2 Estimation

2.1 Data and questions

Data set 2.1 *Wages in the USA*

These data are taken from Dyer, D. (1981), *Canadian Journal of Statistics*, **9**. They comprise annual wages (in multiples of US$100) of 30 production line workers in a large American firm.

Table 2.1 *US wages*

112	123	128	116	119
154	103	132	140	111
119	115	107	108	101
108	107	151	105	157
112	125	103	158	112
156	119	104	104	115

□

Data set 2.2 *Severe ideopathic respiratory distress syndrome*

The data were published by van Vliet, P.K. and Gupta, J.M. (1973), Sodium bicarbonate in ideopathic respiratory distress syndrome, *Arch. Diseases in Childhood*, **48**, 249-255. The condition, known by its acronym SIRDS, is very serious and can result in the death. The data comprise birth weights (kg) of 50 infants who displayed SIRDS

Table 2.2 Birth weights (kg) of infants with severe ideopathic respiratory distress syndrome

1.050*	2.500*	1.890*	1.760	2.830
1.175*	1.030*	1.940*	1.930	1.410
1.230*	1.100*	2.200*	2.015	1.715
1.310*	1.185*	2.270*	2.090	1.720
1.500*	1.225*	2.440*	2.600	2.040
1.600*	1.262*	2.560*	2.700	2.200
1.720*	1.295*	2.730*	2.950	2.400
1.750*	1.300*	1.130	3.160	2.550
1.770*	1.550*	1.575	3.400	2.570
2.275*	1.820*	1.680	3.640	3.005

* child died

There are important questions to be asked about these data. Can we identify children at risk quickly and accurately? Is it possible to relate the risk of death to birthweight?
□

Data set 2.3 *Silver content of Byzantine coins*

A number of coins from the reign of King Manuel I, Comnemus (1143 - 80) were discovered in Cyprus. They arise from four different coinages at intervals throughout his reign. The question of interest is whether there is any significant difference in their silver content with the passage of time; there is a suspicion that it was deliberately and steadily reduced. The data give the silver content (%Ag) of the coins.

Table 2.3 Silver content of coins

First	Second	Third	Fourth
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8
6.9			
6.2			

On the face of it the suspicion could be correct in that the fourth coinage would seem to have lower silver content than, say, the first coinage, but there is a need for firm statistical evidence if it is to be confirmed.

□

Data set 2.4 *Radiocarbon dating*

Radiocarbon datings have given historians a powerful tool for determining more precisely the periods during which ancient civilisations flourished. The following set of data came from the Lake Lamoka site and were contributed by S.R. Wilson from the Australian National University to Andrews, D.F. and Herzberg, A.M. (1985), *Data*, Springer-Verlag: New York.

Table 2.4 Radiocarbon dating

Sample number	Radiocarbon age determination
C-288	2419
M-26	2485
C-367	3433
M-195	2575
M-911	2521
M-912	2451
Y-1279	2550
Y-1280	2540

□

2.2 Desirable properties of estimators

Suppose we have $X = X_1, X_2, \ldots, X_n$ drawn from a distribution with some parameter θ. Often $X_1, X_2, \ldots, X_n$ form a *sample.*

Definition 2.0 *Estimators*

An *estimator* $\widehat{\theta}_n$ of θ is just a function of the observed data which (we hope) forms a useful approximation to the parameter:

$$\widehat{\theta}_n = g(X_1, X_2, \ldots, X_n).$$

Note that $\widehat{\theta}_n$ can depend only on the observed data, and not on any unknown parameters. You have already met a number of estimators in Mods, such as. the sample mean and sample variance. The estimator is a function of random variables, so is itself a random variable, with a distribution, mean, and variance, etc.

Here is a familiar property which you have already met.

Definition 2.1 *Unbiasedness*
$\widehat{\theta}_n$ is said to be *unbiased* for θ if

$$E\left(\widehat{\theta}_n\right) = \theta, \quad \forall \theta \in \Theta.$$

□

Next a new property, but one which appeals to common-sense. The idea is that, the larger the amount of data, the closer the estimate should be to the parameter to be estimated. This is expressed in terms of the estimator converging in probability to the parameter value.

Convergence in probability (borrowed from probability course)
A sequence of random variables $Z_1, Z_2, \ldots, Z_n, ...$ is said to *converge in probability* to a constant z if for any $\epsilon > 0$, as $n \to \infty$

$$P(|Z_n - z| > \epsilon) \to 0.$$

Definition 2.2 *Consistency*
$\widehat{\theta}_n$ is said to be *consistent* for θ if

$$\widehat{\theta}_n \xrightarrow{P} \theta.$$

□

Definition 2.3 *Efficiency*
$\widehat{\theta}_A$ is said to be more *efficient* than $\widehat{\theta}_B$ if

$$V\left(\widehat{\theta}_A\right) < V\left(\widehat{\theta}_B\right), \quad \forall \theta \in \Theta.$$

□

Again, this appeals to common-sense.

2.3 Revision and extension of maximum likelihood estimation

The basic idea starts with the joint distribution of $X = X_1, X_2, \ldots, X_n$ depending upon a parameter θ,

$$f(\mathbf{x}; \theta) = f(x_1, x_2, \ldots, x_n; \theta).$$

For fixed θ, probability statements can be made about X. If we have observations, x, but θ is unknown, we regard information about θ as being contained in the likelihood

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta),$$

where L is regarded as a function of θ with $\mathbf{x}$ fixed.

Example 2.1

Suppose $X = X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables with parameter $\theta \in [0, 1]$.

$$\textit{i.e.}\quad P(X_i = 1) = \theta,\ P(X_i = 0) = 1 - \theta.$$

Observations are $x = (1, 0, 0, 1, 0, 1, 1)$ and

$$\begin{aligned} L(\theta; \mathbf{x}) &= \prod_{i=1}^{7} f(x_i; \theta) \\ &= \prod_{i=1}^{7} \theta^{x_i} (1 - \theta)^{1 - x_i} \\ &= \theta^4 (1 - \theta)^3. \end{aligned}$$

In general, for a sample size n,

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i} = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}.$$

■

Maximum likelihood estimation

The value of θ which maximises $L(\theta)$ is called the *maximum likelihood estimate* of θ.

Example 2.1 (continued)

In the previous example, we can find this by differentiating $L(\theta)$, or equivalently by differentiating $l(\theta) = \log L(\theta)$.

$$\begin{aligned} l(\theta) &= \textstyle\sum_i x_i \log\theta + (n - \sum_i x_i) \log(1 - \theta) \\ \frac{\partial l(\theta)}{\partial \theta} &= \textstyle\sum_i x_i / \theta - (n - \sum_i x_i) / (1 - \theta) \end{aligned}$$

Putting $\frac{\partial l(\theta)}{\partial \theta} = 0$ we obtain

$$\theta = \sum\nolimits_i x_i \Big/ n.$$

This leads us to $\widehat{\theta}_n$, the *maximum likelihood estimator* based on a sample of size n

$$\widehat{\theta}_n = \sum\nolimits_i X_i \Big/ n.$$

■

Example 2.2

Suppose $X = X_1, X_2, \ldots, X_n$ are independent normal random variables, $N(\mu, \sigma^2)$.

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \end{aligned}$$

and

$$\begin{aligned} l(\mu, \sigma^2) &= \log L(\mu, \sigma^2) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

Differentiating,

$$\begin{aligned} \frac{\partial l(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \\ \frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

Equating these derivatives to zero results in

$$\widehat{\mu} = \overline{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

■

2.4 Order statistics

To expand our armoury of useful and descriptive statistics we shall need to know about *order statistics.*

Definition 2.4 *Order statistics*

Suppose that X_1 , X_2, . . , X_n is a random sample then the full set of *order statistics* is the re-ordering $X_{(1)}$, $X_{(2)}$, . . , $X_{(n)}$ such that $X_{(1)} \leq X_{(2)} \leq . . \leq X_{(n)}$. In particular the minimum order statistic is $X_{(1)} = \min X_i$ and the maximum order statistic is $X_{(n)} = \max\ X_i$. If the random sample is derived from a continuous distribution then all the inequalities will be strict.
□

Example 2.3 *US wages*

The data in Table 2.1 are shown below, with the data having been ordered.

Table 2.1: *US wages*

101	107	112	119	140
103	107	112	119	151
103	108	115	123	154
104	108	115	125	156
104	111	116	128	157
105	112	119	132	158

You can see that the realisation of the minimum order statistic is 101 and the realisation of the maximum order statistic is 158.
■

The distributions of $X_{(1)} = \min\{X_i\}$ and $X_{(n)} = \max\{X_i\}$ are straightforward to derive. The c.d.f. of $X_{(1)}$ is

$$\begin{aligned} F_{(1)}(x) &= 1 - P\left(X_{(1)} > x\right)\right) = 1 - P\left(X_1 > x, \ldots, X_n > x\right) \\ &= 1 - [1 - F(x)]^n \end{aligned}$$

so that, differentiating,

$$f_{(1)}(x) = F'_{(1)}(x) = n\,[1 - F(x)]^{n-1}\,f(x).$$

Similarly

$$F_{(n)}(x) = P\left(X_{(n)} \leq x\right) = F(x)^n$$

giving

$$f_{(n)}(x) = nF(x)^{n-1}f(x).$$

Example 2.4

Suppose $X_i \sim U(0,\theta)$. Then

$$F(x) = \begin{cases} 0, & x < 0, \\ \dfrac{x}{\theta}, & 0 \leq x < \theta, \\ 1, & x \geq \theta, \end{cases} \qquad \text{and} \quad f(x) = \begin{cases} \dfrac{1}{\theta}, & x \in (0,\theta), \\ 0, & \text{otherwise.} \end{cases}$$

The p.d.f.'s of $X_{(1)} = \min\{X_i\}$ and $X_{(n)} = \max\{X_i\}$ are given by

$$f_{(1)}(x) = \frac{n\,(\theta - x)^{n-1}}{\theta^n}, \quad x \in (0,\theta)$$

and

$$f_{(n)}(x) = \frac{nx^{n-1}}{\theta^n}, \quad x \in (0,\theta).$$

■

For a continuous random sample, each with density function $f(x)$ and cumulative distribution function $F(x)$, the joint density function of the order statistics is

$$f(x_{(1)}, x_{(2)}, .., x_{(n)}) = n! \prod_{i=1}^{n} f(x_{(i)}), \quad x_{(1)} < x_{(2)} < \cdots < x_{(n)},$$

and the marginal distribution for $X_{(r)}$ is

$$f_{(r)}(x) = \frac{n!}{(r-1)!(n-r)!} F(x)^{r-1}(1-F(x))^{n-r} f(x).$$

Lemma 2.1 The p.d.f. of $X_{(r)}$ is given by

$$f_{(r)}(x) = \frac{n!}{(r-1)!(n-r)!} F(x)^{r-1} \left[1 - F(x)\right]^{n-r} f(x).$$

□

Proof Putting $r = 1$ and $r = n$ in the expression gives

$$f_{(1)}(x) = n\left[1 - F(x)\right]^{n-1} f(x), \quad f_{(n)}(x) = nF(x)^{n-1} f(x).$$

The c.d.f. $F_{(r)}$ of $X_{(r)}$ is

$$F_{(r)}(x) = P\left(X_{(r)} \leq x\right) = \sum_{j=r}^{n} \binom{n}{j} \left[F(x)\right]^j \left[1 - F(x)\right]^{n-j}$$

i.e. the probability that at least r of the X_i's are less than or equal to x.

Therefore

$$F_{(r)}(x) - F_{(r+1)}(x) = \binom{n}{r} \left[F(x)\right]^r \left[1 - F(x)\right]^{n-r}$$

and, by differentiation,

$$\begin{aligned} f_{(r+1)}(x) &= f_{(r)}(x) - \binom{n}{r} [F(x)]^{r-1} [1-F(x)]^{n-r-1} [r - nF(x)] f(x) \\ &= \frac{n!}{r!(n-r)!} F(x)^{r-1} [1-F(x)]^{n-r-1} (n-r) F(x) f(x) \\ &= \frac{n!}{r!\,(n-(r+1))!} F(x)^{(r+1)-1} [1-F(x)]^{n-(r+1)} f(x), \end{aligned}$$

so the proposition is proved inductively.
■

There are many uses of the order statistics in terms of analysing and describing data, three in particular being the *median*, the *lower quartile* and the *upper quartile*.

Definition 2.5 *Median*

The *median* of a random sample is defined by

$$X_m = X_{\left(\frac{1}{2}(n+1)\right)}.$$

This is the central order statistic if the size of the sample is odd and is the average of the two central order statistics if it is even, so that

$$\text{if} \quad n = 2k+1, \text{ then } X_m = X_{(k+1)}, \quad \text{and if} \quad n = 2k, \text{ then } X_m = \frac{1}{2}\left(X_{(k)} + X_{(k+1)}\right).$$

□

Example 2.3 (revisited) *US wages*
For the wages data, the realisation is $x_m = \frac{1}{2}(115 + 115)$. Thus 50% of the data have a smaller value than the median and 50% of the data have a higher value.
■

Definition 2.6 *Quartiles*

The *lower quartile* is defined by

$$q_L = X_{\left(\frac{1}{4}(n+1)\right)},$$

and the *upper quartile* is defined by

$$q_U = X_{\left(\frac{3}{4}(n+1)\right)}.$$

□

Example 2.3 (revisited) *US wages*

The wages data comprises 30 values and $31/4 = 7\frac{3}{4}$, so the realisation of the lower quartile is $q_L = x_{\left(7\frac{3}{4}\right)} = 107 + \frac{3}{4}(107 - 107) = 107$. The realisation of the upper quartile is $q_U = x_{\left(23\frac{1}{4}\right)} = 128 + \frac{1}{4}(132 - 128) = 129$.
■

You will see one of the main uses of order statistics in the next sub-section, where we look at how to produce useful plots of the data.

2.4.1 Plotting the data

For univariate data there are many useful plots each of which attempts to give a graphical representation of the data. These include histograms, boxplots and plots of empirical cumulative distribution functions (see *Rice, Chapter* 10 for examples and descriptions of these plots). For paired data, scatter plots are a very quick way of visually checking for association between the paired observations.

Boxplots

Boxplots are particularly useful in giving an intuitive feel for the way the data are distributed. They are very easy to construct, but their main advantage is that all computer packages can produce them on-screen instantly. Let us look at how they are constructed.

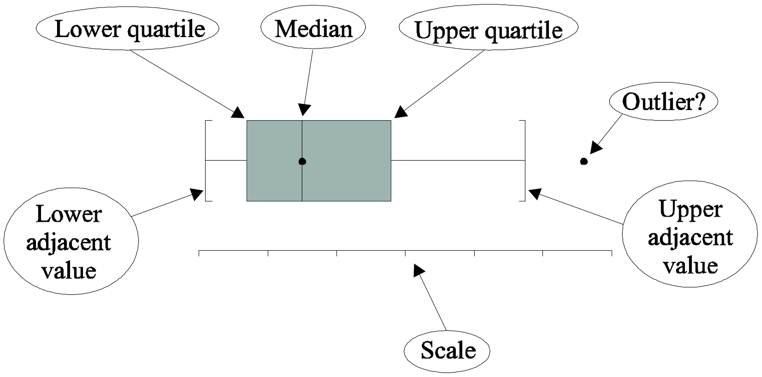


Figure 2.1 Boxplot construction

With one exception the plot is largely self-explanatory; that exception is the idea of adjacent values. These are calculated by first calculating the inter-quartile range $q_U - q_L$; then the upper adjacent value is the largest value not exceeding that distance from the upper quartile, the lower adjacent value is the smallest value not less than that distance from the lower quartile. Any point which lies outside these boundaries is depicted as a point and considered as a possible outlier (i.e. a value which is non-typical for some reason or other). Thus we have a picture which, in some sense, represents the way the data are distributed.

Example 2.3 (revisited) *US wages*
A boxplot of the wage data, produced using the R package, is shown in Figure 2.2.

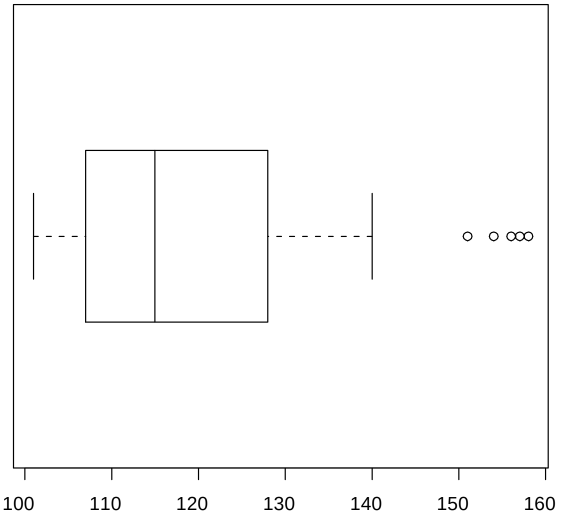


Figure 2.2 Boxplot of US wages data

It shows some skewness in the data and some possible outliers, but overall you can see that most of the workers at that time earned between $10,000 and $14,000 with a median wage around $11,500 and 50% of workers earning between just over $10,500 and just under $13,000.

■

But boxplots are *really* useful when it comes to comparison.

Example 2.5 *Infants with SIRDS*

The data set in Table 2.2 is reproduced below.

Table 2.2 Birth weights (kg) of infants with severe ideopathic respiratory distress syndrome

1.050*	2.500*	1.890*	1.760	2.830
1.175*	1.030*	1.940*	1.930	1.410
1.230*	1.100*	2.200*	2.015	1.715
1.310*	1.185*	2.270*	2.090	1.720
1.500*	1.225*	2.440*	2.600	2.040
1.600*	1.262*	2.560*	2.700	2.200
1.720*	1.295*	2.730*	2.950	2.400
1.750*	1.300*	1.130	3.160	2.550
1.770*	1.550*	1.575	3.400	2.570
2.275*	1.820*	1.680	3.640	3.005

* child died

One way of comparing the birthweights of infants who died with the birthweights of those who survived is to use a graph with the boxplots compared against the same scale.

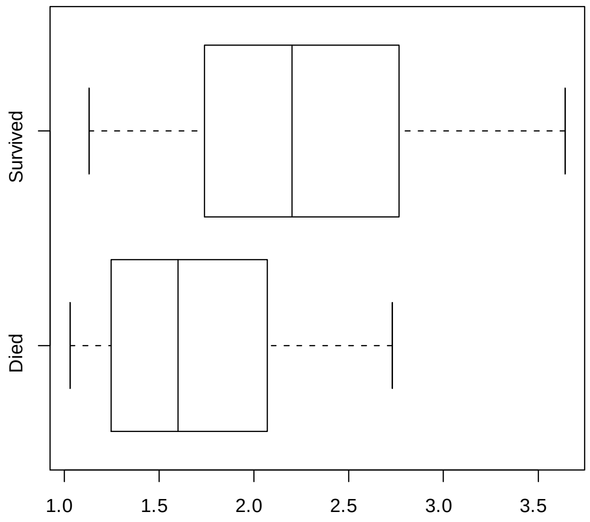


Figure 2.3 Boxplots of SIRDS data

The first thing you notice about this plot is that the median value of birtheweight for children who died is less than the lower quartile for those who survived. The plot makes it immediately clear that the two groups have typically different birthweights.

■

Probability Plots

These are sometimes called *quantile-quantile plots.* They can be used to examine the plausibility of a density function from which a set of data could have been drawn. If such a known distribution is found then analysis of the data is performed more readily, particularly if this distribution should turn out to be normal. We need some theory before we develop the graphical method.

Definition 2.7

Let X be a continuous random variable with cumulative distribution function $F_X(x)$, then the *probability integral transform* is

$$Y = F_X(X).$$

□

The transformed random variable Y has a particularly useful property.

Lemma 2.2 For $F_X(x)$ strictly monotonic increasing on (a, b) , X taking values only in (a, b) , then if $Y = F_X(X)$,

$$Y \sim U(0, 1).$$

Y so defined is uniformly distributed on $(0, 1)$.
□

Proof For $0 \le y \le 1$,

$$P\left(Y \le y\right) = P\left(F_X(X) \le y\right) = P\left(X \le F_X^{-1}(y)\right) = F_X\left(F_X^{-1}(y)\right) = y,$$

giving the cumulative distribution function of the uniform distribution on $(0, 1)$.
■

We shall also need a result concerning the expected values of order statistics from a uniform distribution.

Lemma 2.3 If $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ are the order statistics of a random sample drawn from a uniform distribution $U(0, 1)$, then

$$E\left(Y_{(k)}\right) = \frac{k}{n+1} = P\left(Y_i \le \frac{k}{n+1}\right).$$

[Proof left as an exercise]
□

We conclude that, for the uniform distribution, $E\left(Y_{(1)}\right), E\left(Y_{(2)}\right), \ldots, E(Y_{(n)})$ splits the interval $(0, 1)$ into equal bites of probability, each of length $\frac{1}{n+1}$.
Using the results of the two lemmas, it is now possible to give the method of construction for a probability plot. The question we wish to address is:
'Is the data set plausibly drawn from a distribution with cdf $F(x)$?'
To construct a probability plot, you take the following steps:

(i) Arrange the data as $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. From Lemma 2.2, if the supposition is correct, $F(x_{(1)}), F(x_{(2)}), \ldots, F(x_{(n)})$ are realisations of order statistics from $U(0, 1)$.

(ii) Solve
$$F(z_{(k)}) = \frac{k}{n+1}$$
for $z_{(k)}$: remember that $E\left(F(X_{(k)}\right) = \frac{k}{n+1}$.

(iii) Plot the ordered pairs $\left(x_{(k)}, z_{(k)}\right)$.

If the supposition is reasonable we should get an approximate straight line, $z = x$.

If the distribution contains some unknown parameters we need to adapt the method, if possible.

Example 2.6

Suppose we wish to check whether a distribution is exponential with mean μ. Then we note $F(x) = 1 - e^{-x/\mu}$, so that

$$\frac{k}{n+1} \approx 1 - \mathrm{e}^{-x_{(k)}/\mu}.$$

Re-arranging

$$x_{(k)} \approx -\mu \log\left(1 - \frac{k}{n+1}\right).$$

Hence, if the distribution is plausibly exponential, plot $\left(x_{(k)}, \log\left(1 - \frac{k}{n+1}\right)\right)$ and expect to see an approximate straight line. The slope will provide an estimate of μ.

■

Example 2.3 (revisited) *US wages*

Many financial modellers assume that wages have a *Pareto distribution* with c.d.f.

$$F\left(x\right) = 1 - \left(\frac{\alpha}{x}\right)^{\theta}, \quad x > \alpha,$$

where α represents the minimum wage and both α and θ are unknown parameters. Writing

$$\frac{k}{n+1} \approx 1 - \left(\frac{\alpha}{x_{(k)}}\right)^{\theta}$$

and re-arranging,

$$x_{(k)} \approx \alpha\left(1 - \frac{k}{n+1}\right)^{-1/\theta}.$$

Clearly we cannot produce a probability plot because we do not know the value of θ, but taking logs of both sides gives

$$\log x_{(k)} \approx \log \alpha - \frac{1}{\theta} \log\left(1 - \frac{k}{n+1}\right),$$

so plotting ordered pairs

$$\left(\log x_{(k)}, \log\left(1 - \frac{k}{n+1}\right)\right)$$

should produce a straight line if the distribution is plausible as a model for the US wages data.

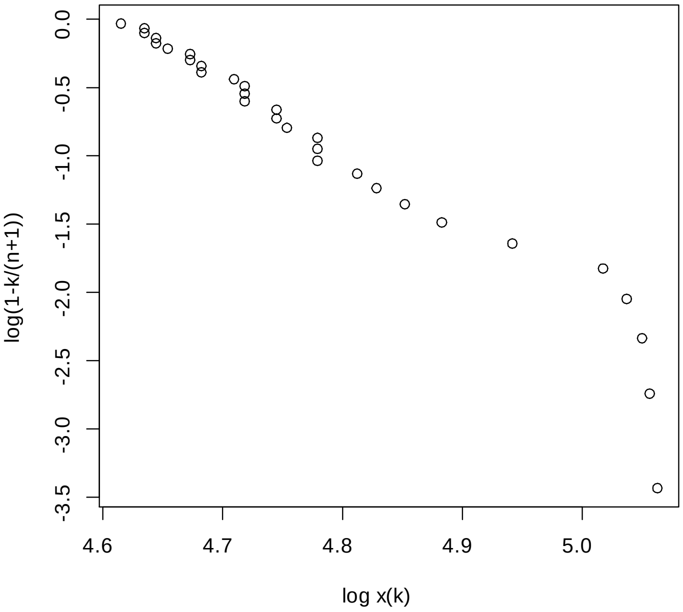


Figure 2.4 Pareto probability plot for US wages data

As you can see, a Pareto probability model doesn't look very convincing.
■

In practice the distribution we wish to consider is frequently the normal distribution, which has two unknown parameters. We do a little more adjustment.

Definition 2.8 *Normal scores*

The numbers

$$z_{(k)} = \Phi^{-1}\left(\frac{k}{n+1}\right),$$

where Φ is the c.d.f. of $Z \sim N(0,1)$, are called the *normal scores*. They are the $\dfrac{k}{n+1}$ quantiles of the standard normal distribution.
□

We use the normal scores to construct a *normal probability plot*. Suppose that we wish to check whether or not the data could have been drawn from a normal distribution $N\left(\mu, \sigma^2\right)$. As above take the ordered data $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. Note that, if $X \sim N\left(\mu, \sigma^2\right)$, then

$$\frac{X-\mu}{\sigma} \sim N(0,1).$$

Plot the pairs$\big(x_{(k)}, z_{(k)}\big)$. If the data are plausibly normal then we should have an approximate straight line with slope $1/\sigma$ and intercept μ/σ.

In practice you will find that most statistical computer packages will generate normal scores and produce probability plots: in fact, most packages will do this for a variety of distributions. The plots in the examples which follow were generated using the R package.

Example 2.7 *Silver content of Byzantine coins*

The data are reproduced below.

Table 2.3 Silver content of coins

First	Second	Third	Fourth
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8
6.9			
6.2			

The graph shows a normal probability plot of the silver content of the first coinage.

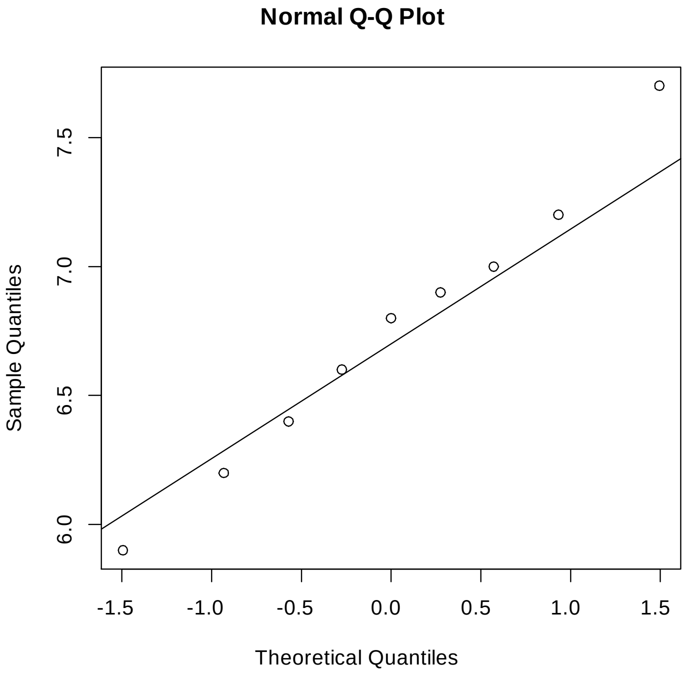


Figure 2.5 Silver content of coins

It looks as if a normal distribution is appropriate.
■

Example 2.8 *Radio-carbon dating*

The data from Table 2.4 are reproduced below.

Table 2.4 Radiocarbon dating

Sample number	Radiocarbon age determination
C-288	2419
M-26	2485
C-367	3433
M-195	2575
M-911	2521
M-912	2451
Y-1279	2550
Y-1280	2540

Radioactive-carbon dating was undertaken on 8 samples from a single early site. There is one rather obvious outlier which is easily seen from the plot.

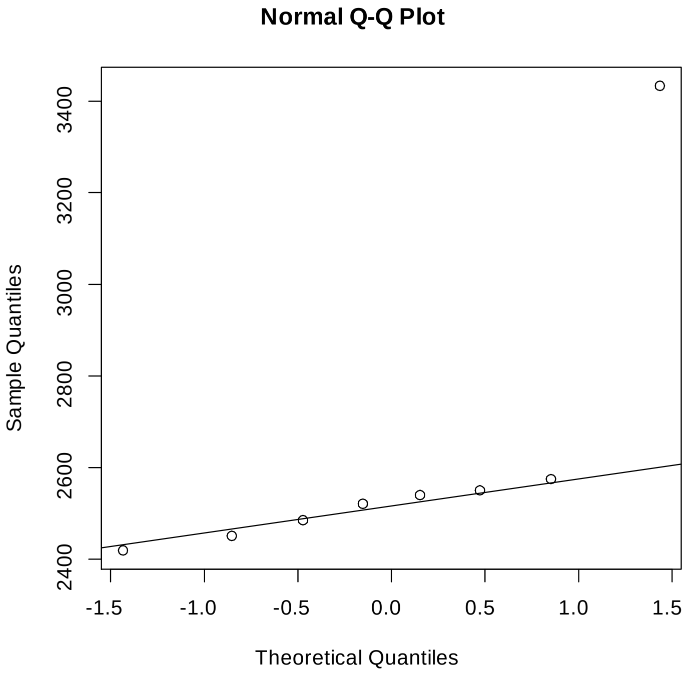


Figure 2.6 Radiocarbon dating

If possible we would check out the outlier with the laboratory and the collector but, since this is not possible, we will drop it and consider whether the remaining data points could have been drawn from a normal distribution.

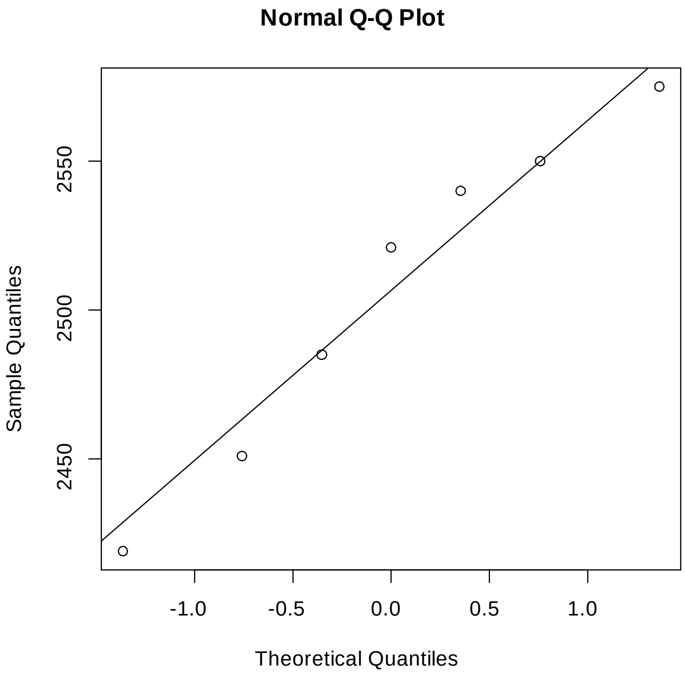


Figure 2.7 Radiocarbon dating

As can be seen from Figure 2.7, the 7 remaining data points do give an approximate straight line on a normal probability plot.
■

2.5 Comparing estimators

Remember that we began with three desirable properties of estimators. These were

Unbiasedness $E\left(\widehat{\theta}_n\right) = \theta, \quad \forall \theta \in \Theta.$

Consistency $\widehat{\theta}_n \xrightarrow{P} \theta.$

Efficiency $\widehat{\theta}_A$ is more *efficient* than $\widehat{\theta}_B$ if $V\left(\widehat{\theta}_A\right) < V\left(\widehat{\theta}_B\right), \quad \forall \theta \in \Theta.$

Even if we assume unbiasedness and consistency to be desirable, it is possible to have more than one such estimator.

Example 2.9

Consider the linear estimator

$$\widehat{\theta}_n = \sum_{i=1}^{n} a_i X_i$$

where $E\left(X_i\right)=\theta,\ V\left(X_i\right)=\sigma^2$ for $1\leq i\leq n$.

$$E\left(\widehat{\theta}_n\right)=\sum_{i=1}^n a_i E\left(X_i\right)=\theta\sum_{i=1}^n a_i,$$

so the estimator is unbiased provided

$$\sum_{i=1}^n a_i=1.$$

For i.i.d. random variables,

$$V\left(\widehat{\theta}_n\right)=\sum_{i=1}^n a_i^2\sigma^2.$$

Now

$$\begin{aligned} \sum_{i=1}^n\left(a_i-\frac{1}{n}\right)^2 &\geq 0\\ \Rightarrow\quad \sum_{i=1}^n a_i^2-\frac{2}{n}\sum_{i=1}^n a_i+\frac{1}{n} &\geq 0\\ \text{and, if }\ \textstyle\sum_{i=1}^n a_i=1,\qquad \Rightarrow\quad \sum_{i=1}^n a_i^2 &\geq \frac{1}{n}. \end{aligned}$$

Equality occurs iff $a_i=\dfrac{1}{n},\ 1\leq i\leq n$.

The conclusion then is that, if $\widehat{\theta}_n$ is a linear unbiased estimator of the form $\sum_{i=1}^n a_iX_i$ and if $\overline{X}=\dfrac{1}{n}\sum_{i=1}^n X_i$, then

$$V\left(\overline{X}\right)\leq V\left(\widehat{\theta}_n\right).$$

Of all linear unbiased estimators, $\overline{X}$ is most efficient.

■

2.5.1 Food for thought

1. Is the sample mean the “best” estimator of the distribution mean?

2. If $\overline{X}$ is unbiased for θ, is $g\left(\overline{X}\right)$ unbiased for $g\left(\theta\right)$?

3. Is there any reason to doubt the principle of seeking the unbiased estimator with the minimum variance?

4. Can a biased estimator be “better” (whatever that means!) than an unbiased estimator?

Is the sample mean the "best" estimator of the distribution mean?

Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$.

$$E\left(\overline{X}\right) = \theta, \quad V\left(\overline{X}\right) = \frac{1}{12n}.$$

Now consider $\widehat{\theta}_n = \frac{1}{2}\left(X_{(1)} + X_{(n)}\right)$. Obviously $\widehat{\theta}_n$ is symmetrically distributed about θ so

$$E\left(\widehat{\theta}_n\right) = \theta.$$

Now

$$\begin{aligned} f_{X_{(1)}X_{(n)}}(u,v) &= n(n-1)\left[F(v) - F(u)\right]^{n-2} f(u)f(v), \\ &\qquad \theta - \tfrac{1}{2} < u < v < \theta + \tfrac{1}{2}. \end{aligned}$$

where

$$F(x) = \begin{cases} 0, & x < \theta - \frac{1}{2}, \\ x + \frac{1}{2} - \theta, & \theta - \frac{1}{2} \le x < \theta + \frac{1}{2}, \\ 1, & x \ge \theta + \frac{1}{2}, \end{cases}$$

and

$$f(x) = F'(x),$$

so that

$$\begin{aligned} f_{X_{(1)}X_{(n)}}(u,v) &= n(n-1)\left[v - u\right]^{n-2}, \\ &\qquad \theta - \tfrac{1}{2} < u < v < \theta + \tfrac{1}{2}. \end{aligned}$$

You have my word as a gentleman that, after a lot of boring slog,

$$V\left(\widehat{\theta}_n\right) = \frac{1}{2(n+1)(n+2)}$$

and

$$V\left(\widehat{\theta}_n\right) < V\left(\overline{X}\right) \quad \text{for} \quad n > 2.$$

Question: Does this mean we can sometimes do better by ignoring some of the data values?

If $\overline{X}$ is unbiased for θ, is $g\left(\overline{X}\right)$ unbiased for $g(\theta)$?

Let

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x \ge 0.$$

Then

$$E\left(\overline{X}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \theta, \quad V\left(\overline{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(X_i\right) = \frac{\theta^2}{n}.$$

Now consider $g(x) = x^2$.

$$\begin{aligned} E\left[g\left(\overline{X}\right)\right] &= E\left(\overline{X}^2\right) \\ &= V\left(\overline{X}\right) + E\left(\overline{X}\right)^2 \\ &= \frac{\theta^2}{n} + \theta^2 = \theta^2\left(1 + \frac{1}{n}\right). \end{aligned}$$

Is there any reason to doubt the principle of seeking the unbiased estimator with the minimum variance?
Finals question 1979 − IV − 15, part (iv).

$X_1, X_2, \ldots, X_n$ is a random sample from a Poisson distribution with mean λ. Hence

$$Y = \sum_{i=1}^{n} X_i \sim Poisson\,(n\lambda)\,.$$

Part (iv) asks for an unbiased minimum variance estimator of $e^{-2n\lambda}$.

Let this be $g(Y)$. Then

$$E\left[g(Y)\right] = \sum_{y=0}^{\infty} g(y) \frac{e^{-n\lambda}(n\lambda)^y}{y!} = e^{-2n\lambda}$$

or

$$\sum_{y=0}^{\infty} g(y) \frac{(n\lambda)^y}{y!} = e^{-n\lambda}.$$

But

$$e^{-x} = \sum_{k=0}^{\infty} (-1)^k \frac{x^k}{k!}$$

so that

$$g(y) = (-1)^y\,, \quad y = 0, 1, \ldots$$

i.e.

$$g(y) = \begin{cases} -1, & y \text{ odd}, \\ 1, & y \text{ even}. \end{cases}$$

THIS IS COMPLETE AND UTTER RUBBISH!

Can a biased estimator be "better" than an unbiased estimator?
Look at the distributions of the following two estimators of θ.

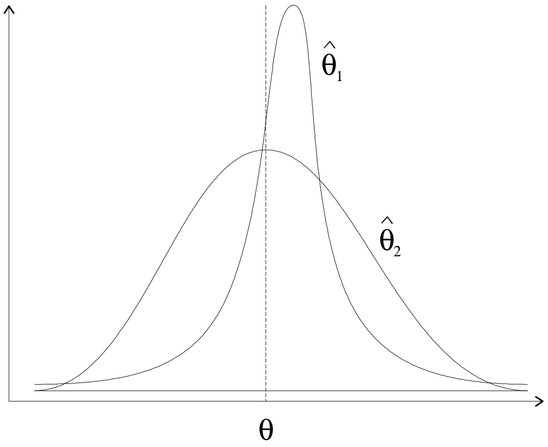


Figure 2.8

Which do you prefer?

Even though $\widehat{\theta}_1$ is biased,

$$E\left[\left(\widehat{\theta}_1 - \theta\right)^2\right] < V\left(\widehat{\theta}_2\right).$$

$\widehat{\theta}_1$ has smaller *mean squared error.*

Are m.l.e.'s the answer?
The maximum likelihood principle is intuitively appealing and does not involve worries about bias.

(i) Maximum likelihood estimators are asymptotically unbiased.

(ii) If $\widehat{\theta}$ is the m.l.e. of θ, then $g\left(\widehat{\theta}\right)$is the m.l.e. of $g\left(\theta\right)$.

This is the *invariance property* of maximum likelihood estimators.

Are there *any* problems with m.l.e.'s? Consider $X_1, X_2, \ldots, X_n$ where $X_i \sim U(\theta - \frac{1}{2}, \theta + \frac{1}{2}),\ 1 \leq i \leq n$, and note that $\theta - \frac{1}{2} \leq x_{(1)} < x_{(n)} \leq \theta + \frac{1}{2}$.

The likelihood function is

$$L(\theta; \mathbf{x}) = \begin{cases} 1, & x_{(n)} - \frac{1}{2} \leq \theta \leq x_{(1)} + \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

Clearly

$$\max_\theta \{L(\theta; \mathbf{x})\} = 1 \quad \forall \theta \in \left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$$

i.e. the likelihood function is maximised for *every* statistic $\widehat{\theta}\,(X_1, X_2, \ldots, X_n)$ such that $\widehat{\theta}(\mathbf{X}) \in \left[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}\right]$.

Conclusion: An m.l.e. may not be unique.

2.6 More about likelihood

2.6.1 Invariance property of m.l.e.'s

Lemma 2.4 If $\widehat{\theta}$ is an m.l.e. of θ and if g is a function, then $g\left(\widehat{\theta}\right)$ is an m.l.e. of $g(\theta)$.
□

Proof If g is one-to-one, then

$$L(\theta) = L\left(g^{-1}\left(g(\theta)\right)\right)$$

are both maximised by $\widehat{\theta}$, so

$$\begin{aligned} \widehat{\theta} = &\ g^{-1}\left(\widehat{g(\theta)}\right) \\ \text{or} \qquad & \\ g\left(\widehat{\theta}\right) = &\ \widehat{g(\theta)}. \end{aligned}$$

If g is many-to-one, then $\widehat{\theta}$ which maximises $L(\theta)$ still corresponds to $g\left(\widehat{\theta}\right)$, so $g\left(\widehat{\theta}\right)$ still corresponds to the maximum of $L(\theta)$
■

Example 2.10

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution $B(1, \theta)$. Consider m.l.e.'s of the mean, θ, and variance, $\theta(1-\theta)$.

Note, by the way, that $\theta(1-\theta)$ is not a 1-1 function of θ.

The log-likelihood is

$$\begin{aligned} l(\theta) = &\ \textstyle\sum_i x_i \log\theta + (n - \sum_i x_i)\log(1-\theta) \\ \text{and} \qquad & \\ \frac{dl(\theta)}{d\theta} = &\ \textstyle\sum_i x_i/\,\theta - (n - \sum_i x_i)/\,(1-\theta) \end{aligned}$$

so it is easily shown that the m.l.e. of θ is $\widehat{\theta} = \overline{X}$.

Putting $\nu = \theta(1-\theta)$,

$$\frac{dl(\nu)}{d\nu} = \frac{dl\left(\nu(\theta)\right)}{d\theta} \cdot \frac{d\theta}{d\nu}$$

so it is easily seen that, since $\frac{d\theta}{d\nu}$ is not, in general, equal to zero,

$$\widehat{\nu} = \nu\left(\widehat{\theta}\right) = \overline{X}\left(1-\overline{X}\right).$$

■

2.6.2 Relative likelihood

If $\sup_\theta L(\theta) < \infty$, the *relative likelihood* is

$$RL(\theta) = \frac{L(\theta)}{\sup_\theta L(\theta)}; \quad 0 \le RL(\theta) \le 1.$$

Relative likelihood is invariant to known 1-1 transformations of x, for if y is a 1-1 function of x,

$$f_Y(y;\theta) = f_X\left(x(y);\theta\right)\left|\frac{dx}{dy}\right|.$$

$\left|\frac{dx}{dy}\right|$ is independent of θ, so

$$RL_X(\theta) = RL_Y(\theta).$$

2.6.3 Likelihood summaries

Realistic statistical problems often have many parameters. These cause problems because it can be hard to visualise $L(\theta)$, and it becomes necessary to use summaries.

Key idea

In large samples, log-likelihoods are often approximately quadratic near the maximum.

Example 2.11

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ.

i.e.

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \ge 0.$$

Then

$$l(\lambda) = n\log\lambda - \lambda\textstyle\sum_i x_i, \qquad \frac{dl(\lambda)}{d\lambda} = n/\,\lambda - \textstyle\sum_i x_i,$$

$$\frac{d^2l(\lambda)}{d\lambda^2} = -\,n/\,\lambda^2, \qquad \frac{d^3l(\lambda)}{d\lambda^3} = 2n/\,\lambda^3.$$

The log-likelihood has a maximum at $\widehat{\lambda} = n \left/ \sum_i x_i \right.$, so

$$\begin{aligned} RL(\lambda) = & \left(\frac{\lambda}{\widehat{\lambda}}\right)^n e^{n-\lambda\sum_i x_i} \\ = & \left(\frac{\lambda}{\widehat{\lambda}} e^{1-\lambda/\widehat{\lambda}}\right)^n, \qquad \lambda > 0. \\ \to & \ 1 \quad \text{as} \quad \lambda \to \widehat{\lambda}. \end{aligned}$$

Now, what does the likelihood look like in the neighbourhood of $\widehat{\lambda}$, as $n \to \infty$?

$$\begin{aligned} \log RL(\lambda) &= l(\lambda) - l(\widehat{\lambda}) \\ &= l(\widehat{\lambda}) + l'(\widehat{\lambda})\left(\lambda - \widehat{\lambda}\right) + \frac{1}{2}l''(\widehat{\lambda})\left(\lambda - \widehat{\lambda}\right)^2 - l(\widehat{\lambda}) \\ &\quad + O(\lambda - \widehat{\lambda})^3 \end{aligned}$$

using Taylor series.

Now $l'(\widehat{\lambda}) = 0$ and $l''(\widehat{\lambda}) = -n \left/ \widehat{\lambda}^2 \right.$, so

$$\log RL(\lambda) \simeq -\frac{n\left(\lambda - \widehat{\lambda}\right)^2}{2\widehat{\lambda}^2} \to -\infty \quad \text{as} \quad n \to \infty$$

unless $\lambda = \widehat{\lambda}$.

Thus, as $n \to \infty$,

$$RL(\lambda) \to \begin{cases} 1, & \lambda = \widehat{\lambda}, \\ 0, & \text{otherwise.} \end{cases}$$

Conclusion

Likelihood becomes more concentrated about the maximum as $n \to \infty$, and values far from the maximum become less and less plausible.

■

In general

We call the value $\widehat{\theta}$ which maximises $L(\theta)$ or, equivalently, $l(\theta) = \log L(\theta)$ the *maximum likelihood estimate*, and

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2}$$

is called the *observed information*.

Usually $J(\theta) > 0$ and $J(\widehat{\theta})$ measures the concentration of $l(\theta)$ at $\widehat{\theta}$. Close to $\widehat{\theta}$, we summarise

$$l(\theta) \simeq l(\widehat{\theta}) - \frac{1}{2}\left(\theta - \widehat{\theta}\right)^2 J(\widehat{\theta}).$$

2.6.4 Information

In a model with log-likelihood $l(\theta)$, the *observed information* is

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2}.$$

When observations are independent, $L(\theta)$ is a product of densities so

$$l(\theta) = \sum_i \log f(x_i; \theta)$$

and

$$J(\theta) = -\sum_i \frac{\partial^2}{\partial \theta^2} \log f(x_i; \theta).$$

Since

$$l(\theta) \simeq l(\widehat{\theta}) - \frac{1}{2}\left(\theta - \widehat{\theta}\right)^2 J(\widehat{\theta}),$$

for θ near to $\widehat{\theta}$, we see that large $J(\widehat{\theta})$ implies that $l(\theta)$ is more concentrated about $\widehat{\theta}$.

This means that the data are less ambiguous about possible values of θ, *i.e.* we have more information about θ.

2.6.5 Expected information

Univariate distributions
Before an experiment is conducted, we have no data so that we cannot evaluate $J(\theta)$.

But we can find its expected value

$$I(\theta) = E\left(-\frac{\partial^2 l(\theta)}{\partial \theta^2}\right).$$

This is called the *expected information* or *Fisher's information.*

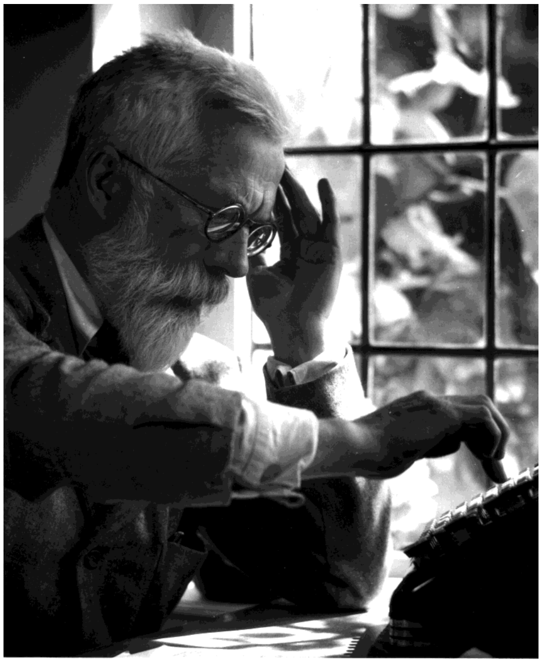
If the observations are a random sample, then the whole sample expected information is

$$I(\theta) = ni(\theta)$$

where

$$i(\theta) = E\left(-\frac{\partial^2}{\partial \theta^2} \log f(X_i; \theta)\right),$$

the single observation Fisher information.



Sir Ronald Aylmer Fisher (1890 - 1962)

Example 2.12

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Poisson distribution with parameter θ.

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!},$$

giving

$$l(\theta) = \log L(\theta) = \sum_i x_i \log \theta - n\theta - \sum_i \log x_i!$$

Thus

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2} = \sum\nolimits_i x_i \Big/ \theta^2.$$

To find $I(\theta)$, we need $E\left(X_i\right) = \theta$ and

$$I(\theta) = \frac{1}{\theta^2} \sum_i E\left(X_i\right) = \frac{n}{\theta}.$$

■

Multivariate distributions

If $\boldsymbol{\theta}$ is a $(p \times 1)$ vector of parameters, then $\mathbf{I}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ are $(p \times p)$ matrices.

$$\{\mathbf{J}(\boldsymbol{\theta})\}_{rs} = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} \quad \text{and} \quad \{\mathbf{I}(\boldsymbol{\theta})\}_{rs} = E\left(-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s}\right).$$

These matrices are obviously symmetric.

We can also write the above as

$$\mathbf{J}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \quad \text{and} \quad \mathbf{I}(\boldsymbol{\theta}) = E\left(-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right).$$

Example 2.13

$X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with parameters μ and σ^2. We have already seen that

$$L\left(\mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum\nolimits_i \left(x_i - \mu\right)^2\right],$$

so

$$l\left(\mu, \sigma^2\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum\nolimits_i \left(x_i - \mu\right)^2.$$

and

$$\begin{aligned}
\frac{\partial l}{\partial\mu} &= \frac{1}{\sigma^2}\sum\nolimits_i \left(x_i - \mu\right), \\
\frac{\partial l}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum\nolimits_i \left(x_i - \mu\right)^2, \\
\frac{\partial^2 l}{\partial\mu^2} &= -\frac{n}{\sigma^2}, \\
\frac{\partial^2 l}{\partial\mu\partial\sigma^2} &= -\frac{1}{\sigma^4}\sum\nolimits_i \left(x_i - \mu\right). \\
\frac{\partial^2 l}{\partial\left(\sigma^2\right)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum\nolimits_i \left(x_i - \mu\right)^2.
\end{aligned}$$

$$\mathbf{J}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum\nolimits_i \left(x_i - \mu\right) \\ \frac{1}{\sigma^4}\sum\nolimits_i \left(x_i - \mu\right) & \frac{1}{\sigma^6}\sum\nolimits_i \left(x_i - \mu\right)^2 - \frac{n}{2\sigma^4} \end{pmatrix}.$$

To find $\mathbf{I}(\mu, \sigma^2)$, use

$$\begin{aligned}
E\left(X_i\right) &= \mu, \\
V\left(X_i\right) &= E\left[\left(X_i - \mu\right)^2\right] = \sigma^2,
\end{aligned}$$

so that

$$\mathbf{I}(\mu, \sigma^2) = E\left(\mathbf{J}(\mu, \sigma^2)\right) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Example 2.14 *Censored exponential data*

Lifetimes of n components, safety devices, etc. are observed for a time c, when r have failed and $(n-r)$ are still OK.

We have two kinds of observation:

(i) Exact failure times x_i observed if $x_i \le c$, so that

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \ge 0;$$

(ii) x_i unobserved if $x_i > c$,

$$P\left(X > c\right) = e^{-\lambda c}.$$

Data are therefore $x_1, \ldots, x_r, \underbrace{c, \ldots, c}_{n-r \text{ times}}$

The $(n-r)$ components, safety devices, etc. which have not failed are said to be *censored.*

The likelihood is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{r} \lambda e^{-\lambda x_i} \prod_{i=r+1}^{n} e^{-\lambda c} \\ &= \lambda^r \exp\left[-\lambda\left(\sum_{i=1}^{r} x_i + (n-r)c\right)\right]. \end{aligned}$$

$$\begin{aligned} l(\lambda) &= r\log\lambda - \lambda\left(\textstyle\sum_{i=1}^{r} x_i + (n-r)c\right) \\ l'(\lambda) &= r/\lambda - \left(\textstyle\sum_{i=1}^{r} x_i + (n-r)c\right) \\ l''(\lambda) &= -r\left/\lambda^2\right. . \end{aligned}$$

Thus $J(\lambda) = r\left/\lambda^2\right. > 0$ if $r > 0$ so we must observe *at least one* exact failure time.

$$I(\lambda) = E\left(r\left/\lambda^2\right.\right) = \frac{1}{\lambda^2} E\left(\#X_i \text{ observed exactly.}\right)$$

Now $P\left(X_i \text{ observed exactly}\right) = P\left(X_i \le c\right) = 1 - e^{-\lambda c}$, so

$$I_c(\lambda) = \frac{n\left(1 - e^{-\lambda c}\right)}{\lambda^2}.$$

No censoring if $c \to \infty$, giving

$$I_\infty(\lambda) = \frac{n}{\lambda^2} > I_c(\lambda)$$

as one might expect.

The asymptotic efficiency when there is censoring at c relative to no censoring is

$$I_c(\lambda)\,/I_\infty(\lambda) = 1 - e^{-\lambda c}.$$

■

Revision Example 2.15 *Events in a Poisson process*

Events are observed for period $(0, T)$.

n events occur at times $0 < t_1 < t_2 < \ldots < t_n < T$

Two observers A and B. A records exact times, B uses an automatic counter and goes to the pub (*i.e.* B merely records how many events there are).

A knows exact times, and times between events are independent and exponentially distributed, so

$$\begin{aligned} L_A(\lambda) &= \lambda e^{-\lambda t_1} \times \lambda e^{-\lambda(t_2 - t_1)} \times \cdots \times \lambda e^{-\lambda(t_n - t_{n-1})} \times e^{-\lambda(T - t_n)} \\ &= \lambda^n e^{-\lambda T}. \end{aligned}$$

B merely observes the event $[N = n]$, where $N \sim Poi(\lambda T)$, so

$$L_B(\lambda) = \frac{(\lambda T)^n\, e^{-\lambda T}}{n!}.$$

Log-likelihoods are

$$\begin{aligned} l_A(\lambda) &= n \log \lambda - \lambda T, \\ l_B(\lambda) &= n \log \lambda + n \log T - \lambda T - \log n! \end{aligned}$$

and

$$J_A(\lambda) = J_B(\lambda) = n\,/\lambda^2\ .$$

$E(N) = \lambda T$, so $I_A(\lambda) = I_B(\lambda) = T\,/\lambda$, and both observers get the same information. As usual, the one who went to the pub did the right thing.

■

2.6.6 Maximum likelihood estimates

The maximum likelihood estimate $\widehat{\theta}$ of θ maximises $L(\theta)$ and often (but not always) satisfies the *likelihood equation*

$$\frac{\partial l}{\partial \theta}\left(\widehat{\theta}\right) = 0,$$

with

$$J(\widehat{\theta}) = -\frac{\partial^2 l}{\partial \theta^2}\left(\widehat{\theta}\right) > 0$$

for a maximum.

In the vector case, $\widehat{\boldsymbol{\theta}}$ solves simultaneously

$$\frac{\partial l}{\partial \theta_r}\left(\widehat{\boldsymbol{\theta}}\right) = 0, \quad r = 1, \ldots, p,$$

with $\mathbf{J}(\widehat{\boldsymbol{\theta}})$ *positive definite* This implies

$$\det \mathbf{J}(\widehat{\boldsymbol{\theta}}) > 0$$

If the likelihood equation has many solutions, we find them all and check $L(\theta)$ for each. Usually, the equation has to be solved numerically. One way is by Newton-Raphson.

Suppose we have a starting value θ_0. Then

$$0 = \frac{\partial l}{\partial \theta}\left(\widehat{\theta}\right) \simeq \frac{\partial l}{\partial \theta}(\theta_0) + \frac{\partial^2 l}{\partial \theta^2}(\theta_0)\left(\widehat{\theta} - \theta_0\right)$$

which may be re-arranged to

$$\widehat{\theta} = \theta_0 + \frac{U(\theta_0)}{J(\theta_0)},$$

where

$$\begin{aligned} U(\theta) &= \frac{\partial l}{\partial \theta} \text{ is the } \textit{score function}, \\ J(\theta) &= -\frac{\partial^2 l}{\partial \theta^2} \text{ is the } \textit{observed information}. \end{aligned}$$

Now we iterate using θ_0 as a starting value and

$$\theta_{n+1} = \theta_n + \frac{U(\theta_n)}{J(\theta_n)}.$$

Example 2.16 *Extreme value (Gumbel) distribution*

This distribution is used to model such things as annual maximum temperature. Data due to Bliss on numbers of beetles killed by exposure to carbon disulphide are fitted by this model. The c.d.f. is

$$F(x) = \exp\left(-e^{-(x-\eta)}\right), \quad x \in \mathbb{R},\ \eta \in \mathbb{R},$$

and the density is

$$f(x) = \exp\left[-(x-\eta) - e^{-(x-\eta)}\right], \quad x \in \mathbb{R},\ \eta \in \mathbb{R}.$$

The sample log-likelihood is

$$l(\eta) = -\sum_i (x_i - \eta) - \sum_i e^{-(x_i-\eta)},$$

so that

$$\begin{aligned} U(\eta) &= n - \sum_i e^{-(x_i-\eta)}, \\ J(\eta) &= \sum_i e^{-(x_i-\eta)}. \end{aligned}$$

Starting at $\eta_0 = \overline{x}$, iterate using

$$\eta_{n+1} = \eta_n + \frac{n - \sum_i e^{-(x_i-\eta_n)}}{\sum_i e^{-(x_i-\eta_n)}.}$$

■

Fisher scoring
This simply involves replacing $J(\theta)$ with $I(\theta)$.

Example 2.17 *Extreme value distribution*
We need

$$\begin{aligned} I(\eta) &= E\left[J(\eta)\right] = \sum_i E\left[e^{-(X_i-\eta)}\right] \\ &= n\int_{-\infty}^{\infty} e^{-(x-\eta)} \exp\left[-(x-\eta) - e^{-(x-\eta)}\right] dx. \end{aligned}$$

Put $u = e^{-(x-\eta)}$ and the integral becomes

$$I(\eta) = n\int_0^{\infty} ue^{-u} du = n,$$

so Fisher scoring gives the iteration

$$\eta_{n+1} = \eta_n + 1 - \frac{1}{n}\sum_i e^{-(x_i-\eta_n)}.$$

■

2.6.7 Sufficient statistics

You have already seen a likelihood which cannot be summarised by a quadratic.

Example 2.18

$$f\left(x_i;\theta\right) = \theta^{-1}, \quad 0 < x_i < \theta,$$

so

$$L(\theta) = \theta^{-n}, \quad 0 < \max\left\{x_i\right\} < \theta.$$

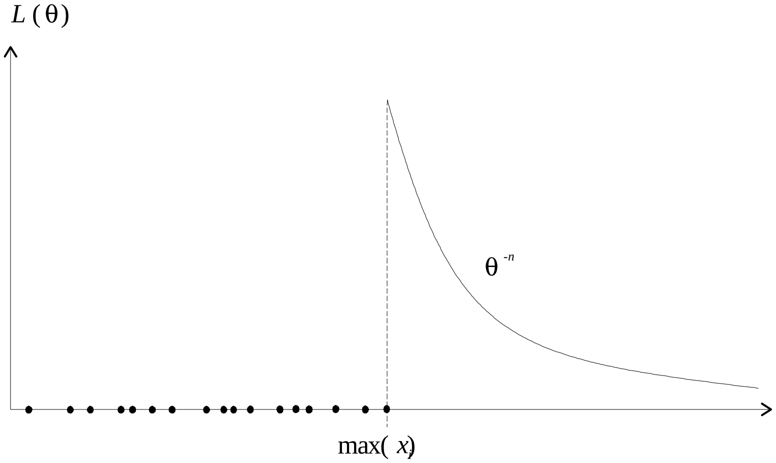


Figure 2.9

Clearly a quadratic approximation is useless here.

Definition 2.9 *Sufficient statistic*

If $S = s(\mathbf{X})$ is such that the conditional density $f_{\mathbf{X}|S}(\mathbf{x}|s;\theta)$ is independent of θ, then S is a *sufficient statistic* for θ.

The important question is:

> Does $s(\mathbf{x})$ reduce the dimensionality of the problem?□

As we will see, the definition is equivalent to saying that the likelihood $L(\theta)$ only depends upon data $\mathbf{x}$ through $s(\mathbf{x})$. So maximum likelihood inference also only depends on $s(\mathbf{x})$.

Example 2.19

Suppose $X_1, X_2 \sim B(n, \theta)$ and consider

$$\begin{aligned}
& P\left(X_1 = x | X_1 + X_2 = r\right) \\
= \; & \frac{P\left(X_1 = x, X_1 + X_2 = r\right)}{P\left(X_1 + X_2 = r\right)} \\
= \; & \frac{P\left(X_1 = x, X_2 = r - x\right)}{P\left(X_1 + X_2 = r\right)} \\
= \; & \frac{\binom{n}{x} \theta^{x} \left(1-\theta\right)^{n-x} \binom{n}{r-x} \theta^{r-x} \left(1-\theta\right)^{n-r+x}}{\binom{2n}{r} \theta^{r} \left(1-\theta\right)^{2n-r}} \\
= \; & \frac{\binom{n}{x}\binom{n}{r-x}}{\binom{2n}{r}}.
\end{aligned}$$

This does not contain θ, so that $X_1 + X_2$ is a sufficient statistic for θ.
■

Theorem 2.1 *Factorization Theorem*

$s(\mathbf{X})$ is a *sufficient statistic* for θ if and only if there exist functions g and h such that

$$f(\mathbf{x}; \theta) = g\left(s(\mathbf{x}); \theta\right) h(\mathbf{x})$$

for all $\mathbf{x} \in \mathbb{R}^n, \;\; \theta \in \Theta$.
□

Proof for discrete random variables

(i) Let $s(\mathbf{x}) = a$ and suppose the factorization condition to be satisfied, so that $f(\mathbf{x}; \theta) = g\left(s(\mathbf{x}); \theta\right) h(\mathbf{x})$.

Then

$$P\left(s(\mathbf{X}) = a\right) = \sum_{\mathbf{y} \in s^{-1}(a)} p(\mathbf{y}) = g(a; \theta) \sum_{\mathbf{y} \in s^{-1}(a)} h(\mathbf{y}).$$

Hence

$$P\left(\mathbf{X} = \mathbf{x} \mid s(\mathbf{X}) = a\right) = \frac{h(\mathbf{x})}{\sum\limits_{\mathbf{y} \in s^{-1}(a)} h(\mathbf{y})}$$

and this does not depend upon θ.

(ii) Let $s(\mathbf{X})$ be a sufficient statistic for θ. Then

$$P(\mathbf{X}=\mathbf{x}) = P\left(\mathbf{X}=\mathbf{x} \mid s(\mathbf{X})=a\right) P\left(s(\mathbf{X})=a\right).$$

But sufficiency $\Rightarrow P\left(\mathbf{X}=\mathbf{x} \mid s(\mathbf{X})=a\right)$ does not depend upon θ, so writing $P\left(s(\mathbf{X})=a\right) = g(a;\theta)$ and $P\left(\mathbf{X}=\mathbf{x} \mid s(\mathbf{X})=a\right) = h(\mathbf{x})$ gives the result.

The proof in the continuous case requires measure theory and is beyond the scope of this course.
■

Example 2.20

$X_1, X_2, \ldots, X_n \sim U(0,\theta)$, so that

$$L(\theta) = \theta^{-n}, \quad 0 < x_1, \ldots, x_n < \theta.$$

$$L(\theta) = \theta^{-n}, \quad \theta > x_{(n)}.$$

This factorizes with $s(\mathbf{x}) = x_{(n)}$ and $h\left(\mathbf{x}\right) = 1$, so that $X_{(n)}$ is a sufficient statistic for θ.■

Example 2.21

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution. Then

$$p(\mathbf{x};\theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}.$$

Trivially this factorizes with $s(\mathbf{x}) = \sum_i x_i$ and $h\left(\mathbf{x}\right) = 1$
■

Example 2.22

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a $N(\mu,\sigma^2)$ distribution, where $(\mu,\sigma^2)^T$ is a vector of unknown parameters. Then

$$\begin{aligned} f(\mathbf{x};\mu,\sigma^2) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right] \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2\right]. \end{aligned}$$

Again this factorizes where $s(\mathbf{x}) = (\overline{x}, \sum_i (x_i-\overline{x})^2)^T$, a vector valued function.
■

2.6.8 Large sample distribution of $\widehat{\theta}$

From the data summary point of view, the m.l.e. $\widehat{\theta}$ and $J\left(\widehat{\theta}\right)$have been thought of in terms of a particular set of data. We now wish to think of $\widehat{\theta}$ in terms of repeated sampling (*i.e.* as a random variable).

Main results

In many situations and subject to regularity conditions

$$\widehat{\theta} \xrightarrow{D} N(\theta, I(\theta)^{-1}),$$

We can apply the approach to obtaining confidence intervals developed in Mods, for the CLT.

Then an approximate 95% confidence interval for θ is given by

$$\widehat{\theta} \pm 1.96 I(\widehat{\theta})^{-1/2}.$$

[or $\widehat{\theta} \pm 1.96 J(\widehat{\theta})^{-1/2}$, regarded by many as better, but not in the books].

In the multivariate case,

$$\widehat{\boldsymbol{\theta}} \xrightarrow{D} N(\boldsymbol{\theta}, \mathbf{I}(\boldsymbol{\theta})^{-1}).$$

Example 2.23 *Exponential distribution*

For an exponential distribution with mean θ,

$$\begin{aligned} L(\theta) &= \theta^{-n} e^{-\sum_i x_i/\theta}, \quad \theta > 0, \\ l(\theta) &= -n \log\theta - \sum_i x_i \left/ \theta \right. , \end{aligned}$$

so that

$$U(\theta) = -\frac{n}{\theta} + \frac{\sum_i x_i}{\theta^2}, \quad J(\theta) = -\frac{n}{\theta^2} + \frac{2\sum_i x_i}{\theta^3}.$$

Thus

$$\widehat{\theta} = \overline{x}, \quad J(\widehat{\theta}) = \frac{n}{\overline{x}^2}$$

and an approximate 95% confidence interval is

$$\overline{x} \pm 1.96 \overline{x} \left/ \sqrt{n} \right.$$

■

Example 2.24 *Normal distribution*

For a normal random sample,

$$\widehat{\mu} = \overline{x}, \quad \widehat{\sigma}^2 = n^{-1} \sum_i (x_i - \overline{x})^2,$$

$$J\left(\mu, \sigma^2\right) = \left(\begin{array}{cc} n/\sigma^2 & \sigma^{-4}\sum_i (x_i - \mu) \\ \sigma^{-4}\sum_i (x_i - \mu) & \sigma^{-6}\sum_i (x_i - \mu)^2 - n/2\sigma^4 \end{array} \right).$$

Therefore

$$I\left(\widehat{\mu}, \widehat{\sigma}^2\right) = \left(\begin{array}{cc} n/\widehat{\sigma}^2 & 0 \\ 0 & n/2\widehat{\sigma}^4 \end{array} \right).$$

An approximate 95% confidence interval for μ is

$$\overline{x} \pm 1.96\widehat{\sigma}/\sqrt{n},$$

and for σ^2 is

$$\widehat{\sigma}^2 \pm 1.96\widehat{\sigma}^2\sqrt{\frac{2}{n}}.$$

Note that the estimators $\widehat{\mu}$ and $\widehat{\sigma}^2$ are asymptotically uncorrelated.
■

Lemma 2.5 If $\widehat{\theta}$ is the maximum likelihood estimator of a parameter θ based on a random sample, under suitable regularity conditions

$$\widehat{\theta} \xrightarrow{D} N(\theta, I(\theta)^{-1}),$$

where $I(\theta)$ is Fisher's information for the sample.
□

Sketch Proof Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with p.d.f. $f(x;\theta)$. Then the log-likelihood, score and observed information are

$$\begin{aligned} l(\theta) &= \sum_i \log f(x_i;\theta), \\ U(\theta) &= \sum_i \frac{\partial}{\partial\theta} \log f(x_i;\theta), \\ J(\theta) &= -\sum_i \frac{\partial^2}{\partial\theta^2} \log f(x_i;\theta). \end{aligned}$$

Let $U_i(\theta)$ be the random variable $U_i(\theta) = \dfrac{\partial}{\partial\theta}\log f(X_i;\theta)$, and, provided that conditions are such that integration and differentiation are interchangeable,

$$\begin{aligned} E\left[U_i(\theta)\right] &= \int f(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta)dx \\ &= \int \frac{\partial}{\partial\theta} f(x;\theta)dx \\ &= \frac{\partial}{\partial\theta}\int f(x;\theta)dx = \frac{\partial}{\partial\theta}1 = 0 \end{aligned}$$

and

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int f(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta)dx \\ &= \int f(x;\theta)\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)dx + \int \frac{\partial}{\partial\theta}f(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta)dx \\ &= E\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] + \int f(x;\theta)\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 dx. \end{aligned}$$

So

$$0 = -i(\theta) + E\left[U_i(\theta)^2\right]$$

and, therefore, $V\left[U_i(\theta)\right] = i(\theta)$.

It follows that $E\left[U(\theta)\right] = 0$, $V\left[U(\theta)\right] = ni(\theta) = I(\theta)$, and the Central Limit Theorem shows that

$$U(\theta) \xrightarrow{D} N\left(0, I(\theta)\right).$$

Now the m.l.e. is a solution of $U(\widehat{\theta}) = 0$, so that, Taylor expanding about θ,

$$U(\theta) + U'(\theta)(\widehat{\theta} - \theta) \simeq 0$$

or

$$U(\theta) - J(\theta)(\widehat{\theta} - \theta) \simeq 0.$$

Re-arranging,

$$\sqrt{I(\theta)}(\widehat{\theta} - \theta) \simeq U(\theta)\frac{\sqrt{I(\theta)}}{J(\theta)} = \frac{U(\theta)}{\sqrt{I(\theta)}}\Bigg/\frac{J(\theta)}{I(\theta)}.$$

From the CLT,

$$\frac{U(\theta)}{\sqrt{I(\theta)}} \xrightarrow{D} N(0,1)$$

and (from WLLN which you will meet in probability)

$$\frac{J(\theta)}{I(\theta)} \xrightarrow{P} 1.$$

Slutsky's Theorem states that, if $S_n \xrightarrow{D} S$ and $U_n \xrightarrow{P} k$, where k is a constant, then

$$S_n + U_n \xrightarrow{D} S + k, \quad S_n U_n \xrightarrow{D} Sk.$$

It therefore results in

$$\sqrt{I(\theta)}(\widehat{\theta} - \theta) \xrightarrow{D} N(0,1)$$

or

$$\widehat{\theta} \xrightarrow{D} N(\theta, I(\theta)^{-1}).$$

■

Requirements of this proof

(i) The true value of θ is interior to the parameter space.

(ii) Differentiation under the integral is valid, so that $E\left[U(\theta)\right] = 0$ and $V\left[U(\theta)\right] = ni(\theta)$. This allows a central limit theorem to apply to $U(\theta)$.

(iii) Taylor expansions are valid for the derivatives of the log-likelihood, so that higher order terms may be neglected.

(iv) A weak law of large numbers applies to $J(\theta)$.

2.7 The δ-method

We often want the asymptotic distribution of some function of a random variable. The basic method of doing this is the *δ-method.*

Let $X_1, X_2, \ldots, X_n$, be a random sample from a distribution with mean μ, variance σ^2.

Then $E\left(\overline{X} - \mu\right) = 0,\ V\left(\overline{X}\right) = \sigma^2/n.$

By Taylor's Theorem

$$\begin{aligned} g\left(\overline{X}\right) &= g\left(\mu + (\overline{X} - \mu)\right) \\ &= g(\mu) + (\overline{X} - \mu)g'(\mu) + \tfrac{1}{2}(\overline{X} - \mu)^2 g''(\delta_n), \\ &\qquad\qquad 0 < \left|\delta_n - \mu\right| < \left|\overline{X} - \mu\right|, \end{aligned}$$

and as n becomes large

$$\begin{aligned} E\left[g\left(\overline{X}\right)\right] &= g(\mu) + O\left(n^{-1}\right), \\ V\left[g\left(\overline{X}\right)\right] &= n^{-1}\sigma^2 g'(\mu)^2 + O\left(n^{-3/2}\right). \end{aligned}$$

Example 2.25 *Exponential distribution*

If

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

then

$$E(X) = \lambda^{-1},\ V(X) = \lambda^{-2},$$

so that

$$E(\overline{X}) = \lambda^{-1},\ V(\overline{X}) = 1/\,n\lambda^2.$$

Suppose we want $E(Y)$ and $V(Y)$, where $Y = \log \overline{X}$.

Then

$$g(x) = \log x,\ g'(x) = 1/\,x$$

and

$$E(Y) \simeq -\log \lambda, \quad V(Y) \simeq \frac{1}{n\lambda^2}\left(\frac{1}{x}\right)^2\bigg|_{x=1/\lambda} = \frac{1}{n}.$$

■

Example 2.26 *Variance stabilising transformations*

The aim of a variance stabilising transformation is to find a transformation such that $g\left(\overline{X}\right)$ has a variance which is approximately constant,
i.e.

$$\sigma^2 g'(\mu)^2 \simeq c, \text{ a constant.}$$

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ with mean μ, variance $V(\mu)$. Then

$$\begin{aligned} g'(\mu) &= cV(\mu)^{-1/2} \\ \Rightarrow \quad g(\mu) &= \int^{\mu} V(u)^{-1/2} du. \end{aligned}$$

In practice, the transformation is usually applied to the sample data directly to produce transformed data $g(X_1), g(X_2), \ldots, g(X_n)$.
For a Poisson distribution, $E(X) = \mu,\ V(X) = \mu$ so that $g(\mu) = \mu^{1/2}$.
Thus, if $X \sim \mathit{Poisson}(\mu)$, then

$$Y = \sqrt{X}$$

has a variance which is approximately constant.

For an exponential distribution, $V(\mu) = \mu^2$, so

$$g(\mu) = \log \mu.$$

■