Chapter 2

The Maximum Likelihood Estimator

We start this chapter with a few "quirky examples", based on estimators we are already familiar with and then we consider classical maximum likelihood estimation.

2.1 Some examples of estimators

Example 1

Let us suppose that $\{X_i\}_{i=1}^n$ are iid normal random variables with mean μ and variance σ^2. The "best" estimators unbiased estimators of the mean and variance are $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ respectively. To see why recall that $\sum_i X_i$ and $\sum_i X_i^2$ are the sufficient statistics of the normal distribution and that $\sum_i X_i$ and $\sum_i X_i^2$ are complete minimal sufficient statistics. Therefore, since $\bar{X}$ and s^2 are functions of these minimally sufficient statistics, by the Lehmann-Scheffe Lemma, these estimators have minimal variance.

Now let us consider the situation where the mean is μ and the variance is μ^2. In this case we have only one unknown parameter μ but the minimally sufficient statistics are $\sum_i X_i$ and $\sum_i X_i^2$. Moreover, it is not complete since both

$$\left(\frac{n}{n+1}\right)\bar{X}^2 \text{ and } s^2 \tag{2.1}$$

are unbiased estimators of μ^2 (to understand why the first estimator is an unbiased estimator we use that $\mathrm{E}[X^2] = 2\mu^2$). Thus violating the conditions of completeness. Furthermore

any convex linear combination of these estimators

$$\alpha\left(\frac{n}{n+1}\right)\bar{X}^2+(1-\alpha)s^2 \qquad 0\leq\alpha\leq 1$$

is an unbiased estimator of μ. Observe that this family of distributions is incomplete, since

$$\mathrm{E}\left[\left(\frac{n}{n+1}\right)\bar{X}^2-s^2\right]=\mu^2-\mu^2,$$

thus there exists a non-zero function $Z(S_x,S_{xx})$ Furthermore

$$\left(\frac{n}{n+1}\right)\bar{X}^2-s^2=\frac{1}{n(n+1)}S_x^2-\frac{1}{n-1}\left(S_{xx}-\frac{1}{n}S_x\right)=Z(S_x,S_{xx}).$$

Thus there exists a non-zero function $Z(\cdot)$ such that $\mathrm{E}[Z(S_x,S_{xx})]=0$, impying the minimal sufficient statistics are not complete.

Thus for all sample sizes and μ, it is not clear which estimator has a minimum variance. We now calculate the variance of both estimators and show that there is no clear winner for all n. To do this we use the normality of the random variables and the identity (which applies only to normal random variables)

$$\begin{aligned}\mathrm{cov}\left[AB,CD\right] &= \mathrm{cov}[A,C]\mathrm{cov}[B,D]+\mathrm{cov}[A,D]\mathrm{cov}[B,C]+\mathrm{cov}[A,C]\mathrm{E}[B]\mathrm{E}[D]+\\ &\quad \mathrm{cov}[A,D]\mathrm{E}[B]\mathrm{E}[C]+\mathrm{E}[A]\mathrm{E}[C]\mathrm{cov}[B,D]+\mathrm{E}[A]\mathrm{E}[D]\mathrm{cov}[B,C]\end{aligned}$$

[12]. Using this result we have

$$\begin{aligned}\mathrm{var}\left[\frac{n}{n+1}\bar{X}^2\right] &= \left(\frac{n}{n+1}\right)^2\mathrm{var}[\bar{X}^2]=\left(\frac{n}{n+1}\right)^2\left\{2\mathrm{var}[\bar{X}]^2+4\mu^2\mathrm{var}[\bar{X}]\right\}\\ &= \left(\frac{n}{n+1}\right)^2\left[\frac{2\mu^4}{n^2}+\frac{4\mu^4}{n}\right]=\frac{2\mu^4}{n}\left(\frac{n}{n+1}\right)^2\left(\frac{1}{n}+4\right).\end{aligned}$$

[1]Observe that this identity comes from the general identity

$$\begin{aligned}&\mathrm{cov}\left[AB,CD\right]\\ =\ &\mathrm{cov}[A,C]\mathrm{cov}[B,D]+\mathrm{cov}[A,D]\mathrm{cov}[B,C]+\mathrm{E}[A]\mathrm{cum}[B,C,D]+\mathrm{E}[B]\mathrm{cum}[A,C,D]\\ &+\mathrm{E}[D]\mathrm{cum}[A,B,C]+\mathrm{E}[C]\mathrm{cum}[A,B,D]+\mathrm{cum}[A,B,C,D]\\ &+\mathrm{cov}[A,C]\mathrm{E}[B]\mathrm{E}[D]+\mathrm{cov}[A,D]\mathrm{E}[B]\mathrm{E}[C]+\mathrm{E}[A]\mathrm{E}[C]\mathrm{cov}[B,D]+\mathrm{E}[A]\mathrm{E}[D]\mathrm{cov}[B,C]\end{aligned}$$

recalling that cum denotes cumulant and are the coefficients of the cumulant generating function (`https://en.wikipedia.org/wiki/Cumulant`), which applies to non-Gaussian random variables too

[2]Note that $\mathrm{cum}(A,B,C)$ is the coefficient of $t_1t_2t_3$ in the series expansion of $\log\mathrm{E}[e^{t_1A+t_2B+t_3B}]$ and can be obtained with $\frac{\partial^3\log\mathrm{E}[e^{t_1A+t_2B+t_3B}]}{\partial t_1\partial t_2\partial t_3}\rfloor_{t_1,t_2,t_3=0}$

On the other hand using that s^2 has a chi-square distribution with $n-1$ degrees of freedom (with variance $2(n-1)^2$) we have

$$\text{var}\left[s^2\right] = \frac{2\mu^4}{(n-1)}.$$

Altogether the variance of these two difference estimators of μ^2 are

$$\text{var}\left[\frac{n}{n+1}\bar{X}^2\right] = \frac{2\mu^4}{n}\left(\frac{n}{n+1}\right)^2\left(4+\frac{1}{n}\right) \text{ and var}\left[s^2\right] = \frac{2\mu^4}{(n-1)}.$$

There is no estimator which clearly does better than the other. And the matter gets worse, since any convex combination is also an estimator! This illustrates that Lehman-Scheffe theorem does not hold in this case; we recall that Lehman-Scheffe theorem states that under completeness any unbiased estimator of a sufficient statistic has minimal variance. In this case we have two different unbiased estimators of sufficient statistics neither estimator is uniformly better than another.

Remark 2.1.1 *Note, to estimate μ one could use $\bar{X}$ or $\sqrt{s^2} \times sign(\bar{X})$ (though it is unclear to me whether the latter is unbiased).*

Exercise 2.1 *Calculate (the best you can)* $\mathrm{E}[\sqrt{s^2} \times sign(\bar{X})]$.

Example 2

Let us return to the censored data example considered in Sections 1.2 and 1.6.4, Example (v). $\{X_i\}_{i=1}^n$ are iid exponential distributed random variables, however we do not observe X_i we observe a censored version $Y_i = \min(X_i, c)$ (c is assumed known) and $\delta_i = 0$ if $Y_i = X_i$ else $\delta_i = 1$.

We recall that the log-likelihood of (Y_i, δ_i) is

$$\begin{aligned}\mathcal{L}_n(\theta) &= \sum_i (1-\delta_i)\left\{-\theta Y_i + \log\theta\right\} - \sum_i \delta_i c\theta \\ &= -\sum_i \theta Y_i - \log\theta \sum_i \delta_i + n\log\theta,\end{aligned}$$

since $Y_i = c$ when $\delta_i = 1$. hence the minimal sufficient statistics for θ are $\sum_i \delta_i$ and $\sum_i Y_i$. This suggests there may be several different estimators for θ.

(i) $\sum_{i=1}^{n} \delta_i$ gives the number of observations which have been censored. We recall that $P(\delta_i = 1) = \exp(-c\theta)$, thus we can use $n^{-1}\sum_{i=1}^{n} \delta_i$ as an estimator of $\exp(-c\theta)$ and solve for θ.

(ii) The non-censored observations also convey information about θ. The likelihood of a non-censored observations is

$$\mathcal{L}_{nC,n}(\theta) = -\theta \sum_{i=1}^{n} (1-\delta_i) Y_i + \sum_{i=1}^{n} (1-\delta_i) \left\{ \log\theta - \log(1-e^{-c\theta}) \right\}.$$

One could maximise this to obtain an estimator of θ

(iii) Or combine the censored and non-censored observations by maximising the likelihood of θ given (Y_i, θ_i) to give the estimator

$$\frac{\sum_{i=1}^{n}(1-\delta_i)}{\sum_{i=1}^{n} Y_i}.$$

The estimators described above are not unbiased (hard to take the expectation), but they do demonstrate that often there is often no unique best method for estimating a parameter.

Though it is usually difficult to find an estimator which has the smallest variance for all sample sizes, in general the maximum likelihood estimator "asymptotically" (think large sample sizes) usually attains the Cramer-Rao bound. In other words, it is "asymptotically" efficient.

Exercise 2.2 (Two independent samples from a normal distribution) *Suppose that $\{X_i\}_{i=1}^{m}$ are iid normal random variables with mean μ and variance σ_1^2 and $\{Y_i\}_{i=1}^{m}$ are iid normal random variables with mean μ and variance σ_2^2. $\{X_i\}$ and $\{Y_i\}$ are independent, calculate their joint likelihood.*

(i) Calculate their sufficient statistics.

(ii) Propose a class of estimators for μ.

2.2 The Maximum likelihood estimator

There are many different parameter estimation methods. However, if the family of distributions from the which the parameter comes from is known, then the maximum likelihood

estimator of the parameter θ, which is defined as

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} L_n(\underline{X};\theta) = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta),$$

is the most commonly used. Often we find that $\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}_n} = 0$, hence the solution can be obtained by solving the derivative of the log likelihood (the derivative of the log-likelihood is often called the *score function*). However, if θ_0 lies on the boundary of the parameter space this will *not* be true. In general, the maximum likelihood estimator will not be an unbiased estimator of the parameter.

We note that the likelihood is invariant to bijective transformations of the data. For example if X has the density $f(\cdot;\theta)$ and we define the transformed random variable $Z = g(X)$, where the function g has an inverse, then it is easy to show that the density of Z is $f(g^{-1}(z);\theta)\frac{\partial g^{-1}(z)}{\partial z}$. Therefore the likelihood of $\{Z_i = g(X_i)\}$ is

$$\prod_{i=1}^{n} f(g^{-1}(Z_i);\theta)\frac{\partial g^{-1}(z)}{\partial z}\rfloor_{z=Z_i} = \prod_{i=1}^{n} f(X_i;\theta)\frac{\partial g^{-1}(z)}{\partial z}\rfloor_{z=Z_i}.$$

Hence it is proportional to the likelihood of $\{X_i\}$ and the maximum of the likelihood in terms of $\{Z_i = g(X_i)\}$ is the same as the maximum of the likelihood in terms of $\{X_i\}$.

Example 2.2.1 (The uniform distribution) *Consider the uniform distribution, which has the density $f(x;\theta) = \theta^{-1}I_{[0,\theta]}(x)$. Given the iid uniform random variables $\{X_i\}$ the likelihood (it is easier to study the likelihood rather than the log-likelihood) is*

$$L_n(\underline{X}_n;\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n} I_{[0,\theta]}(X_i).$$

Using $L_n(\underline{X}_n;\theta)$, the maximum likelihood estimator of θ is $\widehat{\theta}_n = \max_{1\le i\le n} X_i$ (you can see this by making a plot of $L_n(\underline{X}_n;\theta)$ against θ).

To derive the properties of $\max_{1\le i\le n} X_i$ we first obtain its distribution. It is simple to see that

$$P(\max_{1\le i\le n} X_i \le x) \;=\; P(X_1 \le x,\ldots,X_n \le x) = \prod_{i=1}^{n} P(X_i \le x) = \big(\frac{x}{\theta}\big)^n I_{[0,\theta]}(x),$$

and the density of $\max_{1\le i\le n} X_i$ is $f_{\widehat{\theta}_n}(x) = nx^{n-1}/\theta^n$.

Exercise 2.3 *(i) Evaluate the mean and variance of $\widehat{\theta}_n$ defined in the above example.*

(ii) Is the estimator biased? If it is, find an unbiased version of the estimator.

Example 2.2.2 (Weibull with known α) *$\{Y_i\}$ are iid random variables, which follow a Weibull distribution, which has the density*

$$\frac{\alpha y^{\alpha-1}}{\theta^\alpha}\exp(-(y/\theta)^\alpha) \qquad \theta,\alpha>0.$$

Suppose that α is known, but θ is unknown. Our aim is to fine the MLE of θ.

The log-likelihood is proportional to

$$\begin{aligned}\mathcal{L}_n(\underline{X};\theta) &= \sum_{i=1}^n\left(\log\alpha+(\alpha-1)\log Y_i-\alpha\log\theta-\left(\frac{Y_i}{\theta}\right)^\alpha\right)\\ &\propto \sum_{i=1}^n\left(-\alpha\log\theta-\left(\frac{Y_i}{\theta}\right)^\alpha\right).\end{aligned}$$

The derivative of the log-likelihood wrt to θ is

$$\frac{\partial\mathcal{L}_n}{\partial\theta}=-\frac{n\alpha}{\theta}+\frac{\alpha}{\theta^{\alpha+1}}\sum_{i=1}^n Y_i^\alpha=0.$$

Solving the above gives $\widehat{\theta}_n=(\frac{1}{n}\sum_{i=1}^n Y_i^\alpha)^{1/\alpha}$.

Example 2.2.3 (Weibull with unknown α) *Notice that if α is given, an explicit solution for the maximum of the likelihood, in the above example, can be obtained. Consider instead the case that both α and θ are unknown. Now we need to find α and θ which maximise the likelihood i.e.*

$$\arg\max_{\theta,\alpha}\sum_{i=1}^n\left(\log\alpha+(\alpha-1)\log Y_i-\alpha\log\theta-(\frac{Y_i}{\theta})^\alpha\right).$$

The derivative of the likelihood is

$$\begin{aligned}\frac{\partial\mathcal{L}_n}{\partial\theta} &= -\frac{n\alpha}{\theta}+\frac{\alpha}{\theta^{\alpha+1}}\sum_{i=1}^n Y_i^\alpha=0\\ \frac{\partial\mathcal{L}_n}{\partial\alpha} &= \frac{n}{\alpha}-\sum_{i=1}^n\log Y_i-n\log\theta-\frac{n\alpha}{\theta}+\sum_{i=1}^n\log(\frac{Y_i}{\theta})\times(\frac{Y_i}{\theta})^\alpha=0.\end{aligned}$$

It is clear that an explicit expression to the solution of the above does not exist and we need to find alternative methods for finding a solution (later we show how profiling can be used to estimate α).

2.3 Maximum likelihood estimation for the exponential class

Typically when maximising the likelihood we encounter several problems (i) for a given likelihood $\mathcal{L}_n(\theta)$ the maximum may lie on the boundary (even if in the limit of $\mathcal{L}_n$ the maximum lies with in the parameter space) (ii) there are several local maximums (so a numerical routine may not capture the true maximum) (iii) $\mathcal{L}_n$ may not be concave, so even if you are close the maximum the numerical routine just cannot find the maximum (iv) the parameter space may not be convex (ie. $(1-\alpha)\theta_1 + \alpha\theta_2$ may lie outside the parameter space even if θ_1 and θ_2 are in the parameter space) again this will be problematic for numerically maximising over the parameter space. When there is just one unknown parameters these problems are problematic, when the number of unknown parameters is p this becomes a nightmare. However for the full rank exponential class of distributions we now show that everything behaves, in general, very well. First we heuristically obtain its maximum likelihood estimator, and later justify it.

2.3.1 Full rank exponential class of distributions

Suppose that $\{X_i\}$ are iid random variables which has a the natural exponential representation and belongs to the family $\mathcal{F} = \{f(x;\theta) = \exp[\sum_{j=1}^p \theta_j s_j(x) - \kappa(\theta) + c(x)]; \theta \in \Theta\}$ and $\Theta = \{\theta; \kappa(\theta) = \log \int \exp(\sum_{j=1}^p \theta_j s_j(x) + c(x))dx < \infty\}$ (note this condition defines the parameter space, if $\kappa(\theta) = \infty$ the density is no longer defined). Therefore the log likelihood function is

$$\mathcal{L}_n(\underline{X};\theta) = \theta \sum_{i=1}^n \mathbf{s}(X_i) - n\kappa(\theta) + \sum_{i=1}^n c(X_i),$$

where $\sum_{i=1}^n \mathbf{s}(X_i) = (\sum_{i=1}^n s_1(X_i), \ldots, \sum_{i=1}^n s_p(X_i))$ are the sufficient statistics. By the Rao-Blackwell theorem the unbiased estimator with the smallest variance will be a function of $\sum_{i=1}^n \mathbf{s}(X_i)$. We now show that the maximum likelihood estimator of θ is a function of $\sum_{i=1}^n \mathbf{s}(X_i)$ (though there is no guarantee it will be unbiased);

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} \big\{\theta \sum_{i=1}^n \mathbf{s}(X_i) - n\kappa(\theta) + \sum_{i=1}^n c(X_i)\big\}.$$

The natural way to obtain $\widehat{\theta}_n$ is to solve

$$\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta=\widehat{\theta}_n} = 0.$$

However, this equivalence will only hold if the maximum lies *within* the interior of the parameter space (we show below that in general this will be true). Let us suppose this is true, then differentiating $\mathcal{L}_n(\underline{X};\theta)$ gives

$$\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta} = \sum_{i=1}^{n} \mathbf{s}(X_i) - n\kappa'(\theta) = 0.$$

To simplify notation we often write $\kappa'(\theta) = \mu(\theta)$ (since this is the mean of the sufficient statistics). Thus we can invert back to obtain the maximum likelihood estimator

$$\widehat{\theta}_n = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i)\right). \tag{2.2}$$

Because the likelihood is a concave function, it has a unique maximum. But the maximum will only be at $\widehat{\theta}_n = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i)\right)$ *if* $\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i) \in \mu(\Theta)$. If $\mu^{-1}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i))$ takes us outside the parameter space, then clearly this cannot be an estimator of the parameter[3]. Fortunately, in most cases (specifically, if the model is said to be "steep"), $\mu^{-1}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i))$ will lie in the interior of the parameter space. In other words,

$$\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(X_i)\right) = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta).$$

In the next section we define steepness and what may happen if this condition is not satisfied. But first we go through a few examples.

Example 2.3.1 (Normal distribution) *For the normal distribution, the log-likelihood is*

$$\mathcal{L}(\underline{X};\sigma^2,\mu) = \frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n} X_i^2 - 2\mu\sum_{i=1}^{n} X_i + n\mu^2\right) - \frac{n}{2}\log\sigma^2,$$

note we have ignored the 2π constant. Differentiating with respect to σ^2 and μ and setting to zero gives

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \widehat{\mu})^2.$$

[3]For example and estimator of the variance which is negative, clearly this estimator has no meaning

This is the only solution, hence it must the maximum of the likelihood.

Notice that $\widehat{\sigma}^2$ is a slightly biased estimator of σ^2.

Example 2.3.2 (Multinomial distribution) *Suppose $\underline{Y} = (Y_1, \ldots, Y_q)$ (with $n = \sum_{i=1}^q Y_i$) has a multinomial distribution where there are q cells. Without any constraints on the parameters the log likelihood is proportional to (we can ignore the term $c(\underline{Y}) = \log \binom{n}{Y_1,\ldots,Y_q}$)*

$$\mathcal{L}_n(\underline{Y}; \pi) = \sum_{j=1}^{q-1} Y_i \log \pi_i + Y_q \log(1 - \sum_{i=1}^{q-1} \pi_i).$$

The partial derivative for each i is

$$\frac{\mathcal{L}(\underline{Y}; \pi)}{\partial \pi_i} = \frac{Y_i}{\pi} - \frac{Y_p}{1 - \sum_{i=1}^{p-1} \pi_i}.$$

Solving the above we get one solution as $\widehat{\pi}_i = Y_i/n$ (check by plugging it in).

Since there is a diffeomorphism between $\{\pi_i\}$ and its natural parameterisation $\theta_i = \log \pi_i/(1 - \sum_{j=1}^{q-1} \pi_i)$ and the Hessian corresponding to the natural paramerisation is negative definite (recall the variance of the sufficient statistics is $\kappa''(\theta)$), this implies that the Hessian of $\mathcal{L}_n(\underline{Y}; \pi)$ is negative definite, thus $\widehat{\pi}_i = Y_i/n$ is the unique maximum of $\mathcal{L}(\underline{Y}; \pi)$.

Example 2.3.3 ($2 \times 2 \times 2$ Contingency tables) *Consider the example where for n individuals three binary variables are recorded; Z =gender (here, we assume two), X =whether they have disease A (yes or no) and Y =whether they have disease B (yes or no). We assume that the outcomes of all n individuals are independent.*

Without any constraint on variables, we model the above with a multinomial distribution with $q = 3$ i.e. $P(X = x, Y = y, Z = z) = \pi_{xyz}$. In this case the likelihood is proportional to

$$\begin{aligned}
\mathcal{L}(\underline{Y}; \pi) &= \sum_{x=0}^{1} \sum_{y=0}^{1} \sum_{z=0}^{1} Y_{xyz} \log \pi_{xyz} \\
&= Y_{000}\pi_{000} + Y_{010}\pi_{010} + Y_{000}\pi_{001} + Y_{100}\pi_{100} + Y_{110}\pi_{110} + Y_{101}\pi_{101} \\
&\quad Y_{011}\pi_{011} + Y_{111}(1 - \pi_{000} - \pi_{010} - \pi_{001} - \pi_{100} - \pi_{110} - \pi_{101} - \pi_{011}).
\end{aligned}$$

Differentiating with respect to each variable and setting to one it is straightfoward to see that the maximum is when $\widehat{\pi}_{xyz} = Y_{xyz}/n$; which is intuitively what we would have used as the estimator.

However, suppose the disease status of X and Y are independent conditioned on gender. i.e. $P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$ then $P(X = x, Y = y, Z = z) = \pi_{X=x|Z=z}\pi_{Y=y|Z=z}\pi_{Z=z}$, since these are binary variables we drop the number of unknown parameters from 7 to 5. This is a curved exponential model (though in this case the constrained model is simply a 5-dimensional hyperplane in 7 dimensional space; thus the parameter space is convex). The log likelihood is proportional to

$$\begin{aligned}\mathcal{L}(\underline{Y};\pi) &= \sum_{x=0}^{1}\sum_{y=0}^{1}\sum_{z=0}^{1} Y_{xyz}\log \pi_{x|z}\pi_{y|z}\pi_z \\ &= \sum_{x=0}^{1}\sum_{y=0}^{1}\sum_{z=0}^{1} Y_{xyz}\left(\log \pi_{x|z} + \log \pi_{y|z} + \log \pi_z\right).\end{aligned}$$

Thus we see that the maximum likelihood estimators are

$$\widehat{\pi}_z = \frac{Y_{++z}}{n} \quad \widehat{\pi}_{x|z} = \frac{Y_{x+z}}{Y_{++z}} \quad \widehat{\pi}_{y|z} = \frac{Y_{+yz}}{Y_{++z}}.$$

Where in the above we use the standard notation $Y_{+++} = n$, $Y_{++z} = \sum_{x=0}^{1}\sum_{y=0}^{1} Y_{xyz}$ etc. We observe that these are very natural estimators. For example, it is clear that Y_{x+z}/n is an estimator of the joint distribution of X and Z and Y_{++z}/n is an estimator of the marginal distribution of Z. Thus Y_{x+z}/Y_{++z} is clearly an estimator of X conditioned on Z.

Exercise 2.4 *Evaluate the mean and variance of the numerator and denominator of (2.4). Then use the continuous mapping theorem to evaluate the limit of $\widehat{\theta^{-1}}$ (in probability).*

Example 2.3.4 (The beta distribution) *Consider the family of densities defined by*

$$\mathcal{F} = \left\{f(x;\alpha,\beta) = B(\alpha,\beta)^{-1}x^{\alpha-1}(1-x)^{\beta-1}; \alpha \in (0,\infty), \beta \in (0,\infty)\right\}$$

and $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$. This is called the family of beta distributions.

The log likelihood can be written as

$$\begin{aligned}\mathcal{L}_n(\underline{X};\theta) &= \alpha\sum_{i=1}^{n}\log X_i + \beta\sum_{i=1}^{n}\log(1-X_i) - n\left[\log(\Gamma(\alpha)) + \log(\Gamma(\beta) - \log(\Gamma(\alpha+\beta))\right] - \\ &\quad \sum_{i=1}^{n}[\log X_i - \log(1-X_i)].\end{aligned}$$

Thus $\theta_1 = \alpha$, $\theta_2 = \beta$ and $\kappa(\theta_1, \theta_2) = \log(\theta_1) + \log(\theta_2) - \log(\theta_1 + \theta_2)$.

Taking derivatives and setting to zero gives

$$\frac{1}{n}\begin{pmatrix} \sum_{i=1}^n \log X_i \\ \sum_{i=1}^n \log(1 - X_i) \end{pmatrix} = \begin{pmatrix} \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} \\ \frac{\Gamma'(\beta)}{\Gamma(\beta)} - \frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} \end{pmatrix}.$$

To find estimators for α and β we need to numerically solve for the above. But will the solution lie in the parameter space?

Example 2.3.5 (Inverse Gaussian distribution) *Consider the inverse Gaussian distribution defined as*

$$f(x; \theta_1, \theta_2) = \frac{1}{\pi^{1/2}} x^{-3/2} \exp\left(\theta_1 x - \theta_2 x^{-1} + [-2(\theta_1\theta_2)^{1/2} - \frac{1}{2}\log(-\theta_2)]\right),$$

where $x \in (0, \infty)$. Thus we see that $\kappa(\theta_1, \theta_2) = [-2(\theta_1\theta_2)^{1/2} - \frac{1}{2}\log(-\theta_2)]$. In this case we observe that for $\theta_1 = 0$ $\kappa(0, \theta_2) < \infty$ thus the parameter space is not open and $\Theta = (-\infty, 0] \times (-\infty, 0)$. Taking derivatives and setting to zero gives

$$\frac{1}{n}\begin{pmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i^{-1} \end{pmatrix} = \begin{pmatrix} -\left(\frac{\theta_2}{\theta_1}\right)^{1/2} \\ -\frac{\theta_1^{1/2}}{\theta_2^{1/2}} + \frac{1}{\theta_2}. \end{pmatrix}.$$

To find estimators for α and β we need to numerically solve for the above. But will the solution lie in the parameter space?

Example 2.3.6 (The inflated zero Poisson distribution) *Using the natural parameterisation of the infltaed zero Poisson distribution we have*

$$\begin{aligned} \mathcal{L}(\underline{Y}; \theta_1, \theta_2) &= \theta_1 \sum_{i=1}^n I(Y_i \neq 0) + \theta_2 \sum_{i=1}^n I(Y_i \neq 0) Y_i \\ &\quad - \log\left(\frac{e^{\theta_1} - \theta_2^{-1}}{1 - \theta_2^{-1}}(1 - e^{-e^{\theta_2}}) + e^{-e^{\theta_2}}\right). \end{aligned}$$

where the parameter space is $\Theta = (-\infty, 0] \times (-\infty, \infty)$, which is not open (note that 0 corresponds to the case $p = 0$, which is the usual Poisson distribution with no inflation).

To find estimators for θ and p we need to numerically solve for the above. But will the solution lie in the parameter space?

2.3.2 Steepness and the maximum of the likelihood

The problem is that despite the Hessian $\nabla^2\mathcal{L}(\theta)$ being non-negative definite, it could be that the maximum is at the boundary of the likelihood. We now state some results that show that in most situations, this does not happen and usually (2.2) maximises the likelihood. For details see Chapters 3 and 5 of `http://www.jstor.org/stable/pdf/4355554.pdf?acceptTC=true` (this reference is mathematically quite heavy) for a maths lite review see Davidson (2004) (page 170). Note that Brown and Davidson use the notation $\mathcal{N}$ to denote the parameter space Θ.

Let $\mathcal{X}$ denote the range of the sufficient statistics $\mathbf{s}(X_i)$ (i.e. what values can $s(X)$ take). Using this we define its convex hull as

$$C(\mathcal{X}) = \{\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2; \quad \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, 0 \leq \alpha \leq 1\}.$$

Observe that $\frac{1}{n}\sum_i \mathbf{s}(X_i) \in C(\mathcal{X})$, even when $\frac{1}{n}\sum_i \mathbf{s}(X_i)$ does not belong to the observation space of the sufficient statistic $\mathcal{X}$. For example X_i may be counts from a Binomial distribution $\text{Bin}(m, p)$ but $C(\mathcal{X})$ would be the reals between $[0, m]$.

Example 2.3.7 (Examples of $C(\mathcal{X})$) *(i) The normal distribution*

$$C(\mathcal{X}) = \left\{\alpha(x, x^2) + (1-\alpha)(y, y^2); \quad x, y \in \mathbb{R}, 0 \leq \alpha \leq 1\right\} = (-\infty, \infty)(0, \infty).$$

(ii) The β-distribution

$$C(\mathcal{X}) = \{\alpha(\log x, \log(1-x)) + (1-\alpha)(\log y, \log(1-y)); \quad x, y \in [0,1], 0 \leq \alpha \leq 1\} = (\mathbb{R}^{-1})^2$$

(iii) The exponential with censoring (see 2.3)

$$C(\mathcal{X}) = \{\alpha(y_1, \delta_1) + (1-\alpha)(y_2, \delta_2); \quad y_1 \in [0, c], \delta_1, \delta_2 = \{0, 1\}; 0 \leq \alpha \leq 1\} = \textit{triangle}.$$

(iv) The binomial distribution $Y \sim Bin(n, \pi)$. Then

$$C(\mathcal{X}) = \{\alpha x + (1-\alpha)y; 0 \leq \alpha \leq 1, y = 0, \ldots, m\} = [0, m].$$

Now we give conditions under which $\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right)$ maximises the likelihood within the parameter space Θ. Define the parameter space $\Theta = \{\theta; \kappa(\theta) < \infty\} \subset \mathbb{R}^q$. Let $int(\Theta)$ denote the interior of a set, which is the largest open set in Θ. Next we define the notion of *steep*.

Definition 2.3.1 *Let $\kappa : \mathbb{R}^p \to (-\infty, \infty)$ be a convex function (so $-\kappa$ is concave). κ is called steep if for all $\theta_1 \in B(\Theta)$ and $\theta_0 \in int(\Theta)$, $\lim_{\rho\to\infty}(\theta_1 - \theta_0)\frac{\partial\kappa(\theta)}{\partial\theta}\rfloor_{\theta=\theta_0+\rho(\theta_1-\theta_0)} = \infty$. This condition is equivalent to $\lim_{\theta\to B(\Theta)} |\kappa'(\theta)| \to \infty$. Intuitively, steep simply means the function is very steep at the boundary.*

- *Regular exponential family*

  If the parameter space is open (such as $\Theta = (0, 1)$ or $\Theta = (0, \infty)$) meaning the density is not defined on the boundary, then the family of exponentials is called a regular family.

  In the case that Θ is open (the boundary does not belong to Θ), then κ is not defined at the boundary, in which case κ is steep.

  Note, at the boundary $\lim_{\theta\to B(\Theta)} \log f(x;\theta)$ will approach $-\infty$, since $\{\log f(x;\theta)\}$ is convex over θ this means that its maximum will be within the interior of the parameter space (just what we want!).

- *Non-regular exponential family*

  If the parameter space is closed, this means at the boundary the density is defined, then we require that at the boundary of the parameter space $\kappa(\cdot)$ is *steep*. This condition needs to be checked by considering the expectation of the sufficient statistic at the boundary or equivalently calculating $\kappa'(\cdot)$ at the boundary.

If $\kappa(\theta)$ is steep we have the following result. Brown (1986), Theorem 3.6 shows that there is a homeomorphism[4] between $int(\Theta)$ and $int(C(\mathcal{X}))$.

Most *importantly* Brown (1986), Theorem 5.5 shows that if the density of X_i belongs to a full rank exponential family (using the natural parameterisation) $f(x;\theta) = \exp[\sum_{j=1}^p \theta_j s_j(x) - \kappa(\theta) + c(x)]$ with $\theta = (\theta_1, \ldots, \theta_p) \in \Theta$, where $\kappa(\cdot)$ is steep and for a given data set $\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i) \in C(\mathcal{X})$, then

$$\widehat{\theta}_n = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right) = \arg\max_{\theta\in int(\Theta)} \big\{\theta\sum_{i=1}^n \mathbf{x}_i - n\kappa(\theta) + \sum_{i=1}^n c(X_i)\big\}.$$

In most situations the full rank exponential family will have a parameter space which is open and thus steep.

[4] A homeomorphism between two spaces means there is a bijection between two spaces and the f and f^{-1} which maps between the two spaces is continuous.

Example 2.3.8 (Binomial distribution and observations that lie on the boundary) *Suppose that $\{Y_i\}_{i=1}^n$ are iid Binomially distributed random variables $Y_i \sim Bin(m, \pi_i)$. The log likelihood of Y_i is $Y_i \log(\frac{\pi}{1-\pi}) + m(1-\pi)$. Thus the log likelihood of the sample is proportional to*

$$\mathcal{L}_n(\underline{Y}; \pi) = \sum_{i=1}^n Y_i \log \pi + \sum_{i=1}^n (m - Y_i) \log(1-\pi) = \theta \sum_{i=1}^n Y_i - nm \log(1 + e^\theta),$$

where $\theta \in (-\infty, \infty)$. The theory states above that the maximum of the likelihood lies within the interior of $(-\infty, \infty)$ if $\sum_{i=1}^n Y_i$ lies within the interior of $C(\mathcal{Y}) = (0, nm)$.

On the other hand, there is a positive probability that $\sum_{i=1}^n Y_i = 0$ or $\sum_{i=1}^n Y_i = nm$ (i.e. all successes or all failures). In this case, the above result is not informative. However, a plot of the likelihood in this case is very useful (see Figure 2.1). More precisely, if $\sum_i Y_i = 0$, then $\widehat{\theta}_n = -\infty$ (corresponds to $\widehat{p} = 0$), if $\sum_i Y_i = nm$, then $\widehat{\theta}_n = \infty$ (corresponds to $\widehat{p} = 1$). Thus even when the sufficient statistics lie on the boundary of $C(\mathcal{Y})$ we obtain a very natural estimator for θ.

Example 2.3.9 (Inverse Gaussian and steepness) *Consider the log density of the inverse Gaussian, where X_i are iid positive random variables with log likelihood*

$$\mathcal{L}(\underline{X}; \theta) = \theta_1 \sum_{i=1}^n X_i + \theta_2 \sum_{i=1}^n X_i^{-1} - n\kappa(\theta_1, \theta_2) - \frac{3}{2} \sum_{i=1}^n \log X_i - \frac{1}{2} \log \pi,$$

where $\kappa(\theta_1, \theta_2) = -2\sqrt{\theta_1 \theta_2} - \frac{1}{2} \log(-2\theta_2)$. Observe that $\kappa(0, \theta_2) < \infty$ hence $(\theta_1, \theta_2) \in (-\infty, 0] \times (-\infty, 0)$.

However, at the boundary $\frac{\partial \kappa(\theta_1, \theta_2)}{\partial \theta_1} \rfloor_{\theta_1 = 0} = -\infty$. Thus the inverse Gaussian distribution is steep but non-regular. Thus the MLE is $\mu^{-1}(\cdot)$.

Example 2.3.10 (Inflated zero Poisson) *Recall*

$$\begin{aligned} \mathcal{L}(\underline{Y}; \theta_1, \theta_2) &= \theta_1 \sum_{i=1}^n I(Y_i \neq 0) + \theta_2 \sum_{i=1}^n I(Y_i \neq 0) Y_i \\ &\quad - \log \left(\frac{e^{\theta_1} - \theta_2^{-1}}{1 - \theta_2^{-1}} (1 - e^{-e^{\theta_2}}) + e^{-e^{\theta_2}} \right). \end{aligned}$$

where the parameter space is $\Theta = (-\infty, 0] \times (-\infty, \infty)$, which is not open (note that 0 corresponds to the case $p = 0$, which is the usual Poisson distribution with no inflation).

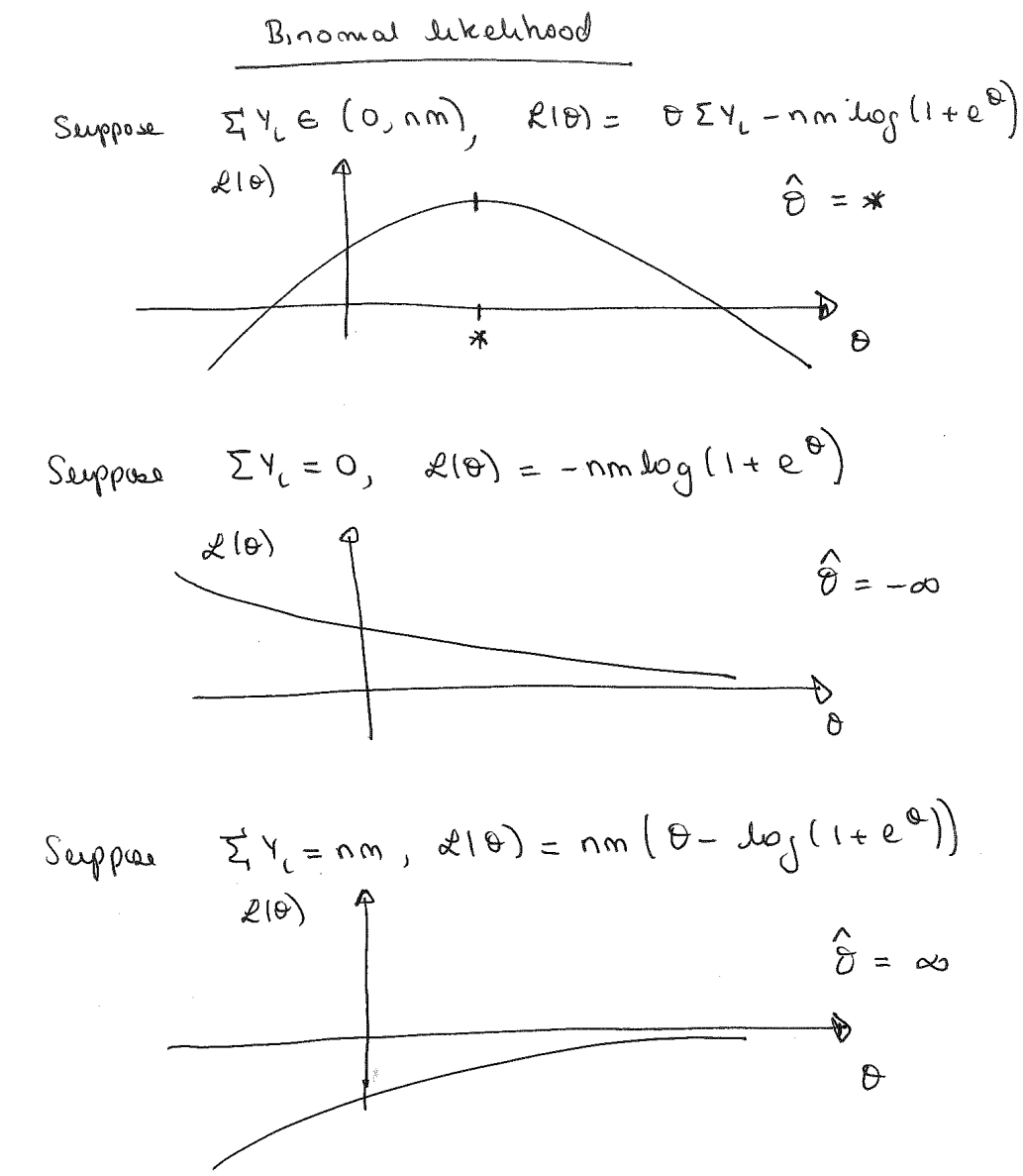

Figure 2.1: Likelihood of Binomial for different scenarios

However, the derivative $\frac{\partial \kappa(\theta_1,\theta_2)}{\partial \theta_1}$ *is finite at* $\theta_1 = 0$ *(for* $\theta_2 \in \mathbb{R}$*). Thus* $\kappa(\cdot)$ *is not steep and care needs to be taken in using* μ^{-1} *as the MLE.*

$\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i))$ *may lie outside the parameter space. For example,* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i))$ *may give an estimator of* θ_1 *which is greater than zero; this corresponds to the probability* $p < 0$*, which makes no sense. If* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i))$ *lies out the parameter space we need to search on the boundary for the maximum.*

Example 2.3.11 (Constraining the parameter space) *If we place an "artifical" constraint on the parameter space then a maximum may not exist within the interior of the parameter space. For example, if we model survival times using the exponential distribution* $f(x;\theta) = \theta\exp(-\theta x)$ *the parameter space is* $(0,\infty)$*, which is open (thus with probability one the likelihood is maximised at* $\widehat{\theta} = \mu^{-1}(\bar{X}) = 1/\bar{X}$*). However, if we constrain the parameter space* $\widetilde{\Theta} = [2,\infty)$*,* $1/\bar{X}$ *may lie outside the parameter space and we need to use* $\widehat{\theta} = 2$*.*

Remark 2.3.1 (Estimating ω**)** *The results above tell us if* $\kappa(\cdot)$ *is steep in the parameter space and* $\mathcal{L}_n(\theta)$ *has a unique maximum and there is a diffeomorphism between* θ *and* ω *(if the exponential family is full rank), then* $\mathcal{L}_n(\theta(\omega))$ *will have a unique maximum. Moreover the Hessian of the likelihood of both parameterisations will be negative definite. Therefore, it does not matter if we maximise over the natural parametrisation or the usual parameterisation*

$$\widehat{\omega}_n = \eta^{-1}\left(\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\right)\right).$$

Remark 2.3.2 (Minimum variance unbiased estimators) *Suppose* X_i *has a distribution in the natural exponential family, then the maximum likelihood estimator is a function of the sufficient statistic* $s(\underline{X})$*. Moreover if the exponential is full and* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i)$ *is an* **unbiased** *estimator of* θ*, then* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i)$ *is the minumum variance unbiased estimator of* θ*. However, in general* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i)$ *will not be an unbiased estimator. However, by invoking the continuous mapping theorem (*`https://en.wikipedia.org/wiki/Continuous_mapping_theorem`*), by the law of large numbers* $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \xrightarrow{a.s.} \mathrm{E}[\mathbf{x}]_i$*, then* $\mu^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i) \xrightarrow{a.s.} \mu^{-1}(\mathrm{E}(\mathbf{x})) = \mu^{-1}[\kappa'(\theta)] = \theta$*. Thus the maximum likelihood estimator converges to* θ*.*

2.3.3 The likelihood estimator of the curved exponential

Example 2.3.12 (Normal distribution with constraint) *Suppose we place the constraint on the parameter space $\sigma^2 = \mu^2$. The log-likelihood is*

$$\mathcal{L}(\underline{X};\mu) = \frac{-1}{2\mu^2}\sum_{i=1}^{n} X_i^2 + \frac{1}{\mu}\sum_{i=1}^{n} X_i - \frac{n}{2}\log\mu^2.$$

Recall that this belongs to the curved exponential family and in this case the parameter space is not convex. Differentiating with respect to μ gives

$$\frac{\partial\mathcal{L}(\underline{X};\sigma^2,\mu)}{\partial\mu} = \frac{1}{\mu^3}S_{xx} - \frac{S_x}{\mu^2} - \frac{1}{\mu} = 0.$$

Solving for μ leads to the quadratic equation

$$p(\mu) = \mu^2 + S_x\mu - S_{xx} = 0.$$

Clearly there will be two real solutions

$$\frac{-S_x \pm \sqrt{S_x^2 + 4S_{xx}}}{2}.$$

We need to plug them into the log-likelihood to see which one maximises the likelihood.

Observe that in this case the Hessian of the log-likelihood cannot be negative (unlike the full normal). However, we know that a maximum exists since a maximum exists on for the full Gaussian model (see the previous example).

Example 2.3.13 (Censored exponential) *We recall that the likelihood corresponding to the censored exponential is*

$$\mathcal{L}_n(\theta) = -\theta\sum_{i=1}^{n} Y_i - \log\theta\sum_{i=1}^{n}\delta_i + \log\theta. \tag{2.3}$$

We recall that $\delta_i = 1$ if censoring takes place. The maximum likelihood estimator is

$$\widehat{\theta} = \frac{n - \sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n} Y_i} \in (0,\infty)$$

Basic calculations show that the mean of the exponential is $1/\theta$, therefore the estimate of the mean is

$$\widehat{\theta^{-1}} = \frac{\sum_{i=1}^{n} Y_i}{\underbrace{n - \sum_{i=1}^{n}\delta_i}_{\text{no. not censored terms}}}. \tag{2.4}$$

If the exponential distribution is curved (number of unknown parameters is less than the number of minimally sufficient statistics), then the parameter space $\Omega = \{\omega = (\omega_1, \ldots, \omega_d); (\theta_1(\omega), \ldots, \theta_q(\omega)) \in \Theta\} \subset \Theta$ (hence it is a curve on Θ). Therefore, by differentiating the likelihood with respect to ω, a maximum within the parameter space must satisfy

$$\nabla_\omega \mathcal{L}_n(\theta(\omega)) = \frac{\partial \theta(\omega)}{\partial \omega} \left(\sum_{i=1}^n \mathbf{x}_i - n \frac{\partial \kappa(\theta)}{\partial \theta} \rfloor_{\theta=\theta(\omega)} \right) = 0. \tag{2.5}$$

Therefore, either (a) there exists an $\omega \in \Omega$ such that $\theta(\omega)$ is the global maximum of $\{\mathcal{L}_n(\theta); \theta \in \Theta\}$ (in this case $\sum_{i=1}^n \mathbf{x}_i - n \frac{\partial \kappa(\theta)}{\partial \theta} \rfloor_{\theta=\theta(\omega)} = 0$) or (b) there exists an $\omega \in \Omega$ such that $\frac{\partial \theta(\omega)}{\partial \omega}$ and $\sum_{i=1}^n \mathbf{x}_i - n \frac{\partial \kappa(\theta)}{\partial \theta} \rfloor_{\theta=\theta(\omega)}$ are orthogonal. Since $\mathcal{L}_n(\underline{X}, \theta)$ for $\theta \in \Theta$ and $\sum_{i=1}^n \mathbf{x}_i \in int(\mathcal{X})$ has a global maximum a simple illustration this means that $\mathcal{L}_n(\theta(\omega))$ will have a maximum. In general (2.5) will be true. As far as I can see the only case where it may not hold is when $\theta(\omega)$ lies on some contour of $\mathcal{L}_n(\theta)$. This suggests that a solution should in general exist for the curved case, but it may not be unique (you will need to read Brown (1986) for full clarification). Based this I suspect the following is true:

- If Ω is a curve in Θ, then $\frac{\partial \mathcal{L}_n(\omega)}{\partial \omega} = 0$ may have multiple solutions. In this case, we have to try each solution $\mathcal{L}_n(\widehat{\omega})$ and use the solution which maximises it (see Figure 2.2).

Exercise 2.5 *The aim of this question is to investigate the MLE of the inflated zero Poisson parameters λ and p. Simulate from a inflated zero poisson distribution with (i) $p = 0.5$, $p = 0.2$ and $p = 0$ (the class is when there is no inflation), use $n = 50$. Evaluate the MLE (over 200 replications) make a Histogram and QQplot of the parameter estimators (remember if the estimator of p is outside the parameter space you need to locate the maximum on the parameter space).*

Exercise 2.6 *(i) Simulate from the model defined in Example 2.3.12 (using $n = 20$) using R. Calculate and maximise the likelihood over 200 replications. Make a QQplot of the estimators and calculate the mean squared error.*

For one realisation make a plot of the log-likelihood.

(ii) Sample from the inverse Gamma distribution (using $n = 20$) and obtain its maximum likelihood estimator. Do this over 200 replications and make a table summarizing its bias and average squared error. Make a QQplot of the estimators.

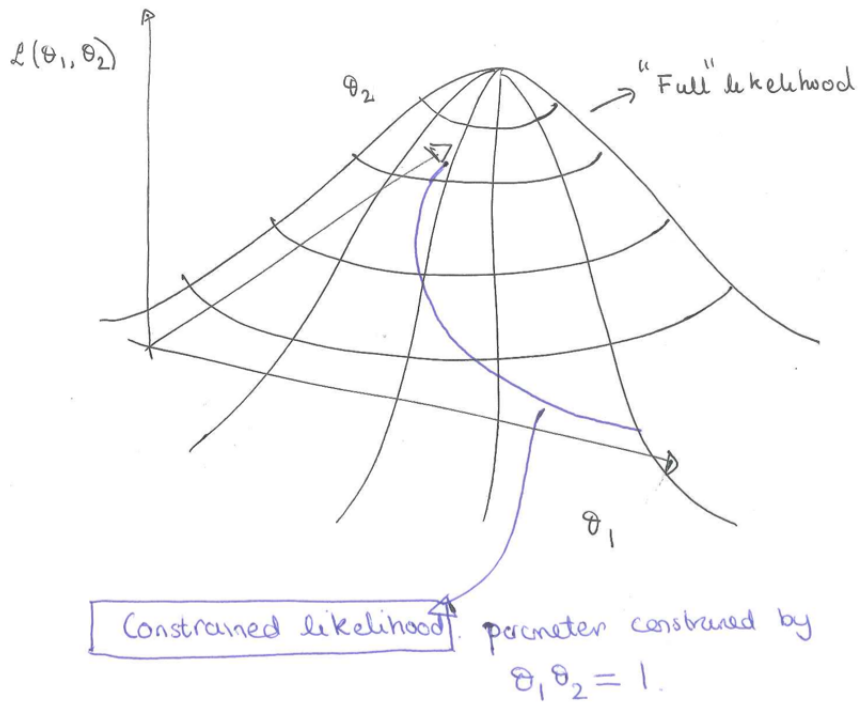

Figure 2.2: Likelihood of 2-dimension curved exponential

(iii) Consider the exponential distribution described in Example 2.3.11 where the parameter space is constrained to $[2, \infty]$*. For samples of size* $n = 50$ *obtain the maximum likelihood estimator (over 200 replications). Simulate using the true parameter*

(a) $\theta = 5$ *(b)* $\theta = 2.5$ *(c)* $\theta = 2$*.*

Summarise your results and make a QQplot (against the normal distribution) and histogram of the estimator.

2.4 The likelihood for dependent data

We mention that the likelihood for dependent data can also be constructed (though often the estimation and the asymptotic properties can be a lot harder to derive). Suppose $\{X_t\}_{t=1}^n$ is a time series (a sequence of observations over time where there could be dependence). Using Bayes rule (ie. $P(A_1, A_2, \ldots, A_n) = P(A_1) \prod_{i=2}^n P(A_i | A_{i-1}, \ldots, A_1)$)

we have

$$L_n(\underline{X};\theta) = f(X_1;\theta)\prod_{t=2}^{n} f(X_t|X_{t-1},\ldots,X_1;\theta).$$

Under certain conditions on $\{X_t\}$ the structure above $\prod_{t=2}^{n} f(X_t|X_{t-1},\ldots,X_1;\theta)$ can be simplified. For example if X_t were Markovian then X_t conditioned on the past on depends only on the recent past, i.e. $f(X_t|X_{t-1},\ldots,X_1;\theta) = f(X_t|X_{t-1};\theta)$ in this case the above likelihood reduces to

$$L_n(\underline{X};\theta) = f(X_1;\theta)\prod_{t=2}^{n} f(X_t|X_{t-1};\theta). \tag{2.6}$$

We apply the above to a very simple time series. Consider the AR(1) time series

$$X_t = \phi X_{t-1} + \varepsilon_t, \qquad t \in \mathbb{Z},$$

where ε_t are iid random variables with mean zero and variance σ^2. In order to ensure that the recurrence is well defined for all $t \in \mathbb{Z}$ we assume that $|\phi| < 1$ in this case the time series is called stationary[5].

We see from the above that the observation X_{t-1} has a linear influence on the next observation and it is Markovian; conditioned on X_{t-1}, X_{t-2} and X_t are independent (the distribution function $P(X_t \leq x|X_{t-1},X_{t-2}) = P(X_t \leq x|X_{t-1})$). Therefore by using (2.6) the likelihood of $\{X_t\}_t$ is

$$L_n(\underline{X};\phi) = f(X_1;\phi)\prod_{t=2}^{n} f_\varepsilon(X_t - \phi X_{t-1}), \tag{2.7}$$

where f_ε is the density of ε and $f(X_1;\phi)$ is the marginal density of X_1. This means the likelihood of $\{X_t\}$ only depends on f_ε and the marginal density of X_t. We use $\widehat{\phi}_n = \arg\max L_n(\underline{X};\phi)$ as the mle estimator of a.

We now derive an explicit expression for the likelihood in the case that ε_t belongs to the exponential family. We focus on the case that $\{\varepsilon_t\}$ is Gaussian; since X_t is the sum of Gaussian random variables $X_t = \sum_{j=0}^{\infty}\phi^j\varepsilon_{t-j}$ (almost surely) X_t is also Gaussian. It can be shown that if $\varepsilon_t \sim \mathcal{N}(0,\sigma^2)$, then $X_t \sim \mathcal{N}(0,\sigma^2/(1-\phi^2))$. Thus the log likelihood

[5]If we start the recursion at some finite time point t_0 then the time series is random walk and is called a unit root process it is not stationary.

for Gaussian "innovations" is

$$\begin{aligned}\mathcal{L}_n(\phi,\sigma^2) &= -\frac{1}{2\sigma^2}\underbrace{\sum_{t=2}^{n}X_t^2}_{=\sum_{t=2}^{n-1}X_t^2-X_n^2}+\frac{\phi}{\sigma^2}\sum_{t=2}^{n}X_tX_{t-1}-\frac{\phi^2}{2\sigma^2}\underbrace{\sum_{t=2}^{n}X_{t-1}^2}_{=\sum_{t=2}^{n-1}X_t^2-X_1^2}-\frac{n-1}{2}\log\sigma^2\\ &\quad -\frac{(1-\phi^2)}{2\sigma^2}X_1^2-\frac{1}{2}\log\frac{\sigma^2}{1-\phi^2}\\ &= -\frac{1-\phi^2}{2\sigma^2}\sum_{t=1}^{n-1}X_t^2+\frac{\phi}{\sigma^2}\sum_{t=2}^{n}X_tX_{t-1}-\frac{1}{2\sigma^2}(X_1^2+X_n^2)-\frac{n-1}{2}\log\sigma^2-\frac{1}{2}\log\frac{\sigma^2}{1-\phi^2},\end{aligned}$$

see Efron (1975), Example 3. Using the factorisation theorem we see that the sufficient statistics, for this example are $\sum_{t=1}^{n-1}X_t^2$, $\sum_{t=2}^{n}X_tX_{t-1}$ and $(X_1^2+X_n^2)$ (it almost has two sufficient statistics!). Since the data is dependent some caution needs to be applied before ones applies the results on the exponential family to dependent data (see Küchler and Sørensen (1997)). To estimate ϕ and σ^2 we maximise the above with respect to ϕ and σ^2. It is worth noting that the maximum can lie on the boundary -1 or 1.

Often we ignore the term the distribution of X_1 and consider the *conditional log-likelihood*, that is $X_2,\ldots,X_n$ conditioned on X_1. This gives the conditional log likelihood

$$\begin{aligned}Q_n(\phi,\sigma^2;X_1) &= \log\prod_{t=2}^{n}f_\varepsilon(X_t-\phi X_{t-1})\\ &= -\frac{1}{2\sigma^2}\sum_{t=2}^{n}X_t^2+\frac{\phi}{\sigma^2}\sum_{t=2}^{n}X_tX_{t-1}-\frac{\phi^2}{2\sigma^2}\sum_{t=2}^{n}X_{t-1}^2-\frac{n-1}{2}\log\sigma^2,\end{aligned}\tag{2.8}$$

again there are three sufficient statistics. However, it is interesting to note that if the maximum of the likelihood lies within the parameter space $\phi\in[-1,1]$ then $\widehat{\phi}_n=\sum_{t=2}^{n}X_tX_{t-1}/\sum_{t=2}^{n}X_{t-1}^2$ (the usual least squares estimator).

2.5 Evaluating the maximum: Numerical Routines

In an ideal world an explicit closed form expression would exist for the maximum of a (log)-likelihood. In reality this rarely happens.

Usually, we have to use a numerical routine to maximise the likelihood. It is relative straightforward to maximise the likelihood of random variables which belong to the exponential family (since they typically have a negative definite Hessian). However, the story

becomes more complicated if the likelihood does not belong to the exponential family, for example mixtures of exponential family distributions.

Let us suppose that $\{X_i\}$ are iid random variables which follow the classical normal mixture distribution

$$f(y;\theta) = pf_1(y;\theta_1) + (1-p)f_2(y;\theta_2),$$

where f_1 is the density of the normal with mean μ_1 and variance σ_1^2 and f_2 is the density of the normal with mean μ_2 and variance σ_2^2. The log likelihood is

$$\mathcal{L}_n(\underline{Y};\theta) = \sum_{i=1}^{n} \log\left(p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{1}{2\sigma_1^2}(X_i-\mu_1)^2\right] + (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left[-\frac{1}{2\sigma_2^2}(X_i-\mu_2)^2\right]\right).$$

Studying the above it is clear there does not explicit solution to the maximum. Hence one needs to use a numerical algorithm to maximise the above likelihood.

We discuss a few such methods below.

The Newton Raphson Routine The Newton-Raphson routine is the standard method to numerically maximise the likelihood, this can often be done automatically in `R` by using the `R` functions `optim` or `nlm`. To apply Newton-Raphson, we have to assume that the derivative of the likelihood exists (this is not always the case - think about the ℓ_1-norm based estimators!) and the maximum lies inside the parameter space such that $\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}_n} = 0$. We choose an initial value $\theta_n^{(1)}$ and apply the routine

$$\theta_n^{(j)} = \theta_n^{(j-1)} - \left(\frac{\partial^2\mathcal{L}_n(\theta)}{\partial\theta^2}\rfloor_{\theta_n^{(j-1)}}\right)^{-1}\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta_n^{(j-1)}}.$$

This routine can be derived from the Taylor expansion of $\frac{\partial\mathcal{L}_n(\theta_{n-1})}{\partial\theta}$ about θ_0 (see Section 2.6.3). A good description is given in `https://en.wikipedia.org/wiki/Newton%27s_method`. We recall that $-\frac{\partial^2\mathcal{L}_n(\theta)}{\partial\theta^2}\rfloor_{\theta_n^{(j-1)}}$ is the observed Fisher information matrix. If the algorithm does not converge, sometimes we replace $-\frac{\partial^2\mathcal{L}_n(\theta)}{\partial\theta^2}\rfloor_{\theta_n^{(j-1)}}$ with its expectation (the Fisher information matrix); since this is positive definite it may give better results (this is called Fisher scoring).

If the likelihood has just one global maximum and is concave, then it is quite easy to maximise. If on the other hand, the likelihood has a few local maximums and the initial value θ_1 is not chosen close enough to the true maximum, then the

routine may converge to a local maximum. In this case it may be a good idea to do the routine several times for several different initial values θ_n^*. For each candidate value $\widehat{\theta}_n^*$ evaluate the likelihood $\mathcal{L}_n(\widehat{\theta}_n^*)$ and select the value which gives the largest likelihood. It is best to avoid these problems by starting with an informed choice of initial value.

Implementing a Newton-Rapshon routine without much thought can lead to estimators which take an incredibly long time to converge. If one carefully considers the likelihood one can shorten the convergence time by rewriting the likelihood and using faster methods (often based on the Newton-Raphson).

Iterative least squares This is a method that we shall describe later when we consider Generalised linear models. As the name suggests the algorithm has to be interated, however at each step weighted least squares is implemented (see later in the course).

The EM-algorithm This is done by the introduction of dummy variables, which leads to a new 'unobserved' likelihood which can easily be maximised (see later in the course).

2.6 Statistical inference

2.6.1 A quick review of the central limit theorem

In this section we will not prove the central limit theorem. Instead we summarise the CLT and generalisations of it. The purpose of this section is not to lumber you with unnecessary mathematics but to help you understand when an estimator is close to normal (or not).

Lemma 2.6.1 (The famous CLT) *Let us suppose that* $\{X_i\}_{i=1}^n$ *are iid random variables, let* $\mu = \mathrm{E}(X_i) < \infty$ *and* $\sigma^2 = \mathrm{var}(X_i) < \infty$. *Define* $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. *Then we have*

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2).$$

Heuristically, we can write $(\bar{X} - \mu) \approx \mathcal{N}(0, \frac{\sigma^2}{n})$.

What this means that if we have a large enough sample size and made a quantile plot against the normal distribution the points should lie roughly on the $x = y$ line (though there will be less matching in the tails).

Remark 2.6.1 *(i) The above lemma appears to be 'restricted' to just averages. However, it can be used in several different contexts. Averages arise in several different situations. It is not just restricted to the average of the observations. By judicious algebraic manipulations, one can show that several estimators can be rewritten as an average (or approximately as an average). At first appearance, the MLE does not look like an average, however, in Section 2.6.3 we show that it can be approximated by a "useable" average.*

(ii) The CLT can be extended in several ways.

(a) To random variables whose variance are not all the same (ie. independent but identically distributed random variables).

(b) Dependent random variables (so long as the dependency 'decays' in some way).

(c) Weighted averages can also be asymptotically normal; so long as the weights are 'distributed evenly' over all the random variables.

* *Suppose that $\{X_i\}$ are iid non-normal random variables, $Y = \sum_{j=0}^{M} \phi^j X_j$ ($|\phi| < 1$) will never be normal (however large M).*
* *However, $Y = \frac{1}{n}\sum_{i=1}^{n} \sin(2\pi i/12) X_i$ is asymptotically normal.*

- *There exists several theorems which one can use to prove normality. But really the take home message is, look at your estimator and see whether asymptotic normality it looks plausible. Always check through simulations (even if asymptically it is normal, it may require a very large sample size for it to be close to normal).*

Example 2.6.1 (Some problem cases) *A necessary condition is that the second moment of X_i should exist. If it does not the CLT will not hold. For example if $\{X_i\}$ follow a t-distribution with 2 degrees of freedom*

$$f(x) = \frac{\Gamma(3/2)}{\sqrt{2\pi}} \left(1 + \frac{x^2}{2}\right)^{-3/2},$$

then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ will not have a normal limit.

We apply can immediately apply the above result to the MLE in the full rank exponential class.

2.6.2 Sampling properties and the full rank exponential family

In Section 2.3.1 we showed that if $\{X_i\}_{i=1}^n$ belonged to the exponential family and the maximum of the likelihood lay inside the parameter space (satisfied if the distribution is "steep") then

$$\widehat{\theta}_n = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right)$$

is the maximum likelihood estimator. Since we have an "explicit" expression for the estimator it is straightforward to derive the sampling properties of $\widehat{\theta}_n$. By using the law of large numbers

$$\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i) \stackrel{\text{a.s.}}{\to} \mathrm{E}[\mathbf{s}(X)] \qquad n \to \infty$$

then by the continuous mapping theorem

$$\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right) \stackrel{\text{a.s.}}{\to} \mu^{-1}\left(\mathrm{E}[\mathbf{s}(X)]\right) = \theta \qquad n \to \infty.$$

Thus the maximum likelihood estimator is a consistent estimator of θ. By using the CLT we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\mathbf{s}(X_i) - \mathrm{E}[\mathbf{s}(X_i)]\right) \stackrel{\mathcal{D}}{\to} \mathcal{N}\left(0, \mathrm{var}[\mathbf{s}(X_i)]\right) \qquad n \to \infty$$

where we recall that $\mathrm{var}[\mathbf{s}(X_i)] = \kappa''(\theta)$. Now by using that

$$\begin{aligned}\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right) - \mu^{-1}\left(\mu(\theta)\right) &\approx \frac{\partial \mu^{-1}(x)}{\partial x}\rfloor_{x=\theta}\left(\frac{1}{n}\sum_{i=1}^n \left[\mathbf{s}(X_i) - \mathrm{E}(\mathbf{s}(X_i))\right]\right) \\ &= \left(\frac{\partial \mu(x)}{\partial x}\right)\rfloor_{x=\theta}^{-1}\left(\frac{1}{n}\sum_{i=1}^n \left[\mathbf{s}(X_i) - \mathrm{E}(\mathbf{s}(X_i))\right]\right).\end{aligned}$$

Thus by using the above, the continuous mapping theorem and the CLT for averages we have

$$\begin{aligned}\sqrt{n}\left[\mu^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{s}(X_i)\right) - \mu^{-1}\left(\mu(\theta)\right)\right] &\stackrel{\mathcal{D}}{\to} \mathcal{N}\left(0, \left(\frac{\partial \mu(x)}{\partial x}\right)\rfloor_{x=\theta}^{-1}\kappa''(\theta)\left(\frac{\partial \mu(x)}{\partial x}\right)\rfloor_{x=\theta}^{-1}\right) \\ &= \mathcal{N}\left(0, \kappa''(\theta)^{-1}\right).\end{aligned}$$

We recall that $\kappa''(\theta)$ is the Fisher information of θ based on X_1.

Thus we have derived the sampling properties of the maximum likelihood estimator for the exponential class. It is relatively straightfoward to derive. Interesting we see that the limiting variance is the inverse of the Fisher information. So asymptotically the MLE estimator attains the Cramer-Rao lower bound (though it is not really a variance). However, the above derivation apply only to the full exponential class, in the following section we derive a similar result for the general MLE.

2.6.3 The Taylor series expansion

The Taylor series is used all over the place in statistics. It can be used to prove consistency of an estimator, normality (based on the assumption that averages converge to a normal distribution), obtaining the limiting variance of an estimator etc. We start by demonstrating its use for the log likelihood.

We recall that the mean value (in the univariate case) states that

$$f(x) = f(x_0) + (x - x_0)f'(\bar{x}_1) \text{ and } f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2}f''(\bar{x}_2),$$

where $\bar{x}_1 = \alpha x_0 + (1 - \alpha)x$ and $\bar{x}_2 = \beta x + (1 - \beta)x_0$ (for some $0 \leq \alpha, \beta \leq 1$). In the case that $f : \mathbb{R}^q \to \mathbb{R}$ we have

$$\begin{aligned} f(\underline{x}) &= f(\underline{x}_0) + (\underline{x} - \underline{x}_0)\nabla f(\underline{x})\rfloor_{\underline{x}=\bar{\underline{x}}_1} \\ f(\underline{x}) &= f(\underline{x}_0) + (\underline{x} - \underline{x}_0)'\nabla f(\underline{x})\rfloor_{\underline{x}=\underline{x}_0} + \frac{1}{2}(\underline{x} - \underline{x}_0)'\nabla^2 f(\underline{x})\rfloor_{\underline{x}=\bar{\underline{x}}_2}(\underline{x} - \underline{x}_0), \end{aligned}$$

where $\bar{\underline{x}}_1 = \alpha x_0 + (1 - \alpha)x$ and $\bar{\underline{x}}_2 = \beta x + (1 - \beta)x_0$ (for some $0 \leq \alpha, \beta \leq 1$). In the case that $f(\underline{x})$ is a vector, then the mean value theorem does not directly work, i.e. the following *is not true*

$$\underline{f}(\underline{x}) = \underline{f}(\underline{x}_0) + (\underline{x} - \underline{x}_0)'\nabla \underline{f}(\underline{x})\rfloor_{\underline{x}=\bar{\underline{x}}_1},$$

where $\bar{\underline{x}}_1$ lies between $\underline{x}$ and $\underline{x}_0$. However, it is quite straightforward to overcome this inconvience. The mean value theorem does hold pointwise, for every element of the vector $\underline{f}(\underline{x}) = (f_1(\underline{x}), \ldots, f_p(\underline{x}))$, ie. for every $1 \leq j \leq p$ we have

$$f_j(\underline{x}) = f_i(\underline{x}_0) + (\underline{x} - \underline{x}_0)\nabla f_j(\underline{y})\rfloor_{\underline{y}=\alpha\underline{x}+(1-\alpha)\underline{x}_0},$$

where $\bar{\underline{x}}_j$ lies between $\underline{x}$ and $\underline{x}_0$. Thus if $\nabla f_j(\underline{x})\rfloor_{\underline{x}=\bar{\underline{x}}_j} \rightarrow \nabla f_j(\underline{x})\rfloor_{\underline{x}=\underline{x}_0}$, we do have that

$$\underline{f}(\underline{x}) \approx \underline{f}(\underline{x}_0) + (\underline{x} - \underline{x}_0)'\nabla \underline{f}(\underline{x}).$$

We use the above below.

- Application 1: An expression for $\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0)$ in terms of $(\widehat{\theta}_n - \theta_0)$.

  The expansion of $\mathcal{L}_n(\widehat{\theta}_n)$ about θ_0 (the true parameter)

  $$\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0) = -\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n}(\widehat{\theta}_n - \theta_0) - \frac{1}{2}(\widehat{\theta}_n - \theta_0)'\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n}(\widehat{\theta}_n - \theta_0)$$

  where $\bar{\theta}_n = \alpha\theta_0 + (1-\alpha)\widehat{\theta}_n$. If $\widehat{\theta}_n$ lies in the interior of the parameter space (this is an extremely important assumption here) then $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n} = 0$. Moreover, if it can be shown that $|\widehat{\theta}_n - \theta_0| \xrightarrow{\mathcal{P}} 0$ and $n^{-1}\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}$ converges uniformly to $\mathrm{E}(n^{-1}\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\theta_0})$ (see Assumption 2.6.1(iv), below), then we have

  $$\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n} \xrightarrow{\mathcal{P}} \mathrm{E}\left(\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\theta_0}\right) = -I_n(\theta_0). \tag{2.9}$$

  This altogether gives

  $$2(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(\theta_0)) \approx (\widehat{\theta}_n - \theta_0)' I_n(\theta_0)(\widehat{\theta}_n - \theta_0). \tag{2.10}$$

- Application 2: An expression for $(\widehat{\theta}_n - \theta_0)$ in terms of $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0}$

  The expansion of the p-dimension vector $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n}$ pointwise about θ_0 (the true parameter) gives (for $1 \leq j \leq d$)

  $$\frac{\partial \mathcal{L}_{j,n}(\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n} = \frac{\partial \mathcal{L}_{j,n}(\theta)}{\partial \theta}\rfloor_{\theta_0} + \frac{\partial^2 \mathcal{L}_{j,n}(\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_{j,n}}(\widehat{\theta}_n - \theta_0),$$

  where $\bar{\theta}_{j,n} = \alpha_j \bar{\theta}_{j,n} + (1-\alpha_j)\theta_0$. Using the same arguments as in Application 1 and equation (2.9) we have

  $$\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} \approx I_n(\theta_0)(\widehat{\theta}_n - \theta_0).$$

  We mention that $U_n(\theta_0) = \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0}$ is often called the *score or U statistic*. And we see that the asymptotic sampling properties of U_n determine the sampling properties of $(\widehat{\theta}_n - \theta_0)$.

Remark 2.6.2 *(i) In practice $I_n(\theta_0)$ is unknown and it is approximated by the Hessian evaluated at the estimated parameter $\widehat{\theta}_n$, $-\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\widehat{\theta}_n}$. A discussion on the quality of this approximation is given in Efron and Hinkley (1978).*

(ii) Bear in mind that $\nabla^2_\theta \mathcal{L}_n(\theta)$ is not necessarily negative definite, but its limit is the negative Fisher information matrix $-I_n(\theta)$ (non-negative definite over $\theta \in \Theta$). Therefore for "large n $\nabla^2_\theta \mathcal{L}_n(\theta)$ will be negative definite".

(iii) The quality of the approximation (2.9) depends on the the second order efficiency measure $I_n(\widehat{\theta}_n) - I_n(\theta_0)$ (this term was coined by C.R.Rao and discussed in Rao (1961, 1962, 1963)). Efron (1975), equation (1.1) shows this difference depends on the so called curvature of the parameter space.

Example 2.6.2 (The Weibull) *Evaluate the second derivative of the likelihood given in Example 2.2.3, take the expection on this, $I_n(\theta, \alpha) = \mathrm{E}(\nabla^2 \mathcal{L}_n)$ (we use the ∇ to denote the second derivative with respect to the parameters α and θ).*

Application 2 implies that the maximum likelihood estimators $\widehat{\theta}_n$ and $\widehat{\alpha}_n$ (recalling that no explicit expression for them exists) can be written as

$$\begin{pmatrix} \widehat{\theta}_n - \theta \\ \widehat{\alpha}_n - \alpha \end{pmatrix} \approx I_n(\theta, \alpha)^{-1} \begin{pmatrix} \sum_{i=1}^n \left(-\frac{\alpha}{\theta} + \frac{\alpha}{\theta^{\alpha+1}} Y_i^\alpha \right) \\ \sum_{i=1}^n \left(\frac{1}{\alpha} - \log Y_i - \log \theta - \frac{\alpha}{\theta} + \log(\frac{Y_i}{\theta}) \times (\frac{Y_i}{\theta})^\alpha \right) \end{pmatrix}$$

2.6.4 Sampling properties of the maximum likelihood estimator

We have shown that under certain conditions the maximum likelihood estimator can often be the minimum variance unbiased estimator (for example, in the case of the normal distribution). However, in most situations for finite samples the mle may not attain the Cramer-Rao lower bound. Hence for finite sample $\mathrm{var}(\widehat{\theta}_n) > I_n(\theta)^{-1}$. However, it can be shown that asymptotically the "variance" (it is not the true variance) of the mle attains the Cramer-Rao lower bound. In other words, for large samples, the "variance" of the mle is close to the Cramer-Rao bound. We will prove the result in the case that $\mathcal{L}_n$ is the log likelihood of independent, identically distributed random variables. The proof can be generalised to the case of non-identically distributed random variables.

We first state sufficient conditions for this to be true.

Assumption 2.6.1 *Suppose $\{X_i\}$ be iid random variables with density $f(X;\theta)$.*

(i) The conditions in Assumption 1.3.1 hold. In particular:

(a)

$$\mathrm{E}_{\theta_0}\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0}\right) = \int \frac{\partial f(x;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0}dx = \frac{\partial}{\partial\theta}\int \frac{\partial f(x;\theta)}{\partial\theta}dx\rfloor_{\theta=\theta_0} = 0.$$

(b)

$$\mathrm{E}_{\theta_0}\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta=\theta_0}\right)^2\right] = \mathrm{E}_{\theta_0}\left[-\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\rfloor_{\theta=\theta_0}\right].$$

(ii) Almost sure uniform convergence of the likelihood:

$\sup_{\theta\in\Theta}\frac{1}{n}|\mathcal{L}_n(\underline{X};\theta) - \mathrm{E}(\mathcal{L}_n(\underline{X};\theta))| \xrightarrow{a.s.} 0$ *as* $n\to\infty$.

We mention that directly verifying uniform convergence can be difficult. However, it can be established by showing that the parameter space is compact, point wise convergence of the likelihood to its expectation and almost sure equicontinuity in probability.

(iii) Model identifiability:

For every $\theta\in\Theta$, there does not exist another $\tilde{\theta}\in\Theta$ such that $f(x;\theta) = f(x;\tilde{\theta})$ for all x in the sample space.

(iv) Almost sure uniform convergence of the second derivative of the likelihood (using the notation ∇_θ): $\sup_{\theta\in\Theta}\frac{1}{n}|\nabla^2_\theta\mathcal{L}_n(\underline{X};\theta) - \mathrm{E}(\nabla^2_\theta\mathcal{L}_n(\underline{X};\theta))| \xrightarrow{a.s.} 0$ *as* $n\to\infty$.

This can be verified by using the same method described in (ii).

We require Assumption 2.6.1(ii,iii) to show consistency and Assumptions 1.3.1 and 2.6.1(iii-iv) to show asymptotic normality.

Theorem 2.6.1 *Supppose Assumption 2.6.1(ii,iii) holds. Let θ_0 be the true parameter and $\widehat{\theta}_n$ be the mle. Then we have $\widehat{\theta}_n \xrightarrow{a.s.} \theta_0$ (consistency).*

PROOF. First define $\ell(\theta) = \mathrm{E}[\log f(X;\theta)]$ (the limit of the expected log-likelihood). To prove the result we first need to show that the expectation of the maximum likelihood is

maximum at the true parameter and that this is the unique maximum. In other words we need to show that $\mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta) - \frac{1}{n}\mathcal{L}_n(\underline{X};\theta_0)) \leq 0$ for all $\theta \in \Theta$. To do this, we have

$$\begin{aligned}\ell(\theta) - \ell(\theta_0) = \mathrm{E}\left(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta) - \mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta_0)\right) &= \int \log \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx \\ &= \mathrm{E}\left(\log \frac{f(X;\theta)}{f(X;\theta_0)}\right).\end{aligned}$$

Now by using Jensen's inequality (since log is a concave function) we have

$$\mathrm{E}\left(\log \frac{f(X;\theta)}{f(X;\theta_0)}\right) \leq \log \mathrm{E}\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right) = \log \int \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx = \log \int f(x;\theta)dx = 0,$$

since $\theta \in \Theta$ and $\int f(x;\theta)dx = 1$. Thus giving $\mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta)) - \mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta_0)) \leq 0$. To prove that $\mathrm{E}(\frac{1}{n}[\mathcal{L}_n(\underline{X};\theta)) - \mathcal{L}_n(\underline{X};\theta_0)]) = 0$ only when θ_0, we use the identifiability condition in Assumption 2.6.1(iii), which means that $f(x;\theta) = f(x;\theta_0)$ for all x *only* when θ_0 and no other function of f gives equality. Hence only when $\theta = \theta_0$ do we have

$$\mathrm{E}\left(\log \frac{f(X;\theta)}{f(X;\theta_0)}\right) = \log \int \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx = \log \int f(x;\theta)dx = 0,$$

thus $\mathrm{E}(\frac{1}{n}\mathcal{L}_n(\underline{X};\theta))$ has a unique maximum at θ_0.

Finally, we need to show that $\widehat{\theta}_n \stackrel{\text{a.s.}}{\to} \theta_0$. To simplify notation for the remainder of this proof we assume the likelihood has been standardized by n i.e.

$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta).$$

We note that since $\ell(\theta)$ is maximum at θ_0 if $|\mathcal{L}_n(\widehat{\theta}_n) - \ell(\theta_0)| \stackrel{\text{a.s.}}{\to} 0$, then $\widehat{\theta}_n \stackrel{\text{a.s.}}{\to} \theta_0$. Thus we need only prove $|\mathcal{L}_n(\widehat{\theta}_n) - \ell(\theta_0)| \stackrel{\text{a.s.}}{\to} 0$. We do this using a sandwich argument.

First we note for every mle $\widehat{\theta}_n$

$$\mathcal{L}_n(\underline{X};\theta_0) \leq \mathcal{L}_n(\underline{X};\widehat{\theta}_n) \stackrel{\text{a.s.}}{\to} \ell(\widehat{\theta}_n) \leq \ell(\theta_0), \tag{2.11}$$

where we are treating $\widehat{\theta}_n$ as if it were a non-random fixed value in Θ. Returning to $|\mathrm{E}(\mathcal{L}_n(\underline{X};\theta_0)) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)|$ (they swapped round) we note that the difference can be written as

$$\ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n) = \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\} + \left\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\right\} + \left\{\mathcal{L}_n(\underline{X};\theta_0) - \ell(\widehat{\theta}_n)\right\}.$$

Now by using (2.11) we have

$$\begin{aligned}\ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n) &\leq \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\} + \Big\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\Big\} + \left\{\underbrace{\mathcal{L}_n(\underline{X};\widehat{\theta}_n)}_{\geq \mathcal{L}_n(\underline{X};\theta_0)} - \ell(\widehat{\theta}_n)\right\} \\ &= \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\}\end{aligned}$$

and

$$\begin{aligned}\ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n) &\geq \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\} + \Big\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\Big\} + \left\{\mathcal{L}_n(\underline{X};\theta_0) - \underbrace{\ell(\theta_0)}_{\geq \ell(\widehat{\theta}_n)}\right\} \\ &= \Big\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\Big\}\end{aligned}$$

Thus

$$\Big\{\ell(\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)\Big\} \leq \ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n) \leq \{\ell(\theta_0) - \mathcal{L}_n(\underline{X};\theta_0)\}.$$

The above also immediately follows from (2.11). This is easily seen in Figure 2.3, which Reza suggested. Thus we have sandwiched the difference $\mathrm{E}(\mathcal{L}_n(\underline{X};\theta_0)) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)$. Therefore, under Assumption 2.6.1(ii) we have

$$|\ell(\theta_0) - \mathcal{L}_n(\underline{X};\widehat{\theta}_n)| \leq \sup_{\theta\in\Theta} |\ell(\theta) - \mathcal{L}_n(\underline{X};\theta)| \stackrel{\text{a.s.}}{\to} 0.$$

Since $\mathrm{E}[\mathcal{L}_n(\underline{X};\theta)]$ has a unique maximum at $\mathrm{E}[\mathcal{L}_n(\underline{X};\theta_0)]$ this implies $\widehat{\theta}_n \stackrel{\text{a.s.}}{\to} \theta_0$.

□

Hence we have shown consistency of the mle. It is important to note that this proof is not confined to just the likelihood it can also be applied to other contrast functions. We now show asymptotic normality of the MLE.

Theorem 2.6.2 *Suppose Assumption 2.6.1 is satisfied (where θ_0 is the true parameter). Let*

$$I(\theta_0) = \mathrm{E}\left(\left[\frac{\partial \log f(X_i;\theta)}{\partial\theta}\rfloor_{\theta_0}\right]^2\right) = \mathrm{E}\left(-\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}\rfloor_{\theta_0}\right).$$

(i) Then the score statistic is

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial\theta}\rfloor_{\theta_0} \stackrel{\mathcal{D}}{\to} \mathcal{N}\Big(0, I(\theta_0)\Big). \tag{2.12}$$

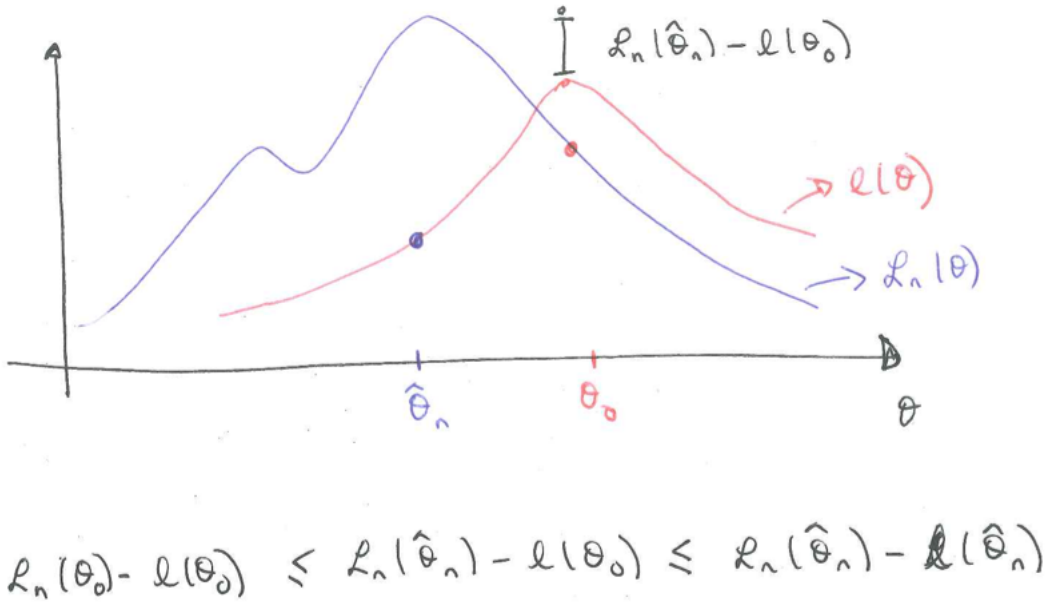

Figure 2.3: Difference between likelihood and expectation.

(ii) Then the mle is

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, I(\theta_0)^{-1}\Big).$$

(iii) The log likelihood ratio is

$$2\Big(\mathcal{L}_n(\underline{X}; \widehat{\theta}_n) - \mathcal{L}_n(\underline{X}; \theta_0)\Big) \xrightarrow{\mathcal{D}} \chi^2_p$$

(iv) The square MLE

$$n(\widehat{\theta}_n - \theta_0)' I_n(\theta_0)(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \chi^2_p.$$

PROOF. First we will prove (i). We recall because $\{X_i\}$ are iid random variables, then

$$\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\underline{X}; \theta)}{\partial \theta} \rfloor_{\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta)}{\partial \theta} \rfloor_{\theta_0},$$

is the sum of independent random variables. We note that under Assumption 2.6.1(i) we have

$$\mathrm{E}\left(\frac{\partial \log f(X_i; \theta)}{\partial \theta} \rfloor_{\theta_0}\right) = \int \frac{\partial \log f(x; \theta)}{\partial \theta} \rfloor_{\theta_0} f(x; \theta_0) dx = 0,$$

thus $\frac{\partial \log f(X_i;\theta)}{\partial \theta}\rfloor_{\theta_0}$ is a zero mean random variable and its variance is $I(\theta_0)$.

Hence $\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta_0}$ is the sum of iid random variables with mean zero and variance $I(\theta_0)$. Therefore, by the CLT for iid random variables we have (2.12).

We use (i) and Taylor (mean value) theorem to prove (ii). We first note that by the mean value theorem we have

$$\underbrace{\frac{1}{n}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\widehat{\theta}_n}}_{=0} = \frac{1}{n}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta_0} + (\widehat{\theta}_n - \theta_0)\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n}. \tag{2.13}$$

Using the consistency result in Theorem 2.6.1 ($\widehat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$, thus $\bar{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$) and Assumption 2.6.1(iv) we have

$$\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n} \xrightarrow{\text{a.s.}} \frac{1}{n}\mathrm{E}\left(\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\theta_0}\right) = \mathrm{E}\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\rfloor_{\theta_0}\right) = -I(\theta_0). \tag{2.14}$$

Substituting the above in (2.15) we have

$$\frac{1}{n}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta_0} - I(\theta_0)(\widehat{\theta}_n - \theta_0) + \underbrace{\left(\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n} - I(\theta_0)\right)(\widehat{\theta}_n - \theta_0)}_{\text{small}} = 0 \tag{2.15}$$

Multiplying the above by $\sqrt{n}$ and rearranging gives

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \;=\; I(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\underline{X};\theta)}{\partial \theta}\rfloor_{\theta_0} + o_p(1).$$

[6] Hence by substituting the (2.12) into the above we have (ii).

To prove (iii) we use (2.10), which we recall is

$$2\left(\mathcal{L}_n(\underline{X};\widehat{\theta}_n) - \mathcal{L}_n(\underline{X};\theta_0)\right) \approx (\widehat{\theta}_n - \theta_0)' n I(\theta_0)(\widehat{\theta}_n - \theta_0)'.$$

Now by using that $\sqrt{n}I(\theta_0)^{-1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I)$ (see (i)) and substituting this into the above gives (iii).

The proof of (iv) follows immediately from (ii). □

This result tells us that asymptotically the mle attains the Cramer-Rao bound. Furthermore, if $\widehat{\theta}$ is a p-dimension random vector and $I(\theta_0)$ is diagonal, then the elements of

[6] We mention that the proof above is for univariate $\frac{\partial^2 \mathcal{L}_n(\underline{X};\theta)}{\partial \theta^2}\rfloor_{\bar{\theta}_n}$, but by redo-ing the above steps pointwise it can easily be generalised to the multivariate case too

$\widehat{\theta}$ will be asymptotically independent (for example the sample mean and sample variance estimator for the normal distribution). However if $I(\theta_0)$ is not diagonal, then off-diagonal elements in $I(\theta_0)^{-1}$ measure the degree of correlation between the estimators. See Figure 2.4.

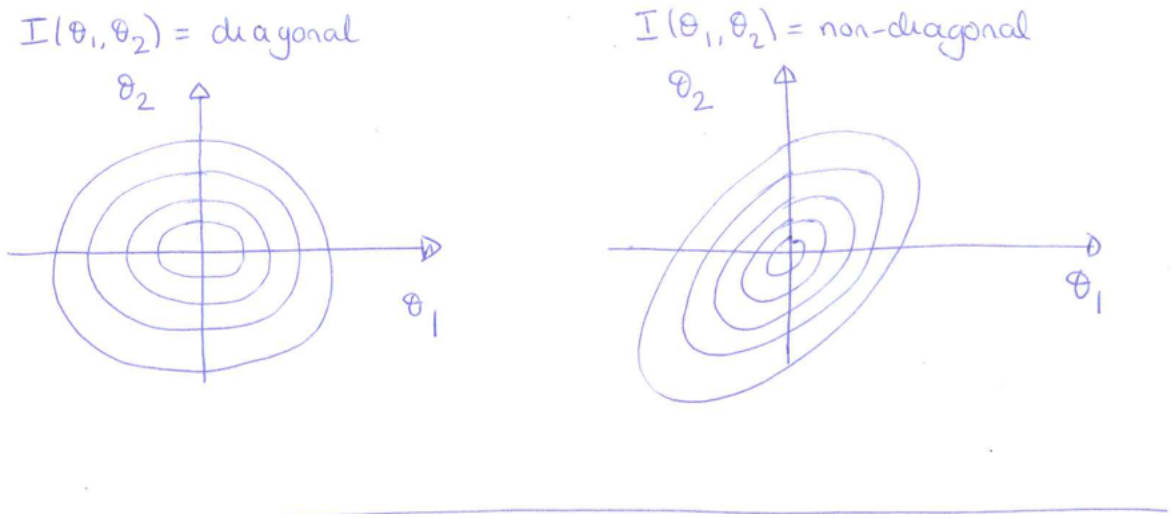

Figure 2.4: Contour plot of two dimensional normal distribution of two parameter estimators with diagonal and non-diagonal information matrix

Example 2.6.3 (The Weibull) *By using Example 2.6.2 we have*

$$\begin{pmatrix} \widehat{\theta}_n - \theta \\ \widehat{\alpha}_n - \alpha \end{pmatrix} \approx I_n(\theta, \alpha)^{-1} \begin{pmatrix} \sum_{i=1}^n \left(-\frac{\alpha}{\theta} + \frac{\alpha}{\theta^{\alpha+1}} Y_i^{\alpha} \right) \\ \sum_{i=1}^n \left(\frac{1}{\alpha} - \log Y_i - \log \theta - \frac{\alpha}{\theta} + \log(\frac{Y_i}{\theta}) \times (\frac{Y_i}{\theta})^{\alpha} \right) \end{pmatrix}.$$

Now we observe that RHS consists of a sum iid random variables (this can be viewed as an average). Since the variance of this exists (you can show that it is $I_n(\theta, \alpha)$), the CLT can be applied and we have that

$$\sqrt{n} \begin{pmatrix} \widehat{\theta}_n - \theta \\ \widehat{\alpha}_n - \alpha \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I(\theta, \alpha)^{-1}\right),$$

where $I(\theta, \alpha) = \mathrm{E}[(\nabla \log f(X; \theta, \alpha))^2]$.

Remark 2.6.3 *(i) We recall that for iid random variables that the Fisher information for sample size n is*

$$I_n(\theta_0) = \mathrm{E}\Big\{ \frac{\partial \log L_n(X;\theta)}{\partial \theta}\rfloor_{\theta_0} \Big\}^2 = n\mathrm{E}\Big(\frac{\partial \log f(X;\theta)}{\partial \theta}\rfloor_{\theta_0} \Big)^2 = nI(\theta_0)$$

Therefore since

$$\begin{aligned} (\widehat{\theta}_n - \theta_0) &\approx I_n(\theta_0)^{-1} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} = \Big[\frac{1}{n} I(\theta_0)\Big]^{-1} \frac{1}{n} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} \\ \Rightarrow \sqrt{n}(\widehat{\theta}_n - \theta_0) &\approx \sqrt{n} I_n(\theta_0)^{-1} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} = \Big[\frac{1}{n} I(\theta_0)\Big]^{-1} \frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0} \end{aligned}$$

and $\mathrm{var}\big(\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0}\big) = n^{-1}\mathrm{E}\Big[\big(\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0}\big)^2\Big] = I(\theta_0)$*, it can be seen that* $|\widehat{\theta}_n - \theta_0| = O_p(n^{-1/2})$*.*

(ii) Under suitable conditions a similar result holds true for data which is not iid.

(iii) These results only apply when θ_0 *lies* **inside** *the parameter space* Θ*.*

We have shown that under certain regularity conditions the mle will asymptotically attain the Fisher information bound. It is reasonable to ask how one can interprete this bound.

(i) Situation 1. $I_n(\theta_0) = \mathrm{E}\Big(- \frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\theta_0} \Big)$ is large (hence variance of the mle will be small) then it means that the gradient of $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}$ is large. Hence even for small deviations from θ_0, $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}$ is likely to be far from zero. This means the mle $\widehat{\theta}_n$ is likely to be in a close neighbourhood of θ_0.

(ii) Situation 2. $I_n(\theta_0) = \mathrm{E}\Big(-\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2}\rfloor_{\theta_0} \Big)$ is small (hence variance of the mle will large). In this case the gradient of the likelihood $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}$ is flatter and hence $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \approx 0$ for a large neighbourhood about the true parameter θ. Therefore the mle $\widehat{\theta}_n$ can lie in a large neighbourhood of θ_0.

Remark 2.6.4 (Lagrange Multipliers) *Often when maximising the likelihood it has to be done under certain constraints on the parameters. This is often achieved with the use of Lagrange multipliers (a dummy variable), which enforces this constraint.*

For example suppose the parameters in the likelihood must sum to one then we can enforce this constraint by maximising the criterion

$$\mathcal{L}_n(\theta, \lambda) = \underbrace{\mathcal{L}_n(\theta)}_{likelihood} + \lambda[\sum_{j=1}^{q} \theta_j - 1]$$

with respect to θ and *the dummy variable* λ.

2.7 Some questions

Exercise 2.7 *Suppose* $X_1, \ldots, X_n$ *are i.i.d. observations. A student wants to test whether each* X_i *has a distribution in the parametric family* $\{f(x;\alpha) : \alpha \in \Theta\}$ *or the family* $\{g(x;\beta) : \beta \in \Gamma\}$. *To do this he sets up the hypotheses*

$$H_0 : X_i \sim f(\cdot\,;\alpha_0) \quad vs. \quad H_A : X_i \sim g(\cdot\,;\beta_0),$$

where α_0 *and* β_0 *are the unknown true parameter values. He constructs the log-likelihood ratio statistic*

$$L = \max_{\beta\in\Gamma} \mathcal{L}_g(\boldsymbol{X};\beta) - \max_{\alpha\in\Theta} \mathcal{L}_f(\boldsymbol{X};\alpha) = \mathcal{L}_g(\boldsymbol{X};\hat{\beta}) - \mathcal{L}_f(\boldsymbol{X};\hat{\alpha}),$$

where

$$\mathcal{L}_g(\boldsymbol{X};\beta) = \sum_{i=1}^{n} \log g(X_i;\beta), \quad \mathcal{L}_f(\boldsymbol{X};\alpha) = \sum_{i=1}^{n} \log f(X_i;\alpha),$$

$\hat{\alpha} = \arg\max_{\alpha\in\Theta} \mathcal{L}_f(\boldsymbol{X};\alpha)$ *and* $\hat{\beta} = \arg\max_{\beta\in\Gamma} \mathcal{L}_g(\boldsymbol{X};\beta)$. *The student applies what he believe he learned in class to* L *and assumes that the distribution of* L *under the null hypothesis (asymptotically) follows a chi-squared distribution with one-degree of freedom. He does the test at the 5% level using the critical value* $\chi^2 = 3.84$, *rejecting the null in favor of the alternative if* $L > 3.84$.

(a) Using well known results, derive the asymptotic distribution of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \log \frac{g(X_i;\beta_0)}{f(X_i;\alpha_0)}$$

under the null and the alternative.

(b) *Is the distribution of L chi-squared? If not, derive the asymptotic distribution of L.*

Hint: You will need to use your answer from (a).

Note: This part is tough; but fun (do not be disillusioned if it takes time to solve).

(c) *By using your solution to parts (a) and (b), carefully explain what the actual type error I of the student's test will be (you do not need to derive the Type I error, but you should explain how it compares to the 5% level that the student uses).*

(d) *By using your solution to parts (a) and (b), carefully explain what the power of his test will be (you do not have to derive an equation for the power, but you should explain what happens to the sample size grows, giving a precise justification).*

(e) *Run some simulations to illustrate the above.*

Exercise 2.8 *Find applications where likelihoods are maximised with the use of Lagrange multipliers. Describe the model and where the Lagrange multiplier is used.*

2.8 Applications of the log-likelihood theory

We first summarise the results in the previous section (which will be useful in this section). For convenience, we will assume that $\{X_i\}_{i=1}^n$ are iid random variables, whose density is $f(x;\theta_0)$ (though it is relatively simple to see how this can be generalised to general likelihoods - of not necessarily iid rvs). Let us suppose that θ_0 is the true parameter that we wish to estimate. Based on Theorem 2.6.2 we have

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, I(\theta_0)^{-1}\bigg), \tag{2.16}$$

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n}{\partial \theta}\rfloor_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, I(\theta_0)\bigg) \tag{2.17}$$

and

$$2\left[\mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\theta_0)\right] \xrightarrow{\mathcal{D}} \chi^2_p, \tag{2.18}$$

where p are the number of parameters in the vector θ and $I(\theta_0) = \mathrm{E}[(\frac{\partial \log f(X;\theta)}{\partial \theta}\rfloor_{\theta_0})^2] = n^{-1}\mathrm{E}[(\frac{\partial \log \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta_0})^2]$. It is worth keeping in mind that by using the usual Taylor expansion the log-likelihood ratio statistic is asymptotically equivalent to

$$2\Big[\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}(\theta_0)\Big] \stackrel{\mathcal{D}}{=} -ZI(\theta_0)Z,$$

where $Z \sim \mathcal{N}(0, I(\theta_0))$.

Note: There are situations where the finite sampling distributions of the above are known, in which case there is no need to resort to the asymptotic sampling properties.

2.8.1 Constructing confidence sets using the likelihood

One the of main reasons that we show asymptotic normality of an estimator (it is usually not possible to derive normality for finite samples) is to construct confidence intervals/sets and to test.

In the case that θ_0 is a scaler (vector of dimension one), it is easy to use (2.16) to obtain

$$\sqrt{n}I(\theta_0)^{1/2}\big(\hat{\theta}_n - \theta_0\big) \stackrel{\mathcal{D}}{\to} N(0,1). \tag{2.19}$$

Based on the above the 95% CI for θ_0 is

$$\left[\widehat{\theta}_n - \frac{1}{\sqrt{n}}I(\theta_0)z_{\alpha/2}, \hat{\theta}_n + \frac{1}{\sqrt{n}}I(\theta_0)z_{\alpha/2}\right].$$

The above, of course, requires an estimate of the (standardised) expected Fisher information $I(\theta_0)$, typically we use the (standardised) observed Fisher information evaluated at the estimated value $\widehat{\theta}_n$.

The CI constructed above works well if θ is a scalar. But beyond dimension one, constructing a CI based on (2.16) (and the p-dimensional normal) is extremely difficult. More precisely, if θ_0 is a p-dimensional vector then the analogous version of (2.19) is

$$\sqrt{n}I(\theta_0)^{1/2}\big(\hat{\theta}_n - \theta_0\big) \stackrel{\mathcal{D}}{\to} N(0, I_p).$$

However, this does not lead to a simple set construction. One way to construct the confidence interval (or set) is to 'square' $\big(\hat{\theta}_n - \theta_0\big)$ and use

$$n\big(\hat{\theta}_n - \theta_0\big)'I(\theta_0)\big(\hat{\theta}_n - \theta_0\big) \stackrel{\mathcal{D}}{\to} \chi^2_p. \tag{2.20}$$

Based on the above a 95% CI is

$$\left\{\theta; \big(\hat{\theta}_n - \theta\big)' n \mathrm{E}\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\rfloor_{\theta_0}\right)^2 \big(\hat{\theta}_n - \theta\big) \leq \chi^2_p(0.95)\right\}. \tag{2.21}$$

Note that as in the scalar case, this leads to the interval with the smallest length. A disadvantage of (2.21) is that we have to (a) estimate the information matrix and (b) try to find all θ such the above holds. This can be quite unwieldy. An alternative method, which is asymptotically equivalent to the above but removes the need to estimate the information matrix is to use (2.18). By using (2.18), a $100(1-\alpha)\%$ confidence set for θ_0 is

$$\left\{\theta; 2\big(\mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\theta)\big) \leq \chi^2_p(1-\alpha)\right\}. \tag{2.22}$$

The above is not easy to calculate, but it is feasible.

Example 2.8.1 *In the case that θ_0 is a scalar the 95% CI based on (2.22) is*

$$\left\{\theta; \mathcal{L}_n(\theta) \geq \mathcal{L}_n(\hat{\theta}_n) - \frac{1}{2}\chi^2_1(0.95)\right\}.$$

See Figure 2.5 which gives the plot for the confidence interval (joint and disjoint).

Both the 95% confidence sets in (2.21) and (2.22) will be very close for relatively large sample sizes. However one advantage of using (2.22) instead of (2.21) is that it is easier to evaluate - no need to obtain the second derivative of the likelihood etc.

A feature which differentiates (2.21) and (2.22) is that the confidence sets based on (2.21) is symmetric about $\widehat{\theta}_n$ (recall that $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$) is symmetric about $\bar{X}$, whereas the symmetry condition may not hold for sample sizes when constructing a CI for θ_0 using (2.22). Using (2.22) there is no guarantee the confidence sets consist of only one interval (see Figure 2.5). However, if the distribution is exponential with full rank (and is steep) the likelihood will be concave with the maximum in the interior of the parameter space. This will mean the CI constructed using (2.22) will be connected.

If the dimension of θ is large it is difficult to evaluate the confidence set. Indeed for dimensions greater than three it is extremely hard. However in most cases, we are only interested in constructing confidence sets for certain parameters of interest, the other unknown parameters are simply nuisance parameters and confidence sets for them are not of interest. For example, for the normal family of distribution we may only be interested in constructing an interval for the mean, and the variance is simply a nuisance parameter.

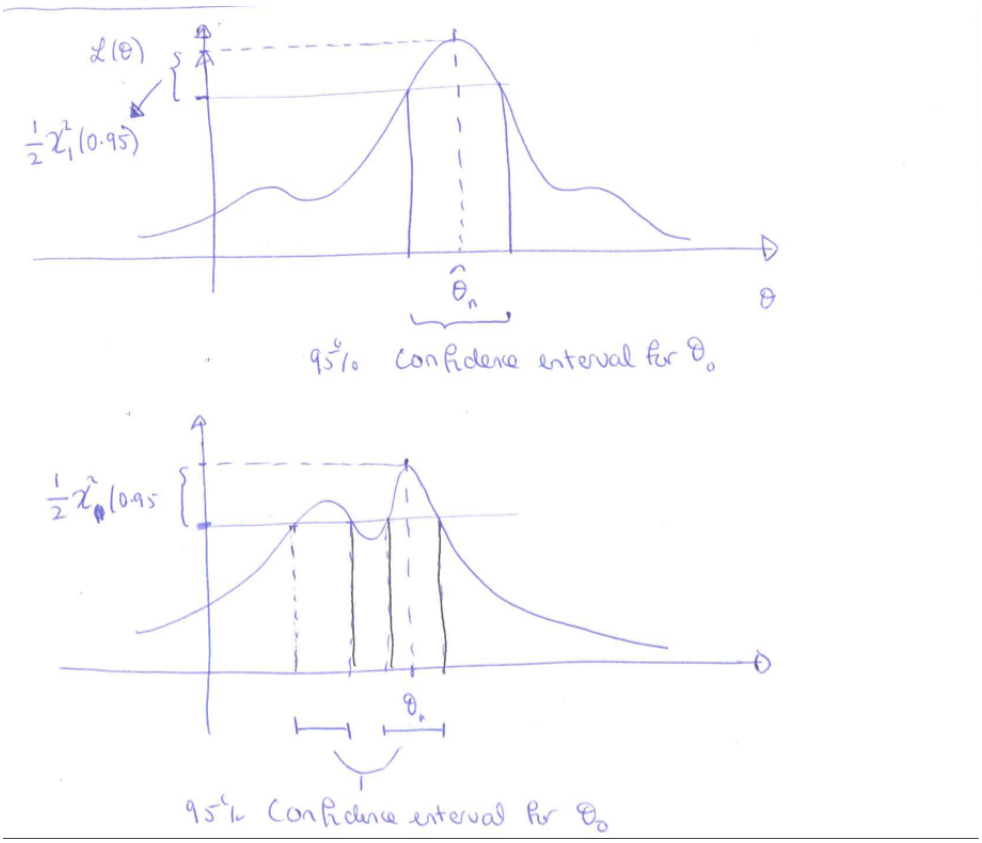

Figure 2.5: Constructing confidence intervals using method (2.22).

2.8.2 Testing using the likelihood

Let us suppose we wish to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_A : \theta \neq \theta_0$. We can use any of the results in (2.16), (2.17) and (2.18) to do the test - they will lead to slightly different p-values, but 'asympototically' they are all equivalent, because they are all based (essentially) on the same derivation.

We now list the three tests that one can use.

The Wald test

The Wald statistic is based on (2.16). We recall from (2.16) that if the null is true, then we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \left\{\mathrm{E}\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\rfloor_{\theta_0}\right)^2\right\}^{-1}\right).$$

Thus we can use as the test statistic

$$T_1 = \sqrt{n} I(\theta_0)^{1/2} \big(\widehat{\theta}_n - \theta_0\big)$$

to test the hypothesis. Under the null

$$T_1 \stackrel{\mathcal{D}}{\to} \mathcal{N}(0,1).$$

We now consider how the test statistics behaves under the alternative $H_A : \theta = \theta_1$. If the alternative were true, then we have

$$\begin{aligned} I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_0) &= I(\theta_0)^{1/2} \Big((\widehat{\theta}_n - \theta_1) + (\theta_1 - \theta_0) \Big) \\ &\approx I(\theta_0)^{1/2} I_n(\theta_1)^{-1} \sum_i \frac{\partial \log f(X_i;\theta_1)}{\partial \theta_1} + I(\theta_0)^{1/2}(\theta_1 - \theta_0) \end{aligned}$$

where $I_n(\theta_1) = \mathrm{E}_{\theta_1}[(\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} \rfloor_{\theta=\theta_1})^2]$.

Local alternatives and the power function

In the case that the alternative is *fixed* (does not change with sample size), it is clear that the power in the test goes to 100% as $n \to \infty$. To see we write

$$\begin{aligned} \sqrt{n} I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_0) &= \sqrt{n} I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_1) + \sqrt{n} I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\ &\approx I(\theta_0)^{1/2} I(\theta_1)^{-1} \frac{1}{\sqrt{n}} \sum_i \frac{\partial \log f(X_i;\theta_1)}{\partial \theta_1} + \sqrt{n} I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\ &\stackrel{\mathcal{D}}{\to} N\left(0, I(\theta_0)^{1/2} I(\theta_1)^{-1} I(\theta_0)^{1/2}\right) + \sqrt{n} I(\theta_0)^{1/2}(\theta_1 - \theta_0). \end{aligned}$$

Using the above calculation we see that

$$P\left(\text{Reject} | \theta = \theta_1\right) = 1 - \Phi\left(\frac{z_{1-\alpha/2} - \sqrt{n}(\theta_1 - \theta_0) I(\theta_0)^{1/2}}{\sqrt{I(\theta_0)^{1/2} I(\theta_1)^{-1} I(\theta_0)^{1/2}}} \right).$$

Thus, we see that as $n \to \infty$, the power gets closer to 100%. However, this calculation does not really tell us how the test performs for θ_1 close to the θ_0.

To check the effectiveness of a given testing method, one lets the alternative get *closer* to the the null as $n \to \infty$. This allows us to directly different statistical tests (and the factors which drive the power).

How to choose the closeness:

- Suppose that $\theta_1 = \theta_0 + \frac{\phi}{n}$ (for fixed ϕ), then the center of T_1 is

$$\begin{aligned}
\sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_0) &= \sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_1) + \sqrt{n}I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\
&\approx I(\theta_0)^{1/2}I(\theta_1)^{-1}\frac{1}{\sqrt{n}}\sum_i \frac{\partial \log f(X_i;\theta_1)}{\partial \theta_1} + \sqrt{n}(\theta_1 - \theta_0) \\
&\xrightarrow{\mathcal{D}} N\left(0, \underbrace{I(\theta_0)^{1/2}I(\theta_1)^{-1}I(\theta_0)^{1/2}}_{\to I}\right) + \underbrace{\frac{I(\theta_0)^{1/2}\phi}{\sqrt{n}}}_{\to 0} \approx N(0, I_p).
\end{aligned}$$

  Thus the alternative is too close to the null for us to discriminate between the null and alternative.

- Suppose that $\theta_1 = \theta_0 + \frac{\phi}{\sqrt{n}}$ (for fixed ϕ), then

$$\begin{aligned}
\sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_0) &= \sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta}_n - \theta_1) + \sqrt{n}I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\
&\approx I(\theta_0)^{1/2}I(\theta_1)^{-1}\frac{1}{\sqrt{n}}\sum_i \frac{\partial \log f(X_i;\theta_1)}{\partial \theta_1} + \sqrt{n}I(\theta_0)^{1/2}(\theta_1 - \theta_0) \\
&\xrightarrow{\mathcal{D}} N\left(0, I(\theta_0)^{1/2}I(\theta_1)^{-1}I(\theta_0)^{1/2}\right) + I(\theta_0)^{1/2}\phi \\
&\approx N\left(I(\theta_0)^{1/2}\phi, I(\theta_0)^{1/2}I(\theta_0 + \phi n^{-1/2})^{-1}I(\theta_0)^{1/2}\right).
\end{aligned}$$

  Therefore, for a given ϕ we can calculate the power at a given level α. Assume for simplicity that $\phi > 0$ and θ is univariate. Then

$$\begin{aligned}
P\left(|T_1| > z_{1-\alpha/2}\right) &\geq P\left(T_1 > z_{1-\alpha/2}\right) = P\left(Z > \frac{z_{1-\alpha/2} - I(\theta_0)^{1/2}\phi}{\sqrt{I(\theta_0)^{1/2}I(\theta_0 + \phi n^{-1/2})^{-1}I(\theta_0)^{1/2}}}\right) \\
&= 1 - \Phi\left(\frac{z_{1-\alpha/2} - I(\theta_0)^{1/2}\phi}{\sqrt{I(\theta_0)^{1/2}I(\theta_0 + \phi n^{-1/2})^{-1}I(\theta_0)^{1/2}}}\right) \\
&\approx 1 - \Phi\left(z_{1-\alpha/2} - \phi I(\theta_0)^{1/2}\right).
\end{aligned}$$

  this gives the power function of the test for a fixed n over ϕ. What we observe is that the power of the test $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$ depends on the size of $I(\theta_0)^{1/2}$. The larger the Fisher information $I(\theta_0)$ the greater the ability of the Wald test to discriminate between the null and the alternative. Based on what we understand about the Fisher information this make sense. The larger the Fisher information the "better" our ability to estimate the true parameter.

In the case that the dimension of θ is $p > 1$, we use the test statistic $\tilde{n}_1 = \big(\hat{\theta}_n - \theta_0\big)\sqrt{n}\mathrm{E}\bigg(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\big(\hat{\theta}_n - \theta_0\big)$ instead of T_1. Noting that the distribution of T_1 is a chi-squared with p-degrees of freedom.

The Score test

The score test is based on the score. Under the null the distribution of the score is

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n}{\partial\theta}\rfloor_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, \bigg\{\mathrm{E}\bigg(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\bigg\}\bigg).$$

Thus we use as the test statistic

$$T_2 = \frac{1}{\sqrt{n}}\bigg\{\mathrm{E}\bigg(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\bigg\}^{-1/2}\frac{\partial \mathcal{L}_n}{\partial\theta}\rfloor_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

An advantage of this test is that the maximum likelihood estimator (under either the null or alternative) does not have to be calculated.

The log-likelihood ratio test

This test is based on (2.18), and the test statistic is

$$T_3 = 2\big(\max_{\theta\in\Theta}\mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\big) \xrightarrow{\mathcal{D}} \chi^2_p.$$

T_3 is often called Wilk's statistic. An advantage of this test statistic is that it is asymptotically *pivotal*, in the sense that it does not depend on any nuisance parameters (we discuss this in the next chapter). However, using the chi-square distribution will only give the p-value corresponding to a "two-sided" hypothesis. This is because the chi-square distribution is based on the approximation

$$T_3 = 2\big(\mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\theta_0)\big) \approx n(\widehat{\theta}_n - \theta_0)^2 I(\theta_0),$$

which assumes that $\widehat{\theta}_n = \arg\max_\theta \mathcal{L}_n(\theta)$ and solves $\frac{\partial \mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}_n} = 0$. However, in a one-sided test $H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0$ parameter space is restricted to $\mu \geq \mu_0$ (it is on the boundary), this means that T_3 will not have a chi-square (though it will be very close) and p-value will be calculated in a slightly diifferent way. This boundary issue is not a problem for Wald test, since for the Wald test we simply calculate

$$P\left(Z \geq T_1\right) = \Phi\left(\sqrt{n}I(\theta_0)^{1/2}(\widehat{\theta} - \theta_0)\right).$$

Indeed, we show in Chapter 4, that the p-value for the one-sided test using the log-likelihood ratio statistic corresponds to that of p-value of the one-sided tests using the Wald statistic.

Exercise 2.9 *What do the score and log-likelihood ratio test statistics look like under the alternative? Derive the power function for these test statistics.*

You should observe that the power function for all three tests is the same.

Applications of the log-likeihood ratio to the multinomial distribution

We recall that the multinomial distribution is a generalisation of the binomial distribution. In this case at any given trial there can arise m different events (in the Binomial case $m = 2$). Let Z_i denote the outcome of the ith trial and assume $P(Z_i = k) = \pi_i$ ($\pi_1 + \ldots + \pi_m = 1$). Suppose that n trials are conducted and let Y_1 denote the number of times event 1 arises, Y_2 denote the number of times event 2 arises and so on. Then it is straightforward to show that

$$P(Y_1 = k_1, \ldots, Y_m = k_m) = \binom{n}{k_1, \ldots, k_m} \prod_{i=1}^{m} \pi_i^{k_i}.$$

If we do not impose any constraints on the probabilities $\{\pi_i\}$, given $\{Y_i\}_{i=1}^m$ it is straightforward to derive the mle of $\{\pi_i\}$ (it is very intuitive too!). Noting that $\pi_m = 1 - \sum_{i=1}^{m-1} \pi_i$, the log-likelihood of the multinomial is proportional to

$$\mathcal{L}_n(\underline{\pi}) \;=\; \sum_{i=1}^{m-1} y_i \log \pi_i + y_m \log(1 - \sum_{i=1}^{m-1} \pi_i).$$

Differentiating the above with respect to π_i and solving gives the mle estimator $\hat{\pi}_i = Y_i/n$! We observe that though there are m probabilities to estimate due to the constraint $\pi_m = 1 - \sum_{i=1}^{m-1} \pi_i$, we only have to estimate $(m-1)$ probabilities. We mention, that the same estimators can also be obtained by using Lagrange multipliers, that is maximising $\mathcal{L}_n(\underline{\pi})$ subject to the parameter constraint that $\sum_{j=1}^{p} \pi_i = 1$. To enforce this constraint, we normally add an additional term to $\mathcal{L}_n(\underline{\pi})$ and include the dummy variable λ. That is we define the constrained likelihood

$$\tilde{\mathcal{L}}_n(\underline{\pi}, \lambda) \;=\; \sum_{i=1}^{m} y_i \log \pi_i + \lambda(\sum_{i=1}^{m} \pi_i - 1).$$

Now if we maximise $\tilde{\mathcal{L}}_n(\underline{\pi}, \lambda)$ with respect to $\{\pi_i\}_{i=1}^m$ and λ we will obtain the estimators $\hat{\pi}_i = Y_i/n$ (which is the same as the maximum of $\mathcal{L}_n(\underline{\pi})$).

To derive the limiting distribution we note that the second derivative is

$$-\frac{\partial^2 \mathcal{L}_n(\underline{\pi})}{\partial \pi_i \pi_j} = \begin{cases} \frac{y_i}{\pi_i^2} + \frac{y_m}{(1-\sum_{r=1}^{m-1}\pi_r)^2} & i = j \\ \frac{y_m}{(1-\sum_{r=1}^{m-1}\pi_r)^2} & i \neq j \end{cases}$$

Hence taking expectations of the above the information matrix is the $(k-1) \times (k-1)$ matrix

$$I(\pi) = n \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_m} & \frac{1}{\pi_m} & \cdots & \frac{1}{\pi_m} \\ \frac{1}{\pi_m} & \frac{1}{\pi_2} + \frac{1}{\pi_m} & \cdots & \frac{1}{\pi_m} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\pi_{m-1}} & \cdots & \frac{1}{\pi_{m-1}} + \frac{1}{\pi_m} & \end{pmatrix}.$$

Provided no of π_i is equal to either 0 or 1 (which would drop the dimension of m and make $I(\pi)$) singular, then the asymptotic distribution of the mle the normal with variance $I(\pi)^{-1}$.

Sometimes the probabilities $\{\pi_i\}$ will not be 'free' and will be determined by a parameter θ (where θ is an r-dimensional vector where $r < m$), ie. $\pi_i = \pi_i(\theta)$, in this case the likelihood of the multinomial is

$$\mathcal{L}_n(\underline{\pi}) = \sum_{i=1}^{m-1} y_i \log \pi_i(\theta) + y_m \log(1 - \sum_{i=1}^{m-1} \pi_i(\theta)).$$

By differentiating the above with respect to θ and solving we obtain the mle.

Pearson's goodness of Fit test

We now derive Pearson's goodness of Fit test using the log-likelihood ratio.

Suppose the null is $H_0 : \pi_1 = \tilde{\pi}_1, \ldots, \pi_m = \tilde{\pi}_m$ (where $\{\tilde{\pi}_i\}$ are some pre-set probabilities) and H_A : the probabilities are not the given probabilities. Hence we are testing restricted model (where we do not have to estimate anything) against the full model where we estimate the probabilities using $\pi_i = Y_i/n$.

The log-likelihood ratio in this case is

$$W = 2\big\{ \arg\max_{\pi} \mathcal{L}_n(\pi) - \mathcal{L}_n(\tilde{\pi}) \big\}.$$

Under the null we know that $W = 2\big\{ \arg\max_\pi \mathcal{L}_n(\pi) - \mathcal{L}_n(\tilde{\pi}) \big\} \xrightarrow{\mathcal{D}} \chi^2_{m-1}$ (because we have to estimate $(m-1)$ parameters). We now derive an expression for W and show that the Pearson-statistic is an approximation of this.

$$\begin{aligned} \frac{1}{2}W &= \sum_{i=1}^{m-1} Y_i \log\big(\frac{Y_i}{n}\big) + Y_m \log\frac{Y_m}{n} - \sum_{i=1}^{m-1} Y_i \log\tilde{\pi}_i - Y_m \log\tilde{\pi}_m \\ &= \sum_{i=1}^{m} Y_i \log\big(\frac{Y_i}{n\tilde{\pi}_i}\big). \end{aligned}$$

Recall that Y_i is often called the observed $Y_i = O_i$ and $n\tilde{\pi}_i$ the expected under the null $E_i = n\tilde{\pi}_i$. Then $W = 2\sum_{i=1}^m O_i \log\big(\frac{O_i}{E_i}\big) \xrightarrow{\mathcal{P}} \chi^2_{m-1}$. By making a Taylor expansion of $x\log(xa^{-1})$ about $x = a$ we have $x\log(xa^{-1}) \approx a\log(aa^{-1}) + (x-a) + \frac{1}{2}(x-a)^2/a$. We let $O = x$ and $E = a$, then assuming the null is true and $E_i \approx O_i$ we have

$$W = 2\sum_{i=1}^{m} Y_i \log\big(\frac{Y_i}{n\tilde{\pi}_i}\big) \approx 2\sum_{i=1}^{m} \big((O_i - E_i) + \frac{1}{2}\frac{(O_i-E_i)^2}{E_i}\big).$$

Now we note that $\sum_{i=1}^m E_i = \sum_{i=1}^m O_i = n$ hence the above reduces to

$$W \approx \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{\mathcal{D}} \chi^2_{m-1}.$$

We recall that the above is the Pearson test statistic. Hence this is one methods for deriving the Pearson chi-squared test for goodness of fit.

Remark 2.8.1 (Beyond likelihood) *In several applications in statistics we cannot articulate the question of interest in terms of the parameters of a distribution. However, we can often articulate it in terms of some parameters, ϕ. Indeed, whether ϕ is zero or will tell us something about the data. For example:*

(i) Are some parameters in a linear regression zero?

(ii) Is there correlation between two variables?

(iii) Is there an interaction between two categorical variables in a regression?

(iv) In my own area of research on detecting nonstationarities, transforming the time series can yield more information then the original data. For example, nonstationarities imply correlations in the transformed data. The list goes on.

None of the above requires us to place distributional assumptions on the data. However, we can still test $H_0 : \phi = 0$ against $H_A : \phi \neq 0$. If we can estimate this quantity and obtain its limiting distribution under the null and show that under the alternative it "shifts", using $\widehat{\phi}$ we can construct a test statistic which has some power (though it may not be the most powerful test).

2.9 Some questions

Exercise 2.10 *A parameterisation of a distribution is identifiably if there does not exist another set of parameters which can give the same distribution.* `https://en.wikipedia.org/wiki/Identifiability`. *Recall this assumption was used when deriving the sampling properties of the maximum likelihood estimator.*

Suppose X_i are iid random variables which come from a mixture of distributions. The density of X_i is

$$f(x; \pi, \lambda_1, \lambda_2) = \pi\lambda_1 \exp(-\lambda_1 x) + (1 - \pi)\lambda_2 \exp(-\lambda_2 x)$$

where $x > 0$, $\lambda_1, \lambda_2 > 0$ and $0 \leq \pi \leq 1$.

(i) Are the parameters identifiable?

(ii) Does standard theory apply when using the log-likelihood ratio test to test $H_0 : \pi = 0$ vs $H_A : \pi \neq 0$.

(iii) Does standard theory apply when using the log-likelihood to estimate π when $\lambda_1 = \lambda_2$.