Chapter 1

The Likelihood

In this chapter we review some results that you may have came across previously. We define the likelihood and construct the likelihood in slightly non-standard situations. We derive properties associated with the likelihood, such as the Crámer-Rao bound and sufficiency. Finally we review properties of the exponential family which are an important parametric class of distributions with some elegant properties.

1.1 The likelihood function

Suppose $\underline{x} = \{X_i\}$ is a realized version of the random vector $\underline{X} = \{X_i\}$. Suppose the density f is unknown, however, it is known that the true density belongs to the density class $\mathcal{F}$. For each density in $\mathcal{F}$, $f_{\underline{X}}(\underline{x})$ specifies how the density changes over the sample space of $\underline{X}$. Regions in the sample space where $f_{\underline{X}}(\underline{x})$ is "large" point to events which are more likely than regions where $f_{\underline{X}}(\underline{x})$ is "small". However, we have in our hand $\underline{x}$ and our objective is to determine which distribution the observation $\underline{x}$ may have come from. In this case, it is useful to turn the story around. For a given realisation $\underline{x}$ and each $f \in \mathcal{F}$ one evaluates $f_{\underline{X}}(\underline{x})$. This "measures" the *likelihood* of a particular density in $\mathcal{F}$ based on a realisation $\underline{x}$. The term likelihood was first coined by Fisher.

In most applications, we restrict the class of densities $\mathcal{F}$ to a "parametric" class. That is $\mathcal{F} = \{f(\underline{x};\theta); \theta \in \Theta\}$, where the form of the density $f(\underline{x};\cdot)$ is known but the finite dimensional parameter θ is unknown. Since the aim is to make decisions about θ based on a realisation $\underline{x}$ we often write $L(\theta;\underline{x}) = f(\underline{x};\theta)$ which we call the *likelihood*. For convenience, we will often work with the log-likelihood $\mathcal{L}(\theta;\underline{x}) = \log f(\underline{x};\theta)$. Since the logarithm is a monotonic transform the maximum of the likelihood and log-likelihood will be the same. This preservation of maximum is very important.

Let us consider the simplest case that $\{X_i\}$ are iid random variables with probability function (or probability density function) $f(x;\theta)$, where f is known but the parameter θ is unknown. The likelihood function of θ based on $\{X_i\}$ is

$$L(\theta;\underline{X}) = \prod_{i=1}^{n} f(X_i;\theta) \qquad (1.1)$$

and the log-likelihood turns product into sum

$$\log L(\theta;\underline{X}) = \mathcal{L}(\theta;\underline{X}) = \sum_{i=1}^{n} \log f(X_i;\theta). \qquad (1.2)$$

We now consider some simple examples.

Example 1.1.1 *(i) Suppose that $\{X_i\}$ are iid normal random variables with mean μ and variance σ^2 the log likelihood is*

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

Observe that the parameters and random variables are "separable".

(ii) Suppose that $\{X_i\}$ are iid binomial random variables $X_i \sim$Bin(m,π). We assume m is known, then the log likelihood for π is

$$\begin{aligned}\mathcal{L}_n(\pi;\underline{X}) &= \sum_{i=1}^{n}\log\binom{m}{X_i} + \sum_{i=1}^{n}\Big(X_i\log\pi + (m-X_i)\log(1-\pi)\Big)\\ &= \sum_{i=1}^{n}\log\binom{m}{X_i} + \sum_{i=1}^{n}\Big(X_i\log\big(\frac{\pi}{1-\pi}\big) + m\log(1-\pi)\Big).\end{aligned}$$

Observe that the parameters and random variables are "separable".

(iii) Suppose that $\{X_i\}$ are independent random variables which give the number of "successes" out of m_i. It seems reasonable to model $X_i \sim$Bin(m_i,π_i). It is believed that the the regressors z_i influence the chance of success π_i. We try to model this influence with the nonlinear transform

$$\pi_i = g(e^{\beta' z_i}) = \frac{e^{\beta' z_i}}{1+e^{\beta' z_i}},$$

where β are the unknown parameters of interest. Then the log likelihood is

$$\mathcal{L}_n(\beta;\underline{X}) = \sum_{i=1}^{n}\log\binom{m_i}{X_i} + \sum_{i=1}^{n}\left(X_i\log\left(\frac{g(\beta'z_i)}{1-g(\beta'z_i)}\right) + m_i\log(1-g(\beta'z_i))\right).$$

(iv) *Modelling categorical data in a contingency table. Suppose a continency table contains C cells, where each cell gives the number for the corresponding event. Let $1 \leq \ell \leq C$, at each "trial" probability of being placed in cell ℓ is π_ℓ. If we do not make any assumptions on the probabilities (except that each trial are iid random variables) then we model the number of counts in each cell using a multinomial distribution. Suppose the total number of counts is n and the number of counts observed in cell ℓ is X_ℓ, then the distribution is $P(X_1 = x_1, \ldots, X_C = x_c) = \binom{n}{x_1,\ldots,x_C}\pi_1^{x_1}\ldots\pi_C^{x_C}$, which has the log-likelihood*

$$\mathcal{L}_n(\pi_1,\pi_2,\ldots,\pi_{C-1};X_1,\ldots,X_C) = \log\binom{n}{X_1,\ldots,X_C} + \sum_{i=1}^{C}X_i\log\pi_i$$
$$= \log\binom{n}{X_1,\ldots,X_C} + \sum_{i=1}^{C-1}X_i\log\frac{\pi_i}{1-\sum_{j=1}^{C-1}\pi_j} + n\log(1-\sum_{j=1}^{C-1}\pi_j).$$

Observe that the parameters and random variables are "separable".

(v) *Suppose X is a random variable that only takes integer values, however, there is no upper bound on the number of counts. When there is no upper bound on the number of counts, the Poisson distribution is often used as an alternative to the Binomial. If X follows a Poisson distribution then $P(X = k) = \lambda^k\exp(-\lambda)/k!$. The log-likelihood for the iid Poisson random variables $\{X_i\}$ is*

$$\mathcal{L}(\lambda;\underline{X}) = \sum_{i=1}^{n}\left(X_i\log\lambda - \lambda - \log X_i!\right).$$

Observe that the parameters and random variables are "separable".

(vi) *Suppose that $\{X_i\}$ are independent exponential random variables which have the density $\theta^{-1}\exp(-x/\theta)$. The log-likelihood is*

$$\mathcal{L}_n(\theta;\underline{X}) = \sum_{i=1}^{n}\left(-\log\theta - \frac{X_i}{\theta}\right).$$

(vii) A generalisation of the exponential distribution which gives more flexibility in terms of shape of the distribution is the Weibull. Suppose that $\{X_i\}$ are independent Weibull random variables which have the density $\frac{\alpha x^{\alpha-1}}{\theta^{\alpha}}\exp(-(x/\theta)^{\alpha})$ where $\theta, \alpha > 0$ (in the case that $\alpha = 0$ we have the regular exponential) and x is defined over the positive real line. The log-likelihood is

$$\mathcal{L}_n(\alpha, \theta; \underline{X}) = \sum_{i=1}^{n} \left(\log \alpha + (\alpha - 1)\log X_i - \alpha \log \theta - \left(\frac{X_i}{\theta}\right)^{\alpha} \right).$$

Observe that the parameters and random variables are not *"separable". In the case, that α is known, but θ is unknown the likelihood is proportional to*

$$\mathcal{L}_n(\theta; \underline{X};) \propto \sum_{i=1}^{n} \left(-\alpha \log \theta - \left(\frac{X_i}{\theta}\right)^{\alpha} \right),$$

observe the other terms in the distribution are fixed and do not vary, so are omitted. If α is known, the unknown parameter and random variables are "separable".

Often I will exchange $\mathcal{L}(\theta; \underline{X}) = \mathcal{L}(\underline{X}; \theta)$, but they are the *same*.

Look closely at the log-likelihood of iid random variables, what does its average

$$\frac{1}{n}\mathcal{L}(\underline{X}; \theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i; \theta) \tag{1.3}$$

converge to as $n \to \infty$?

1.2 Constructing likelihoods

Constructing the likelihood for the examples given in the previous section was straightforward. However, in many real situations, half the battle is finding the correct distribution and likelihood.

Many of the examples we consider below depend on using a dummy/indicator variable that we treat as a Bernoulli random variables. We recall if δ is a Bernoulli random variable that can take either 0 or 1, where $P(\delta = 1) = \pi$ and $P(\delta = 0) = 1 - \pi$, then $P(\delta = x) = (1 - \pi)^{1-x}\pi^{x}$. We observe that the log-likelihood for π given δ is $(1 - \delta)\log(1 - \pi) + \delta \log \pi$. Observe after the log transform, that the random variable and the parameter of interest are "separable".

Mixtures of distributions

Suppose Y is a mixture of two subpopulations, with densities $f_0(x;\theta)$ and $f_1(x;\theta)$ respectively. The probability of belonging to density 0 is $1-p$ and probability of belonging to density 1 is p. Based this information, we can represent the random variable $Y = \delta U + (1-\delta)V$, where U, V, δ are independent random variables; U has density f_1, V has density f_0 and $P(\delta = 1) = p$ and $P(\delta = 0) = 1-p$. The density of Y is

$$f_Y(x;\theta) = f_Y(x|\delta = 0, \theta)P(\delta = 0) + f_Y(x|\delta = 1, \theta)P(\delta = 1) = (1-p)f_0(x;\theta) + pf_1(x;\theta).$$

Thus the log likelihood of θ and p given $\{Y_i\}$ is

$$\mathcal{L}(\{Y_i\};\theta,p) = \sum_{i=1}^{n} \log\left[(1-p)f_0(Y_i;\theta) + pf_1(Y_i;\theta)\right].$$

Observe that the random variables and parameters of interest are not separable.

Suppose we not only observe Y but we observe the mixture the individual belongs to. Not only do we have more information about our parameters, but also estimation becomes easier. To obtain the joint likelihood, we require the joint distribution of (Y, δ), which is a mixture of density and point mass. To derive this we note that by using limiting arguments

$$\begin{aligned}
\lim_{\epsilon\to 0} \frac{P(Y \in [y-\epsilon/2, y+\epsilon/2], \delta = x;\theta,p)}{\epsilon} &= \lim_{\epsilon\to 0} \frac{F_x(y+\epsilon/2) - F_x(y-\epsilon/2)}{\epsilon} P(\delta = x; p) \\
&= f_x(y;\theta)P(\delta = x;p) \\
&= f_1(y;\theta)^x f_0(y;\theta)^{1-x} p^x (1-p)^{1-x}.
\end{aligned}$$

Thus the log-likelihood of θ and p given the joint observations $\{Y_i, \delta_i\}$ is

$$\mathcal{L}(Y_i,\delta_i;\theta,p) = \sum_{i=1}^{n} \left\{\delta_i \log f_1(Y_i;\theta) + (1-\delta_i)\log f_0(Y_i;\theta) + \delta_i \log p + (1-\delta_i)\log(1-p)\right\}. \quad (1.4)$$

The parameters and random variables are separable in this likelihood.

Of course in reality, we do not observe δ_i, but we can predict it, by conditioning on what is observed Y_i. This is effectively constructing the expected log-likelihood of $\{Y_i, \delta_i\}$ conditioned on $\{Y_i\}$. This is not a log-likelihood per se. But for reasons that will become clear later in the course, in certain situations it is useful to derive the expected log-likelihood when conditioned on random variables of interest. We now construct the expected log-likelihood of $\{Y_i, \delta_i\}$ conditioned on $\{Y_i\}$. Using (1.4) and that $\{Y_i, \delta_i\}$ are independent over i we have

$$\mathrm{E}[\mathcal{L}(Y_i, \delta_i; \theta, p)|\{Y_i\}] = \sum_{i=1}^{n} \{\mathrm{E}[\delta_i|Y_i, \theta, p] \left(\log f_1(Y_i; \theta) + \log p\right) + \mathrm{E}[(1-\delta_i)|Y_i] \left(\log f_0(Y_i; \theta) + \log(1-p)\right)\}$$

$\mathrm{E}[\delta_i|Y_i, \theta, p] = P[\delta_i = 1|Y_i, \theta, p]$, hence it measures the probability of the mixture 1 being chosen when Y_i is observed and is

$$P[\delta_i = 1|Y_i, \theta, p] = \frac{P[\delta_i = 1, Y_i, \theta, p]}{P[Y_i, \theta, p]} = \frac{P[Y_i|\delta_i = 1, \theta, p]P(\delta_i = 1, \theta, p)}{P[Y_i, \theta, p]} = \frac{pf_1(Y_i; \theta)}{pf_1(Y_i; \theta) + (1-p)f_0(Y_i; \theta)}.$$

Similarly

$$P[\delta_i = 0|Y_i, \theta, p] = \frac{(1-p)f_0(Y_i; \theta)}{pf_1(Y_i; \theta) + (1-p)f_0(Y_i; \theta)}.$$

Substituting these in the the above gives the expected log-likelihood conditioned on $\{Y_i\}$;

$$\begin{aligned} \mathrm{E}[\mathcal{L}(Y_i, \delta_i; \theta, p)|\{Y_i\}] &= \sum_{i=1}^{n} \Bigg\{ \left(\frac{pf_1(Y_i; \theta)}{pf_1(Y_i; \theta) + (1-p)f_0(Y_i; \theta)}\right) \left(\log f_1(Y_i; \theta) + \log p\right) + \\ &\quad \left(\frac{(1-p)f_0(Y_i; \theta)}{pf_1(Y_i; \theta) + (1-p)f_0(Y_i; \theta)}\right) \left(\log f_0(Y_i; \theta) + \log(1-p)\right) \Bigg\}. \end{aligned}$$

Observe that this is not in terms of δ_i.

The censored exponential distribution

Suppose $X \sim Exp(\theta)$ (density of X is $f(x; \theta) = \theta \exp(-x\theta)$), however X is censored at a known point c and Y is observed where

$$Y = \begin{cases} X & X \leq c \\ c & X > c \end{cases} \tag{1.5}$$

It is known if an observation is censored. We define the censoring variable

$$\delta = \begin{cases} 0 & X \leq c \\ 1 & X > c \end{cases}$$

The only unknown is θ and we observe (Y, δ). Note that δ is a Bernoulli variable (Binomial with $n = 1$) with $P(\delta = 0) = 1 - \exp(-c\theta)$ and $P(\delta = 1) = \exp(-c\theta)$. Thus the likelihood of θ based only δ is $L(\delta; \theta) = (1-\pi)^{1-\delta}\pi^{\delta} = (1-e^{-c\theta})^{1-\delta}(e^{-c\theta})^{1-\delta}$.

Analogous to the example above, the likelihood of (Y, δ) is a mixture of a density and a point mass. Thus the likelihood θ based on (Y, δ) is

$$\begin{aligned} L(Y, \delta; \theta) &= \begin{cases} f(Y|\delta = 0)P(\delta = 0) & \delta = 0 \\ f(Y|\delta = 1)P(\delta = 1) & \delta = 1 \end{cases} \\ &= [f(Y|\delta = 0)P(\delta = 0)]^{1-\delta}[f(Y|\delta = 1)P(\delta = 1)]^{\delta} \\ &= [\exp(-\theta Y + \log\theta)]^{1-\delta}[\exp(-c\theta)]^{\delta}. \end{aligned}$$

This yields the log-likelihood of θ given $\{Y_i, \delta_i\}$

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \left\{(1 - \delta_i)\left[-\theta Y_i + \log\theta\right] - \delta_i c\theta\right\}. \tag{1.6}$$

The inflated zero Poisson distribution

The Possion distribution is commonly used to model count data. However, there arises many situations where the proportion of time zero occurs is larger than the proportion one would expect using a Poisson distribution. One often models this “inflation” using a mixture distribution. Let U be a Poisson distributed random variable where $P(U = k) = \lambda^k \exp(-\lambda)/k!$. We see that $P(U = 0) = \exp(-\lambda)$. We can boost this chance by defining a new random variable Y, where

$$Y = \delta U$$

and δ is a Bernoulli random variable taking zero or one with $P(\delta = 0) = p$ and $P(\delta = 1) = (1 - p)$. It is clear that

$$\begin{aligned} P(Y = 0) &= P(Y = 0|\delta = 0)P(\delta = 0) + P(Y = 0|\delta = 1)P(\delta = 1) \\ &= 1 \times p + P(U = 0)(1 - p) = p + (1 - p)e^{-\lambda} \geq e^{-\lambda} = P(U = 0). \end{aligned}$$

Thus, in situations where there are a large number of zeros, the inflated zero Poisson seems appropriate. For $k > 1$, we have

$$\begin{aligned} P(Y = k) &= P(Y = k|\delta = 0)P(\delta = 0) + P(Y = k|\delta = 1)P(\delta = 1) \\ &= P(U = k)(1 - p) = (1 - p)\frac{\lambda^k e^{-\lambda}}{k!}. \end{aligned}$$

Thus altogether the distribution of Y is

$$P(Y = k) = \left\{p + (1 - p)e^{-\lambda}\right\}^{I(k=0)} \left\{(1 - p)\frac{\lambda^k e^{-\lambda}}{k!}\right\}^{I(k\neq 0)},$$

where $I(\cdot)$ denotes the indicator variable. Thus the log-likelihood of λ, p given $\underline{Y}$ is

$$\mathcal{L}(\underline{Y};\lambda,p) = \sum_{i=1}^{n} I(Y_i = 0)\log\left(p + (1-p)e^{-\lambda}\right) + \sum_{i=1}^{n} I(Y_i \neq 0)\left(\log(1-p) + \log\frac{\lambda^{Y_i}e^{-\lambda}}{Y_i!}\right).$$

Exercise 1.1 *Let us suppose that X and Z are independent random variables with densities f_X and f_Z respectively. Assume that X is positive.*

(i) Derive the density function of $1/X$.

(ii) Show that the density of XZ is

$$\int \frac{1}{x} f_Z(\frac{y}{x}) f_X(x)dx \tag{1.7}$$

(or equivalently $\int c^{-1}f_Z(cy)f_X(c^{-1})dc$).

(iii) Consider the linear regression model

$$Y_i = \underline{\alpha}'\underline{x}_i + \sigma_i\varepsilon_i$$

where the regressors x_i is observed, ε_i follows a standard normal distribution (mean zero and variance 1) and σ_i^2 follows a Gamma distribution

$$f(\sigma^2;\lambda) = \frac{\sigma^{2(\kappa-1)}\lambda^{\kappa}\exp(-\lambda\sigma^2)}{\Gamma(\kappa)}, \quad \sigma^2 \geq 0,$$

with $\kappa > 0$.

Derive the log-likelihood of Y_i (assuming the regressors are observed).

Exercise 1.2 *Suppose we want to model the average amount of daily rainfall in a particular region. Empirical evidence suggests that it does not rain on many days in the year. However, if it does rain on a certain day, the amount of rain follows a Gamma distribution.*

(i) Let Y denote the amount of rainfall in a particular day and based on the information above write down a model for Y.

Hint: Use the ideas from the inflated zero Poisson model.

(ii) Suppose that $\{Y_i\}_{i=1}^n$ is the amount of rain observed n consecutive days. Assuming that $\{Y_i\}_{i=1}^n$ are iid random variables with the model given in part (ii), write down the log-likelihood for the unknown parameters.

(iii) Explain why the assumption that $\{Y_i\}_{i=1}^n$ are independent random variables is tenuous.

1.3 Bounds for the variance of an unbiased estimator

So far we have iid observations $\{X_i\}$ with from a known parametric family i.e. the distribution of X_i comes from $\mathcal{F} = \{f(x;\theta);\theta \in \Theta\}$, where θ is a finite dimension parameter however the true θ is unknown. There are an infinite number of estimators of θ based on an infinite number of decision rules. Which estimator do we choose? We should choose the estimator which is "closest" to the true parameter. There are several different distance measures, but the most obvious is the mean square error. As the class of all estimators is "too large" we restrict ourselves to unbiased estimators, $\widetilde{\theta}(\underline{X})$ (where mean of estimator is equal to the true parameter) and show that the mean squared error

$$\mathrm{E}\left(\widetilde{\theta}(\underline{X})-\theta\right)^2 = \mathrm{var}\left(\widetilde{\theta}(\underline{X})\right) + \left(\mathrm{E}[\widetilde{\theta}(\underline{X})]-\theta\right)^2 = \mathrm{var}\left(\widetilde{\theta}(\underline{X})\right)$$

is bounded below by the inverse of the Fisher information (this is known as the Cramer-Rao bound). To show such a bound we require the regularity assumptions. We state the assumptions and in the case that θ is a scalar, but they can easily be extended to the case that θ is a vector.

Assumption 1.3.1 (Regularity Conditions 1) *Let us suppose that $L_n(\cdot;\theta)$ is the likelihood.*

(i) $\frac{\partial}{\partial\theta}\int L_n(\underline{x};\theta)d\underline{x} = \int \frac{\partial L_n(\underline{x};\theta)}{\partial\theta}d\underline{x} = 0$ *(for iid random variables (rv) this is equivalent to checking if* $\int \frac{\partial f(x;\theta)}{\partial\theta}dx = \frac{\partial}{\partial\theta}\int f(x;\theta)dx$*).*

Observe since a by definition a density integrates to one, then $\frac{\partial}{\partial\theta}\int L_n(\underline{x};\theta)d\underline{x} = 0$.

(ii) For any function g not a function of θ, $\frac{\partial}{\partial\theta}\int g(\underline{x})L_n(\underline{x};\theta)d\underline{x} = \int g(\underline{x})\frac{\partial L_n(\underline{x};\theta)}{\partial\theta}d\underline{x}$.

(iii) $\mathrm{E}\big(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\big)^2 > 0$.

To check Assumption 1.3.1(i,ii) we need to apply Leibniz's rule `https://en.wikipedia.org/wiki/Leibniz_integral_rule`

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} g(x)f(x;\theta)dx = \int_{a(\theta)}^{b(\theta)} g(x)\frac{\partial f(x,\theta)}{\partial\theta}dx + f(b(\theta),\theta)g(b(\theta))b'(\theta) - f(a(\theta),\theta)g(a(\theta))a'(\theta)\text{(1.8)}$$

Therefore Assumption 1.3.1(i,ii) holds if $f(b(\theta),\theta)g(b(\theta))b'(\theta) - f(a(\theta),\theta)g(a(\theta))a'(\theta) = 0$.

Example 1.3.1 *(i) If the support of the density does not depend on θ it is clear from (1.8) that Assumption 1.3.1(i,ii) is satisfied.*

(ii) If the density is the uniform distribution $f(x;\theta) = \theta^{-1}I_{[0,\theta]}(x)$ then the conditions are not satisfied. We know that $\theta^{-1}\int_0^\theta dx = 1$ (thus it is independent of θ) hence $\frac{d\theta^{-1}\int_0^\theta dx}{d\theta} = 0$. However,

$$\int_0^\theta \frac{d\theta^{-1}}{d\theta}dx = \frac{-1}{\theta} \text{ and } f(b(\theta),\theta)b'(\theta) - f(a(\theta),\theta)a'(\theta) = \theta^{-1}.$$

Thus we see that Assumption 1.3.1(i) is not satisfied. Therefore, the uniform distribution does not satisfy the standard regularity conditions.

(iii) Consider the density

$$f(x;\theta) = \frac{1}{2}(x-\theta)^2\exp[-(x-\theta)]I_{[\theta,\infty)}(x).$$

The support of this estimator depends on θ, however, it does satisfy the regularity conditions. This is because $f(x;\theta) = 0$ at both $x = \theta$ and $x = \infty$. This means that for any θ

$$f(b(\theta),\theta)g(b(\theta))b'(\theta) - f(a(\theta),\theta)g(a(\theta))a'(\theta) = 0.$$

Therefore from the Leibnitz rule we have

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} g(x)f(x;\theta)dx = \int_{a(\theta)}^{b(\theta)} g(z)\frac{\partial f(x,\theta)}{\partial\theta}dx.$$

Thus Assumption 1.3.1 is satisfied.

We now state the Cramer-Rao bound, which gives the minimal attaining variance bound for a large class of estimators. We will use the matrix inequality $A \geq B$ to mean that $A - B$ is a non-negative definite matrix (or equivalently positive semi-definite).

Theorem 1.3.1 (The Cramér-Rao bound) *Suppose the likelihood $L_n(\underline{X};\theta)$ satisfies the regularity conditions given in Assumption 1.3.1. Let $\widetilde{\theta}(\underline{X})$ be an unbiased estimator of θ, then*

$$\text{var}\left[\widetilde{\theta}(\underline{X})\right] \geq \left[\text{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2\right]^{-1}.$$

PROOF. We prove the result for the univariate case. Recall that $\widetilde{\theta}(X)$ is an unbiased estimator of θ therefore

$$\int \widetilde{\theta}(\underline{x}) L_n(\underline{x};\theta) d\underline{x} = \theta.$$

Differentiating both sides wrt to θ, and taking the derivative into the integral (allowed under the regularity condition) gives

$$\int \widetilde{\theta}(\underline{x}) \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} d\underline{x} = 1.$$

By Assumption 1.3.1(i) $\frac{d \int L_n(\underline{x};\theta) d\underline{x}}{d\theta} = \int \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} d\underline{x} = 0$. Thus adding $\theta \int \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} d\underline{x}$ to both sides of the above we have

$$\int \left\{ \widetilde{\theta}(\underline{x}) - \theta \right\} \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} d\underline{x} = 1.$$

Multiplying and dividing by $L_n(\underline{x};\theta)$ gives

$$\int \left\{ \widetilde{\theta}(\underline{x}) - \theta \right\} \frac{1}{L_n(\underline{x};\theta)} \frac{\partial L_n(\underline{x};\theta)}{\partial \theta} L_n(\underline{x};\theta) dx = 1. \tag{1.9}$$

Hence (since $L_n(\underline{x};\theta)$ is the distribution of $\underline{X}$) we have

$$\mathrm{E}\bigg(\left\{ \widetilde{\theta}(\underline{X}) - \theta \right\} \frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta} \bigg) = 1.$$

Recalling that the Cauchy-Schwartz inequality is $\mathrm{E}(UV) \leq \mathrm{E}(U^2)^{1/2} \mathrm{E}(V^2)^{1/2}$ (where equality only arises if $U = aV + b$ (where a and b are constants)) and applying it to the above we have

$$\mathrm{var}\left[\widetilde{\theta}(\underline{X}) \right] \mathrm{E}\left[\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta} \right)^2 \right] \geq 1 \qquad \Rightarrow \mathrm{var}\left[\widetilde{\theta}(\underline{X}) \right] \geq \mathrm{E}\left[\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta} \right)^2 \right]^{-1}.$$

Thus giving the Cramer-Rao inequality. □

Corollary 1.3.1 (Estimators which attain the Cramér-Rao bound) *Suppose Assumption 1.3.1 is satisfied. Then the estimator $\widetilde{\theta}(\underline{X})$ attains the Cramer-Rao bound only if it can be written as*

$$\hat{\theta}(\underline{X}) = a(\theta) + b(\theta) \frac{\partial \log L_n(\underline{X};\theta)}{\partial \theta}$$

for some functions $a(\cdot)$ and $b(\cdot)$ of θ[1].

[1] Of course, in most cases it makes no sense to construct an estimator of θ, which involves θ.

PROOF. The proof is clear and follows from when the Cauchy-Schwartz inequality is an equality in the derivation of the Cramer-Rao bound. □

We next derive an equivalent expression for $\mathrm{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2$ (called the Fisher information).

Lemma 1.3.1 *Suppose the likelihood $L_n(\underline{X};\theta)$ satisfies the regularity conditions given in Assumption 1.3.1 and for all $\theta \in \Theta$, $\frac{\partial^2}{\partial\theta^2}\int g(\underline{x})L_n(\underline{x};\theta)d\underline{x} = \int g(\underline{x})\frac{\partial^2 L_n(\underline{x};\theta)}{\partial\theta^2}d\underline{x}$, where g is any function which is not a function of θ (for example the estimator of θ). Then*

$$\mathrm{var}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right) = \mathrm{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2 = -\mathrm{E}\left(\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial\theta^2}\right).$$

PROOF. To simplify notation we focus on the case that the dimension of the vector θ is one. To prove this result we use the fact that L_n is a density to obtain

$$\int L_n(\underline{x};\theta)d\underline{x} = 1.$$

Now by differentiating the above with respect to θ gives

$$\frac{\partial}{\partial\theta}\int L_n(\underline{x};\theta)d\underline{x} = 0.$$

By using Assumption 1.3.1(ii) we have

$$\int \frac{\partial L_n(\underline{x};\theta)}{\partial\theta}d\underline{x} = 0 \Rightarrow \quad \int \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}L_n(\underline{x};\theta)d\underline{x} = 0$$

Differentiating again with respect to θ and taking the derivative inside gives

$$\begin{aligned}
& \int \frac{\partial^2 \log L_n(\underline{x};\theta)}{\partial\theta^2}L_n(\underline{x};\theta)d\underline{x} + \int \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\frac{\partial L_n(\underline{x};\theta)}{\partial\theta}d\underline{x} = 0 \\
\Rightarrow \quad & \int \frac{\partial^2 \log L_n(\underline{x};\theta)}{\partial\theta^2}L_n(\underline{x};\theta)d\underline{x} + \int \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\frac{1}{L_n(\underline{x};\theta)}\frac{\partial L_n(\underline{x};\theta)}{\partial\theta}L_n(\underline{x};\theta)d\underline{x} = 0 \\
\Rightarrow \quad & \int \frac{\partial^2 \log L_n(\underline{x};\theta)}{\partial\theta^2}L_n(\underline{x};\theta)d\underline{x} + \int \left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)d\underline{x} = 0
\end{aligned}$$

Thus

$$-\mathrm{E}\left(\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial\theta^2}\right) = \mathrm{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2.$$

The above proof can easily be generalized to parameters θ, with dimension larger than 1. This gives us the required result.

Note in all the derivations we are evaluating the second derivative of the likelihood at the *true parameter*. □

We mention that there exists distributions which do not satisfy Assumption 1.3.1. These are called *non-regular distributions*. The Cramer-Rao lower bound does hold for such distributions.

Definition 1.3.1 (The Fisher information matrix) *The matrix*

$$I(\theta) = \left[\mathrm{E}\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2\right] = -\left[\mathrm{E}\left(\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial\theta^2}\right)\right],$$

whose inverse forms the lower bound of Cramér-Rao bound is called the Fisher information matrix. It plays a critical role in classical inference.

Essentially $I(\theta)$ tells us how much "information" the data $\{X_i\}_{i=1}^n$ contains about the true *parameter θ.*

Remark 1.3.1 *Define the quantity*

$$\begin{aligned} I_{\theta_0}(\theta) &= -\int \left(\frac{\partial^2 \log L_n(\underline{x};\theta)}{\partial\theta^2}\right) L_n(\underline{x};\theta_0)d\underline{x} \\ &= -\left[\mathrm{E}_{\theta_0}\left(\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial\theta^2}\right)\right]. \end{aligned}$$

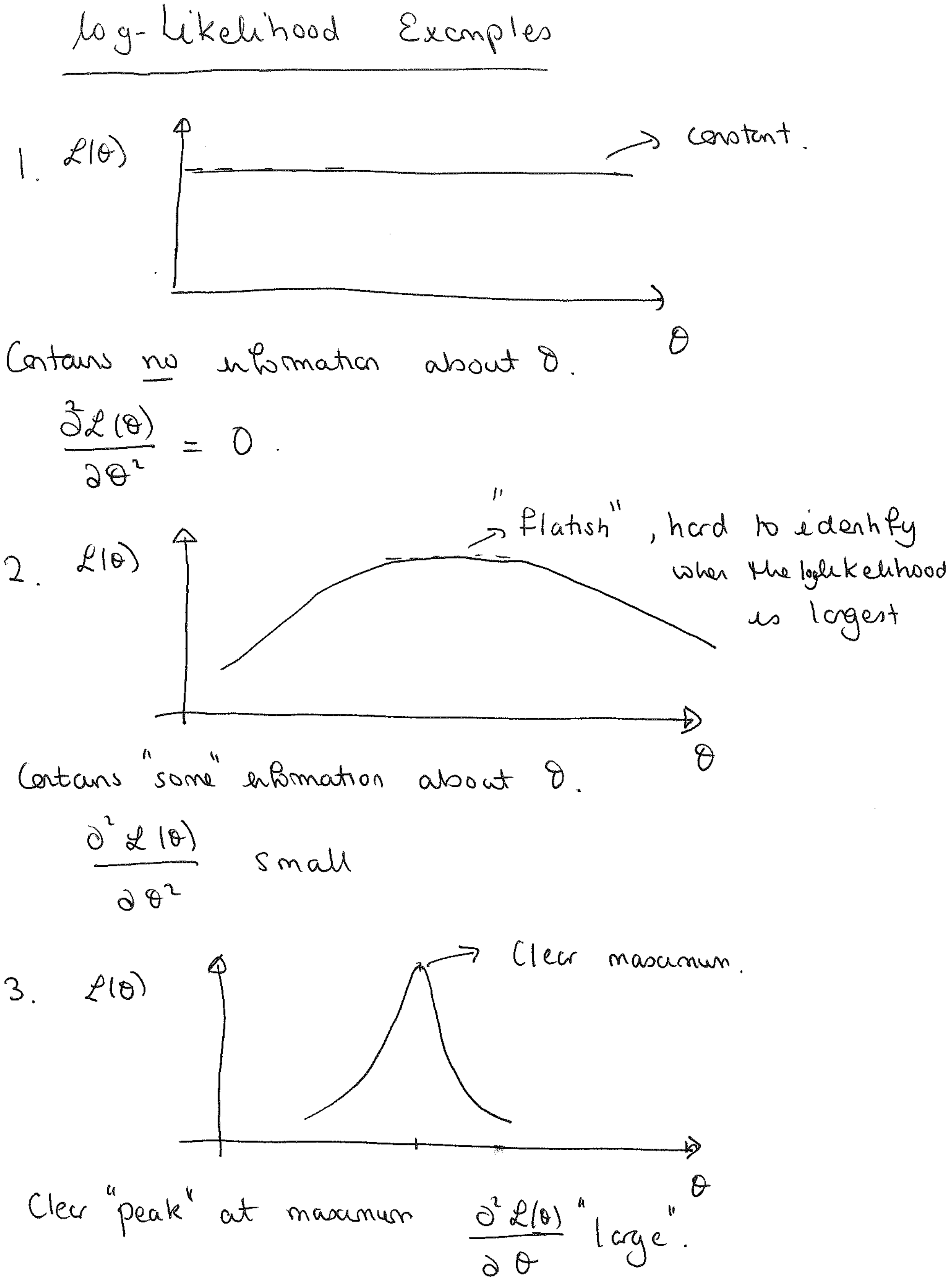
This quantity evaluates the negative expected second derivative of the log-likelihood over θ, but the expectation is taken with respect to the "true" density $L_n(\underline{x};\theta_0)$. This quantity will not be positive for all θ. However, by the result above we evaluate $I_{\theta_0}(\theta)$ at $\theta = \theta_0$, then

$$I_{\theta_0}(\theta_0) = \mathrm{var}_{\theta_0}\left(\frac{\partial \log L_n(\underline{x};\theta_0)}{\partial\theta}\right).$$

In other words, when the expectation of the negative second derivative of log-likelihood is evaluated at the true *parameter this is the Fisher information which is positive.*

Exercise 1.3 *Suppose $\{X_i\}$ are iid random variables with density $f(x;\theta)$ and the Fisher information for θ based on $\{X_i\}$ is $I(\theta)$.*

Let $Y_i = g(X_i)$ where $g(\cdot)$ is a bijective diffeomorphism (the derivatives of g and its inverse exist). Intuitive when one makes such a transformation no "information" about θ should be lost or gained. Show that the Fisher information matrix of θ based on $\{Y_i\}$ is $I(\theta)$.



Contains no information about θ.

$$\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2} = 0.$$

Contains "some" information about θ.

$$\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2} \quad \text{small}$$

Clear "peak" at maximum $\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta}$ "large".

Figure 1.1: Interpretation of the Fisher information

Example 1.3.2 *Consider the example of censored data given in Section 1.2. Both the observations and the censored variables, $\{Y_i\}$ and $\{\delta_i\}$, where*

$$\delta_i = I(Y_i \geq c)$$

contain information about the parameter θ. However, it seems reasonable to suppose that $\{Y_i\}$ contains more information about θ than $\{\delta_i\}$. We articulate what we mean by this in the result below.

Lemma 1.3.2 *Let us suppose that the log-likelihood of $\underline{X}$, $\mathcal{L}_n(\underline{X};\theta)$ satisfies Assumption 1.3.1. Let $\underline{Y} = B(\underline{X})$ be some statistic (of arbitrary dimension) of the original data. Let $\mathcal{L}_{B(\underline{X})}(\underline{Y};\theta)$, and $\mathcal{L}(\underline{X}|\underline{Y};\theta)$ denote the log-likelihood of $\underline{Y} = B(\underline{X})$ and conditional likelihood of $\underline{X}|\underline{Y}$ (we assume these satisfy Assumption 2.6.1, however I think this is automatically true). Then*

$$I_{\underline{X}}(\theta) \geq I_{B(\underline{X})}(\theta)$$

where

$$I_{\underline{X}}(\theta) = \mathrm{E}\left(\frac{\partial \mathcal{L}_{\underline{X}}(\underline{X};\theta)}{\partial \theta}\right)^2 \text{ and } I_{B(\underline{X})}(\theta) = \mathrm{E}\left(\frac{\partial \mathcal{L}_{B(\underline{X})}(\underline{Y};\theta)}{\partial \theta}\right)^2.$$

In other words the original Fisher information contains the most information about the parameter. In general, most transformations of the data will lead to a loss in information. We consider some exceptions in Lemma 1.4.1.

PROOF. Writing the conditional density of $\underline{X}$ given $B(\underline{X})$ as the ratio of a joint density of $\underline{X}, B(\underline{X})$ and marginal density of $B(\underline{X})$ we have

$$f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y}) = \frac{f_{\underline{X},B(\underline{X})}(\underline{x},\underline{y};\theta)}{f_{B(\underline{X})}(\underline{y};\theta)} \Rightarrow f_{\underline{X},B(\underline{X})}(\underline{x},\underline{y};\theta) = f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y})f_{B(\underline{X})}(\underline{y};\theta),$$

where $f_{\underline{X}|B(\underline{X})}$ denotes the density of $\underline{X}$ conditioned on $B(\underline{X})$ and $f_{\underline{X},B(\underline{X})}$ the joint density of $\underline{X}$ and $B(\underline{X})$. Note that if $B(\underline{x}) = y$, then the joint density $f_{\underline{X},B(\underline{X})}(\underline{x},y;\theta)$ is simply the density of $f_{\underline{X}}(\underline{x};\theta)$ with the constraint that $\underline{y} = B(\underline{x})$ i.e. $f_{\underline{X},B(\underline{X})}(\underline{x},\underline{y};\theta) = f_{\underline{X}}(\underline{x};\theta)\delta(B(\underline{x}) = \underline{y})$, where δ denotes the indicator variable[2]. Thus we have

$$f_{\underline{X}}(\underline{x};\theta)\delta(B(\underline{x}) = \underline{y}) = f_{\underline{X}|B(\underline{X})}(\underline{x}|y,\theta)f_{B(\underline{X})}(\underline{y};\theta).$$

[2] To understand why, consider the joint density of $X, Y = B(X)$ the density ie not defined over $\mathbb{R}^2$ but over the curve $(x, B(x))$ $f_{X,B(X)}(x,y) = f_X(x)\delta(y = B(x))$

Having written the likelihood in this way, the derivative of the log likelihood is

$$\begin{aligned}
\frac{\partial \log f_{\underline{X}}(\underline{x};\theta)}{\partial\theta} &= \frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y}) f_{B(\underline{X})}(\underline{y};\theta)}{\partial\theta} \\
&= \frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y},\theta)}{\partial\theta} + \frac{\partial \log f_{B(\underline{X})}(\underline{y};\theta)}{\partial\theta}.
\end{aligned}$$

Therefore

$$\begin{aligned}
I_{\underline{X}}(\theta) = \mathrm{var}\left(\frac{\partial \log f_{\underline{X}}(\underline{X};\theta)}{\partial\theta}\right) &= \mathrm{var}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\right) + \underbrace{\mathrm{var}\left(\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right)}_{=I_{B(\underline{X})}(\theta)} + \\
& 2\mathrm{cov}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}, \frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right). \qquad (1.10)
\end{aligned}$$

Under the stated regularity conditions, since $f_{B(\underline{X})}$, is a density it is clear that

$$\mathrm{E}\left(\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right) = 0$$

and

$$\mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\big|B(\underline{X})\right) = \int \frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y},\theta)}{\partial\theta} f_{\underline{X}|B(\underline{X})}(\underline{x}|\underline{y},\theta)d\underline{x} = 0. \qquad (1.11)$$

Thus using the law of iterated expectation $\mathrm{E}(A) = \mathrm{E}(\mathrm{E}[A|B])$, then $\mathrm{E}[\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}] = 0$. Returning to (1.10), since the mean is zero this implies that

$$I_{\underline{X}}(\theta) = I_{B(\underline{X})}(\theta) + \mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\right)^2 + 2\mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\frac{\partial \log f_{B(\underline{X})}(\underline{Y};\theta)}{\partial\theta}\right)$$

Finally we show that the above covariance is zero. To do so we use that $\mathrm{E}(XY) = \mathrm{E}(X\mathrm{E}[Y|X])$ (by the law of iterated expectation) then by using (1.11) we have

$$\begin{aligned}
& \mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right) \\
= & \mathrm{E}\left(\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\underbrace{\mathrm{E}\left[\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\middle|\frac{\partial \log f_{B(\underline{X})}(B(\underline{X});\theta)}{\partial\theta}\right]}_{=0 \text{ by } (1.11)}\right) = 0.
\end{aligned}$$

Thus

$$I_{\underline{X}}(\theta) = I_{B(\underline{X})}(\theta) + \mathrm{E}\left(\frac{\partial \log f_{\underline{X}|B(\underline{X})}(\underline{X}|B(\underline{X}),\theta)}{\partial\theta}\right)^2.$$

As all the terms are positive, this immediately implies that $I_{\underline{X}}(\theta) \geq I_{B(\underline{X})}(\theta)$. □

Definition 1.3.2 (Observed and Expected Fisher Information) *(i) The observed Fisher information matrix is defined as*

$$I(\underline{X};\theta) = -\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial\theta^2}.$$

(ii) The expected Fisher information matrix is defined as

$$I(\theta) = \mathrm{E}\left(-\frac{\partial^2 \log L_n(\underline{X};\theta)}{\partial\theta}\right)$$

These will play an important role in inference for parameters.

Often we want to estimate a function of θ, $\tau(\theta)$. The following corollary is a generalization of the Cramer-Rao bound.

Corollary 1.3.2 *Suppose Assumption 1.3.1 is satisfied and $T(\underline{X})$ is an unbiased estimator of $\tau(\theta)$. Then we have*

$$\mathrm{var}\left[T(\underline{X})\right] \geq \frac{\tau'(\theta)^2}{\mathrm{E}\left[\left(\frac{\partial \log L_n(\underline{X};\theta)}{\partial\theta}\right)^2\right]}.$$

Exercise 1.4 *Prove the above corollary.*

In this section we have learnt how to quantify the amount of information the data contains about a parameter and show that for the majority of transformations of data (with the exception of bijections) we loose information. In the following section we define a transformation of data, where in some certain situations, will substantially reduce the dimension of the data, but will not result in a loss of information.

1.4 Sufficient statistics

We start with a simple example from introductory statistics.

Example 1.4.1 *Samples of size 10 and 15 are drawn from two different distributions. How to check if the two samples come from the same distribution? The data is given in Figure 1.2. If the distributions are known to come from the Gaussian family of distributions with, for the sake of argument, standard deviation one, then all the information about the unknown parameter, is characteristized in terms of the sample means $\bar{X}_A$ and*

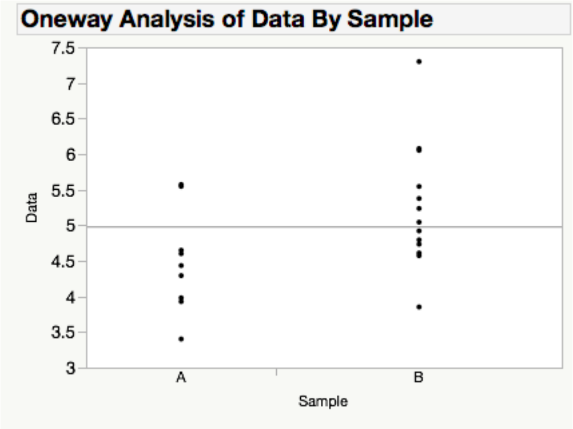


Figure 1.2: Samples from two population

$\bar{X}_B$ *(in this example,* 4.6 *and* 5.2 *respectively). The sample mean is* sufficient *for describing all the information about the unknown mean, more precisely the data conditioned on sample mean is free of any information about μ.*

On the other hand, if the data comes from the Cauchy family of distributions $\{f_\theta(x) = [\pi(1 + (x-\theta)^2)]^{-1}\}$ there does not exist a lower dimensional transformations of the data which contains all the information about θ. The observations conditioned on any lower transformation will still contain information about θ.

This example brings us to a formal definition of sufficiency.

Definition 1.4.1 (Sufficiency) *Suppose that $\underline{X} = (X_1, \ldots, X_n)$ is a random vector. A statistic $s(\underline{X})$ is said to be sufficient for the family $\mathcal{F}$ of distributions, if the conditional density $f_{\underline{X}|S((\underline{X}))}(y|s)$ is the same for all distributions in $\mathcal{F}$.*

This means in a parametric class of distributions $\mathcal{F} = \{f(\underline{x};\theta); \theta \in \Theta\}$ the statistic $s(\underline{X})$ is sufficient for the parameter θ, if the conditional distribution of $\underline{X}$ given $s(\underline{X})$ is not a function of θ.

Example 1.4.2 (Order statistics) *Suppose that $\{X_i\}_{i=1}^n$ are iid random variables with density $f(x)$. Let $X_{(1)}, \ldots, X_{(n)}$ denote the ordered statistics (i.e. $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$). We will show that the order statistics $X_{(1)}, \ldots, X_{(n)}$ is the sufficient statistic over the family of all densities $\mathcal{F}$.*

To see why, note that it can be shown that the joint density of the order statistics

$X_{(1)}, \ldots, X_{(n)}$ is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n! \prod_{i=1}^n f(x_i) & x_1 \leq \ldots, \leq x_n \\ 0 & otherwise \end{cases} \tag{1.12}$$

Clearly the density of the $X_1, \ldots, X_n$ is $\prod_{i=1}^n f(X_i)$. Therefore the density of $X_1, \ldots, X_n$ given $X_{(1)}, \ldots, X_{(n)}$ is

$$f_{X_1,\ldots,X_n|X_{(1)},\ldots,X_{(n)}} = \frac{1}{n!},$$

which is simply the chance of selecting the ordering $X_1, \ldots, X_n$ from a sequence $X_{(1)}, \ldots, X_{(n)}$. Clearly this density does not depend on any distribution.

This example is interesting, but statistically not very useful. In general we would like the number of sufficient statistic to be a lower dimension than the data itself (sufficiency is a form of compression).

Exercise 1.5 *Show (1.12).*

Usually it is extremely difficult to directly obtain a sufficient statistic from its definition. However, the factorisation theorem gives us a way of obtaining the sufficient statistic.

Theorem 1.4.1 (The Fisher-Neyman Factorization Theorem) *A necessary and sufficient condition that $s(\underline{X})$ is a sufficient statistic is that the likelihood function, L (not log-likelihood), can be factorized as $L_n(\underline{X}; \theta) = h(\underline{X})g(s(\underline{X}); \theta)$, where $h(\underline{X})$ is not a function of θ.*

Example 1.4.3 (The uniform distribution) *Let us suppose that $\{X_i\}$ are iid uniformly distributed random variables with density $f_\theta(x) = \theta^{-1} I_{[0,\theta]}(x)$. The likelihood is*

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

Since $L_n(\underline{X}; \theta)$ is only a function of $\max_i X_i$, it is immediately clear that $s(\underline{X}) = \max_i X_i$ is a sufficient.

Example 1.4.4 (The normal distribution) *Let $\{X_i\}_{i=1}^n$ be iid normal random variables. The likelihood is*

$$\begin{aligned} L_n(\underline{X};\mu,\sigma^2) &= \frac{1}{(2\pi\sigma)^n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2\right] = \frac{1}{(2\pi\sigma)^n}\exp\left[-\frac{1}{2\sigma^2}\left(S_{xx}-2S_x\mu+\mu^2\right)\right] \\ &= g(S_x,S_{xx};\mu,\sigma^2) \end{aligned}$$

where $S_x = \sum_{i=1}^n X_i$ and $S_{xx} = \sum_{i=1}^n X_i^2$. We see immediately from the factorisation theorem that the density is a function of two sufficient statistics S_x and S_{xx}. Thus S_x and S_{xx} are the sufficient statistics for μ and σ^2.

Suppose we treat σ^2 as known, then by using

$$\begin{aligned} L_n(\underline{X};\mu,\sigma^2) &= \frac{1}{(2\pi\sigma)^n}\exp\left[-\frac{1}{2\sigma^2}S_{xx}\right]\exp\left[-\frac{S_x\mu}{\sigma^2}+\frac{\mu^2}{\sigma^2}\right] \\ &= g_1(S_{xx};\sigma^2)g_2(S_x;\mu,\sigma^2) \end{aligned}$$

we see that the sufficient statistic for the mean, μ, is $S_x = \sum_{i=1}^n X_i$. I.e. any function of $\{X_i\}$ conditioned on S_x contains no information about the mean μ. This includes the S_{xx}. However, S_x contains information about the both μ and σ^2. We can explicitly see this because $S_x \sim N(n\mu, n\sigma^2)$.

Note that alternative sufficient statistics for the normal distribution are $S_x = \sum_i X_i$ and $S'_{xx} = \sum_i (X_i - n^{-1}S_x)^2$. Sufficient statistics are not unique!

Example 1.4.5 (The exponential family) *The exponential family of distributions, characterized by*

$$f(x;\omega) = \exp\left[s(x)\eta(\omega) - b(\omega) + c(x)\right], \tag{1.13}$$

is broad class of distributions which includes the normal distributions, binomial, exponentials etc. but not the uniform distribution. Suppose that $\{X_i\}_{i=1}^n$ are iid random variables which have the form (1.13) We can write and factorize the likelihood as

$$\begin{aligned} L_n(\underline{X};\omega) &= \exp\left[\eta(\omega)\sum_{i=1}^n s(X_i) - nb(\omega)\right]\exp\left[\sum_{i=1}^n c(X_i)\right] \\ &= g(\sum_{i=1}^n s(X_i);\omega)h(X_1,\ldots,X_n). \end{aligned}$$

We immediately see that $\sum_{i=1}^n s(X_i)$ is a sufficient statistic for ω.

The above example is for the case that the number of parameters is one, however we can generalize the above to the situation that the number of parameters in the family is p

$$f(x;\omega) = \exp\left[\sum_{j=1}^{p} s_j(x)\eta_j(\omega) - b(\omega) + c(x)\right],$$

where $\omega = (\omega_1, \ldots, \omega_p)$. The sufficient statistics for the p-dimension is $(\sum_{i=1}^{n} s_1(X_i), \ldots, \sum_{i=1}^{n} s_p(X_i))$. Observe, we have not mentioned, so far, about this being in anyway minimal, that comes later.

For example, the normal distribution is parameterized by two parameters; mean and variance. Typically the number of sufficient statistics is equal to the the number of unknown parameters. However there can arise situations where the number of sufficient statistics is more than the number of unknown parameters.

Example 1.4.6 *Consider a mixture model, where we know which distribution a mixture comes from. In particular, let $g_0(\cdot;\theta)$ and $g_1(\cdot;\theta)$ be two different densities with unknown parameter θ. Let δ be a Bernoulli random variables which takes the values 0 or 1 and the probability $P(\delta = 1) = 1/2$. The random variables (X, δ) have the joint "density"*

$$\begin{aligned} f(x,\delta;\theta) &= \begin{cases} \frac{1}{2}g_0(x;\theta) & \delta = 0 \\ \frac{1}{2}g_1(x;\theta) & \delta = 1 \end{cases} \\ &= (1-\delta)\frac{1}{2}g_0(x;\theta) + \delta\frac{1}{2}g_1(x;\theta) = (\frac{1}{2}g_0(x;\theta))^{1-\delta}(\frac{1}{2}g_1(x;\theta))^{\delta}. \end{aligned}$$

Example; the population of males and females where we observe the gender and height of an individual. Both (X, δ) are the sufficient statistics for θ. Observe that X by itself is not sufficient because

$$P(\delta|X = x) = \frac{g_1(x;\theta)}{g_0(x;\theta) + g_1(x;\theta)}.$$

Hence conditioned on just X, the distribution of δ contains information about θ, implying X by itself is not sufficient.

Remark 1.4.1 (Ancillary variables) *The above example demonstrates the role of an* ancillary variable. *If we observe only X, since the marginal density of X is*

$$\frac{1}{2}g_0(x;\theta) + \frac{1}{2}g_1(x;\theta),$$

then X contains information about θ. On the other hand, if we only observe δ, it contains no information about θ (the marginal distribution of δ is half). This means that θ is an ancillary variable (since its marginal distribution contains no information about θ).

Furthermore, since (X, δ) are the sufficient statistics for θ, δ is an ancillary variable and δ in conjuction with X does contain information about θ then δ is called an ancillary complement.

We already came across an ancillary variable. We recall that for the normal distribution one version of the sufficient statistics is $S_x = \sum_i X_i$ and $S'_{xx} = \sum_i (X_i - n^{-1}S_x)^2$. Now we see that $S'_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2\chi^2_{n-1}$, hence it is an ancillary variable for the mean, since its marginal distribution does not depend on μ. However, it is not an ancillary complement for S_x since S'_{xx} conditioned on S_x does not depend on μ in fact they are independent! So S'_{xx} conditioned or otherwise contains no information whatsoever about the mean μ.

From the examples above we immediately see that the sufficient statistic is not unique. For example, for the Gaussian family of distributions the order statistics $X_{(1)}, \ldots, X_{(n)}$, S_{xx}, S_x and S_x, S'_{xx} are all sufficient statistics. But it is clear that S_{xx}, S_x or S_x, S'_{xx} is "better" than $X_{(1)}, \ldots, X_{(n)}$, since it "encodes" the data in fewer terms. In other words, it drops the dimension of the data from n to two. This brings us to the notion of minimal sufficiency.

Definition 1.4.2 (Minimal sufficiency) *A statistic $S(\underline{X})$ is minimal sufficient if (a) it is sufficient and (b) if $T(\underline{X})$ is sufficient statistic there exists an f such that $S(\underline{X}) = f(T(\underline{X}))$.*

Note that the minimal sufficient statistic of a family of distributions is not unique.

The minimal sufficient statistic corresponds to the coarsest sufficient partition of sample space, whereas the data generates the finest partition. We show in Lemma 1.6.4 that if a family of distributions belong to the exponential family and the sufficient statistics are linearly independent, then these sufficient statistics are minimally sufficient.

We now show that the minimal sufficient statistics of the exponential class of distributions are quite special.

Theorem 1.4.2 (Pitman-Koopman-Darmois theorem) *Suppose that $\mathcal{F}$ is a parametric class of distributions whose domain* does not depend on the parameter*, this as-*

sumption includes the Cauchy family, Weibull distributions and exponential families of distributions but not *the uniform family. Only in the case that distribution belongs to the exponential family will the number of minimal sufficient statistic* not *depend on sample size.*

The uniform distribution has a finite number of sufficient statistics ($\max X_i$), which does not depend on the sample size and it does not belong the exponential family. However, the Pitman-Koopman-Darmois theorem does not cover the uniform distribution since its domain depends on the parameter θ.

Example 1.4.7 (Number of sufficient statistics is equal to the sample size) *(i) Consider the Cauchy family of distributions*

$$\mathcal{F} = \left\{ f_\theta; f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2} \right\}.$$

the joint distribution of $\{X_i\}_{i=1}^n$ where X_i follow a Cauchy is

$$\prod_{i=1}^n \frac{1}{\pi(1+(x_i-\theta)^2)}.$$

We observe that the parameters cannot be separated from any of the variables. Thus we require all the data to characterize the parameter θ.

(ii) The Weibull family of distributions

$$\mathcal{F} = \left\{ f_\theta; f_{\phi,\alpha}(x) = \left(\frac{\alpha}{\phi}\right)\left(\frac{x}{\phi}\right)^{\alpha-1} \exp[-(x/\phi)^\alpha] \right\}$$

Example 1.4.8 (The truncated exponential) *Suppose that X is an exponentially distributed random variable but is truncated at c. That is*

$$f(x;\theta) = \frac{\theta\exp(-\theta x)}{1-e^{-c\theta}} I(x \leq c).$$

However, the truncation point c is the point which cuts the exponential distribution in half, that is $1/2 = e^{-c\theta} = 1 - e^{-c\theta}$. Thus $c = \theta^{-1}\log 2$. Thus the boundary of the distribution depends on the unknown parameter θ (it does not belong to the exponential family).

Suppose $\{X_i\}$ are iid random variables with distribution $f(x;\theta) = 2\theta\exp(-x\theta)I(x \leq \theta^{-1}\log 2)$ where $\theta \in \Theta = (0,\infty)$. The likelihood for θ is

$$\begin{aligned} L(\theta;\underline{X}) &= 2^n\theta^n\exp(-\theta\sum_{i=1}^{n}X_i)\prod_{i=1}^{n}I_{[0,\theta^{-1}\log 2]}(X_i) \\ &= 2^n\theta^n\exp(-\theta\sum_{i=1}^{n}X_i)I_{[0,\theta^{-1}\log 2]}(\max X_i), \end{aligned}$$

thus we see there are two sufficient statistics for θ, $s_1(\underline{X}) = \sum_i X_i$ and $s_2(\underline{X}) = \max_i X_i$.

We recall that from Lemma 1.3.2 that most transformations in the data will lead to a loss in information about the parameter θ. One important exception are sufficient statistics.

Lemma 1.4.1 (The Fisher information matrix and sufficient statistics) *Suppose Assumption 1.3.1 holds and $S(\underline{X})$ is a sufficient statistic for a parametric family of distributions $\mathcal{F} = \{f_\theta;\theta \in \Theta\}$. Let $I_{\underline{X}}(\theta)$ and $I_{S(\underline{X})}(\theta)$ denote the Fisher information of $\underline{X}$ and $S(\underline{X})$ respectively. Then for all $\theta \in \Theta$*

$$I_{S(\underline{X})}(\theta) = I_{\underline{X}}(\theta).$$

PROOF. From the proof of Lemma 1.3.2 we have

$$I_{\underline{X}}(\theta) = I_{S(\underline{X})}(\theta) + \mathrm{E}\left(\frac{\partial \log f_{\underline{X}|S(\underline{X})}(\underline{X}|S(\underline{X}),\theta)}{\partial\theta}\right)^2. \tag{1.14}$$

By definition of a sufficient statistic

$$f_{\underline{X}|S(\underline{X})}(\underline{x}|\underline{y},\theta)$$

does not depend on θ. This means that $\frac{\partial \log f_{\underline{X}|S(\underline{X})}(\underline{X}|S(\underline{X}),\theta)}{\partial\theta} = 0$, consequently the second term on the right hand side of (1.14) is zero, which gives the required result. □

Remark 1.4.2 *It is often claimed that only transformations of the data which are sufficient statistics have the same information as the original data. This is not necessarily true, sufficiency is not a necessary condition for Lemma 1.4.1 to hold. `http://arxiv.org/pdf/1107.3797v2.pdf` gives an example where a statistic that is* not *a sufficient statistic of the data has the same Fisher information as the Fisher information of the data itself.*

1.4.1 The Fisher information and ancillary variables

We defined the notion of ancillary in the previous section. Here we give an application. Indeed we have previously used the idea of an ancillary variable in regression even without thinking about it! I discuss this example below

So let us start with an example. Consider the problem of simple linear regression where $\{Y_i, X_i\}_{i=1}^n$ are iid bivariate Gaussian random variables and

$$Y_i = \beta X_i + \varepsilon_i,$$

where $\mathrm{E}[\varepsilon_i] = 0$, $\mathrm{var}[\varepsilon_i] = 1$ and X_i and ε_i are independent and β is the unknown parameter of interest. We observe $\{Y_i, X_i\}$. Since X_i contains no information about β it seems logical to look at the conditional log-likelihood of Y_i conditioned on X_i

$$\mathcal{L}(\beta; \underline{Y}|\underline{X}) = \frac{-1}{2}\sum_{i=1}^{n}(Y_i - \beta X_i).$$

Using the factorisation theorem we see that sufficient statistics for β are $\sum_{i=1}^n Y_i X_i$ and $\sum_{i=1}^n X_i^2$. We see that the distribution of $\sum_{i=1}^n X_i^2$ contains no information about β. Thus it is an ancillary variable. Furthermore, since the conditional distribution of $\sum_{i=1}^n X_i^2$ conditioned on $\sum_{i=1}^n X_i Y_i$ does depend on β it is an ancillary complement (I have no idea what the distribution is).

Now we calculate the Fisher information matrix. The second derivative of the likelihood is

$$\frac{\partial^2 \mathcal{L}(\beta; \underline{Y}|\underline{X})}{\partial \beta^2} = -\sum_{i=1}^{n} X_i^2 \Rightarrow -\frac{\partial^2 \mathcal{L}(\beta; \underline{Y}|\underline{X})}{\partial \beta^2} = \sum_{i=1}^{n} X_i^2.$$

To evaluate the Fisher information, do we take the expectation with respect to the distribution of $\{X_i\}$ or not? In other words, does it make sense to integrate influence of the observed regressors (which is the ancillary variable) or not? Typically, in regression one does not. We usually write that the variance of the least squares estimator of a simple linear equation with no intercept is $(\sum_{i=1}^n X_i^2)$.

We now generalize this idea. Suppose that (X, A) are sufficient statistics for the parameter θ. However, A is an ancillary variable, thus the marginal distribution contains no information about θ. The joint log-likelihood can be written as

$$\mathcal{L}(\theta; X, A) = \mathcal{L}(\theta; X|A) + \mathcal{L}(A)$$

where $\mathcal{L}(\theta; X|A)$ is the conditional log-likelihood of X conditioned on A and $\mathcal{L}(A)$ is the marginal log distribution of A which does not depend on A. Clearly the second derivative of $\mathcal{L}(\theta; X, A)$ with respect to θ is

$$-\frac{\partial^2 \mathcal{L}(\theta; X, A)}{\partial \theta^2} = -\frac{\partial^2 \mathcal{L}(\theta; X|A)}{\partial \theta^2}.$$

The Fisher information is the expectation of this quantity. But using the reasoning in the example above it would seem reasonable to take the expectation conditioned on the ancillary variable A.

1.5 Sufficiency and estimation

It is clear from the factorisation theorem that the sufficient statistic contains all the "ingredients" about the parameter θ. In the following theorem we show that by projecting any unbiased estimator of a parameter onto its sufficient statistic we reduce its variance (thus improving the estimator).

Theorem 1.5.1 (The Rao-Blackwell Theorem) *Suppose $s(\underline{X})$ is a sufficient statistic and $\widetilde{\theta}(\underline{X})$ is an unbiased estimator of θ then if we define the new unbiased estimator $\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X})]$, then $\mathrm{E}[\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X})]] = \theta$ and*

$$\mathrm{var}\left[\mathrm{E}\left(\widetilde{\theta}(\underline{X})|s(\underline{X})\right)\right] \leq \mathrm{var}\left[\widetilde{\theta}(\underline{X})\right].$$

PROOF. Using that the distribution of $\underline{X}$ conditioned on $s(\underline{X})$ *does not* depend on θ, since $s(\underline{X})$ is sufficient (very important, since our aim is to estimate θ) we have

$$\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X}) = y] = \int \widetilde{\theta}(\underline{x}) f_{\underline{X}|s(\underline{X})=y}(\underline{x}) d\underline{x}$$

is only a function of $s(\underline{X}) = y$ (and not θ).

We know from the theory of conditional expectations that since $\sigma(s(\underline{X})) \subset \sigma(X_1, \ldots, X_n)$, then $\mathrm{E}[\mathrm{E}(X|\mathcal{G})] = \mathrm{E}[X]$ for any sigma-algebra $\mathcal{G}$. Using this we immediately we have $\mathrm{E}[\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X})]] = \mathrm{E}[\widetilde{\theta}(\underline{X})] = \theta$. Thus $\mathrm{E}[\widetilde{\theta}(\underline{X})|s(\underline{X})]$ is an unbiased estimator.

To evaluate the variance we use the well know equality $\mathrm{var}[X] = \mathrm{var}[\mathrm{E}(X|Y)] + \mathrm{E}[\mathrm{var}(X|Y)]$. Clearly, since all terms are positive $\mathrm{var}[X] \geq \mathrm{var}[\mathrm{E}(X|Y)]$. This immediately gives the Rao-Blackwell bound. □

Example 1.5.1 *Suppose $\{X_i\}_{i=1}^n$ are iid normal random variable with mean μ and variance σ^2. We know that $S_x = \sum_{i=1}^n X_i$ is a sufficient statistic for μ. We also know that X_1 is an unbiased estimator of μ, but it is not sufficient. It is clear that $\mathrm{var}[\widetilde{\theta}] = \mathrm{var}[X_1] = \sigma^2$. To improve the estimator we condition X_1 on S_x, that is define $\widehat{\theta} = \mathrm{E}[X_1|S_x]$, by the Rao-Blackwell theorem this has a smaller variance than X_1. To show that this is true for this example, we use that $X_1, \ldots, X_n$ are jointly normal then $\mathrm{E}[X_1|S_x]$ is the best linear predictor of X_1 given S_x*

$$\mathrm{E}[X_1|S_x] = \frac{\mathrm{cov}[X_1, S_x]}{\mathrm{var}[S_x]} S_x = \frac{\sigma^2}{n\sigma^2} S_x = \bar{X},$$

which is not a surprise.

Is this the best estimator amongst all unbiased estimator? The Lehmann-Scheffe theorem shows that it is.

The Rao-Blackwell theorem tells us that estimators with the smallest variance must be a function of a sufficient statistic. Of course, one can ask is there a unique estimator with the minumum variance. For this we require completeness of the sufficient statistic. Uniqueness immediately follows from the idea of completeness.

Definition 1.5.1 (Completeness) *Let $s(\underline{X})$ be a minimally sufficient statistic for all $\theta \in \Theta$. Suppose $Z(\cdot)$ is a function of $s(\underline{X})$ such that $\mathrm{E}_\theta[Z(s(\underline{X}))] = 0$. $s(\underline{X})$ is a complete sufficient statistic if and only if $\mathrm{E}[Z(s(\underline{X}))] = 0$ implies $Z(t) = 0$ for all t and all $\theta \in \Theta$.*

Example 1.5.2 *If the exponential family has full rank, that is the number of unknown parameters is equal to the dimension of the exponential family (and the parameter space Θ is an open set, as yet I cannot give a good condition for this) then it is complete (see Lehmann (1986), Section 4.3, Theorem 1).*

Examples include the fully parameterized normal distribution, exponential distribution, binomial distribution etc.

Example 1.5.3 (The constrained normal) *Suppose that $X \sim \mathcal{N}(\mu^2, \mu^2)$. Then $S_x = \sum_{i=1}^n X_i$ and $S'_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$ are still the sufficient statistics for μ^2. To see why consider the conditional distribution of $S'_{xx}|S_x$, we know that S'_{xx} and S_x are independent thus it is the marginal distribution of S'_{xx} which is $\mu^2\chi^2_{n-1}$. Clearly this still depends on the parameter μ^2. Hence we cannot reduce the number of sufficient statistics when we constrain the parameters.*

However, S_x and S_{xx} are not complete sufficient statistics for μ^2, since there exists a non-zero function $Z(S_x, S_{xx})$ such that

$$\mathrm{E}\left(Z(S_x, S_{xx})\right) = \mathrm{E}\left(\bar{X} - s^2\right) = \mu^2 - \mu^2 = 0.$$

Example 1.5.4 (The uniform distribution $f(x;\theta) = \theta^{-1}I_{[0,\theta]}(x)$) *Given the random variables $\{X_i\}_{i=1}^n$ we recall that the sufficient statistic is $\max(X_i)$, we now show that it is complete. Since $P(\max(X_i) \leq x) = (x/\theta)^n I_{[0,\theta]}(x)$ the density is $f_{\max}(x) = nx^{n-1}/\theta^n I_{[0,\theta]}(x)$. We now look for functions Z (which do not depend on θ) where*

$$\mathrm{E}_\theta(Z(\max_i X_i)) = \frac{n}{\theta^n}\int_0^\theta Z(x)x^{n-1}dx.$$

It is "clear" that there cannot exist a function X where the above is zero for all $\theta \in (0,\infty)$ (I can't think of a cute mathematical justification). Thus $\max_i(X_i)$ is a complete minimal sufficient statistic for $\{X_i\}$.

Theorem 1.5.2 (Lehmann-Scheffe Theorem) *Suppose that $\{S_1(\underline{X}), \ldots, S_p(\underline{X})\}$ is a complete minimally sufficient statistic for the parametric family $\mathcal{F} = \{f_\theta; \theta \in \Theta\}$ and for all $\theta \in \Theta$ $T(\underline{X})$ is an unbiased estimator estimator of θ then $\widehat{\theta}[\underline{X}] = \mathrm{E}\left[T(\underline{X})|s(\underline{X})\right]$ is the unique minimum variance unbiased estimator (UMVUE) for all $\theta \in \Theta$.*

PROOF. Suppose $\phi[s(\underline{X})]$ is an unbiased estimator of θ with a smaller variance than $\widehat{\theta}[s(\underline{X})]$ then taking differences it is clear by unbiasedness that

$$\mathrm{E}\left(\widehat{\theta}[s(\underline{X})] - \widehat{\phi}[s(\underline{X})]\right) = 0.$$

However, completeness immediately implies that $\widehat{\phi}[s(\underline{x})] - \widehat{\theta}[s(\underline{x})] = 0$ almost surely. Thus proving the result. □

This theorem tells us if the conditions are satisfied, then for every $\theta \in \Theta$, the estimator $T(\underline{X})$ will give the smallest variance amongst all estimators which are unbiased. The condition that the comparison is done over all *unbiased* estimators is very important. If we drop the relax the condition to allow biased estimators then improvements are possible.

Remark 1.5.1 *Consider the example of the truncated exponential in Example 1.4.8. In this example, there are two sufficient statistics, $s_1(\underline{X}) = \sum_{i=1}^n X_i$ and $s_2(\underline{X}) = \max_i X_i$ for the unknown parameter θ, neither are ancillary in the sense that their marginal distributions depend on θ. Thus both sufficient statistics can be used to estimate θ.*

In general if there are two sufficient statistics for one parameter, θ, and neither of the sufficient statistics are ancillary, then usually one can use either sufficient statistic as a means of constructing an estimator of θ.

Exercise 1.6 *In the above remark, calculate the expectation of $\max_i X_i$ and $\sum_i X_i$ and use this to propose two different estimators for θ.*

Example 1.5.5 *For the curious, `http://www.tandfonline.com/doi/abs/10.1080/00031305.2015.1100683?journalCode=utas20` give an example of minimal sufficient statistics which are not complete and use the Rao-Blackwell theorem to improve on the estimators (though the resulting estimator does not have minimum variance for all θ in the parameter space).*

1.6 The exponential family of distributions

We now expand a little on the exponential family described in the previous section. In a nutshell the exponential family is where the parameters of interest and the random variables of the log-likelihood are separable. As we shall see below, this property means the number of minimal sufficient statistics will always be finite and estimation relatively straightforward.

1.6.1 The natural/canonical exponential family

We first define the one-dimension natural exponential family

$$f(x;\theta) = \exp\left(s(x)\theta - \kappa(\theta) + c(x)\right), \qquad (1.15)$$

where $\kappa(\theta) = \log \int \exp(s(x)\theta + c(x))d\nu(x)$ and $\theta \in \Theta$ (which define below). If the random variable is continuous, then typically $\nu(x)$ is the Lebesgue measure, on the other hand if it is discrete then $\nu(x)$ is the point mass, for example for the Poisson distribution $d\nu(x) = \sum_{k=0}^{\infty} \delta_k(x)dx$.

Example 1.6.1 *We now give an example of a distribution which immediately has this parameterisation. The exponential distribution has the pdf is $f(x;\lambda) = \lambda \exp(-\lambda x)$, which can be written as*

$$\log f(x;\lambda) = (-x\lambda + \log\lambda) \qquad \lambda \in (0,\infty)$$

Therefore $s(x) = -x$ *and* $\kappa(\lambda) = -\log \lambda$.

The parameter space for this family is defined as

$$\Theta = \left\{\theta; \int \exp\left(s(x)\theta + c(x)\right) d\nu(x) < \infty\right\},$$

in other words all parameters where this integral is finite and thus gives a well defined density. The role of $\kappa(\theta)$ is as a normaliser and ensures that density integrates to one i.e

$$\int f(x;\theta) d\nu(x) = \int \exp\left(s(x)\theta - \kappa(\theta) + c(x)\right) d\nu(x) = \exp(-\kappa(\theta)) \int \exp\left(s(x)\theta + c(x)\right) d\nu(x) = 1$$

we see that

$$\kappa(\theta) = \log \int \exp\left(s(x)\theta + c(x)\right) d\nu(x)$$

By using the factorisation theorm, we can see that $\sum_{i=1}^{n} s(X)$ is the sufficient statistic for the family $\mathcal{F} = \{f(x;\theta); \theta \in \Theta\}$. The one-dimensional natural exponential is only a function of one-parameter. The p-dimensional natural exponential generalisation is defined as

$$f(x;\theta) = \exp\left[\mathbf{s}(x)'\theta - \kappa(\theta) + c(x)\right]. \tag{1.16}$$

where $\mathbf{s}(x) = (s_1(x), \dots, s_p(x))$ is a vector which is a function of x and $\theta = \{\theta_1, \dots, \theta_p\}$ is a p-dimension parameter. The parameter space for this family is defined as

$$\Theta = \left\{\theta; \int \exp\left(\mathbf{s}(x)'\theta + c(x)\right) d\nu(x) < \infty\right\},$$

again $\kappa(\theta)$ is such that

$$\kappa(\theta) = \log \int \exp\left(\sum_{j=1}^{p} s_j(x)\theta_j + c(x)\right) d\nu(x)$$

and ensures that the density integrates to one.

Lemma 1.6.1 *Consider the* p*-dimension family* $\mathcal{F}$ *of densities where* $\mathcal{F} = \{f(x;\theta); \theta = (\theta_1, \dots, \theta_p) \in \Theta\}$ *with*

$$f(x;\theta) = \exp\left[\mathbf{s}(x)'\theta - \kappa(\theta) + c(x)\right].$$

By using the Factorisation theorem it can be seen that $\{\sum_{i=1}^{n} s_1(X_i), \dots, \sum_{i=1}^{n} s_p(X_i)\}$ *are the sufficient statistics for* $\mathcal{F}$.

However, once one goes beyond dimension one, there can arise redundancy in the representation. For example, consider the two-dimensional exponential family defined by

$$\mathcal{F} = \left\{ f(x;\theta_1,\theta_2) = \exp\left(\alpha s(x)\theta_1 + \beta s(x)\theta_2 - \kappa(\theta_1,\theta_2) + c(x)\right); (\theta_1,\theta_2) \in \Theta \right\},$$

since $f(x;\theta_1,\theta_2)$ is a density, then

$$\kappa(\theta_1,\theta_2) = \log\left(\int \exp\left[(\theta_1\alpha + \theta_2\beta)s(x) + c(x)\right] d\nu(x)\right).$$

We see that $\kappa(\theta_1,\theta_2)$ is the same for all θ_1,θ_2 such that $(\theta_1\alpha + \theta_2\beta)$ is constant. Thus for all parameters

$$(\theta_1,\theta_2) \in \Theta_C = \{(\theta_1,\theta_2); (\theta_1,\theta_2) \in \Theta, (\theta_1\alpha + \theta_2\beta) = C\}$$

the densities $f(x;\theta_1,\theta_2)$ are the same. This means the densities in $\mathcal{F}$ are *not identifiable.*

Definition 1.6.1 *A class of distributions/model* $\mathcal{F} = \{f(x;\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ *is non-identifiable if there exists a* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ *such that* $f(x;\boldsymbol{\theta}_1) = f(x;\boldsymbol{\theta}_2)$ *for all* $x \in \mathbb{R}$.

Non-identifiability of a model can be hugely problematic in estimation. If you cannot identify the parameter, then a likelihood can have several maximums, the limit of the estimator is no longer well defined (it can be estimating several different estimators).

In the above example, a minimal representation of the above function is the one-dimensional exponential family

$$\mathcal{F} = \left\{ f(x;\theta) = \exp\left[\theta s(x) - \kappa(\theta) + c(x)\right]; \theta \in \Theta \right\}.$$

Therefore to prevent this over parameterisation and lack of identifiability we assume that the functions $\{s_j(x)\}_{j=1}^{p}$ in the canonical representation are linear independent i.e. there does not exist constants $\{\alpha_j\}_{j=1}^{p}$ and C such that

$$\sum_{j=1}^{p} \alpha_j s_j(x) = C$$

for all x in the domain of X. This representation is called minimal. As can be seen from the example above, if there is linear dependence in $\{s_i(x)\}_{i=1}^{p}$, then it is easy to find an alternative representation which is of a lower dimension and canonical.

Lemma 1.6.2 *If $\{X_i\}_{i=1}^n$ are iid random variables, which belong to the p-dimensional exponential family that has the form*

$$\begin{aligned}\mathcal{F} &= \left\{f(x;\theta) = \exp\left[\sum_{j=1}^{p}\theta_j s_j(x) - \kappa(\theta) + c(x)\right];\theta\in\Theta\right\}\\ \textit{where } \Theta &= \left\{\theta;\int \exp\left[\sum_{j=1}^{p}\theta_j s_j(x) + c(x)\right]d\nu(x) < \infty\right\}\end{aligned}$$

and this is a minimal representation. Then the minimal sufficient statistics are $\{\sum_{i=1}^n s_1(X_i), \ldots, \sum_{i=1}^n s_p(X_i)\}$.

If the parameter space Θ is an open set, then the family of distributions $\mathcal{F}$ is called *regular*. The importance of this will become clear in the next chapter. The parameter space Θ is often called the *natural* parameter space. Note that the the natural parameter space is convex. This means if $\theta_1, \theta_2 \in \mathcal{N}$ then for any $0 \leq \alpha \leq 1$ $\alpha\theta_1 + (1-\alpha)\theta_2 \in \mathcal{N}$. This is proved by using Hölder's inequality and that $\kappa(\theta_1), \kappa(\theta_2) < \infty$ and $e^{\kappa(\theta)} = \int \exp(\theta'\mathbf{s}(x) + c(x))d\nu(x)$.

Remark 1.6.1 *Convexity of the parameter space basically mean if $\theta_1, \theta_2 \in \mathbb{R}^d$ and both of them are such that give a well defined density then for any convex combination (think a line between the two points) will also yield a well defined density.*

1.6.2 Moments of the canonical representation

In this section we derive the moments of the canonical exponential family using some cute tricks. To simplify the exposition we focus on canonical exponential families of dimension one, though the same result holds for higher dimensions.

Definition 1.6.2 (Cumulant generating function) *The cumulant generating function (for a univariate random variable) is defined as $C_X(t) = \log \mathrm{E}[e^{tX}]$. The power series expansion of the cumulant generating function is*

$$C_X(t) = \log \mathrm{E}[e^{tX}] = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!},$$

where $\kappa_n = C_X^{(n)}(0)$ (analogous to the moment generating function). Note that $\kappa_1(X) = \mathrm{E}[X]$, $\kappa_2(X) = \mathrm{var}[X]$ and $\kappa_j = \kappa_j(X, \ldots, X)$. X is a Gaussian random variable iff $\kappa_j = 0$ for $j \geq 3$.

We use the above in the lemma below.

Lemma 1.6.3 *[Moment generating functions] Suppose that X is a random variable with density*

$$f(x;\theta) = \exp\left(s(x)\theta - \kappa(\theta) + c(x)\right), \theta \in \Theta \tag{1.17}$$

where

$$\Theta = \left\{\theta; \int \exp\left(s(x)\theta - \kappa(\theta) + c(x)\right) d\nu(x) < \infty\right\},$$

. If $\theta \in int(\Theta)$ (the interior of θ, to ensure that it is an open set),

(i) Then the moment generating function of $s(X)$ is

$$\mathrm{E}\left[\exp(s(X)t)\right] = M_{s(X)}(t) = \exp\left[\kappa(t+\theta) - \kappa(\theta)\right]$$

(ii) The cumulant generating function is

$$\log \mathrm{E}\left[\exp(s(X)t)\right] = C_{s(X)}(t) = \kappa(t+\theta) - \kappa(\theta).$$

(iii) Furthermore $\mathrm{E}_\theta[s(X)] = \kappa'(\theta) = \mu(\theta)$ and $\mathrm{var}_\theta[s(X)] = \kappa''(\theta)$.

(iv) $\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} = -\kappa''(\theta)$, thus $\log f(x;\theta)$ has a negative definite Hessian.

This result easily generalizes to p-order exponential families.

PROOF. We choose t sufficiently small such that $(\theta + t) \in int(\Theta)$, since $(\theta + t)$ belongs to the parameter space, then $f(y;(\theta+t))$ is a valid density/distribution. The moment generating function of $s(X)$ is

$$M_{s(X)}(t) = \mathrm{E}\left[\exp(ts(X))\right] \;\; = \;\; \int \exp(ts(x))\exp(\theta s(x) - \kappa(\theta) + c(x))d\nu(x).$$

Taking $\exp(-\kappa(\theta))$ out of the integral and adding and subtracting $\exp(\kappa(\theta+t))$ gives

$$\begin{aligned} M_{s(X)}(t) &= \exp(\kappa(\theta+t) - \kappa(\theta)) \int \exp((\theta+t)s(x) - \kappa(\theta+t) + c(x))d\nu(x) \\ &= \exp(\kappa(\theta+t) - \kappa(\theta)), \end{aligned}$$

since $\int \exp((\theta+t)y - \kappa(\theta+t) + c(y))dy = \int f(y;(\theta+t))dy = 1$. To obtain the moments we recall that the derivatives of the cumulant generating function at zero give the cumulant of the random variable. In particular $C'_{s(X)}(0) = \mathrm{E}[s(X)]$ and $C''_{s(X)}(0) = \mathrm{var}[s(X)]$. Which immediately gives the result. □

1.6.3 Reparameterisations and examples

We recall that a distribution belongs to the exponential family $\mathcal{F}$ if $f \in \mathcal{F}$ can be written as

$$f(x;\omega) = \exp\left(\sum_{j=1}^{p}\phi_j(\omega)s_j(x) - A(\omega) + c(x)\right),$$

where $\omega = (\omega_1, \ldots, \omega_q)$ are the q-dimensional parameters. Since this family of distributions is parameterized by ω and not θ it is not in natural form. With the exponential distribution there are very few distributions which immediately have a canonical/natural exponential representation. However, it can be seen (usually by letting $\theta_j = \phi_j(\omega)$) that all exponential families of distributions can be reparameterized such that it has a canonical/natural representation. Moreover by making sufficient transformations, to ensure the sufficient statistics do not satisfy any linear constraints, the representation will be minimal (see the monograph `http://www.jstor.org/stable/pdf/4355554.pdf?acceptTC=true`, Lawrence Brown (1986), Proposition 1.5, for the precise details). Let $\Phi(\omega) = (\phi_1(\omega), \ldots, \phi_p(\omega))$ and Ω denote the parameter space of ω. Then we see that $\Phi : \Omega \to \Theta$, where Θ is the natural parameter space defined by

$$\Theta = \left\{\theta; \int \exp\left(\sum_{j=1}^{p}\theta_j s_j(x) + c(x)\right)dx < \infty\right\}.$$

Thus Φ is an injection (one-to-one) mapping from Ω to Θ. Often the mapping is a bijection (injective and surjective), in which case $p = q$. In such cases, the exponential family is said to have *full rank* (technically, full rank requires that $\mathcal{N}$ is an open set; when it is closed strange things can happen on the boundary of the set).

If the image of Φ, $\Phi(\Omega)$, is not a linear subset of $\mathcal{N}$, then the exponential family $\mathcal{F}$ is called a curved exponential.

Recall that θ is a function of the d-dimension parameters ω if

(i) If $p = d$ then the exponential family is said to have full rank. In this case the sufficient statistics are complete.

(i) If $p > d$ then the exponential family is said to be a curved exponential family. This means the image $\Phi(\Omega)$ (the parameter space of ω onto θ) is not a linear subset of Θ For curved exponential families there are nonlinear constraints between the unknown parameters.

When the exponential family is curved it is *not complete* (see Exercise 1.6.2). The implication of this is that there is no unique unbiased estimator (in terms of the sufficient statistics), which will give the minimal variance for all parameters in the parameter space. See Brown (1986), Theorem 1.9 (page 13) for details on the above.

Lemma 1.6.4 *If a distribution belongs to the exponential family, and the sufficient statistics are linearly independent then the sufficient statistics are minimally sufficient.*

Example 1.6.2 (The normal distribution) *We recall that S_{xx}, S_x are the sufficient statistics of the normal family of distributions, where $(S_{xx}, S_x) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$. It is clear that (S_{xx}, S_x) are linearly independent (i.e. for no linear combination $\alpha S_{xx} + \beta S_x = 0$ for all S_x and S_{xx}), thus by Lemma 1.6.4 they are minimally sufficient.*

Exercise 1.7 *Suppose that $\{X_i\}_{i=1}^n$ are iid normal random variables where the ratio between mean and standard deviation $\gamma = \sigma/\mu$ is known. What are the minimal sufficient statistics?*

1.6.4 Examples

By making appropriate transformations, we show that the below well known distributions can be written in natural form.

(i) The exponential distribution is already in natural exponential form and the parameter space is $\Theta = (0, \infty)$.

(ii) For the binomial distribution where $X \sim Bin(n, p)$ we note

$$\log f(x; p) = x \log p + (n - x) \log(1 - p) + \log \binom{n}{x}.$$

One natural parameterisation is to let $\theta_1 = \log p$, $\theta_2 = \log(1 - p)$ with sufficient statistics x and $(n - x)$. This a two-dimensional natural exponential representation. However we see that the sufficient statistics are subject to a linear constraint, namely $s_1(x) + s_2(x) = x + (n - x) = n$. Thus this representation is not minimal. Instead we rearrange $\log f(x; p)$

$$\log f(x; p) = x \log \frac{p}{1 - p} + n \log(1 - p) + \log \binom{n}{x}.$$

Let $\theta = \log(\frac{p}{1-p})$, since $\theta(p) = \log(\frac{p}{1-p})$ is invertible this gives the natural representation

$$\log f(x;\theta) = \left[x\theta - n\log\left(1+\exp(\theta)\right) + \log\binom{n}{x}\right].$$

Hence the parameter of interest, $p \in (0,1)$, has been transformed, to $\theta \in (-\infty,\infty)$. The natural parameter space is $\Theta = (-\infty,\infty)$. The sufficient statistic is $\sum_i X_i$. $d\nu(x) = dx$, the Lebesgue measure.

(iii) The normal family of distributions can be written as

$$\log f(x;\mu,\sigma) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\log 2\pi. \qquad (1.18)$$

In this case the natural exponential parametrisation is $\mathbf{x} = (-\frac{1}{2}x^2, x)$, $\theta = (\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}) = (\theta_1,\theta_2)$ and $\kappa(\theta_1,\theta_2) = \theta_2^2/(2\theta_1) - 1/2\log(\theta_1)$. In this case $\Theta = (0,\infty)\times(-\infty,\infty)$. The sufficient statistics are $\sum_i X_i$ and $\sum_i X_i^2$. $d\nu(x) = dx$, the Lebesgue measure.

(iv) The multinomial distribution can be written as

$$\begin{aligned}\log f(x_1,\ldots,x_p;\pi) &= \sum_{i=1}^{p} x_i \log\pi_i + \log n! - \sum_{i=1}^{p} x_i! \\ &= \sum_{i=1}^{p-1} x_i \log\frac{\pi_i}{\pi_p} + n\log\pi_p + \log n! - \sum_{i=1}^{p} x_i!.\end{aligned}$$

For $1 \le i \le p-1$ let $\theta_i = \log\pi_i/\pi_p$ then the natural representation is

$$\log f(x_1,\ldots,x_p;\pi) = \sum_{i=1}^{p-1}\theta_i x_i - n\log\left(1+\sum_{i=1}^{p-1}\exp(-\theta_i)\right) + \log n! - \sum_{i=1}^{p} x_i!$$

and the parameters space is $\mathbb{R}^{p-1}$. The sufficient statistics are $\sum_i X_{i1},\ldots,\sum_i X_{i,p-1}$. The point measure is $d\nu(x) = \sum_{j_1,j_2,\ldots,j_{p-1}=1}^{n} \delta_{j_1}(x_1)\ldots\delta_{j_{p-1}}(x_{j-1})\delta_{[0,n]}(x_1+\ldots+x_{p-1})dx_1\ldots dx_{p-1}$.

Note that one can also write the multinomial as

$$\log f(x_1,\ldots,x_p;\pi) = \sum_{i=1}^{p}\theta_i x_i + \log n! - \sum_{i=1}^{p} x_i!,$$

where $\theta_i = \log\pi_i$. However this is not in minimal form because $n - \sum_{i=1}^{n} x_i = 0$ for all $\{x_i\}$ in the sample space; thus they are not linearly independent.

(v) The censored exponential distribution. $X \sim Exp(\lambda)$ (density of X is $f(x;\lambda) = \exp[-x\lambda + \log\lambda]$), however X is censored at a known point c and Y is observed where

$$Y = \begin{cases} X & X \leq c \\ c & X > c \end{cases}$$

and c is assumed *known*. Suppose we observe $\{Y_i, \delta_i\}$, using (2.3) we have

$$\mathcal{L}(\lambda) = -\sum_{i=1}^{n}(1-\delta_i)\lambda Y_i + (1-\delta_i)\log\lambda - \delta_i c\lambda.$$

We recall that by definition of Y when $\delta = 1$ we have $Y = c$ thus we can write the above as

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

Thus when the sample size is n the sufficient statistics are $s_1(Y,\delta) = \sum_i Y_i$, $s_2(Y,\delta) = \sum_i \delta_i = \sum_i I(Y_i \geq c))$. The natural parameterisation is $\theta_1 = -\lambda$, $\theta_2 = -\log(-\lambda)$ and $\kappa(\theta_1,\theta_2) = \theta_2 = \frac{1}{2}(-\log(-\theta_1) + \theta_2)$ (thus we see that parameters are subject to nonlinear constraints). As $s_1(Y,\delta) = \sum_i Y_i, s_2(Y,\delta) = \sum_i \delta_i$ are not linearly dependent this means that the censored exponential distribution has a 2-dimensional natural exponential representation. The measure is $d\nu(x,\delta) = dx[\delta_0(\delta)d\delta + \delta_1(\delta)d\delta]$

However since the parameter space is not the entire natural parameter space $\mathcal{N} = (-\infty,0)\times(-\infty,0)$ (since $\theta_1(\lambda) = \lambda$ and $\theta_2(\lambda) = \log\lambda$) but a subset of it, then the family is curved and thus the sufficient statistics are not complete. This means that there is no unique unbiased estimator with minimal variance.

(vi) The von-Mises distributions are distributions defined on a sphere. The simplest is the von-Mises distribution defined on a 1-d circle

$$f(x;\kappa,\mu) = \frac{1}{2\pi I_0(\kappa)}\exp\left(\kappa\cos(x-\mu)\right) \qquad x \in [0,2\pi],$$

where I_0 is a Bessel function of order zero ($\kappa > 0$ and $\mu \in \mathbb{R}$). We will show that it has a natural 2-dimensional exponential representation

$$\begin{aligned} \log f(x;\kappa,\mu) &= \kappa\cos(x-\mu) - \log 2\pi I_0(\kappa) \\ &= \kappa\cos(x)\cos(\mu) + \kappa\sin(x)\sin(\mu) - \log 2\pi I_0(\kappa). \end{aligned}$$

Let $s_1(x) = \cos(x)$ and $s_2(x) = \sin(x)$ and we use the parameterisation $\theta_1(\kappa,\mu) = \kappa\cos\mu$, $\theta_2(\kappa,\mu) = \kappa\sin\mu$, $\kappa(\theta_1,\theta_2) = -\log 2\pi I_0(\sqrt{\theta_1^2+\theta_2^2})$. The sufficient statistics are $\sum_i \cos(X_i)$ and $\sum_i \sin(X_i)$ and $(\theta_1,\theta_2)\in\mathbb{R}^2$ The measure is $d\nu(x) = dx$.

(vii) Consider the inflated zero Poisson distribution which has the log-likelihood

$$\begin{aligned}
&\mathcal{L}(\underline{Y};\lambda,p)\\
=&\sum_{i=1}^n I(Y_i=0)\log\left(p+(1-p)e^{-\lambda}\right)+\sum_{i=1}^n I(Y_i\neq 0)\left(\log(1-p)+\log\frac{\lambda^{Y_i}e^{-\lambda}}{Y_i!}\right)\\
=&\sum_{i=1}^n [1-I(Y_i\neq 0)]\log\left(p+(1-p)e^{-\lambda}\right)+\log\lambda\sum_{i=1}^n I(Y_i\neq 0)Y_i\\
&+(\log(1-p)-\lambda)\sum_{i=1}^n I(Y_i\neq 0)-\sum_{i=1}^n I(Y_i\neq 0)\log Y!\\
=&\left\{-\log\left(p+(1-p)e^{-\lambda}\right)+(\log(1-p)-\lambda)\right\}\sum_{i=1}^n I(Y_i\neq 0)\\
&+\log\lambda\sum_{i=1}^n I(Y_i\neq 0)Y_i+n\underbrace{\log\left(p+(1-p)e^{-\lambda}\right)}_{-\kappa(\cdot)}-\sum_{i=1}^n I(Y_i\neq 0)\log Y!.
\end{aligned}$$

This has a natural 2-dimension exponential representation. Let

$$\begin{aligned}
\theta_1 &= \left\{-\log\left(p+(1-p)e^{-\lambda}\right)+(\log(1-p)-\lambda)\right\}\\
\theta_2 &= \log\lambda
\end{aligned}$$

with sufficient statistics $s_1(\underline{Y}) = \sum_{i=1}^n I(Y_i\neq 0)$, $s_2(\underline{Y}) = \sum_{i=1}^n I(Y_i\neq 0)Y_i$. The parameter space is $(\theta_1,\theta_2)\in(-\infty,0]\times(-\infty,\infty)$, the 0 end point for θ_1 corresponds to $p=0$. If we allowed $p<0$ (which makes no sense), then the parameter space for θ_1 can possibly be greater than 0, but this makes no sense. If calculated correctly

$$\kappa(\theta_1,\theta_2) = -\log\left(\frac{e^{\theta_1}-\theta_2^{-1}}{1-\theta_2^{-1}}(1-e^{-e^{\theta_2}})+e^{-e^{\theta_2}}\right).$$

The measure is the point mass $d\nu(x) = \sum_{j=0}^\infty \delta_j(x)dx$.

(viii) Suppose (X_i,Y_i) are iid random variables with densities $\theta\exp(-\theta x)$ and $\theta^{-1}\exp(-\theta^{-1}y)$ respectively. Then the joint density is $f(x,y) = \exp(-\theta x-\theta^{-1}y)$. The slight difference here is that there are two random variables at play. But this not change the analysis. The natural exponential parameterisation is

$$f(x,y;\theta_1,\theta_2) = \exp\left(-\theta_1 x-\theta_2 y\right)\quad \theta_1,\theta_2>0$$

subject to tthe constraint $\theta_1\theta_2 = 1$. The log-likelihood is

$$\mathcal{L}_n(\theta) \;=\; -\theta_1 \sum_{i=1}^{n} X_i - \theta_2 \sum_{i=1}^{n} Y_i,$$

thus the minimal sufficient statistics are $s_1(\underline{X}, \underline{Y}) = \sum_i^n X_i$ and $s_2(\underline{X}, \underline{Y}) = \sum_i^n Y_i$. However, the parameter space is $(\theta, 1/\theta)$ which is not a linear subset in $(R^+)^2$, thus it is not complete. This is a curved exponential. The measure is $d\nu(x,y) = dxdy$.

1.6.5 Some additional properties of the exponential family

We first state some definitions which we use later.

Definition 1.6.3 (Concave, convex functions and the Hessian)

- *A function is said to be concave if*

$$f\left(y + \alpha(x-y)\right) = f\left(\alpha x + (1-\alpha)y\right) \geq \alpha f(x) + (1-\alpha)f(y) = f(y) + \alpha\left[f(x) - f(y)\right].$$

and strictly concave if

$$f\left(y + \alpha(x-y)\right) = f\left(\alpha x + (1-\alpha)y\right) > \alpha f(x) + (1-\alpha)f(y) = f(y) + \alpha\left[f(x) - f(y)\right].$$

For $d = 1$ this can be seen as the curve of f lying above the tangent between the points $(x, f(x))$ and $(y, f(y))$. This immediately implies that if $y > x$, then

$$f(y) - f(x) \;<\; \frac{f\left(x + \alpha(y-x)\right) - f(x)}{\alpha} \Rightarrow \frac{f(y) - f(x)}{y - x} < \frac{f\left(x + \alpha(y-x)\right) - f(x)}{\alpha(y-x)}$$

for all $0 < \alpha < 1$. Thus

$$\frac{f(y) - f(x)}{y - x} < f'(x).$$

- *The Hessian of a function of p variables $f : \mathbb{R}^p \to \mathbb{R}$ is its second derivative $\nabla^2_\theta f(\theta) = \left\{\frac{\partial^2 f(\theta)}{\partial\theta_i \partial\theta_j}; 1 \leq i,j \leq p\right\}$.*

- *Examples of concave functions are $f(x) = -x^2$ (for $x \in (-\infty, \infty)$) and $f(x) = \log x$ for $x \in (0, \infty)$. Observe that $-x^2$ is maximised at $x = 0$, whereas the maximum of $\log x$ lies outside of the interval $(0, \infty)$.*

A function is a concave function if and only if the Hessian, $\nabla^2_\theta f$, is negative semi-definite.

We now show consider the properties of the log likelihood of the natural exponential.

(i) We now show that second derivative of log-likelihood of a function from the natural exponential family has a negative definite Hessian. It is straightforward to show that the second derivative of the log-likelihood is

$$\nabla^2_\theta \mathcal{L}_n(\theta) = -\sum_{i=1}^{n} \nabla^2_\theta \kappa(\theta) = -n\nabla^2_\theta \kappa(\theta).$$

From Lemma 1.6.3 we see that for all $\theta \in \Theta$ $\nabla^2_\theta \kappa(\theta)$ corresponds to the variance of a random variable X_θ with density f_θ. This implies that $\nabla^2_\theta \kappa(\theta) \geq 0$ for all $\theta \in \Theta$ and thus the Hessian $\nabla^2_\theta \mathcal{L}_n(\theta)$ is semi-negative definite. We will later show that this means that $\mathcal{L}_n(\underline{X};\theta)$ can easily be maximised. Thus for the natural exponential family the observed and expected Fisher information are the same.

Examples of different concave likelihoods are given in Figure 1.3. Observe that the maximum may not always lie within the interior of the parameter space.

(ii) We recall that θ is a function of the parameters ω. Therefore the Fisher information for ω is related, but not equal to the Fisher information for θ. More precisely, in the case of the one-dimension exponential family the likelihood is

$$\mathcal{L}_n(\theta(\omega)) = \theta(\omega)\sum_{i=1}^{n} s(X_i) - n\kappa(\theta(\omega)) + n\sum_{i=1}^{n} c(X_i).$$

Therefore the second derivative with respect to ω is

$$\frac{\partial^2 \mathcal{L}_n[\theta(\omega)]}{\partial \omega^2} = -n\frac{\partial \theta'}{\partial \omega}\frac{\partial^2 \kappa(\theta)}{\partial \theta^2}\frac{\partial \theta}{\partial \omega} + \left(\sum_{i=1}^{n} X_i - n\frac{\partial \kappa(\theta)}{\partial \theta}\right)\frac{\partial^2 \theta}{\partial \omega^2}.$$

Recall that $\mathrm{E}[\left(\sum_{i=1}^{n} X_i - n\frac{\partial \kappa(\theta)}{\partial \theta}\right) = n\mathrm{E}[X_i] - n\kappa'(\theta) = 0$. Using this we have

$$I(\omega) \;=\; -\mathrm{E}\left(\frac{\partial^2 \mathcal{L}_n[\theta(\omega)]}{\partial \omega^2}\right) = n\frac{\partial \theta'}{\partial \omega}\frac{\partial^2 \kappa(\theta)}{\partial \theta^2}\frac{\partial \theta}{\partial \omega}.$$

In this case the observed and expected Fisher information matrices are not the same.

However, if there is a diffeomorphism between the space of θ and ω, negative definite $\nabla^2_\theta \mathcal{L}_n(\theta) = \frac{\partial^2 \kappa(\theta)}{\partial \theta^2}$ implies negative definite $\nabla^2_\omega \mathcal{L}_n(\theta(\omega))$. This is because when there is a diffeomorphism (a continuous invertible mapping between two spaces), the

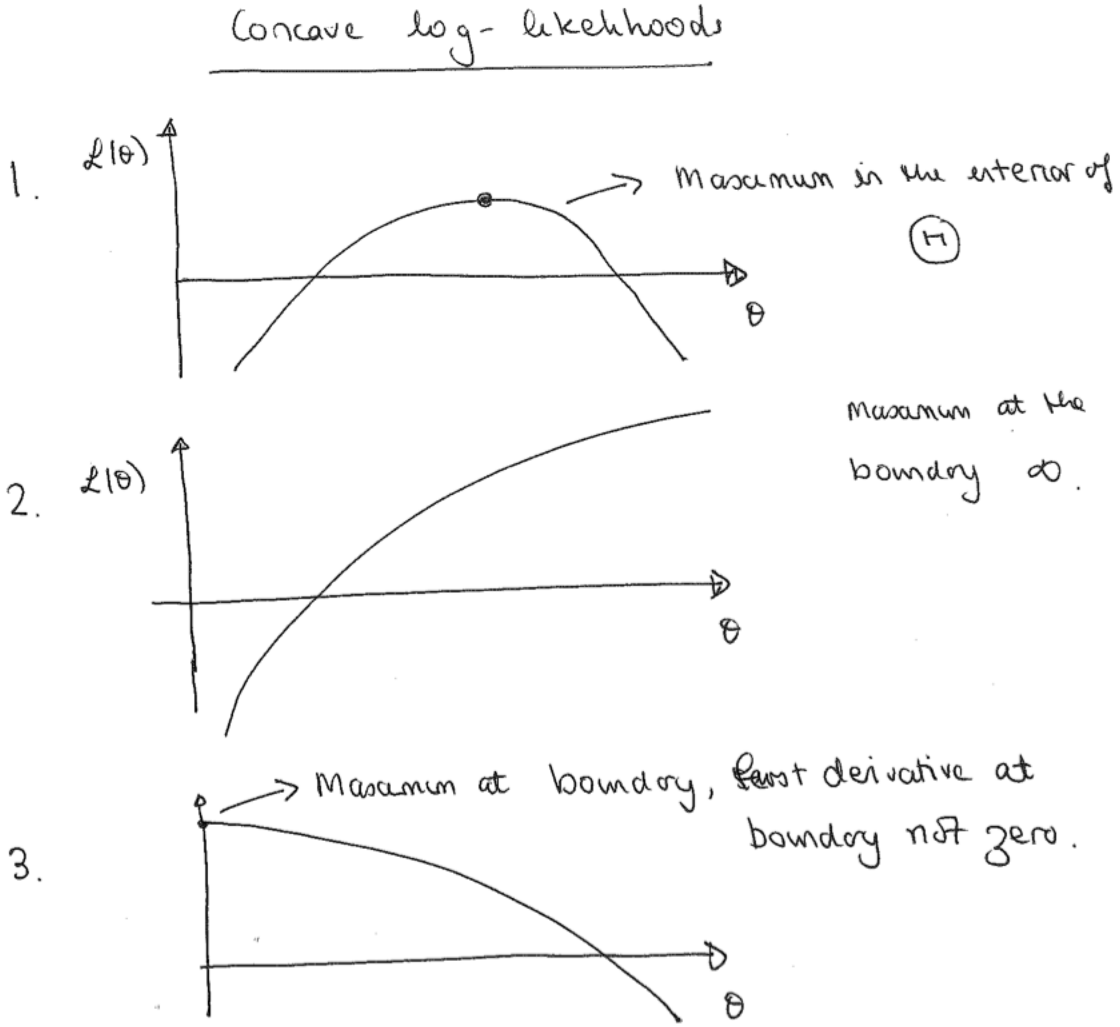

Concave log- likelihoods
1.
L(θ)
Maximum in the interior of Θ
θ
2.
L(θ)
Maximum at the boundary ∞.
θ
3.
Maximum at boundary, first derivative at boundary not zero.
θ

Figure 1.3: Examples of different concave likelihoods

eigen-values of the corresponding Hessian matrices will change, however *the signs will not*. Therefore if $\nabla^2_{\theta}\mathcal{L}_n(\theta)$ is negative definite then so is its reparametrisation $\nabla^2_{\omega}\mathcal{L}_n(\theta(\omega))$.

(ii) The natural parameter space $\mathcal{N}$ is convex, this means if $\theta_1, \theta_2 \in \Theta$ then $\alpha\theta_1 + (1-\alpha)\theta_2 \in \Theta$ for $0 \leq \alpha \leq 1$ (easily proved using Hölder's inequality).

(iii) The function $\kappa(\theta)$ is convex (easily proved using that $\kappa(\theta) = \log \int \exp(\theta s(x) + c(x))d\nu(x)$ and Hölder's inequality).

1.7 The Bayesian Cramer-Rao inequality

The classical Cramér-Rao inequality is useful for assessing the quality of a given estimator. But from the derivation we can clearly see that it only holds if the estimator is unbiased.

As far as I am aware no such inequality exists for the *mean squared error* of estimators that are biased. For example, this can be a problem in nonparametric regression, where estimators in general will be biased. How does one access the estimator in such cases? To answer this question we consider the Bayesian Cramer-Rao inequality. This is similar to the Cramer-Rao inequality but does not require that the estimator is unbiased, so long as we place a prior on the parameter space. This inequality is known as the Bayesian Cramer-Rao or van-Trees inequality (see [?] and [?]).

Suppose $\{X_i\}_{i=1}^n$ are random variables with distribution function $L_n(\underline{X};\theta)$. Let $\tilde{\theta}(\underline{X})$ be an estimator of θ. We now Bayesianise the set-up by placing a prior distribution on the parameter space Θ, the density of this prior we denote as λ. Let $\mathrm{E}[g(\underline{x})|\theta] = \int g(\underline{x})L_n(\underline{x}|\theta)d\underline{x}$ and E_{λ} denote the expectation over the density of the parameter λ. For example

$$\mathrm{E}_{\lambda}\mathrm{E}[\widetilde{\theta}(X)|\theta] = \int_a^b \int_{\mathbb{R}^n} \widetilde{\theta}(\underline{x})L_n(\underline{x}|\theta)d\underline{x}\lambda(\theta)d\theta.$$

Now we place some assumptions on the prior distribution λ.

Assumption 1.7.1 *θ is defined over the compact interval $[a,b]$ and $\lambda(x) \to 0$ as $x \to a$ and $x \to b$ (so $\lambda(a) = \lambda(b) = 0$).*

Theorem 1.7.1 *Suppose Assumptions 1.3.1 and 1.7.1 hold. Let $\widetilde{\theta}(\underline{X})$ be an estimator of θ. Then we have*

$$\mathrm{E}_{\lambda}\left[\mathrm{E}_{\theta}\left\{\left(\widetilde{\theta}(\underline{X})-\theta\right)^{2}\middle|\theta\right\}\right] \geq \left[\mathrm{E}_{\lambda}[I(\theta)]+I(\lambda)\right]^{-1}$$

where

$$\begin{aligned}\mathrm{E}_{\lambda}[I(\theta)] &= \int\int\left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\right)^{2} L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\ \textit{and}\quad I(\lambda) &= \int\left(\frac{\partial\log\lambda(\theta)}{\partial\theta}\right)^{2}\lambda(\theta)d\theta.\end{aligned}$$

PROOF. First we derive a few equalities. We note that under Assumption 1.7.1 we have

$$\int_a^b \frac{dL_n(\underline{x};\theta)\lambda(\theta)}{d\theta}d\theta = L_n(\underline{x};\theta)\lambda(\theta)\Big]_a^b = 0,$$

thus

$$\int_{\mathbb{R}^n}\widetilde{\theta}(\underline{x})\int_a^b\frac{\partial L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}d\theta d\underline{x} = 0. \qquad (1.19)$$

Next consider $\int_{\mathbb{R}^n}\int_a^b\theta\frac{\partial L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}d\theta d\underline{x}$. Using integration by parts we have

$$\begin{aligned}&\int_{\mathbb{R}^n}\int_a^b\theta\frac{dL_n(\underline{x};\theta)\lambda(\theta)}{d\theta}d\theta d\underline{x} = \int_{\mathbb{R}^n}\left(\theta L_n(\underline{x};\theta)\lambda(\theta)\Big]_a^b\right)d\underline{x} - \int_{\mathbb{R}^n}\int_a^b L_n(\underline{x};\theta)\lambda(\theta)d\theta dx \\ &= -\int_{\mathbb{R}^n}\int_a^b L_n(\underline{x};\theta)\lambda(\theta)d\theta d\underline{x} = -1. \qquad (1.20)\end{aligned}$$

Subtracting (1.20) from (1.19) we have

$$\int_{\mathbb{R}^n}\int_a^b\left(\widetilde{\theta}(\underline{x})-\theta\right)\frac{\partial L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}d\theta d\underline{x} = \int_{\mathbb{R}^n}\int_a^b L_n(\underline{x};\theta)\lambda(\theta)d\theta d\underline{x} = 1.$$

Multiplying and dividing the left hand side of the above by $L_n(\underline{x};\theta)\lambda(\theta)$ gives

$$\begin{aligned}&\int_{\mathbb{R}^n}\int_a^b\left(\widetilde{\theta}(\underline{x})-\theta\right)\frac{1}{L_n(\underline{x};\theta)\lambda(\theta)}\frac{dL_n(\underline{x};\theta)\lambda(\theta)}{d\theta}L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta = 1. \\ \Rightarrow &\int_{\mathbb{R}^n}\int_a^b\left(\widetilde{\theta}(\underline{x})-\theta\right)\frac{d\log L_n(\underline{x};\theta)\lambda(\theta)}{d\theta}\underbrace{L_n(\underline{x};\theta)\lambda(\theta)}_{\text{measure}}d\underline{x}d\theta = 1\end{aligned}$$

Now by using the Cauchy-Schwartz inequality we have

$$1 \leq \underbrace{\int_a^b\int_{\mathbb{R}^n}\left(\widetilde{\theta}(\underline{x})-\theta\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta}_{\mathrm{E}_{\lambda}\left(\mathrm{E}((\widetilde{\theta}(X)-\theta)^2|\theta)\right)}\int_a^b\int_{\mathbb{R}^n}\left(\frac{d\log L_n(\underline{x};\theta)\lambda(\theta)}{d\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta.$$

Rearranging the above gives

$$\mathrm{E}_\lambda\big[\mathrm{E}_\theta(\widetilde{\theta}(X)-\theta)^2\big] \geq \left[\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta\right]^{-1}.$$

Finally we want to show that the denominator of the RHS of the above can equivalently written as the information matrices:

$$\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta = \mathrm{E}_\lambda(I(\theta)) + I(\lambda).$$

We use basic algebra to show this:

$$\begin{aligned}
&\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\
= &\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta} + \frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\
= &\underbrace{\left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta}_{\mathrm{E}_\lambda(I(\theta))} + 2\int_a^b \int_{\mathbb{R}^n} \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\frac{\partial \log \lambda(\theta)}{\partial\theta} L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\
&+\underbrace{\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta}_{I(\lambda)}.
\end{aligned}$$

We note that

$$\int_a^b \int_{\mathbb{R}^n} \frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\frac{\partial \log \lambda(\theta)}{\partial\theta} dx d\theta = \int \frac{\partial \log \lambda(\theta)}{\partial\theta} \underbrace{\int \frac{\partial L_n(\underline{x};\theta)}{\partial\theta} d\underline{x}}_{=0} d\theta = 0.$$

and $\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta = \int_a^b \left(\frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 \lambda(\theta)d\theta$. Therefore we have

$$\begin{aligned}
&\int \int \left(\frac{\partial \log L_n(\underline{x};\theta)\lambda(\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta \\
= &\underbrace{\int_a^b \int_{\mathbb{R}^n} \left(\frac{\partial \log L_n(\underline{x};\theta)}{\partial\theta}\right)^2 L_n(\underline{x};\theta)\lambda(\theta)d\underline{x}d\theta}_{\mathrm{E}_\lambda(I(\theta))} + \int_{\mathbb{R}^n} L_n(\underline{x};\theta) \underbrace{\int_a^b \left(\frac{\partial \log \lambda(\theta)}{\partial\theta}\right)^2 \lambda(\theta)d\theta}_{I(\lambda)} d\underline{x}.
\end{aligned}$$

Since $\int_{\mathbb{R}^n} L_n(\underline{x};\theta)d\underline{x} = 1$ we obtain the required result. □

We will consider applications of the Bayesian Cramer-Rao bound in Section **??** for obtaining lower bounds of nonparametric density estimators.

1.8 Some questions

Exercise 1.8 *The distribution function of the random variable X_i is $F(x) = 1-\exp(-\lambda x)$.*

(i) Give a transformation of $\{X_i\}_i$, such that the transformed variable is uniformly distributed on the interval $[0,1]$.

(ii) Suppose that $\{X_i\}$ are iid random variables. Use your answer in (i), to suggest a method for checking that $\{X_i\}$ has the distribution $F(x) = 1-\exp(-\lambda x)$ (by checking I mean a graphical tool)?

Exercise 1.9 *Find the Fisher information matrix of*

(i) The normal distribution with unknown mean μ and variance σ^2.

(ii) The normal distribution with unknown mean μ and variance μ^2.

(iii) Let $g : \mathbb{R} \to \mathbb{R}$ be density. Show that $\frac{1}{\rho}g\left(\frac{x-\mu}{\rho}\right)$ is a density function. This is known as the location-scale model.

Define the family of distributions

$$\mathcal{F} = \left\{ f(x;\mu,\rho) = \frac{1}{\rho}g\left(\frac{x-\mu}{\rho}\right); \mu \in \mathbb{R}, \rho \in (0,\infty) \right\}.$$

Suppose that μ and ρ is unknown, obtain the corresponding expected Fisher information (make your derivation neat, explaining which terms depend on parameters and which don't); compare your result to (i) when are they similar?

Exercise 1.10 *Construct a distribution which does not belong to the exponential family but has only a finite number of sufficient statistics (the minimal number of sufficient statistics does not grow with n).*

Exercise 1.11 *Suppose that Z is a Weibull random variable with density $f(x;\phi,\alpha) = (\frac{\alpha}{\phi})(\frac{x}{\phi})^{\alpha-1}\exp(-(x/\phi)^{\alpha})$. Show that*

$$\mathrm{E}(Z^r) = \phi^r \Gamma\left(1 + \frac{r}{\alpha}\right).$$

Hint: Use

$$\int x^a \exp(-x^b)dx = \frac{1}{b}\Gamma\left(\frac{a}{b} + \frac{1}{b}\right) \qquad a, b > 0.$$

This result will be useful in some of the examples used later in this course.

Suppose we have two different sampling schemes to estimate a parameter θ, one measure for understanding which method is better able at estimating the parameter is the *relative frequency.* Relative frequency is defined as the ratio between the two corresponding Fisher information matrices. For example, if we have two iid samples from a normal distribution $N(\mu, 1)$ (one of size n and the other of size m), then the relative frequency is $I_n(\mu)/I_m(\mu) = \frac{n}{m}$. Clearly if $n > m$, then $I_n(\mu)/I_m(\mu) = \frac{n}{m} > 1$. Hence the sample of size n contains more information about the parameter μ.

Exercise 1.12 *Consider the censored exponential in equation (1.5), where $\{(Y_i, \delta_i)\}_{i=1}^n$ is observed.*

(i) Calculate the expected Fisher information of the censored likelihood.

(ii) Calculate the expected Fisher information of $\{\delta_i\}$.

(iii) Calculate the expected Fisher information when there is no censoring.

By using the notion of relative efficiency comment on which sampling scheme contains the most and least information about the parameter θ.