ.

The standard inversion shows that

$$\overline{y}_n \pm z_{1-\alpha/2}\text{se}$$

is a $100(1-\alpha)\%$ confidence interval for μ.

The pivotal method for the bootstrap defines

$$R_n(\widehat{\theta}, \theta) = \widehat{\theta}_n - \theta$$

and assumes that it is a pivot with distribution function H. If H is known, standard inversion gives the confidence interval. Since H is unknown, it is estimated from the quantiles of the bootstrap.

In the R function **boot.ci**, this interval is called "basic."

9.6.3 Percentile Interval

Given the bootstrap replicates

$$\widehat{\theta}^*_{n1}, \widehat{\theta}^*_{n2}, \ldots, \widehat{\theta}^*_{nB}$$

The percentile interval is simply defined as

$$\left[\theta^*_{\alpha/2}\ , \theta^*_{1-\alpha/2}\right]$$

where $\theta^*_{\alpha/2}$ is the $\alpha/2$ quantile of the set of bootstrap replicates and $\theta^*_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the set of bootstrap replicates.

In the R function **boot.ci** this interval is called "perc."

The R library **boot** has a wide variety of bootstrap functions.

9.6.4 Parametric Version

There is also a parametric version of the bootstrap in which we

1. Assume the model density is known.
2. Estimate parameters by maximum likelihood or some other methods.
3. Use the estimates to draw random bootstrap samples from the known distribution, substituting the estimated parameter values for the parameters.
4. Use the resulting bootstrap distribution to assess standard errors, confidence limits, etc.
5. This version is particularly useful to check on approximations such as the delta method.

9.6.5 Dangers of the Bootstrap

All you ever learn using the bootstrap, without further modeling assumptions, are properties of $\widehat{F}_n$. Unless you have a way of saying how much and/or in what ways knowledge of $\widehat{F}_n$ can be transformed into knowledge of F, the bootstrap can only tell you about $\widehat{F}_n$, not about F [46].

9.6.6 The Number of Possible Bootstrap Samples

If we have a sample size of n there are only

$$\binom{2n-1}{n-1}$$

possible bootstrap samples. To see this imagine n boxes defined by $n-1$ lines

$$|\,|\ \cdots\ |\,|$$

The first bootstrap observation can be put one of the n boxes, the second into $n+1$ possible positions, the third $n+2, \ldots$, the nth into $2n-1$ positions.

1. The total is

$$n(n-1)\cdots(2n-1) = (2n-1)_{(n-1)} = \frac{(2n-1)!}{(n-1)!}$$

2. The balls can be ordered in $n!$ ways so that the total number of possible samples is

$$\frac{(2n-1)!}{(n-1)!n!} = \binom{2n-1}{n-1}$$

3. Thus it would be possible to enumerate all the possible samples.

   Recalling that (Stirling's Approximation)

$$r! \approx (2\pi r)^{-1/2} r^r e^{-r}$$

we have that

$$\begin{aligned}\binom{2n-1}{n-1} &\approx \frac{[2\pi(2n-1)]^{-1/2}(2n-1)^{2n-1}e^{-(2n-1)}}{[2\pi n]^{1/2} n^n e^n [2\pi(n-1)]^{1/2}(n-1)^{n-1}} \\ &= \left[\frac{2n-1}{2\pi n(n-1)}\right]^{1/2} \frac{\left\{n\left[2-\frac{1}{n}\right]\right\}^{2n-1}}{n^n\left\{n\left[1-\frac{1}{n}\right]\right\}^{n-1}}\end{aligned}$$

$$= \sqrt{\frac{2-\frac{1}{n}}{2\pi n(1-\frac{1}{n})}} \frac{\left[2-\frac{1}{n}\right]^{2n-1}}{\left[1-\frac{1}{n}\right]^{n-1}}$$
$$\approx (\pi n)^{-1/2} 2^{2n-1}$$

Thus we have

n	5	10	15	20	30
Samples	10^2	10^5	10^8	10^{11}	10^{17}

One can also show that the original sample is the most probable of these to occur.