Chapter 23

MSMYS3 Statistical Theory

(23.1) Statistical Inference

(23.1.1) Bayesian & Frequentist Inference

Problems With Bayesian Inference

Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables. Let $\mathbf{x}$ be a random sample, then what information about the distribution of the Xs can be found from $\mathbf{x}$?

Let the distribution of the Xs be determined by a function $f(x, \theta)$ where θ is some parameter.

- In Bayesian inference define $L \colon \Theta \to \mathbb{R}_0^+$ by $L(\theta) = f(\mathbf{x}, \theta)$. The distribution $p(\theta \mid \mathbf{x})$ is then determined by Bayes' formula for which a distribution for θ is required.
- In frequentist inference it is usual to choose the value of θ which maximises the likelihood function L.
- Alternatively define a function $\hat{\theta} = T(\mathbf{x})$ to estimate θ.

Bayesian inference requires a distribution to be specified for θ. It is more generally the case that nothing at all is known about θ—not even its distribution. There is another problem with Bayesian inference. Recall Bayes' formula

$$p(\theta \mid \mathbf{x}) = \frac{L(\theta, \mathbf{x})\pi(\theta)}{\int_\Theta L(\theta', \mathbf{x})\pi(\theta') \, \mathrm{d}\theta'}$$

where $\pi(\theta)$ is the prior distribution for θ. Observe this expression is uneffected when L is multiplied by anything not dependent on θ—a function of the data c, say.

Let E and E' be experiments for which $L(\theta, \mathbf{x}) = cL'(\theta, \mathbf{x}')$. Then $p(\theta \mid \mathbf{x}) \equiv p(\theta \mid \mathbf{x}')$. The same inference is made from each experiment. This is the Strong Likelihood Principle.

The alarming consequences of the strong likelihood principle are easily illustrated. Suppose two experiments have the following results:

- $X \sim \mathrm{Bin}(N, \theta)$ and $X = 1$ is observed. Hence $L(\theta) = N\theta(1-\theta)^{N-1}$.
- Observe a sequence of Bernoulli trials until the first success occurs (geometric distribution). If X is the number of trials conducted then if X is observed to be N, $L'(\theta) = \theta(1-\theta)^{N-1}$.

Consider the example of throwing a coin. In the first experiment one head could appear anywhere in the N throws. In the second experiment the one head must appear last. Clearly this is much less likely, and the factor of N takes this into account. However, N does not depend on θ so the inference process does not take this into account.

Point Estimation

Frequentist inference makes estimation about θ from a single data point $\mathbf{x}$. If $\hat{\theta} = T(\mathbf{x})$ is an estimate for θ then the random variable $T(\mathbf{X})$ is an estimator for θ. As $T(\mathbf{X})$ is a random variable, it has a mean, a variance, and a distribution called the sampling distribution. The sampling distribution of an estimator is very important in assessing the accuracy of an estimator.

Definition 1 *An estimator $\hat{\theta}$ for a parameter θ is unbiassed if $\mathbb{E}\, T(\mathbf{X}) = \theta$ for all possible values of θ.*

The bias of $\hat{\theta}$ is then given by $\mathbb{E}\,(T - \theta) = (\,\mathbb{E}\, T) - \theta$. Although unbiassedness is clearly preferable, it is not of great importance. What is important is the consistency of an estimator.

Definition 2 *Let $\hat{\theta} = T(\mathbf{X})$ be an estimator for a parameter θ. $\hat{\theta}$ is a consistent estimator for θ if*

$$\forall \varepsilon > 0 \quad Pr\{|T(\mathbf{X}) - \theta| > \varepsilon\} \longrightarrow 0 \textit{ as } n \longrightarrow \infty$$

It is usually best to establish consistency using the mean square error,

$$\text{MSE}\left(\hat{\theta}\right) \overset{\text{def}}{=} \mathbb{E}\,(T(\mathbf{X}) - \theta)^2$$

Lemma 3 $\text{MSE}\left(\hat{\theta}\right) = \text{var}\,(T(\mathbf{X})) + (bias(\hat{\theta}))^2$.

Proof. Since θ is a constant, $\text{var}\,(T - \theta) = \text{var}\, T$. Hence

$$\begin{aligned}\text{var}\,(T(\mathbf{X}) - \theta) &= \mathbb{E}\,(T(\mathbf{X}) - \theta)^2 - (\mathbb{E}\,(T(\mathbf{X}) - \theta))^2 \\ &= \text{MSE}\left(\hat{\theta}\right) - \left(\text{bias}(\hat{\theta})\right)^2\end{aligned}$$ □

From this it is immediately obvious that $\text{MSE}\left(\hat{\theta}\right) \longrightarrow 0$ as $n \longrightarrow \infty \Leftrightarrow \text{var}\,(T(\mathbf{X})) \longrightarrow 0$ and $\text{bias}(\hat{\theta}) \longrightarrow 0$.

Lemma 4 *If $\text{MSE}\left(\hat{\theta}\right) \longrightarrow 0$ as $n \longrightarrow \infty$ then $\hat{\theta}$ is a consistent estimator of θ.*

Proof. Let $g(x)$ be the probability density function of $T(\mathbf{X})$. Then

$$\begin{aligned}Pr\{|T(\mathbf{X}) - \theta| > \varepsilon\} &= \int_{|t-\theta|>\varepsilon} g(t)\, \mathrm{d}t \\ &\leqslant \int_{|t-\theta|>\varepsilon} \frac{(t-\theta)^2}{\varepsilon^2} g(t)\, \mathrm{d}t \\ &\leqslant \int_{-\infty}^{\infty} \frac{(t-\theta)^2}{\varepsilon^2} g(t)\, \mathrm{d}t \\ &= \frac{1}{\varepsilon^2}\, \mathbb{E}\, \left(T(\mathbf{X}) - \theta\right)^2 \\ &= \frac{1}{\varepsilon^2}\, \text{MSE}\left(\hat{\theta}\right)\end{aligned}$$ □

From this it is evident that $\hat{\theta}$ is consistent if both its bias and variance tend to 0 as $n \longrightarrow \infty$.

The Logarithm Of The Likelihood Function

When finding a maximum likelihood estimator it is common to work with $l(\theta, \mathbf{x}) = \ln\,(L(\theta, \mathbf{x}))$. Now, $L(\theta, \mathbf{x})$ is the joint pdf of $X_1, X_2, \ldots, X_n$ (and so is itself a probability density function), and since these are independently and identically distributed,

$$L(\theta, \mathbf{X}) = \prod_{i=1}^{n} f(x_i, \theta)$$

Theorem 5 $\mathbb{E}\,\dfrac{\partial l(\theta,\mathbf{X})}{\partial\theta} = 0$

Proof. Since L is a pdf it must have the property

$$\begin{aligned} \int L(\theta,\mathbf{x})\ \mathrm{d}\mathbf{x} &= 1 \quad \forall\theta\in\Theta \\ \text{so } \frac{\partial}{\partial\theta}\int L(\theta,\mathbf{x})\ \mathrm{d}\mathbf{x} &= 0 \end{aligned}$$

Provided the limits of the integral do not depend on θ,

$$\int \frac{\partial L}{\partial\theta}\ \mathrm{d}\mathbf{x} = 0$$

Note that since $l = \ln L$, $\frac{\partial L}{\partial\theta} = L\frac{\partial l}{\partial\theta}$. Hence

$$\int \frac{\partial l}{\partial\theta} L\ \mathrm{d}\mathbf{x} = 0$$

Since L is a pdf and the equation holds for any $\mathbf{x}$ this equation is, by definition, the expected value of $\frac{\partial l}{\partial\theta}$. Hence the proof is complete. □

Note that

$$\frac{\partial l(\theta,\mathbf{x})}{\partial\theta} \overset{\text{def}}{=} u(\theta,\mathbf{x})$$

is called the score function. Differentiating again gives

$$\begin{aligned} \frac{\partial}{\partial\theta}\int \frac{\partial l}{\partial\theta} L\ \mathrm{d}\mathbf{x} &= 0 \\ \int \frac{\partial}{\partial\theta}\left(\frac{\partial l}{\partial\theta}L\right)\ \mathrm{d}\mathbf{x} &= 0 \\ \int L\frac{\partial^2 l}{\partial\theta^2} + \frac{\partial l}{\partial\theta}\frac{\partial L}{\partial\theta}\ \mathrm{d}\mathbf{x} &= 0 \\ \int L\left(\frac{\partial^2 l}{\partial\theta^2} + \left(\frac{\partial l}{\partial\theta}\right)^2\right)\ \mathrm{d}\mathbf{x} &= 0 \end{aligned}$$

Hence using the linearity of the expectation operator and the fact that this holds for all $\mathbf{x}$,

$$\mathbb{E}\left(\frac{\partial l(\theta,\mathbf{X})}{\partial\theta}\right)^2 = -\,\mathbb{E}\,\frac{\partial^2 l(\theta,\mathbf{X})}{\partial\theta^2} \overset{\text{def}}{=} I_\theta$$

I_θ is the expected Fisher information. Observe that since $\mathbb{E}\,u(\theta,\mathbf{x}) = 0$ the above equation may be re-written as $\operatorname{var}(u(\theta,\mathbf{x})) = I_\theta$.

The Cramér-Rao Inequality

The Cramér-Rao inequality provides a lower bound for the variance of an estimator. This is useful because if an estimator is found to have variance equal to this lower bound, a more efficient estimator cannot be found.

Theorem 6 (Cramér-Rao Inequality) *Suppose that T is an unbiassed estimator for $\tau(\theta)$ where θ is some parameter. Then*

$$\operatorname{var} T(\mathbf{X}) \geqslant \frac{\left(\frac{\mathrm{d}\tau(\theta)}{\mathrm{d}\theta}\right)^2}{I_\theta}$$

If T is an unbiassed estimator for θ (rather than $\tau(\theta)$) then clearly $\operatorname{var} T \geqslant \frac{1}{I_\theta}$.

Proof. Using the definition of expectation,

$$\begin{aligned}
\mathbb{E}\, T = \int T(\mathbf{x})L(\theta,\mathbf{x})\,\mathrm{d}\mathbf{x} &= \tau(\theta) \quad \text{now differentiate with respect to } \theta \\
\int T(\mathbf{x})L\frac{\partial l}{\partial \theta}\,\mathrm{d}\mathbf{x} &= \frac{\mathrm{d}\tau}{\mathrm{d}\theta} \\
\mathbb{E}\left(T(\mathbf{X})\frac{\partial l}{\partial \theta}\right) &= \frac{\mathrm{d}\tau}{\mathrm{d}\theta}
\end{aligned}$$

Recall now that

$$\operatorname{cov}(U,V) = \mathbb{E}\left((U - \mathbb{E}\,U)\,(V - \mathbb{E}\,V)\right) = \mathbb{E}\,(UV) - (\mathbb{E}\,U)(\mathbb{E}\,V)$$

From this it is clear that since $\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = 0$

$$\operatorname{cov}\left(T(\mathbf{X}), \frac{\partial l}{\partial \theta}\right) = \frac{\mathrm{d}\tau}{\mathrm{d}\theta}$$

Now using the result $(\operatorname{var} U)(\operatorname{var} V) \geqslant (\operatorname{cov}(U,V))^2$,

$$(\operatorname{var} T(\mathbf{X}))\left(\operatorname{var}\left(\frac{\partial l}{\partial \theta}\right)\right) \geqslant \left(\frac{\mathrm{d}\tau}{\mathrm{d}\theta}\right)^2$$

Hence the result. □

Definition 7 *If T is an unbiassed estimator for θ then the efficiency of T is given by* $\frac{1}{I_\theta(\operatorname{var} T)}$.

Now, since $\operatorname{cov}(U,V) = (\operatorname{var} U)(\operatorname{var} V) \Leftrightarrow U$ and V are linearly related, T must attain the Cramér-Rau lower bound when $\frac{\partial l}{\partial \theta} = aT + b$ where a and b depend on θ but not $\mathbf{X}$. This occurs when T is perfectly efficient, i.e. has efficiency 1. The following situation has arisen

$$\begin{aligned}
\text{Cramér-Rau bound attained} \quad &\Leftrightarrow \quad T \text{ is efficient} \\
&\Leftrightarrow \quad \operatorname{var} T = \frac{1}{I_\theta} \\
&\Leftrightarrow \quad T = a\theta + b \text{ from covariance result in proof of Cramér-Rau}
\end{aligned}$$

Lemma 8 *Let T be an unbiassed estimator for a parameter θ. Then*

$$T \textit{ is efficient} \quad \Leftrightarrow \quad \frac{\partial l}{\partial \theta} = (T(\mathbf{x}) - \theta)I_\theta$$

Proof. Now, T being efficient is equivalent to the Cramér-Rau bound being attained, and so $\frac{\partial l}{\partial \theta} = aT + b$. The proof therefore relies on finding values for a and b.

$$0 = \mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = a\,\mathbb{E}\,T + b = a\theta + b$$

So $b = -a\theta$. Multiplying $\frac{\partial l}{\partial \theta} = aT + b$ by $\frac{\partial l}{\partial \theta}$ and finding the expected value again,

$$I_\theta = \mathbb{E}\left(\frac{\partial l}{\partial \theta}\right)^2 = \mathbb{E}\left(aT\frac{\partial l}{\partial \theta} + b\frac{\partial l}{\partial \theta}\right) = a\,\mathbb{E}\left(T\frac{\partial l}{\partial \theta}\right) + b\,\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = a$$

The last equality holds because since $\mathbb{E}\,\frac{\partial l}{\partial \theta} = 0$, $\mathbb{E}\left(T\frac{\partial l}{\partial \theta}\right) = \operatorname{cov}\left(T, \frac{\partial l}{\partial \theta}\right)$ which by hypothesis is 1. Now substituting for a and b with the values found,

$$\frac{\partial l}{\partial \theta} = I_\theta T - I_\theta \theta = (T - \theta) I_\theta \qquad \square$$

(23.1.2) Sufficiency

Sufficient Statistics

Computing a statistic to estimate a parameter is all very well, but does a statistic make full use of the data? What functions of the data is it required to calculate in order to extract all the (useful) information? This is the question of sufficiency.

Definition 9 *Suppose $X_1, X_2, \ldots, X_n$ have a joint distribution depending on some parameter θ. The statistic $T = T(\mathbf{X})$ is sufficient for θ if the conditional distribution of $\mathbf{X}$ given the value of T obtained is algebraically independent of θ.*

This definition simply says that there is enough information in T to calculate the distribution of $\mathbf{X}$ without having to refer back to the data.

Example 10 *Consider n Bernoulli trials, so $X_1, X_2, \ldots, X_n$ are independently and identically distributed with $Pr\{X_i = x\} = \theta^x(1-\theta)^{1-x}$. Claim that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.*

$$\begin{aligned}
Prob X_1 = x_1 \wedge X_2 = x_2 \wedge \cdots \wedge X_n = x_n \mid T = t &= \frac{Prob X_1 = x_1 \wedge X_2 = x_2 \wedge \cdots \wedge X_n = x_n \wedge T = t}{Prob T = t} \\
&= \begin{cases} \frac{Prob X_1 = x_1 \wedge X_2 = x_2 \wedge \cdots \wedge X_n = x_n}{Prob T = t} & \text{if } \sum_{i=1}^n x_i = t \\ 0 & \text{if } \sum_{i=1}^n x_i \neq t \end{cases} \\
&= \begin{cases} \frac{\theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2}\ldots\theta^{x_n}(1-\theta)^{1-x_n}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} & \text{if } \sum_{i=1}^n x_i = t \\ 0 & \text{if } \sum_{i=1}^n x_i \neq t \end{cases} \\
&= \begin{cases} \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} & \text{if } \sum_{i=1}^n x_i = t \\ 0 & \text{if } \sum_{i=1}^n x_i \neq t \end{cases} \\
&= \begin{cases} \frac{1}{\binom{n}{t}} & \text{if } \sum_{i=1}^n x_i = t \\ 0 & \text{if } \sum_{i=1}^n x_i \neq t \end{cases}
\end{aligned}$$

Clearly this does not depend on θ, so T is verified as a sufficient statistic.

Clearly this process is rather tedious, and it would be a formidable task to do this for anything but the simplest of distributions.

Lemma 11 *T is a sufficient statistic for the parameter θ if and only if the likelihood can be factorised as*

$$L(\theta, \mathbf{x}) = \kappa_1(T(\mathbf{x}), \theta)\kappa_2(\mathbf{x})$$

where κ_1 and κ_2 are non-negative functions.

Proof. $(\Rightarrow)$

$$L(\theta, \mathbf{x}) = Pr\{\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = t\} = Pr\{\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t\}\ Pr\{T(\mathbf{X}) = t\}$$

where $t = T(\mathbf{x})$. Assuming that T is sufficient, $Pr\{\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t\}$ is algebraically independent of θ, hence define

$$\begin{aligned}\kappa_1(T(\mathbf{x}), \theta) &= Pr\{T(\mathbf{X}) = t\} \\ \kappa_2(\mathbf{x}) &= Pr\{\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t\}\end{aligned}$$

($\Leftarrow$) Assume $L(\theta, \mathbf{x}) = \kappa_1(T(\mathbf{x}), \theta)\kappa_2(\mathbf{x})$. Then when $T(\mathbf{x}) = t$,

$$\begin{aligned}Pr\{\mathbf{X} = \mathbf{x} \mid T = t\} &= \frac{Pr\{\mathbf{X} = \mathbf{x} \text{ and } T = t\}}{Pr\{T = t\}} = \frac{Pr\{\mathbf{X} = \mathbf{x}\}}{Pr\{T = t\}} \quad \text{since it was assumed } T(\mathbf{x}) = t \\ &= \frac{L(\theta, \mathbf{x})}{\sum_{T(\mathbf{x}')=t} L(\theta, \mathbf{x}')} \\ &= \frac{\kappa_1(T(\mathbf{x}), \theta)\kappa_2(\mathbf{x})}{\sum_{T(\mathbf{x}')=t} \kappa_1(T(\mathbf{x}), \theta)\kappa_2(\mathbf{x}')} \\ &= \frac{\kappa_2(\mathbf{x})}{\sum_{T(\mathbf{x}')=t} \kappa_2(\mathbf{x}')} \\ Pr\{\mathbf{X} = \mathbf{x} \mid T = t\} &= \begin{cases} \frac{\kappa_2(\mathbf{x})}{\sum_{T(\mathbf{x}')=t} \kappa_2(\mathbf{x}')} & \text{if } T(\mathbf{x}) = t \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

This is algebraically independent of θ, so T is sufficient. □

Minimal Sufficient Statistics

It is quite possible to find many sufficient statistics, but some of them could be better than others.

Definition 12 *T is a minimal sufficient statistic for the parameter θ if*

1. *T is a sufficient statistic for θ.*
2. *If S is any other sufficient statistic for θ then T is a function of S.*

Condition 2 seems a bit peculiar. Think, however, of a function as a bijective mapping—in its strictest form. This condition then means that if $S(\mathbf{x}) = S(\mathbf{y})$ then $T(\mathbf{x}) = T(\mathbf{y})$. This will be of use later.

Suppose that T is a minimal sufficient statistic and that $\mathbf{x}$ and $\mathbf{y}$ are data observed from some experiment. If $T(\mathbf{x}) = T(\mathbf{y})$ then identical inference should be made about θ. This is the sufficiency principle.

Clearly verifying the definition of a minimal sufficient statistic is not desirable. A simpler test is sought instead. Consider a minimal sufficient statistic T, and define

1. the T partition of the sample space Ω,

$$\Omega = \bigcup_t E_t \quad \text{where} \quad E_t = \{\mathbf{x} \mid T(\mathbf{x}) = t\}$$

   This defines the equivalence relation '$\equiv$' by

$$\mathbf{x} \equiv \mathbf{y} \quad \Leftrightarrow \quad T(\mathbf{x}) = T(\mathbf{y})$$

2. the likelihood partition of the sample space Ω,

$$\Omega = \bigcup_\alpha \Lambda_\alpha \quad \text{where} \quad \Lambda_\alpha = \{\mathbf{y} \mid L(\theta, \mathbf{y}) \propto L(\theta, \mathbf{x})\}$$

for some particular $\mathbf{x}$. Note that the constant of proportionality can be a function of $\mathbf{x}$ and $\mathbf{y}$ but not θ. This defines the equivalence relation '$\sim$' by

$$\mathbf{x} \sim \mathbf{y} \quad \Leftrightarrow \quad L(\theta, \mathbf{x}) \propto L(\theta, \mathbf{y})$$

Lemma 13 *Let T be a statistic for which the T partition of the sample space is the same as the likelihood partition. Then T is a minimal sufficient statistic.*

Proof. Since the likelihood partition is the same as the T partition, the value of $L(\theta, \mathbf{x})$ can be determined from the value of $T(\mathbf{x})$ and no factor must depend on θ. Hence

$$L(\theta, \mathbf{x}) = \kappa_1(T(\mathbf{x}), \theta)\kappa_2(\mathbf{x})$$

hence by the factorisation criterion (Lemma 11) T is a sufficient statistic.

Suppose that S is another sufficient statistic. Then by the same argument

$$L(\theta, \mathbf{x}) = \kappa_3(S(\mathbf{y}), \theta)\kappa_4(\mathbf{y})$$

Hence

$$\begin{aligned} S(\mathbf{x}) = S(\mathbf{y}) \quad &\Rightarrow \quad L(\theta, \mathbf{x}) \propto L(\theta, \mathbf{y}) \quad \text{from above} \\ &\Rightarrow \quad \mathbf{x} \sim \mathbf{y} \\ &\Rightarrow \quad \mathbf{x} \equiv \mathbf{y} \quad \text{by hypothesis} \\ &\Rightarrow \quad T(\mathbf{x}) = T(\mathbf{y}) \end{aligned}$$

Hence T "is a function of" S and so must be a minimal sufficient statistic. □

The Rao-Blackwell Theorem

Let X and Y be random variables. $\mathbb{E}(Y \mid X = x)$ is simply a number, but its precise value depends on the value of x. The quantity $\mathbb{E}(Y \mid X)$ is therefore a random variable, and it is straight forward to show that $\mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}\, Y$. More information about conditional expectation can be found in Chapter **??**.

Theorem 14 (Rao-Blackwell) *Let T be a sufficient statistic for the parameter θ and let S be an unbiassed estimator for θ. Define $U = \mathbb{E}(S \mid T)$ then*

1. *U is a statistic.*
2. *U is an unbiassed estimator for θ.*
3. *$\operatorname{var} U \leqslant \operatorname{var} S$.*

Proof.

1. Since T is sufficient, the conditional distribution of $\mathbf{X}$ given T is independent of θ (by definition). Since S is an unbiassed estimator for θ it is certainly independent of θ and hence $\mathbb{E}(S \mid T)$ is independent of θ i.e. is a statistic.
2. $\mathbb{E}\, U = \mathbb{E}(\mathbb{E}(S \mid T)) = \mathbb{E}\, S = \theta$ and hence U is an unbiassed estimator for θ.
3. $\operatorname{var}((S - \theta) \mid T) \geqslant 0$ hence

$$\begin{aligned} \mathbb{E}((S-\theta)^2 \mid T) &\geqslant (\mathbb{E}((S-\theta) \mid T))^2 = (\mathbb{E}(S \mid T) - \theta)^2 = (U - \theta)^2 \\ \mathbb{E}\left(\mathbb{E}((S-\theta)^2 \mid T)\right) &\geqslant \mathbb{E}(U-\theta)^2 \\ \mathbb{E}(S-\theta)^2 = \operatorname{var} S &\geqslant \mathbb{E}(U-\theta)^2 = \operatorname{var} U \end{aligned}$$

□

If S is not a function of T then the inequality of the Rao-Blackwell theorem becomes strict inequality. The theorem then shows that if a minimum variance unbiassed estimator exists, then there is a function of the minimal sufficient statistic that is also a minimum variance unbiassed estimator.

Using this it is possible to produce relatively 'nice' estimators from really quite peculiar ones. Take for example the task of estimating $e^{-\theta}$ on a Poisson distribution. An obvious choice for an estimator is $e^{-\overline{X}}$, but this is biassed. Consider the estimator S defined as

$$S = I_{\{X_1=0\}} = \begin{cases} 1 & \text{if } X_1 = 0 \\ 0 & \text{otherwise} \end{cases}$$

Because of the form of the Poisson pdf, this has expected value $e^{-\theta}$. Now, it is simple to show that $T = \sum_{i=1}^{n} X_i$ is a minimal sufficient statistic for θ and hence the Rao-Blackwell theorem can be used.

$$\begin{aligned}
\mathbb{E}\left(S \mid T = t\right) &= ProbX_1 = 0 \mid \sum_{i=1}^{n} X_i = t \\
&= \frac{ProbX_1 = 0 \text{ and } \sum_{i=1}^{n} X_i = t}{Prob\sum_{i=1}^{n} X_i = t} \\
&= \frac{ProbX_1 = 0Prob\sum_{i=2}^{n} X_i = t}{Prob\sum_{i=1}^{n} X_i = t} \\
&= \frac{\frac{e^{-\theta}\theta^1}{1!}\frac{e^{-(n-1)\theta}((n-1)\theta)^t}{t!}}{\frac{e^{-n\theta}(n\theta)^t}{t!}} \\
&= \left(\frac{n-1}{n}\right)^t
\end{aligned}$$

Hence the estimator

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

is unbiassed for $e^{-\theta}$ and has smaller variance than S.

Finding Minimal Sufficient Statistics

Definition 15 *A probability distribution belongs to a k parameter exponential family if its pdf can be expressed in the form*

$$f(x,`) = C(x)\exp\left(\sum_{i=1}^{k} A_i(`)B_i(x) + D(`)\right)$$

where the functions $1, A_1(`), A_2(`), \ldots, A_k(`)$ *are linearly independent.*

Most common distributions (with the notable exception of the uniform distribution) are exponential families. Take for example the Binomial distribution.

$$\begin{aligned}
f(x,\theta) &= \binom{N}{x}\theta^x(1-\theta)^{N-x} \\
&= \binom{N}{x}\exp\left(x\ln\theta + (N-x)\ln(1-\theta)\right) \\
&= \binom{N}{x}\exp\left(x\ln\frac{\theta}{1-\theta} + N\ln(1-\theta)\right)
\end{aligned}$$

Lemma 16 *If* $X_1, X_2, \ldots, X_n$ *are independently and identically distributed random variables with pdf* $f(x,`)$ *then if* f

belongs to the k parameter exponential family

$$f(x,`) = C(x)\exp\left(\sum_{i=1}^{k} A_i(`)B_i(x) + D(`)\right)$$

then

$$\mathbf{T} = \left(\sum_{i=1}^{n} B_1(x_i), \sum_{i=1}^{n} B_2(x_i), \ldots, \sum_{i=1}^{n} B_k(x_i)\right)$$

is a minimal sufficient statistic for `.

Proof.

$$\begin{aligned} L(`,\mathbf{x}) &= \prod_{i=1}^{n} f(x_i,`) \\ &= \prod_{i=1}^{n}\left(C(x)\exp\left(\sum_{j=1}^{k} A_j(`)B_j(x_i) + D(`)\right)\right) \\ &= \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{k} A_j(`)B_j(x_i) + nD(`)\right)\prod_{i=1}^{n} C(x_i) \end{aligned}$$

Hence where $\mathbf{y}$ is another vector of observations

$$\frac{L(`,\mathbf{x})}{L(`,\mathbf{y})} = \frac{\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{k} A_j(`)B_j(x_i) + nD(`)\right)\prod_{i=1}^{n} C(x_i)}{\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{k} A_j(`)B_j(y_i) + nD(`)\right)\prod_{i=1}^{n} C(y_i)}$$

Now put $T_j(\mathbf{x}) \overset{\text{def}}{=} \sum_{i=1}^{n} B_j(x_i)$ to give

$$\begin{aligned} &= \frac{\exp\left(\sum_{j=1}^{k} A_j(`)T_j(\mathbf{x}) + nD(`)\right)\prod_{i=1}^{n} C(x_i)}{\exp\left(\sum_{j=1}^{k} A_j(`)T_j(\mathbf{y}) + nD(`)\right)\prod_{i=1}^{n} C(y_i)} \\ &= \exp\left(\sum_{j=1}^{k} A_j(`)\left(T_j(\mathbf{x}) - T_j(\mathbf{y})\right)\right)\frac{\prod_{i=1}^{n} C(x_i)}{\prod_{i=1}^{n} C(y_i)} \end{aligned}$$

This expression is independent of ` whenever $T_j(\mathbf{x}) = T_j(\mathbf{y})$ for all j (using the fact that the As are linearly independent) and hence the L partition coincides with the T partition and $\mathbf{T}$ is minimal sufficient. □

(23.1.3) Maximum Likelihood Estimation

Calculating Maximum Likelihood Estimators

A maximum likelihood estimation ˆ is simply the value of ` for which $L(`,\mathbf{x})$ attains a (local) maximum. In the one dimensional case this can be found by solving $\frac{\partial L}{\partial\theta} = 0$ or indeed $\frac{\partial l}{\partial\theta} = 0$. It is good practise to verify the second derivative as being negative.

In the case of more than one dimension the equations $\frac{\partial l}{\partial\theta_1} = 0$ to $\frac{\partial l}{\partial\theta_k} = 0$ can be solved, and verification as maxima is done by computing the Hessian matrix and showing it to be negative definite.

The common distributions have familiar maximum likelihood estimators, with the notable exception of the uniform distribution $\mathcal{U}[0,\theta]$. Of interest is censored data.

Example 17 *Suppose $X_1, X_2, \dots, X_n$ are independently and identically distributed with*

$$f(x, \theta) = \theta e^{-x\theta}$$

for $x > 0$. However, X_i is only observed if $X_i < T$ for some fixed T.

Let M be the number of observations actually made. Now, $Pr\{X_i < T\} = 1 - e^{-\theta T}$ hence $M \sim Bin(n, 1 - e^{-\theta T})$. The distribution of X_i can now be found, given that $X_i < T$.

$$\frac{f(x,\theta)}{Pr\{X_i < T\}} = \frac{\theta e^{-x\theta}}{1 - e^{\theta T}}$$

Now let $\mathbf{y} = (y_1, y_2, \dots, y_m)$ be a random sample, then the likelihood of $\mathbf{y}$ is given by the product of the pdfs for the random variables which were observed, multiplied by the probability that those random variables were observed. This gives

$$\begin{aligned}
L(\theta, \mathbf{y}) &= \binom{n}{m} \left(1 - e^{\theta T}\right)^m e^{-T(n-m)\theta} \prod_{i=1}^{m} \frac{\theta e^{-x\theta}}{1 - e^{\theta T}} \\
&= \binom{n}{m} e^{-T(n-m)\theta} \theta^m e^{-\theta \sum_{i=1}^n y_i} \\
\frac{\partial l}{\partial \theta} &= -T(n-m) + \frac{m}{\theta} - \sum_{i=1}^n y_i \\
\hat{\theta} &= \frac{m}{(n-m)T + \sum_{i=1}^n y_i}
\end{aligned}$$

Theorem 18 *If $\hat{\theta}$ is a maximum likelihood estimator then*

1. *$\hat{\theta}$ is invariant i.e. if ϕ is an injective function then $\hat{\phi} = \phi(\hat{\theta})$ is a maximum likelihood estimation of $\phi(\theta)$.*

2. *$\hat{\theta}$ is consistent.*

3. *$\hat{\theta}$ is sufficient.*

4. *$\hat{\theta}$ is efficient*

Proof.

1. The result clearly holds since the value of the likelihood is unaffected by such a reparameterisation.

2. Proof is omitted.

3. Suppose T is sufficient for θ then the likelihood can be factorised as

$$L(\theta, \mathbf{x}) = \kappa_1(T(\mathbf{x}), \theta)\kappa_2(\mathbf{x})$$

For any particular random sample $\mathbf{x}$ the maximum is determined by the maximum of κ_1 (as no information about θ can be obtained from κ_2) and hence $\hat{\theta}$ is a function of T.
But T is a sufficient statistic so hence $\hat{\theta}$ is also a sufficient statistic.

4. Suppose θ' is an unbiassed efficient estimator for θ. Then

$$\frac{\partial l}{\partial \theta} = I_\theta(\theta' - \theta)$$

from which it is evident that $\theta = \theta'$ is a maximum likelihood estimation for θ. □

Exponential Families & Maximum Likelihood Estimators

Exponential families can be used to find minimal sufficient statistics, and in a similar way then can be used to find maximum likelihood estimators.

Theorem 19 *Suppose $X_1, X_2, \ldots, X_n$ are independently and identically distributed with*

$$f(x, \phi) = C(x) \exp\left(\phi B(x) + D(\phi)\right)$$

then the maximum likelihood estimator of ϕ solves the equation

$$\frac{1}{n} \sum_{i=1}^{n} B(x_i) = \mathbb{E}\, B(X)$$

Proof. Calculating the likelihood,

$$L(\theta, \mathbf{x}) = \exp\left(\phi \sum_{i=1}^{n} B(x_i) + nD(\phi)\right) \prod_{i=1}^{n} C(x_i)$$

$$\text{so } \frac{\partial l}{\partial \phi} = \sum_{i=1}^{n} B(x_i) + n \frac{\mathrm{d}D}{\mathrm{d}\phi} \tag{20}$$

Hence the maximum likelihood estimator obeys the equation

$$\sum_{i=1}^{n} B(x_i) = -n \frac{\mathrm{d}D}{\mathrm{d}\phi} \tag{21}$$

Now, $\mathbb{E} \frac{\partial l}{\partial \phi} = 0$ and hence taking expectations on equation 20 gives

$$n\, \mathbb{E}\, B(X) = -n \frac{\mathrm{d}D}{\mathrm{d}\phi} \tag{22}$$

Hence using equations 21 and 22

$$n\, \mathbb{E}\, B(X) = \sum_{i=1}^{n} B(x_i)$$

□

It is simple to verify that $\frac{\partial^2 l}{\partial \phi^2} < 0$. In the case of k parameters the maximum likelihood estimators satisfy

$$\sum_{i=1}^{n} B_j(x_i) = -n \frac{\partial D}{\partial \phi_j} = n\, \mathbb{E}\, B_j(X)$$

It is worth noting that an analytic solution to $\frac{\partial l}{\partial \theta} =$ does not always exist. However, the use of numerical methods means that in all practical circumstances a value for the maximum likelihood estimator can be found.

(23.1.4) Implications Of The Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ are independently and identically distributed random variables and T is an estimator for the parameter θ it is desirable to find the sampling distribution for T. Most generally this does not exist, but even so the Central Limit Theorem may be of use.

Theorem 23 (The Central Limit Theorem) *Let $Y_1, Y_2, \ldots, Y_n$ be independently and identically distributed random* *variables with* $\mathbb{E}\, Y_i = \mu$ *and* $\operatorname{var} Y_i = \sigma^2$. *Then*

$$\frac{\sqrt{n}}{\sigma}(\overline{Y} - \mu) \to Z \quad as \quad n \to \infty$$

where $Z \sim \mathcal{N}(0, 1)$.

The immediate consequence of this is that if $T = \overline{X}$ then T has the approximate sampling distribution $\mathcal{N}\left(\mathbb{E}\, X, \frac{\operatorname{var} X}{n}\right)$.

Efficient Estimators

Let i_θ be the expected Fisher information from a single observation, so

$$I_\theta = -\,\mathbb{E}\left(\frac{\partial^2 l}{\partial\theta^2}\right) = -\sum_{i=1}^{n} \mathbb{E}\left(\frac{\partial^2 l(x_i, \theta)}{\partial\theta^2}\right) = n i_\theta$$

Lemma 24 *Let* $T(X_1, X_2, \ldots, X_n)$ *be an unbiassed and efficient estimator for* θ. *Then as* $n \to \infty$,

$$\sqrt{n}(T - \theta) \to \mathcal{N}\left(0, \frac{1}{i_\theta}\right)$$

Proof. Let $Y = \frac{\partial \ln f(X_i, \theta)}{\partial\theta}$ then since this represents a log likelihood for $n = 1$, $\mathbb{E}\, Y_i = 0$ and $\operatorname{var} Y_i = i_\theta$. Hence by the Central Limit Theorem

$$\sqrt{\frac{n}{i_\theta}}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \sqrt{\frac{n}{i_\theta}}\left(\frac{1}{n}\frac{\partial l}{\partial\theta}\right) \to \mathcal{N}(0, 1)$$

Now, since T is unbiassed and efficient

$$\frac{\partial l}{\partial\theta} = (T - \theta) I_\theta = (T - \theta) n i_\theta$$

and so using this with equation **??** gives

$$\sqrt{n i_\theta}(T - \theta) \to \mathcal{N}(0, 1)$$

which is equivalent to the result required. □

This result means that whenever T is an unbiassed and efficient estimator, $\mathcal{N}\left(\theta, \frac{1}{\sqrt{i_\theta}}\right)$ is an approximate sampling distribution for T.

In a similar way to an unbiassed efficient estimator, the sampling distribution of a maximum likelihood estimators is also asymptotic to a Normal distribution. This is in fact the same Normal distribution, i.e.

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}\left(0, \frac{1}{i_\theta}\right)$$

Reparameterised Maximum Likelihood Estimators

Perhaps a little surprisingly the effect of reparameterising a distribution has no effect on the asymptotic Normality of the maximum likelihood estimator (for the new parameter). The speed of convergence may differ, so clearly a good reparameterisation can be worthwhile.

Distributions With Multiple Parameters

The proper generalisation of this section is to deal with k rather than just 1 parameter. Examples include the Normal and Γ distributions when all parameters are unknown. It is fairly clear that

$$\mathbb{E}\left(\frac{\partial f}{\partial \theta_j}\right) = 0 \quad \text{and} \quad \mathbb{E}\left(\frac{\partial f}{\partial \theta_j}\frac{\partial f}{\partial \theta_k}\right) = I_{jk}$$

where I is the expected information matrix. A score vector $\mathbf{u}$ can be defined in the obvious way and in fact I is the variance co-variance matrix of $\mathbf{u}$. The asymptotic result in this case is

$$\hat{}_{ML} \to \mathcal{MN}\left(\grave{}, I^{-1}\right)$$

Note that I has an inverse since it is positive definite.

(23.2) Hypothesis Testing

(23.2.1) Construction Of A Hypothesis Test

Let Ω be the set of all possible values of a parameter θ and let $\omega \subset \Omega$. A hypothesis test takes hypothesis

H_0: $\theta \in \omega$, the null hypothesis.

H_1: $\theta \in \Omega \setminus \omega$, the alternative hypothesis.

The test will reject H_0 in favour of H_1 when $\mathbf{x} \in C$ where $\mathbf{x}$ is the data and C is the 'critical region'.

Definition 25 *A hypothesis is said to be simple if it corresponds to a single point of Ω. Otherwise it is said to be composite.*

It is worth noting that $H_0\colon \mu = 0$ is not simple in the case of a Normal distribution. This is because σ is not specified.

Testing Simple Hypotheses

In the case of testing simple hypotheses $H_0\colon \theta = \theta_0$ and $H_1\colon \theta = \theta_1$ where $\Omega = \{\theta_1, \theta_2\}$. Clearly such a test could go wrong in one of two ways.

- H_0 is rejected when it is true. This is a type 1 error and has probability α.
- H_0 is accepted when it is false. This is a type 2 error and has probability β.

The probability of rejecting H_0 when it is false is clearly useful in evaluating a hypothesis test. This quantity, $1 - \beta$, is called the power of the test.

The type 1 error corresponds exactly to the significance level of the test. This is also called the size of the test.

Lemma 26 (Neyman-Pearson) *The most powerful test of size α has a critical region of the form*

$$C = \left\{\mathbf{x} \mid \frac{L(\theta_1, \mathbf{x})}{L(\theta_0, \mathbf{x})} \geqslant A\right\}$$

where A is a constant.

Proof. Let C be the critical region as described and let C' be the critical region of another test with size less than or equal to α. The difference in powers of these tests is then given by

$$\begin{aligned}\Pr\{\mathbf{X} \in C \mid \theta_1\} - \Pr\{\mathbf{X} \in C' \mid \theta_1\} &= \int_C L(\theta_1, \mathbf{x})\, d\mathbf{x} - \int_{C'} L(\theta_1, \mathbf{x})\, d\mathbf{x} \\ &= \int_{C \setminus C'} L(\theta_1, \mathbf{x})\, d\mathbf{x} - \int_{C' \setminus C} L(\theta_1, \mathbf{x})\, d\mathbf{x}\end{aligned}$$

Now, $L(\theta_1, \mathbf{x}) \geqslant AL(\theta_0, \mathbf{x})$ for $\mathbf{x} \in C$ and the reverse inequality holds for $\mathbf{x} \notin C$. Hence

$$\begin{aligned}&\geqslant \int_{C \setminus C'} AL(\theta_0, \mathbf{x})\, d\mathbf{x} - \int_{C' \setminus C} AL(\theta_0, \mathbf{x})\, d\mathbf{x} \\ &= \int_C AL(\theta_0, \mathbf{x})\, d\mathbf{x} - \int_{C'} AL(\theta_0, \mathbf{x})\, d\mathbf{x} \\ &= A\left(\Pr\{\mathbf{X} \in C \mid \theta_0\} - \Pr\{\mathbf{X} \in C' \mid \theta_0\}\right) \\ &\geqslant A(\alpha - \alpha) = 0\end{aligned}$$

Hence the power of the test with critical region C is at least that with critical region C'. The proof need only be given for the test C' with size α but as it also works for size $\leqslant \alpha$ this is done. □

Although apparently complicated, the Neyman-Pearson lemma can reduce to give quite simple results. Consider the case of $\mathcal{N}(\theta, 1)$.

$$\begin{aligned}C &= \left\{ \mathbf{x} \mid \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2}\sum_{i=1}^n (x_i - \theta_1)^2\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2}\sum_{i=1}^n (x_i - \theta_2)^2\right)} \geqslant A \right\} \\ &= \left\{ \mathbf{x} \mid \exp\left(\frac{-1}{2}\sum_{i=1}^n (x_i - \theta_1)^2 - \frac{-1}{2}\sum_{i=1}^n (x_i - \theta_0)^2\right) \geqslant A \right\} \\ &= \left\{ \mathbf{x} \mid e^{\frac{-1}{2}} \exp\left(2(\theta_0 - \theta_1)\sum_{i=1}^n x_i + \frac{n}{2}\left(\theta_1^2 - \theta_0^2\right)\right) \geqslant A \right\} \\ &= \left\{ \mathbf{x} \mid \exp\left((\theta_1 - \theta_0)\sum_{i=1}^n x_i\right) \geqslant A' \right\} \\ &= \left\{ \mathbf{x} \mid (\theta_1 - \theta_0)\sum_{i=1}^n x_i \geqslant A'' \right\} \\ &= \begin{cases} \left\{\mathbf{x} \mid \sum_{i=1}^n \leqslant A'''\right\} & \text{if } \theta_1 - \theta_0 < 0 \\ \left\{\mathbf{x} \mid \sum_{i=1}^n \geqslant A'''\right\} & \text{if } \theta_1 - \theta_0 > 0 \end{cases}\end{aligned}$$

The value of A is determined by the size of the test, it must be chosen such that

$$\Pr\left\{\frac{L(\theta_1, \mathbf{x})}{L(\theta_0, \mathbf{x})} \geqslant A \mid \theta_0\right\} = \alpha$$

To actually calculate this the null distribution for the test statistic must be specified. In the case where the test statistic has a standard Normal distribution, $A = \Phi^{-1}(1 - \alpha)$.

In the case of discrete data it may not be possible to achieve a size of α exactly. In this case a randomised test may be used. For example

- Accept H_0 if $T \leqslant 7$.
- Reject H_0 if $T \geqslant 9$.

- If $T = 8$ then reject H_0 with probability p.

where T is a suitably chosen test statistic.

Composite Alternative Hypotheses

The familiar hypothesis test will have $H_0\colon \theta = \theta_0$ and $H_1\colon \theta \neq \theta_0$, or possibly a one-sided alternative. In such a case the power of the test depends on precisely which $\theta \in \Omega \setminus \omega$ is taken.

Definition 27 *For a hypothesis with critical region C define the power function η by*

$$\eta(\theta) = \Pr\{\mathbf{X} \in C \mid \theta\}$$

The size of the test is $\eta(\theta_0)$.

Definition 28 *A hypothesis test is said to be uniformly most powerful of size α if*

1. $\eta(\theta_0) = \alpha$
2. *where η^* is the power function of any other test of size α then $\eta(\theta) \geqslant \eta^*(\theta)$ for all $\theta \in \Omega \setminus \{\theta_0\}$.*

Lemma 29 *Suppose H_0 is simple and H_1 is composite. If the likelihood ratio test of size α for testing against*

$$H_1'\colon \theta = \theta' \quad \theta' \in \Omega \setminus \omega$$

does not depend on θ' then it is uniformly most powerful.

Proof. By the Neyman-Pearson lemma the likelihood ratio test is the most powerful of size α. As (by hypothesis) there is no dependence on θ' the definition of being uniformly most powerful is satisfied. □

Maximum Likelihood Ratio Tests

Most generally the null and alternative hypotheses are both composite. Say

$$H_0\colon \theta \in \omega \qquad H_1\colon \theta \in \Omega \setminus \omega$$

then define

$$\lambda = \frac{\sup_{\theta\in\omega} L(\theta, \mathbf{x})}{\sup_{\theta\in\Omega} L(\theta, \mathbf{x})} = \frac{L(\hat{\theta}_\omega, \mathbf{x})}{L(\hat{\theta}, \mathbf{x})}$$

A maximum likelihood ratio test then has the critical region $\{\mathbf{x} \mid \lambda \leqslant A\}$. Clearly $0 \leqslant \lambda \leqslant 1$, and if H_0 is true the value should be near 1. Such a test readily reduces to a likelihood ratio test for simple hypotheses since for simple hypotheses

$$\lambda = \frac{L(\theta_0, \mathbf{x})}{\max\{L(\theta_0, \mathbf{x}), L(\theta_1, \mathbf{x})\}} = \min\left\{1, \frac{L(\theta_0, \mathbf{x})}{L(\theta_1, \mathbf{x})}\right\}$$

Since in practise $A < 1$ this always gives the likelihood ratio test. It can also be shown that if a uniformly most powerful test exists for a simple null hypothesis then it s a maximum likelihood ratio test.

Example 30 *If $X_1, X_2, \ldots, X_n$ are independently and identically distributed with distribution $\mathcal{N}\left(\mu, \sigma^2\right)$ where both μ and σ^2 are unknown, test $H_0\colon \mu = \mu_0$ against $H_1\colon \mu \neq \mu_0$.*

Proof. Solution The test has composite null and alternative hypotheses with

$$\Omega = \left\{(\mu, \sigma^2) \mid \mu \in \mathbb{R}\ \sigma^2 \in \mathbb{R}_0^+\right\} \qquad \omega = \left\{(\mu_0, \sigma^2) \mid \sigma^2 \in \mathbb{R}_0^+\right\}$$

Now, for $\mu = \mu_0$

$$\begin{aligned}
L(\mu_0\sigma^2, \mathbf{x}) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu_0}{\sigma}\right)\right)\\
l(\mu_0\sigma^2, \mathbf{x}) &= \frac{-2}{n}\ln\sigma^2 - \frac{2}{n}\ln 2\pi - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu_0}{\sigma}\right)\\
\frac{\partial l}{\partial\sigma^2} &= \frac{-1}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu_0)^2\\
\tilde{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2
\end{aligned}$$

For the denominator, the usual maximum likelihood estimates are

$$\hat{\mu} = \overline{x} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Hence

$$\begin{aligned}
\lambda &= \frac{\left(\frac{1}{\tilde{\sigma}\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}(x_i-\mu_0)\right)}{\left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\overline{x})\right)}\\
&= \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)^{\frac{-n}{2}}\\
&= \left(\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu_0)}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)^{\frac{-n}{2}}\\
&= \left(1+\frac{t^2}{n-1}\right)^{\frac{-n}{2}}
\end{aligned}$$

where $t = \frac{\sqrt{n}(\overline{x}-\mu_0)}{s}$ is the usual t statistic. The maximum likelihood ratio test is therefore the usual two sided t test in disguise. □

Lemma 31 *When H_0 is true and n is large*

$$-2\ln\lambda \sim \chi^2_q$$

where q is the difference between the number of parameters of the distribution of the data and the number of parameters fixed in the null hypothesis.

This result is of use when the exact distribution of λ cannot be determined.

Two-Sided Alternative Hypotheses

So far only one sided alternatives have been considered. Consider testing $H_0\colon \theta = \theta_0$ against $H_1\colon \theta \neq \theta_0$.

The score test uses the score function $u(\theta) = \frac{\partial l}{\partial\theta}$ and since for large n its distribution is approximately Normal, $u(\theta) \sim \mathcal{N}(0, I_\theta)$. Hence use the test statistic

$$W_u = \frac{(u(\theta_0))^2}{I_{\theta_0}}$$

which has null distribution χ^2_1.

Alternatively the Wald test, or maximum likelihood estimator test uses $\hat{\theta}$, the maximum likelihood estimate, and since for large n

$$\hat{\theta} \sim \mathcal{N}\left(\theta, I_{\theta}^{-1}\right)$$

define the test statistic

$$W_e = \left(\hat{\theta} - \theta_0\right)^2 I_{\hat{\theta}}$$

which has (approximately) a χ_1^2 null distribution. An alternative to to use $W_e = \left(\hat{\theta} - \theta_0\right)^2 I_{\theta_0}$ but this is rarely done.

A third alternative is to use the statistic $-2\ln\lambda$, as all of these have an approximate χ_1^2 distribution.

(23.2.2) Hypothesis Testing & Confidence Intervals

Consider testing the hypothesis $H_{\theta_0}\colon \theta = \theta_0$ against the alternative $H_1\colon \theta \neq \theta_0$. If C_{θ_0} is the critical region then

$$S_{\mathbf{x}} = \left\{\theta_0 \mid \mathbf{x} \in C_{\theta_0}^c\right\}$$

is a $100(1-\alpha)\%$ confidence region for θ where α is the size of the test. It is simply the set of θ_0s that would not be rejected by the test.

A confidence interval can be constructed using one of many statistics, as indeed can a hypothesis test. The Wald statistic W_e, the score statistic W_u and the maximum likelihood ratio statistic $-2\ln\lambda$ are of particular interest due to their universal applicability.

Single Parameter Confidence Intervals

Consider testing $H_0\colon \theta = \theta_0$ against the alternative $H_1\colon \theta \neq \theta_0$. The maximum likelihood ratio is then $\lambda = \frac{L(\theta_0)}{L(\hat{\theta})}$, and $-2\ln\lambda$ has (in this case) an approximate χ_1^2 distribution from which a confidence interval can be constructed.

Two Parameter Confidence Intervals

Consider a situation where the unknown parameters are θ and K. The parameter K is of interest and a confidence interval must be constructed for it, whereas θ is just a nuisance. Testing $H_0\colon K = K_0$ against $H_1\colon K \neq K_0$ has a composite null hypothesis so a maximum likelihood ratio test can be used. Let $\hat{K}$ be the maximum likelihood estimate of K and let $\hat{\theta}_K$ be the maximum likelihood estimate of θ when K has a prescribed value. Hence

$$\lambda = \frac{\sup_{\theta} L(\theta, K_0)}{\sup_{(\theta,K)} L(\theta, K)} = \frac{L\left(\hat{\theta}_{K_0}, K_0\right)}{L\left(\hat{\theta}_{\hat{K}}, \hat{K}\right)} = \frac{L_p(K_0)}{L_p(\hat{K})}$$

where $L_p(K) = L(\hat{\theta}_K, K)$ is the likelihood profile* for K. As there are two parameters and one is given in the null case, $-2\ln\lambda$ is approximately distributed as χ_1^2.

Confidence Intervals For k Parameters

First of all consider the simple null hypothesis $H_0\colon ` = `_0$ testing against $H_1\colon ` \neq `_0$.

- The score statistic W_u is calculated as

$$W_u = \mathbf{u}^T(`_0) I_{`_0}^{-1} \mathbf{u}(`_{\mathbf{0}})$$

*In this two parameter case the likelihood is a surface in three dimensions. Fixing θ in this way allows a two dimensional plot to be made of the intersection of the likelihood surface with the plane $\theta = \hat{\theta}_K$.

where $\mathbf{u}(`)$ is the score vector and $I_{`}$ is the information matrix.

- The Wald statistic W_e is calculated as

$$W_e = \left(\hat{} - `_0\right)^T I_{\hat{}} \left(\hat{} - `_0\right)$$

When H_0 is true both these statistics and the maximum likelihood ratio statistic $-2\ln\lambda$ are approximately equal and have approximate distribution χ^2_k.

When the null hypothesis is composite, say $H_0 \colon ` \in \omega$ and $H_1 \colon ` \in \Omega \setminus \omega$ the value $`_0$ is replaced by $\hat{}_\omega$ to give

$$\begin{aligned}
W_u &= \mathbf{u}^T(\hat{}_\omega) I^{-1}_{\hat{}_\omega} \mathbf{u}(\hat{}_\omega) \\
W_e &= \left(\hat{} - \hat{}_\omega\right)^T I_{\hat{}} \left(\hat{} - \hat{}_\omega\right) \\
\lambda &= \frac{L\left(\hat{}_\omega\right)}{L\left(\hat{}\right)}
\end{aligned}$$

In this case the approximate distribution is χ^2_q where q parameters are specified in H_0.

(23.2.3) The Multinomial Distribution

An application of the above theory to the multinomial distribution produces the familiar Pearson's goodness of fit test, as well as an alternative.

Definition 32 *Let n objects be placed at random into k boxes and let p_i be the probability that an object goes into box i, which is the same for all objects. Let X_i be the number of objects in box i then*

$$\Pr\left\{X_1 = x_1, X_2 = x_2, \dots, X_k = x_k\right\} = \frac{n!}{x_1!x_2!,\dots,x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

is said to be a multinomial distribution.

Let $\mathbf{X} \sim \mathcal{M}_n(p_1, p_2, \dots, p_k)$. Since $\sum_{i=1}^k p_i = 1$ only $k-1$ of the ps are 'independent', say $p_k = 1 - \sum_{i=1}^{n-1} p_i$. Now,

$$\begin{aligned}
L &\propto p_1^{x_1} p_2^{x_2} \dots p_{k-1}^{x_{k-1}} \left(1 - \sum_{i=1}^{k-1} x_i \ln p_i\right)^{x_k} \\
l &= c + \sum_{i=1}^{k-1} x_i \ln p_i + x_k \ln\left(1 - \sum_{i=1}^{k-1} x_i \ln p_i\right) \\
\frac{\partial l}{\partial p_j} &= \frac{x_j}{p_j} - \frac{x_k}{p_k} \qquad (33) \\
\hat{p}_j &= \frac{x_j}{x_k}\hat{p}_k \quad \text{now sum over } j\text{, the number of boxes} \\
1 &= \frac{n}{x_k}\hat{p}_k \\
\hat{p}_k &= \frac{x_k}{n}
\end{aligned}$$

But the p chosen to be calculated as 1 take the sum of the rest can be chosen at will, and hence this holds for any $1 \leqslant j \leqslant k$ so that $\hat{p}_j = \frac{x_j}{n}$.

Consider testing the (simple) hypothesis $H_0\colon p_i = \pi_i$ for all $1 \leqslant i \leqslant k$ against $H_1\colon p_i \neq \pi_i$ for some i. It can be shown that $\hat{p}_i = \frac{x_i}{n}$ and hence

$$\lambda = \left(\frac{\pi_1}{\hat{p}_1}\right)^{x_1} \left(\frac{\pi_2}{\hat{p}_2}\right)^{x_2} \cdots \left(\frac{\pi_k}{\hat{p}_k}\right)^{x_k} \tag{34}$$

Since all the p_is are specified in the null hypothesis this means that $-2\ln\lambda$ has an approximate χ^2_{k-1} distribution.

Returning to equation 33 the score vector (of length $k-1$) has jth element $u_j = \frac{x_j}{p_j} - \frac{x_k}{p_k}$ and so

$$-\frac{\partial^2 l}{\partial p_i \partial p_j} = \begin{cases} \frac{x_j}{p_j^2} + \frac{x_k}{p_k^2} & \text{if } i = j \\ \frac{x_k}{p_k^2} & \text{if } i \neq j \end{cases}$$

Replacing x with X and taking expected values will now give the information matrix. Observe that $\mathbb{E}\, X_j = np_j$ and hence

$$I_{ij} = \begin{cases} \frac{n}{p_j} + \frac{n}{p_k} & \text{if } i = j \\ \frac{n}{p_k} & \text{if } i \neq j \end{cases} \quad \text{so} \quad I = n\left(\frac{1}{p_k}\mathbf{1}\mathbf{1}^T + D^{-1}\right)$$

where $\mathbf{1}$ is the $(k-1)\times 1$ vector of 1s and D is a diagonal matrix with $D_{ii} = p_i$. Using this the Wald and score statistics can be calculated. This is a rather lengthy process, particularly for the score statistic. The final results are

$$W_e = n\sum_{i=1}^{k} \frac{(\hat{p}_i - \pi_i)^2}{\hat{p}_i} \tag{35}$$

$$W_u = n\sum_{i=1}^{k} \frac{(\hat{p}_i - \pi_i)^2}{\pi_i} \tag{36}$$

Equation (36) is the familiar ‘goodness of fit’ test statistic—the Pearson χ^2 statistic. However, the Wald statistic, equation (35), and $-2\ln\lambda$ where λ is as in equation (34) are both perfectly good alternatives for goodness of fit testing.