Advanced Statistical Theory I
Spring 2010
Prof. J. Jin[*]

Chris Almost[†]

Contents

[*] jiashun@stat.cmu.edu

[†] cdalmost@cmu.edu

1 Exponential Families

1.1 Introduction and examples

1.1.1 Definition. A family of distributions is said to form an *exponential family* if the distribution P_θ (the parameter θ may be multi-dimensional) has density of the form

$$p_\theta(x) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta)T_i(x) - B(\theta)\right) h(x).$$

This decomposition is not unique, but B is defined uniquely up to constant multiple. Write $\eta_i = \eta_i(\theta)$ and write

$$p(x|\eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x),$$

where $A(\eta) = B(\theta)$.

1.1.2 Examples.

(i) The family of normal distributions has parameters (ξ, σ^2) and density functions

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) = \exp\left(\frac{\xi}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\xi^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}\right)$$

$$\begin{aligned} T_1(x) &= x \\ T_2(x) &= x^2 \\ B(\xi,\sigma^2) &= \frac{\xi^2}{2\sigma^2} + \log\sqrt{2\pi\sigma^2} \\ \eta_1 &= \frac{\xi}{\sigma^2} \\ \eta_2 &= -\frac{1}{2\sigma^2} \\ A(\eta) &= -\frac{\eta_1^2}{4\eta_2} + \log\sqrt{-\frac{\pi}{\eta_2}}. \end{aligned}$$

(ii) For the Poisson distribution,

$$\begin{aligned} p_\theta(x) &= \frac{e^{-\theta}\theta^x}{x!} = \exp\left(x\log\theta - \theta\right)\frac{1}{x!} \\ T(x) &= x \\ h(x) &= \frac{1}{x!} \\ B(\theta) &= \theta \\ \eta(\theta) &= \log\theta \\ A(\eta) &= e^\eta. \end{aligned}$$

(iii) For the binomial distribution,

$$
\begin{aligned}
p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} &= \exp\left(x\log\frac{\theta}{1-\theta} + n\log(1-\theta)\right)\binom{n}{x} \\
T(x) &= x \\
h(x) &= \binom{n}{x} \\
B(\theta) &= -n\log(1-\theta) \\
\eta(\theta) &= \log\frac{\theta}{1-\theta} \\
A(\eta) &= -n\log\frac{1}{e^\eta + 1}
\end{aligned}
$$

Suppose we have n independent samples $X_1,\dots,X_n$ from P_θ in an exponential family. Individually, the density functions are $p_\theta(x_k)$, so the joint density function is

$$
\begin{aligned}
\prod_{k=1}^n p_\theta(x_k) &= \prod_{k=1}^n \exp\left(\sum_{i=1}^s \eta_i(\theta)T_i(x_k) - A(\theta)\right)h(x_k) \\
&= \exp\left(\sum_{i=1}^s \eta_i\left(\sum_{k=1}^n T_i(x_k)\right) - nA(\theta)\right)h(x_1)\cdots h(x_n) \\
&=: \exp\left(n\sum_{i=1}^s \eta_i\tilde{T}_i(x) - nA(\theta)\right)h(x_1)\cdots h(x_n)
\end{aligned}
$$

where $\tilde{T}_i(x) := \frac{1}{n}\sum_{k=1}^n T_i(x_k)$.

1.1.3 Example. If we have n independent samples from a $N(\xi,\sigma^2)$ distribution then $\tilde{T}_1(x) = \overline{x}$ and $\tilde{T}_2(x) = \frac{1}{n}\sum_{k=1}^n x_k^2$, so the joint density is

$$
\exp\left(\frac{n\xi}{\sigma^2}\overline{x} - \frac{n}{2\sigma^2}\tilde{T}_2 - \frac{n\xi^2}{2\sigma^2} - n\log\sqrt{2\pi\sigma^2}\right).
$$

1.2 Moments

Moments are important in large deviation theory. Suppose we have samples $X_1,\dots,X_n$ from a distribution with first two moments μ and σ^2. Then by the central limit theorem $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \xrightarrow{\text{(d)}} N(0,1)$, so

$$
P\left[\frac{\sqrt{n}|\overline{X}-\mu|}{\sigma} \ge t\right] = P\left[|\overline{X}-\mu| \ge \frac{\sigma t}{\sqrt{n}}\right] = 2\Phi(t)
$$

The *Mill's ratio* is $\Phi(t) \sim \phi(t)$ for large t, where $\phi(t) = ce^{-t^2/2}$. A small deviation in this case is on the order $\lambda \equiv \frac{\sigma t}{\sqrt{n}} = O(\frac{1}{\sqrt{n}})$. A moderate deviation is on the order

$\lambda = O(\frac{\sqrt{\log n}}{\sqrt{n}})$ and a large deviation is $\lambda \gg \frac{\sqrt{\log n}}{\sqrt{n}}$. This theory was developed by Hoeffding, Bernstien, *et al.* (see Shorack and Wellner *Empirical processes and applications*.) Where do moments come in? Later we will see that similar results can be derived from the higher moments.

1.2.1 Definition. The *(generalized) moment generating function* of an exponential family is defined in terms of $T_1(x), \dots, T_s(x)$. The m.g.f. is $M_T : \mathbb{R}^s \to \mathbb{R}$ defined by

$$\begin{aligned} M_T(u_1, \dots, u_s) &:= \mathbb{E}[\exp(u_1 T_1(X) + \cdots + u_s T_s(X))] \\ &= \sum_{r_1,\dots,r_s \geq 0} \beta_{r_1,\dots,r_s} \frac{u_1^{r_1} \cdots u_s^{r_s}}{r_1! \cdots r_s!} \end{aligned}$$

It can be shown that if M_T is defined in a neighbourhood of the origin then it is smooth and $\beta_{r_1,\dots,r_s} = \alpha_{r_1,\dots,r_s} := \mathbb{E}[T_1(X)^{r_1} \cdots T_s(X)^{r_s}]$.

1.2.2 Example. Suppose $X \sim N(\xi, \sigma^2)$. Then the k^{th} *centred moment*, or k^{th} *cumulant*, is

$$\mathbb{E}[(X-\xi)^k] = \sigma^k \, \mathbb{E}\left[\left(\frac{X-\xi}{\sigma} \right)^k \right],$$

and the latter is the k^{th} moment of a standard normal distribution, so it suffices to find those quantities. For $Z \sim N(0,1)$, we calculate the usual m.g.f. and read off the moments.

$$M_Z(u) = \int e^{ux} \phi(x) dx = e^{u^2/2} = \sum_{k=1}^{\infty} \frac{u^{2k}}{2^k k!} = \sum_{k=1}^{\infty} (k-1)(k-3) \cdots 1 \frac{u^{2k}}{(2k)!}.$$

From this, $\alpha_k = 0$ if k is odd and is $(k-1)(k-3)\cdots 1$ if k is even.

1.2.3 Definition. Harold Cramer introduced the *(generalized) cumulant generating function*,

$$K_T(u) := \log M_T(u) = \sum_{r_1,\dots,r_s \geq 0} K_{r_1,\dots,r_s} \frac{u_1^{r_1} \cdots u_s^{r_s}}{r_1! \cdots r_s!}$$

The first few cumulants can be found in terms of the first few moments, and *visa versa*, by expanding the power series and equating terms. For an exponential family, the c.g.f. is generally quite simple.

1.2.4 Theorem. *For an exponential family with*

$$p(x|\eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta) \right) h(x),$$

for η in the interior of the set over which it is defined, M_T and K_T both exist in some neighbourhood of the origin and are given by $K_T(u) = A(\eta + u) - A(\eta)$ and $M_T(u) = e^{K_T(u)} = e^{A(\eta+u)}/e^{A(\eta)}$.

1.2.5 Example. The generalized cumulant generating function for the normal family $N(\xi,\sigma^2)$ is as follows. It can be seen that we recover the usual cumulant generating function by setting $u_2 = 0$.

$$\begin{aligned} K(u_1,u_2) &= \log\sqrt{-\eta_2} - \log\sqrt{-\eta_2 - u_2} + \frac{\eta_1^2}{4\eta_2} - \frac{(\eta_1+u_1)^2}{4(\eta_2+u_2)} \\ &= \log\sqrt{\frac{1}{2\sigma^2}} - \log\sqrt{\frac{1}{2\sigma^2} - u_2} - \frac{\xi^2}{2\sigma^2} + \frac{(\frac{\xi}{\sigma^2}+u_1)^2}{4(\frac{1}{2\sigma^2} - u_2)} \end{aligned}$$

Recall that the *characteristic function* of a random variable X is

$$\varphi_X(u) := \int e^{iux} f_X(x)dx.$$

If $\int |\varphi_X(u)|du < \infty$ then the Fourier transform can be inverted

$$f_X(x) = \frac{1}{2\pi}\int \varphi_X(u)e^{-iux}du.$$

The inversion formula for the usual m.g.f. is more complicated. Under some conditions, pointwise,

$$f_X(x) = \lim_{\gamma\to\infty}\int_{x-i\gamma}^{x+i\gamma} M_X(u)e^{xu}du.$$

On the other hand, for an exponential family the associated m.g.f. is $M_T(u) = \mathbb{E}[\exp(u \cdot T(X))]$. For example, for the normal family the m.g.f. is given by

$$\int e^{u_1x+u_2x^2}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\xi)^2}{2\sigma^2}}dx.$$

1.3 Stein's identity

1.3.1 Example (Stein's shrinkage). This example is discussed in two papers, one in 1958 by Stein and one in 1962 by James and Stein. If $X \sim N(\mu, I_p)$ with $p \geq 3$ then the best estimator for μ is not X but is

This is likely incorrect. It does not match with the "Stein's identity" given in a later lecture. One or both may need updating.

$$\hat{\mu} = \frac{X}{1 - c\frac{p-2}{X\cdot X}}$$

for some $0 < c < 1$. In these papers the following useful identity was used.

1.3.2 Lemma. *Let X be distributed as an element of an exponential family*

$$p(x|\eta) = \exp\left(\sum_{i=1}^{s}\eta_i T_i(x) - A(\eta)\right)h(x).$$

If g is any differentiable function such that $\mathbb{E}[g'(X)] < \infty$ *then*

$$\mathbb{E}\left[\left(\frac{h'(X)}{h(X)} + \sum_{i=1}^{s} \eta_i T_i'(X)\right) g(X)\right] = -\mathbb{E}[g'(X)]$$

This is a generalization of the well-known normal-integration-by-parts rule. Indeed, in the normal case the lemma gives

$$\mathbb{E}[g'(X)] = -\mathbb{E}\left[\left(\frac{\xi}{\sigma^2} - \frac{1}{2\sigma^2}(2X)\right) g(X)\right] = \frac{1}{\sigma^2}\,\mathbb{E}[(X-\xi)g(X)].$$

Consider some particular examples. If $g \equiv 1$ then we see $\mathbb{E}[(X-\xi)] = 0$. If $g(x) = x$ then $\mathbb{E}[(X-\xi)X] = \sigma^2$, giving the second moment. Higher moments can be calculated similarly.

Further motivation for finding higher moments is obtained by noting that

$$g(\overline{X}) = g(\mu) + g'(\mu)(\overline{X} - \mu) + \cdots$$

and also in applications of Chebyshev's inequality,

$$\mathbb{P}[|\overline{X} - \mu| \geq t] \leq \frac{\mathbb{E}[|\overline{X} - \mu|^k]}{t^k}.$$

2 Statistics

2.1 Sufficient statistics

2.1.1 Definition. Let $X \sim P_\theta$ for some element of a family of distributions $\{P_\theta : \theta \in \Theta\}$ (not necessarily an exponential family). A *statistic* $T(X)$ (not the same T as for exponential families) is *sufficient* for θ if $P_\theta[X|T(X)]$ does not depend on θ.

There are several concepts of *estimation* for the parameter θ. One is *point estimation*, where we are looking for a specific value of θ. In this case sufficient statistics are very important. When looking for a *confidence interval* for θ sufficient statistics are less important.

2.1.2 Example. For example, if $X_1, \dots, X_n \sim$ Bernoulli(θ) then $T = \sum_i X_i$ is sufficient for θ. Indeed, $T \sim$ Binomial(n, θ), so

$$\begin{aligned} P_\theta(X|T=t) &= \frac{P_\theta(X, T=t)}{P_\theta(T=t)} \\ &= \frac{\prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \quad \text{and} \quad \sum_{i=1}^n X_i = t \\ &= \frac{1}{\binom{n}{t}} \end{aligned}$$

Similarly, if $X_1,\dots,X_n \sim \text{Poisson}(\theta)$ then $T = \sum_i X_i$ is sufficient for θ. Indeed, $T \sim \text{Poisson}(n\theta)$, so

$$\begin{aligned} P_\theta(X|T=t) &= \frac{P_\theta(X,T=t)}{P_\theta(T=t)} \\ &= \frac{\prod_{i=1}^n e^{-\theta}\frac{\theta^{X_i}}{X_i!}}{e^{-n\theta}\frac{(n\theta)^t}{t!}} \quad \text{and} \quad \sum_{i=1}^n X_i = t \\ &= \left(\frac{1}{n}\right)^t \binom{t}{X_1 \cdots X_n} \end{aligned}$$

2.1.3 Example. Suppose that $X_1,\dots,X_n$ are samples from a continuous distribution F. Then the *order statistic* $(X_{(1)},\dots,X_{(n)})$ is sufficient for F.

Warning: the order statistic, for this course, is *reduced*, in that if there are ties then the number of samples at each level is *not* recorded.

For a continuous distribution the probability that any pair of the X_i are equal is zero, so we may assume there are no ties. It follows that

$$\mathbb{P}[X|T] = \mathbb{P}[X_1,\dots,X_n|X_{(1)},\dots,X_{(n)}] = \frac{1}{n!}$$

since we are, effectively, picking a random permutation of the data uniformly. Clearly this does not depend on the distribution F.

2.1.4 Theorem (Neyman-Fisher-Halmos-Savage).
If the distributions $\{P_\theta : \theta \in \Theta\}$ have densities $\{p_\theta : \theta \in \Theta\}$ with respect to a σ-finite measure μ then T is sufficient for θ if and only if $p_\theta(x) = g_\theta(T(x))h(x)$ for some functions g_θ (depending on θ) and h (not depending on θ).

2.1.5 Corollary. *If $T(X)$ is a sufficient statistic and $T(X) = f(U(X))$ (i.e. T is function of another statistic U) then $U(X)$ is also a sufficient statistic.*

PROOF: $p_\theta(x) = g_\theta(T(x))h(x) = g_\theta(f(U(x)))h(x)$ so U is sufficient. □

The motivation for finding sufficient statistics is data reduction, so we are lead naturally to the concept of a *minimal sufficient statistic*. T is said to be a *m.s.s.* if, for any sufficient statistic U, there is a function f such that $T = f(U)$.

2.1.6 Example. For the normal distribution $N(\xi,\sigma^2)$, $(\sum_i X_i, \sum_i X_i^2)$ is a m.s.s. (It follows then that $(\overline{X}_n, S_n^2)$ is also a m.s.s. since there is a one-to-one correspondence taking one pair to the other.) That it is sufficient is clear based on the representation of $N(\xi,\sigma^2)$ as an exponential family. For the k-dimensional multivariate normal $N_k(\xi,\Sigma)$, $(\sum_i X_i, \sum_i X_i X_i^T)$ is a sufficient statistic. This follows from the form of the density function

$$p_\theta(x_1,\dots,x_n) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i-\xi)^T\Sigma^{-1}(x_i-\xi)\right).$$

Expanding this out, a term $\sum_i x_i^T \Sigma^{-1} x_i$ appears. But,

$$x_i^T \Sigma^{-1} x_i = \mathrm{tr}(x_i^T \Sigma^{-1} x_i) = \mathrm{tr}(\Sigma^{-1} x_i x_i^T)$$

so this concludes the proof.

2.1.7 Example (Two-parameter exponential). The one-parameter exponential distributions

$$p_\theta(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \mathbf{1}_{x>0}$$

clearly form an exponential family. However, the two-parameter exponential distributions

$$p_{a,b}(x) = \frac{1}{b} e^{-\frac{(x-a)}{b}} \mathbf{1}_{x>a}$$

do not. Let $X_1, \dots, X_n \sim E(a,b)$. Then

$$\begin{aligned} p_{a,b}(x_1, \dots, x_n) &= \frac{1}{b^n} e^{-\frac{\sum_i (x_i - a)}{b}} \prod_{i=1}^n \mathbf{1}_{x_i > a} \\ &= \frac{1}{b^n} e^{-\frac{(\sum_i x_i) - na}{b}} \mathbf{1}_{\min_i x_i > a} \end{aligned}$$

so $(\sum_i X_i, \min_i X_i)$ is a sufficient statistic for (a,b).

2.2 Ancillary and complete statistics

2.2.1 Definition. Let $X \sim P_\theta$ for some element of a family of distributions $\{P_\theta : \theta \in \Theta\}$. A statistic $V(X)$ is said to be *ancillary* if the distribution of $V(X)$ does not depend on θ. $V(X)$ is said to be *first-order ancillary* if $\mathbb{E}_\theta[V(X)]$ is constant with respect to θ. A statistic $T(X)$ is *complete* if it has the property "for any measurable function f, if $\mathbb{E}_\theta[f(T(X))] = 0$ for all $\theta \in \Theta$ then $f = 0$ a.s."

2.2.2 Theorem (Basu). *If T is a complete sufficient statistic for the family $\{P_\theta : \theta \in \Theta\}$ then any ancillary statistic V is independent of T.*

Proof: If V is ancillary then $p_A := P_\theta[V \in A]$ does not depend on θ, since the distribution of V does not depend on θ. Let $\eta_A(t) := P_\theta[V \in A | T = t]$, which also does not depend on θ since T is sufficient, i.e. the conditional distribution of X (and hence of V) given T does not depend on θ. By the tower property of conditional expectation, for any $\theta \in \Theta$,

$$\begin{aligned} \mathbb{E}_\theta[\eta_A(T) - p_A] &= \mathbb{E}_\theta[P_\theta[V \in A | T]] - p_A \\ &= \mathbb{E}_\theta[\mathbb{E}_\theta[\mathbf{1}_{V \in A} | T]] - p_A \\ &= \mathbb{E}_\theta[\mathbf{1}_{V \in A}] - p_A \\ &= P_\theta[V \in A] - p_A = 0. \end{aligned}$$

Since T is complete, this implies that $\eta_A(t) = p_A = P[V \in A]$ a.s. Because the conditional distribution of V given T is equal to the unconditional distribution of V, V and T are independent. □

Finding m.s.s. is hard, because the definition of minimal is hard to check. Completeness and sufficiency are easier, and fortunately these properties imply minimality.

Warning: Minimal sufficient statistics are not necessarily complete.

2.2.3 Theorem. *Complete sufficient statistics are m.s.s.*

2.2.4 Theorem. *For exponential families in natural form,*

$$p(x|\eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x),$$

the statistic $T = (T_1, \dots, T_s)$ is sufficient. T is complete if the natural parameter space $\Pi_0 \subseteq \mathbb{R}^s$ contains an s-dimensional rectangle.

The natural parameter space is convex, i.e. if η and $\tilde{\eta}$ are parameters for which $p(x|\eta)$ and $p(x|\tilde{\eta})$ are well-defined densities, then $p(x|(1-\lambda)\eta + \lambda\tilde{\eta})$ is a well-defined density. (This is a consequence of Hölder's inequality.) We have seen that, unfortunately, $U(0,\theta)$ (uniform) and $E(a,b)$ (double parameter exponential) are not exponential families.

2.2.5 Examples.

(i) For $X_1, \dots, X_n \sim \text{Bernoulli}(\theta)$, $\overline{X}$ is complete and sufficient.

(ii) For $X_1, \dots, X_n \sim N(\xi, \sigma^2)$, $(\overline{X}, S^2)$ is complete and sufficient.

(iii) For $U(0,\theta)$, $X_{(n)} = \max_i X_i$ is sufficient because $\mathbb{P}_\theta[X|X_{(n)} = t] \sim U(0,t)$ does not depend on θ. It is also complete. Indeed, the density of $X_{(n)}$ is $\frac{n}{\theta^n} t^{n-1} \mathbf{1}_{(0,\theta)}$. Suppose that $\int_0^\theta f(t) \frac{n}{\theta^n} t^{n-1} dt = 0$ for all $\theta > 0$. Then

$$\int_0^\theta f^+(t) t^{n-1} dt = \int_0^\theta f^-(t) t^{n-1} dt$$

for all $\theta > 0$. But because this holds for all θ this implies

$$\int_A f^+(t) t^{n-1} dt = \int_A f^-(t) t^{n-1} dt$$

for all Borel sets $A \subseteq \mathbb{R}_{++}$. Taking $A = \{f > 0\}$ shows that $f^- \equiv 0$ a.s., and similarly $f^+ \equiv 0$ a.s.

(iv) For $E(a,b)$, $(X_{(1)}, \sum_i (X_i - X_{(1)}))$ is complete and sufficient. It can be shown that $X_{(1)} \sim E(a, b/n)$ and $\sum_i (X_i - X_{(1)}) \sim \Gamma(n-1, b)$.

(v) For $U(\theta-\frac{1}{2},\theta+\frac{1}{2})$, $T=(X_{(1)},X_{(n)})$ is sufficient. It can be shown that T is minimal. Clearly $V=X_{(n)}-X_{(1)}$ is ancillary, so T is *not* complete because otherwise T would be independent of V, which is not true. This is an example of a m.s.s. that is not complete.

(vi) We have seen that $(X_{(1)},\dots,X_{(n)})$ is sufficient for a continuous distribution function F. The ranks $(R_1,\dots,R_n)$, where $R_i := \#\{j : X_j \le X_i\}$, are ancillary. Indeed, $\mathbb{P}_F[R=(r_1,\dots,r_n)]=\frac{1}{n!}$ does not depend on F. It can be shown that the order statistic is also complete, so it is independent of R. Given any function h, there is a symmetric function g such that $h((X_{(1)},\dots,X_{(n)})=g(X_1,\dots,X_n)$. Suppose that $\mathbb{E}_F[h(T)]=0$ for all continuous distributions F. Then

$$0=\mathbb{E}_F[h(T)]=\int g(x_1,\dots,x_n)f(x_1)\cdots f(x_n)dx_1\dots dx_n$$

for all densities f. Let $f_1,\dots,f_n$ be densities and let $f=\sum_i \alpha_i f_i$ be a convex combination. Then

$$0=\int g(x_1,\dots,x_n)\prod_{j=1}^{n}\Big(\sum_{i=1}^{n}\alpha_i f_i(x_j)\Big)dx_j.$$

The right hand side is a polynomial of degree n in n variables α_i. Since it is zero, all its coefficients must be zero. In particular, the coefficient of $\alpha_1\cdots\alpha_n$ is zero, so

$$\begin{aligned}0&=\sum_{\pi\in\mathrm{Sym}_n}\int g(x_1,\dots,x_n)f_1(x_{\pi(1)})\cdots f_n(x_{\pi(n)})dx_{\pi(1)}\cdots dx_{\pi(n)}\\&=\sum_{\pi\in\mathrm{Sym}_n}\int g(x_1,\dots,x_n)f_1(x_1)\cdots f_n(x_n)dx_1\cdots dx_n\\&=n!\int g(x_1,\dots,x_n)f_1(x_1)\cdots f_n(x_n)dx_1\cdots dx_n\end{aligned}$$

where we may rearrange the variables arbitrarily because g is symmetric. Therefore $\int g(x_1,\dots,x_n)f_1(x_1)\cdots f_n(x_n)dx_1\cdots dx_n=0$, and from this it follows that g, and hence h, is almost surely zero.

2.2.6 Definition. The *risk function* associated with a *loss function* $L(\theta,a)$ is defined to be $R(\theta,X):=\mathbb{E}_\theta[L(\theta,X)]$. When writing $L(\theta,a)$ for a loss function, θ is intended to be the true value of the parameter, a is the estimator, and $L(\theta,a)$ is the penalty if a is incorrect. Often $L(\theta,\theta)=0$.

Recall *Jensen's inequality*: if ϕ is convex then $\mathbb{E}[\phi(X)]\ge\phi(\mathbb{E}[X])$. If ϕ is strictly convex then the inequality is strict unless X is a.s. constant.

2.2.7 Theorem (Rao-Blackwell). *Let X be distributed as a element of the family $\{P_\theta : \theta \in \Theta\}$, and let $T(X)$ be sufficient for θ. Let $L(\theta, a)$ be a loss function that is convex in a for any $\theta \in \Theta$.*

Let $\delta(X)$ be an estimator for $g(\theta)$ with finite risk. Then $\eta(T) := \mathbb{E}[\delta(X)|T]$ is an estimator for $g(\theta)$ satisfing $R(\theta, \eta) \le R(\theta, \delta)$ for all $\theta \in \Theta$, with equality everywhere only if $\delta(X) = \eta(T)$ a.s.

Proof: Note first that $\eta(T)$ is a well-defined estimator—in that it doesn't depend on θ—because T is sufficient. The risk is computed as follows.

$$\begin{aligned} R(g(\theta), \delta) &= \mathbb{E}[L(g(\theta), \delta(X))] \\ &= \mathbb{E}[\mathbb{E}[L(g(\theta), \delta(X))|T]] && \text{tower property} \\ &\ge \mathbb{E}[L(g(\theta), \mathbb{E}[\delta(X)|T])] && \text{Jensen's inequality} \\ &= \mathbb{E}[L(g(\theta), \eta(T))] = R(g(\theta), \eta) && \square \end{aligned}$$

2.2.8 Examples.

(i) The loss function defined by $L(\theta, a) = 0$ if $\theta = a$ and 1 otherwise is not convex. This is the *hamming distance* or *0-1 loss function*.

(ii) Another non-convex loss function that arises is $L(\theta, a) = \sqrt{|\theta - a|}$.

(iii) Some common convex loss functions are $L(\theta, a) = |\theta - a|^p$ for $p > 0$. When $p = 2$ this is referred to as *squared error loss*.

3 Unbiased Estimators

3.1 Uniform minimum variance unbiased estimators

3.1.1 Definition. For squared error loss, the risk can be decomposed as follows.

$$R(\theta, \delta(X)) = \mathbb{E}[(\theta - \delta(X))^2] = (\theta - \mathbb{E}[\delta(X)])^2 + \mathrm{Var}(\delta(X)).$$

The first term on the right hand side is the *bias* squared. An *unbiased estimator* for $g(\theta)$ is a statistic $\delta(X)$ such that $\mathbb{E}[\delta(X)] = g(\theta)$. We call the unbiased estimator with smallest possible variance the *uniform minimum variance unbiased estimator* or *UMVUE*.

3.1.2 Example (Stein). Unbiased estimators are not necessarily the estimators with smallest risk. Let $X_i \sim N(\mu, I_k)$, where $k \ge 3$. It can be shown that $\overline{X}$ is the UMVUE. But Stein showed that the following is a better estimator.

$$\hat{\mu} := \left(1 - \frac{k-2}{\sum_i X_i^2}\right)\overline{X}$$

Unbiased estimators may not even exist for some problems. Let $g(\theta) := \frac{1}{\theta}$ and $X \sim \text{Binomial}(n, \theta)$. For an unbiased estimator estimator $\delta(X)$ we would need

$$\frac{1}{\theta} = \mathbb{E}_\theta[\delta(X)] = \sum_{k=0}^{n} \binom{n}{k} \delta(k)\theta^k(1-\theta)^{n-k},$$

but this is impossible because no polynomial in θ equals $1/\theta$.

3.1.3 Theorem (Lehmann-Scheffé). *Let T be a complete sufficient statistic for θ and let $\delta(X)$ be an unbiased estimator for $g(\theta)$ with finite variance. Then $\eta(T) := \mathbb{E}[\delta(X)|T]$ is the unique UMVUE.*

Proof: Note that $\eta(T)$ is a well-defined estimator—in that it does not depend on θ—because T is sufficient. By the tower property,

$$\mathbb{E}_\theta[\eta(T)] = \mathbb{E}_\theta[\delta(X)] = g(\theta),$$

so $\eta(T)$ is unbiased. By the Rao-Blackwell theorem, applied with squared error loss, $\text{Var}(\eta(T)) \le \text{Var}(\delta(X))$.

If $\delta'(X)$ is any other unbiased estimator of $g(\theta)$ then $\mathbb{E}_\theta[\eta(T) - \eta'(T)] = 0$ for all θ. Since T is complete, $\eta(T) = \eta'(T)$ a.s. In particular,

$$\text{Var}(\eta(T)) = \text{Var}(\eta'(T)) \le \text{Var}(\delta'(X)),$$

so $\eta(T)$ is a UMVUE. If $\delta'(X)$ is UMVUE then $\text{Var}(\eta'(T)) = \text{Var}(\delta'(X))$, so $\delta'(X) = \eta'(T) = \eta(T)$ a.s. from the Rao-Blackwell theorem and the completeness of T. Therefore $\eta(T)$ is the unique UMVUE. □

How do we actually find UMVUE? On approach is to try to solve the equation $\mathbb{E}_\theta[\delta(T)] = g(\theta)$ directly, for all θ (I don't understand why this would give a UMVUE), and another is to find an unbiased estimator $\delta(X)$ and apply Lehmann-Scheffé.

3.1.4 Example. Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, so $\theta = (\mu, \sigma^2)$. We have seen that $(\overline{X}, S^2)$ is a complete sufficient statistic for θ.

(i) For $g(\theta) = \mu$, $\overline{X}$ is an unbiased estimator. Since this estimator is a function of a sufficient statistic it is UMVUE.

(ii) Assume $\mu = 0$ and σ^2 is unknown. Consider $g(\theta) = \sigma^r$, with $r > 1 - n$. It can be shown that $T = \sum_i X_i^2$ is complete and sufficient for σ^2, and $T/\sigma^2 \sim \chi_n^2(0)$.

$$\mathbb{E}\left[\left(\frac{T}{\sigma^2}\right)^{r/2}\right] = \mathbb{E}[(\chi_n^2(0))^{r/2}] =: K_{n,r}^{-1} = \frac{2^{r/2}\Gamma(\frac{n+r}{2})}{\Gamma(\frac{n}{2})}$$

Whence $K_{n,r}T^{r/2}$ is an unbiased estimator for $g(\theta)$.

(iii) It can be shown that when both μ and σ^2 are unknown, $K_{n-1,r}S^r$ is an unbiased estimator of σ^r.

(iv) Consider $g(\theta) = \mu/\sigma$. It can be shown that $\overline{X}$ and $K_{n-1,-1}S^{-1}$ are independent, so $K_{n-1,-1}S^{-1}\overline{X}$ is an unbiased estimator of $g(\theta)$, and is UMVUE since it is a function of a sufficient statistic.

(v) Assume $\sigma^2 = 1$ and μ is unknown and take $g(\theta) = \mathbb{P}_\mu[X \le x] = \Phi(x - \mu)$. We know that $\overline{X}$ is complete and sufficient for μ. To find a UMVUE we find an unbiased estimator and then condition on $\overline{X}$. $\delta(X) := \mathbf{1}_{X_1 \le x}$ is an unbiased estimator of $g(\theta)$ (but clearly not a great estimator).

$$\begin{aligned} \mathbb{E}[\delta(X)|\overline{X} = \overline{x}] &= \mathbb{P}[X_1 \le x|\overline{X} = \overline{x}] \\ &= \mathbb{P}[X_1 - \overline{X} \le x - \overline{x}|\overline{X} = \overline{x}] \\ &= \mathbb{P}[X_1 - \overline{X} \le x - \overline{x}] && \text{Basu's theorem} \\ &= \Phi\left(\frac{x - \overline{x}}{\sqrt{1 - 1/n}}\right). \end{aligned}$$

(vi) If both μ and σ^2 are unknown then take $g(\theta) = \mathbb{P}_{\mu,\sigma^2}[X \le x] = \Phi(\frac{x-\mu}{\sigma})$. The same sorts of computations yield

$$\mathbb{E}[\delta(X)|T] = \mathbb{P}\left(\frac{X_1 - \overline{X}}{S} \le \frac{x_1 - \overline{x}}{s}\right)$$

and in principle the distribution of the random variable $(X_1 - \overline{X})/S$ can be computed.

3.1.5 Example. Let $X_1, \dots, X_n \sim$ Bernoulli(θ). Since the Bernoulli distribution is an exponential distribution, we see that $T := \sum X_i$ is sufficient and complete for θ. Clearly $\mathbb{E}[T] = n\theta$, so $\overline{X} = T/n$ is UMVUE.

How to we find a UMVUE $\delta(T)$ for $\theta(1-\theta)$? Note that $T \sim$ Binomial(n, θ), so we can compute the expected value.

$$\theta(1-\theta) \overset{\text{set}}{=} \mathbb{E}_\theta[\delta(T)] = \sum_{k=0}^{n} \delta(k)\binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Divide by $(1-\theta)^n$ and change variables to $\rho := \frac{\theta}{1-\theta}$.

$$\begin{aligned} \sum_{k=0}^{n} \delta(k)\binom{n}{k}\theta^k(1-\theta)^{-k} &= \theta(1-\theta)^{1-n} \\ \sum_{k=0}^{n} \delta(k)\binom{n}{k}\rho^k &= \frac{\rho}{1+\rho}\left(\frac{1}{1+\rho}\right)^{1-n} \\ &= \rho(1+\rho)^{n-2} \\ &= \sum_{k=0}^{n-2}\binom{n-2}{k}\rho^{k+1} = \sum_{k=1}^{n-1}\binom{n-2}{k-1}\rho^k \end{aligned}$$

We conclude $\delta(0) = \delta(n) = 0$ and, for $0 < t < n$,

$$\delta(t) = \frac{\binom{n-2}{t-1}}{\binom{n}{t}} = \frac{\frac{(n-2)!}{(t-1)!(n-t-1)!}}{\frac{n!}{t!(n-t)!}} = \frac{t(n-t)}{n(n-1)} = \frac{t}{n-1} - \frac{t^2}{n(n-1)}.$$

Similar things could be done to find UMVUE for θ^r with $r \le n$. But we could also condition T. Note that $\mathbf{1}_{X_1=1,\dots,X_r=1}$ is an unbiased estimator of θ^r since $\mathbb{P}[X_1 = 1,\dots,X_r = 1] = \theta^r$. Now, $\{X_1 = 1,\dots,X_r = 1\}$ is the event that the first r are all ones, and it's not hard to condition on the event that we got a total of t ones, $\{T = t\}$. With $n \ge t \ge r$,

$$\begin{aligned}\mathbb{P}[X_1 = 1,\dots,X_r = 1|T = t] &= \frac{\theta^r\binom{n-r}{t-r}\theta^{t-r}(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^t} \\ &= \frac{\binom{n-r}{t-r}}{\binom{n}{t}} = \frac{t(t-1)\cdots(t-r+1)}{n(n-1)\cdots(n-r+1)}\end{aligned}$$

For $r = 1$ we get T/n and for $r = 2$ we get

$$\frac{T(T-1)}{n(n-1)} = \frac{T^2}{n(n-1)} - \frac{T}{n(n-1)}.$$

Notice that taking the difference between these gives an estimator for $\theta - \theta^2$ that agrees with the one found above.

3.1.6 Example (One-sample exponential). Let $X_1,\dots,X_n \sim E(a,b)$.

$$p_{a,b}(x) = \frac{1}{b}e^{-\frac{x-a}{b}}\mathbf{1}_{x>a}.$$

(i) Assume $b = 1$ is known. If $Y \sim U(0,\theta)$ then $-\log Y \sim E(-\log\theta, 1)$. From the uniform case, if $Y_1,\dots,Y_n \sim U(0,\theta)$ then $Y_{(n)}$ is complete and sufficient for θ. We have $X_i \sim -\log Y_i$ for $i = 1,\dots,n$, so it seems as though $X_{(1)} \sim -\log Y_{(n)}$ and it should be complete and sufficient for $a = -\log\theta$. Sufficiency is easy and completeness follows from the definition (try it).

(ii) From the assignment, $T_1 := X_{(1)}$ and $T_2 := \sum_i(X_i - X_{(1)})$ are sufficient for (a,b). We show now that they are complete for (a,b). Suppose that $\mathbb{E}_{a,b}f(T_1,T_2) = 0$ for all a and b. We must show that $f = 0$ a.s. We have seen that $T_1 \sim E(a,b/n)$ and $T_2 \sim \Gamma(n-1,b) = \frac{b}{2}\chi^2_{2n-2}$ and they are independent. Define $g(t_1,b) := \mathbb{E}_b f(t_1,T_2)$. Then, with b fixed, for all a,

$$0 = \mathbb{E}_{a,b}f(T_1,T_2) = \mathbb{E}_a[g(T_1,b)] = \int_a^\infty g(t_1,b)e^{-nt_1/b}dt_1$$

It follows that $g(t_1,b) = 0$ a.s.-t_1, for any fixed b. (It also follows because when b is known, T_1 is complete for a in $E(a,b)$.) For any b, there is a set N_b of measure zero such that $g(t_1,b) = 0$ for $t_1 \in N_b^c$. Now,

$$0 = \int|g(t_1,b)|dt_1 = \iint|g(t_1,b)|dt_1db = \iint|g(t_1,b)|dbdt_1$$

by Fubini's theorem, since $|g(t_1, b)|$ is a non-negative function. Therefore, for almost all t_1, $g(t_1, b) = 0$ a.s.-b. However, $g(t_1, b) = \mathbb{E}_b f(t_1, T_2)$ and T_2 is χ^2, so $b \mapsto \mathbb{E}_b f(t_1, T_2)$ is a continuous function for all t_1. It follows that, for almost all t_1, for all b,

$$\mathbb{E}_b f(t_1, T_2) = g(t_1, b) = 0.$$

For fixed $t_1 = X_{(1)}$, $T_2 = \sum_i (X_i - t_1)$ is complete for b because $\frac{b}{2}\chi^2_{2n-2}$ is an exponential family. Therefore $f(t_1, t_2) = 0$ a.s.-t_2, for almost all t_1. More precisely, there is a set N of measure zero such that, for all $t_1 \in N^c$, there is a set M_{t_1} of measure zero such that $f(t_1, t_2) = 0$ for all $t_2 \in M^c_{t_1}$. Whence,

$$0 = \int |f(t_1, t_2)| dt_2 \text{ a.s.-}t_1$$

$$\text{so } 0 = \int\int |f(t_1, t_2)| dt_2 dt_1.$$

Therefore $f = 0$ a.s.

Completeness is very important, but can be hard to verify.

3.2 Power series distributions

3.2.1 Definition. A *power series distribution* has a density function or probability mass function of the form

$$p_\theta(x) = \frac{a(x)\theta^x}{c(\theta)}.$$

It is seen that the binomial distribution is a power series distribution by reparameterizing with $\rho = \frac{\theta}{1-\theta}$. The Poisson distribution is already in this form.

If $X_1, \dots, X_n$ are distributed as a power series distribution then it is easy to see that $T := X_1 + \cdots + X_n$ is sufficient and complete, because power series distributions are exponential families. More explicitly,

$$\mathbb{P}[T = t] = \mathbb{P}\left[X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^n x_i = t\right]$$

$$= \sum_{\sum_i x_i = t} \frac{a(x_1)\cdots a(x_n)\theta^{\sum_i x_i}}{c(\theta)^n} =: \frac{A(t,n)\theta^t}{c(\theta)^n}$$

where $A(t,n) := \sum_{x_1+\cdots+x_n=t} a(x_1)\cdots a(x_n)$. Now, $c(\theta) = \sum_x a(x)\theta^x$, so

$$c(\theta)^n = \sum_{t=0}^\infty \left(\sum_{\sum_i x_i = t} a(x_1)\cdots a(x_n) \right) \theta^t = \sum_{t=0}^\infty A(t,n)\theta^t.$$

To find $A(t,n)$ we can expand the coefficient $c(\theta)^n$ into a power series and read off the coefficients. When looking for a UMVUE for θ^r, we look for a function $\delta(T)$ such that

$$\theta^r = \mathbb{E}[\delta(T)] = \sum_{t=0}^{\infty} \delta(t) \frac{A(t,n)\theta^t}{c(\theta)^n}$$

$$\text{so } \sum_{t=0}^{\infty} \delta(t) A(t,n)\theta^{t-r} = c(\theta)^n = \sum_{t=0}^{\infty} A(t,n)\theta^t$$

Therefore $\delta(t) = 0$ for $t < r$ and $\delta(t) = A(t-r,n)/A(t,n)$ for $t \geq r$. It can be seen that these estimators agree with the estimators obtained in the binomial case.

3.2.2 Example. For the Poisson distribution, $c(\theta) = e^\theta$, so

$$\sum_{t=0}^{\infty} A(t,n)\theta^n = c(\theta)^n = e^{n\theta} = \sum_{k=1}^{\infty} \frac{1}{k!}(n\theta)^k.$$

Thus $A(t,n) = n^t/t!$ and it follows that the UMVUE for θ^r is

$$\begin{cases} T(T-1)\cdots(T-r+1)/n^r & \text{it } T \geq r \\ 0 & \text{otherwise.} \end{cases}$$

However, things are not always peachy. Suppose $X \sim \text{Poisson}(\theta)$ and we use this method to find a UMVUE for for $g(\theta) = e^{-a\theta}$.

$$e^{-a\theta} = \mathbb{E}_\theta[\delta(X)] = \sum_{x=0}^{\infty} \delta(x) \frac{e^{-\theta}\theta^x}{x!}$$

$$\text{so } \sum_{x=0}^{\infty} \delta(x) \frac{\theta^x}{x!} = e^{(1-a)\theta} = \sum_{x=0}^{\infty} (1-a)^x \frac{\theta^x}{x!}$$

Thus $\delta(x) = (1-a)^x$. When $a > 2$ this is a very bad estimator, despite being UMVUE. A much better, but biased, estimator could be obtained by expanding the power series for $g(\theta)$ to a few terms and using the estimators for θ^r.

3.3 Non-parametric UMVUE

Let $X_1, \dots, X_n \sim F$, where F is any continuous c.d.f. We have shown that the order statistic is complete and sufficient and the rank statistics are ancillary. Suppose we want to estimate $\xi = \int x dF(x)$. But the first moment may not exist for all F, so we must restrict the class under consideration. This may change the properties of the above statistics, e.g. there is not reason to suspect that the order statistics remain complete and sufficient. However, Hoeffding showed that the order statistics are complete and sufficient for each class $\mathscr{F}_r$, the class of distribution functions having finite r^{th} moment. It follows that any UMVUE must be symmetric in its arguments $x_1, \dots, x_n$, because at its heart it is function of the order statistic. Therefore, in the non-parametric case, to find a UMVUE it suffices to find a *symmetric* unbiased estimator.

3.3.1 Examples.

(i) For $\mathbb{P}[X \le a]$, $\mathbf{1}_{X_1 \le a}$ is an unbiased estimator, so symmetrize to obtain a UMVUE $\frac{1}{n}\sum_i \mathbf{1}_{X_i \le a}$.

(ii) For $\xi(F) = \int x dF(x)$, $\overline{X}$ is symmetric and unbiased, hence a UMVUE.

(iii) S^2 is symmetric and unbiased, hence a UMVUE, for the variance $\sigma^2(F)$ of F (over $\mathscr{F}_2$).

(iv) To find a UMVUE for $(\xi(F))^2$ we have be a bit careful. Notice

$$(\xi(F))^2 = \int x^2 dF(x) - \sigma^2(F),$$

and for the first term $\frac{1}{n}\sum_i X_i^2$ is symmetric and unbiased. Alternatively, $\xi^2 = \xi \cdot \xi$, so $X_1 X_2$ is unbiased. Symmetrizing, $\frac{1}{n(n-1)}\sum_{i,j} X_i X_j$ is UMVUE, and these two estimators can be seen to be the same.

4 Lower Bounds on Risk

Recall that risk is expected loss. More precisely, the risk associated with an estimator T of a parameter θ is $\mathbb{E}_\theta[L(T,\theta)]$, where L is the loss function, a non-negative function that quantifies the error when T is not equal to θ. For this section we will usually consider loss functions of the form $L(T,\theta) = \ell(|T-\theta|)$, where $\ell : [0,\infty) \to [0,\infty)$ is convex and $\ell(0) = 0$. Many common loss functions are of this form, including the very commonly used squared error loss.

Let $X_1,\dots,X_n \sim P_\theta$ where $\theta \in \Theta \subseteq \mathbb{R}^k$. The *minimax risk* is

$$R_n^* = \inf_{T_n} \sup_\theta \mathbb{E}_\theta[L(T_n,\theta)],$$

where the infimum is taken over all statistics $T_n = T_n(X_1,\dots,X_n)$. We would like bounds on this risk, and ideally to find an estimator that attains this risk. Upper bounds can be found by constructing specific statistics T_n^*. With clever constructions the upper bound $R_n^* \le \sup_\theta \mathbb{E}_\theta[L(T_n^*,\theta)]$ can be quite good.

Lower bounds are trickier. The Cramer-Rao lower bound was one of the first lower bounds on risk that was discovered, and is basically a consequence of the Cauchy-Schwarz inequality. A different approach was taken by Lucien Le Cam, and developed by Lucien Birgé, David Donoho, and others, and is discussed in detail in the next section.

To this end, consider the true value θ_0 and an estimate θ_1, presumably close to θ_0, such that the test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ is “indistinguishable.” It was shown in the paper *Rate of geometric convergence* by Donoho and Liu, and going back to earlier works by Lucien Birgé, that if $L(T,\theta) = \ell(|T-\theta|)$ for some convex function ℓ, then R_n^* is bounded below by a multiple of $\ell(|\theta_1 - \theta_0|)$.

What does it mean for two hypotheses to be indistinguishable? Intuitively, if the sum of the type I and type II errors is close to one for any test then the hypotheses cannot be differentiated. We will make this formal and derive estimates based on notions of the distance between hypotheses.

(i) Hellinger distance, via the Hellinger affinity
(ii) χ^2-distance (not actually a distance)
(iii) L^1-distance
(iv) Kullback-Leibler divergence, from information theory.

4.1 Hellinger distance

To simplify the exposition, in this section all measures considered are assumed to have a density with respect to Lebesgue measure on the real line. Most of the results still hold in a more general setting.

4.1.1 Definition. The *Hellinger distance* between two densities p and q is

$$H(p,q) := \sqrt{\frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx} = \sqrt{1 - A(p,q)},$$

where $A(p,q) := \int_{\mathbb{R}} \sqrt{p(x)q(x)}dx$ is the *Hellinger affinity*.

Remark. Note that $H(p,q)^2 = 1 - A(p,q)$, and $0 \le A(p,q) \le 1$ by Cauchy-Schwarz, so $0 \le H(p,q) \le 1$ for all densities p and q.

4.1.2 Definition. Recall that the L^1-distance between densities p and q is

$$\|p - q\|_1 := \int |p(x) - q(x)| dx.$$

The χ^2*-distance* between p and q is

$$\int \frac{(p(x) - q(x))^2}{p(x)} dx.$$

Remark. Note that the χ^2-distance is not symmetric in p and q. When p is close to q, in the sense that $\|(q-p)/p\|_\infty$ is small,

$$\begin{aligned}
H(p,q)^2 &= 1 - \int \sqrt{p(x)q(x)}dx \\
&= 1 - \int \sqrt{\frac{q(x)}{p(x)}} p(x)dx \\
&= 1 - \int \sqrt{1 + \frac{q(x) - p(x)}{p(x)}} p(x)dx \\
&\approx 1 - \int \left(1 + \frac{1}{2}\frac{q(x) - p(x)}{p(x)} - \frac{3}{2}\left(\frac{q(x) - p(x)}{p(x)}\right)^2\right) p(x)dx
\end{aligned}$$

$$= \frac{3}{2} \int \frac{(q(x) - p(x))^2}{p(x)} dx$$

which is 3/2 times the χ^2-distance.

Claim. For $i \geq 0$, let $H_i : \theta = \theta_i$, and $p_{\theta_i,n}$ be the joint density of $X_1, \dots, X_n$ under hypothesis H_i. If the Hellinger distance $H(p_{\theta_0,n}, p_{\theta_n,n})$ is $o(1/\sqrt{n})$ then the sum of the type I and type II errors converges to 1 for any test, i.e. the hypotheses H_0 and H_n become indistinguishable.

The remainder of this section is devoted to substantiating this claim.

4.1.3 Definition. A *test* is a mapping $\phi : \mathbb{R} \to [0,1]$. The *power function* of a test is $\pi : \Theta \to [0,1] : \theta \mapsto \mathbb{E}_\theta[\phi(X)]$.

4.1.4 Example. Let $\phi(x) := \mathbf{1}_R(x)$ for some $R \subseteq \mathbb{R}$, the *rejection region* for this test. The power function is $\pi(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta[X \in R]$. Under the simple setup of testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, $\pi(\theta_0) = \mathbb{P}_{\theta_0}[X \in R]$ is the *level of a test* or the *type I error*, and $\pi(\theta_1) = \mathbb{P}_{\theta_1}[X \in R]$ is said to be the *power of a test*. Higher power tests reject the alternative hypothesis with greater probability.

$$\begin{aligned}\pi(\theta_1) - \pi(\theta_0) &= 1 - \mathbb{P}_{\theta_1}[\text{Accept}] - \mathbb{P}_{\theta_0}[\text{Reject}] \\ &= 1 - (\text{type II error} + \text{type I error})\end{aligned}$$

4.1.5 Lemma. $\pi(\theta_1) - \pi(\theta_0) \leq \frac{1}{2}\|p_{\theta_1} - p_{\theta_0}\|_1$.

PROOF:

$$\begin{aligned}\pi(\theta_1) - \pi(\theta_0) &= \int \phi(x)p_1(x)dx - \int \phi(x)p_0(x)dx \\ &= \int \phi(x)(p_1(x) - p_0(x))dx \\ &= \int_{p_1 \geq p_0} \phi(x)(p_1(x) - p_0(x))dx + \int_{p_1 < p_0} \phi(x)(p_1(x) - p_0(x))dx \\ &\leq \int_{p_1 \geq p_0} (p_1(x) - p_0(x))dx \\ &= \frac{1}{2}\|p - q\|_1 \\ \text{since } \|p - q\|_1 &= \int_{p_1 \geq p_0} (p_1(x) - p_0(x))dx - \int_{p_1 < p_0} (p_1(x) - p_0(x))dx \\ \text{and } 0 &= \int_{p_1 \geq p_0} (p_1(x) - p_0(x))dx + \int_{p_1 < p_0} (p_1(x) - p_0(x))dx\end{aligned}$$

□

Remark. The above lemma implies that

$$\text{type I error} + \text{type II error} \geq 1 - \tfrac{1}{2}\|p_{\theta_1} - p_{\theta_0}\|_1,$$

so if $\|p_{\theta_n} - p_{\theta_0}\|_1 \to 0$ then the hypotheses become indistinguishable.

4.1.6 Lemma. $1 - H(p_{\theta_n,n}, p_{\theta_0,n})^2 = (1 - H(p_{\theta_n}, p_{\theta_0})^2)^n$

Proof: Hellinger affinity has the property $A(p_{\theta_n,n}, p_{\theta_0,n}) = A(p_{\theta_n}, p_{\theta_0})^n$. □

Show that if $\int p \wedge q \to 0$ then $H^2 \to 1$ and if $\int p \wedge q \to 1$ then $\|p-q\|_1 \to 0$ and $H^2 \to 1$(?) Chain: minimax risk lower bound, testing argument, type I error + II error, L^1-distance, Hellinger distance, Hellinger affinity.

4.1.7 Lemma.

$$0 \leq 2H(p,q)^2 \leq \|p-q\|_1 \leq (2 - A(p,q)^2) \wedge 2\sqrt{2}H(p,q)$$

Proof:

$$\begin{aligned}
2H(p,q)^2 &= \int_{\mathbb{R}} (\sqrt{p} - \sqrt{q})^2 dx \\
&\leq \int_{\mathbb{R}} |\sqrt{p} - \sqrt{q}|(\sqrt{p} + \sqrt{q}) dx \\
&= \int_{\mathbb{R}} |p - q| dx = \|p-q\|_1 \\
&= \int_{\mathbb{R}} |\sqrt{p} - \sqrt{q}||\sqrt{p} + \sqrt{q}| dx \\
&\leq \sqrt{\int_{\mathbb{R}} |\sqrt{p} - \sqrt{q}|^2 dx} \sqrt{\int_{\mathbb{R}} |\sqrt{p} + \sqrt{q}|^2 dx} \\
&\leq \sqrt{(2H^2)(4)} = 2\sqrt{2}H
\end{aligned}$$

It can be shown that $2\int_{\mathbb{R}}(p \wedge q)dx = 2 - \|p-q\|_1$ by noting that

$$p + q = p \vee q + p \wedge q \quad \text{and} \quad |p - q| = p \vee q - p \wedge q.$$

$$\begin{aligned}
A(p,q) = \int_{\mathbb{R}} \sqrt{pq}\, dx &= \int_{\mathbb{R}} \sqrt{(p \vee q)(p \wedge q)} dx \\
&\leq \int_{\mathbb{R}} \sqrt{(p + q)(p \wedge q)} dx \\
&\leq \sqrt{\int_{\mathbb{R}} (p+q) dx \int_{\mathbb{R}} (p \wedge q) dx} \\
&= \sqrt{2 - \|p-q\|_1}
\end{aligned}$$

□

It follows from this theorem that $H(p,q) \approx 0$ if and only if $\|p-q\|_1 \approx 0$, and $H(p,q) \approx 1$ if and only if $\|p-q\|_1 \approx 2$. In some sense these distances are "equivalent."

4.1.8 Example. Let p_0 and p_1 be densities, and ν be a functional of densities. Let $\ell : [0,\infty) \to \mathbb{R}$ be convex with $\ell(0) = 0$. If $H(p_0,p_1) < 1$ then

$$\inf_{T_n} \max \{ \mathbb{E}_{n,0}[\ell(|T_n - \nu(p_0)|)], \mathbb{E}_{n,1}[\ell(|T_n - \nu(p_1)|)] \} \\ \geq \ell \left(\frac{1}{4} |\nu(p_1) - \nu(p_0)| \big(1 - H(p_0,p_1)^2\big)^{2n} \right)$$

To prove it, note that

$$\mathbb{E}[\ell(|T_n - \nu(p_0)|)] \geq \ell(\mathbb{E}[|T_n - \nu(p_0)|])$$

by Jensen's inequality, so it suffices to prove the result for the identity function.

$$\begin{aligned}
&\inf_{T_n} \max \{ \mathbb{E}_{n,0}[\ell(|T_n - \nu(p_0)|)], \mathbb{E}_{n,1}[\ell(|T_n - \nu(p_1)|)] \} \\
&\geq \inf_{T_n} \frac{1}{2} (\mathbb{E}_{n,0}[|T_n - \nu(p_0)|] + \mathbb{E}_{n,1}[|T_n - \nu(p_1)|]) \\
&= \frac{1}{2} \inf_{T_n} \int |T_n - \nu(p_0)| p_0(x_1) \cdots p_0(x_n) + |T_n - \nu(p_1)| p_1(x_1) \cdots p_1(x_n) dx_1 \cdots dx_n \\
&\geq \frac{1}{2} \inf_{T_n} \int (|T_n - \nu(p_0)| + |T_n - \nu(p_1)|)(p_0(x_1) \cdots p_0(x_n) \wedge p_1(x_1) \cdots p_1(x_n)) dx_1 \cdots dx_n \\
&\geq \frac{1}{2} \inf_{T_n} \int |\nu(p_1) - \nu(p_0)| (p_0(x_1) \cdots p_0(x_n) \wedge p_1(x_1) \cdots p_1(x_n)) dx_1 \cdots dx_n \\
&= \frac{1}{2} |\nu(p_1) - \nu(p_0)| \int (p_0(x_1) \cdots p_0(x_n) \wedge p_1(x_1) \cdots p_1(x_n)) dx_1 \cdots dx_n \\
&\geq \frac{1}{4} |\nu(p_1) - \nu(p_0)| \big(1 - H(p_{0,n}, p_{1,n})^2\big)^2 \\
&= \frac{1}{4} |\nu(p_1) - \nu(p_0)| \big(1 - H(p_0, p_1)^2\big)^{2n}
\end{aligned}$$

Suppose the family $\{p_\theta, \theta \in \Theta \subseteq \mathbb{R}^k\}$ has a common support that does not depend on θ, and that the p_θ are differentiable with respect to θ. Moreover, suppose there is $\dot{\ell}_{\theta_0}$ (which may usually be taken to be $\nabla_\theta p_\theta(x)/p_\theta(x)$) such that

$$\int_{\mathbb{R}} \left(\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right)^2 dx = o(|\theta - \theta_0|).$$

In the univariate case $(\sqrt{p_\theta})'(\theta - \theta_0) = \frac{1}{2\sqrt{p_{\theta_0}}}(\frac{d}{d\theta} p_\theta|_{\theta=\theta_0})(\theta - \theta_0)$. Essentially, the condition is asking that the linear approximation for $\sqrt{p_\theta}$ is good.

In the claim near the beginning of this section, take $\theta_n := \theta_0 + h/\sqrt{n}$, where h is a fixed vector. Then

$$\begin{aligned} nH(p_{\theta_n}, p_{\theta_0})^2 &= \frac{n}{2} \int (\sqrt{p_{\theta_n}} - \sqrt{p_{\theta_0}})^2 dx \\ &= \frac{1}{2} \int (\sqrt{n}(\sqrt{p_{\theta_n}} - \sqrt{p_{\theta_0}}))^2 dx \\ &= \frac{1}{8} \int h^T \ell_{\theta_0}(x)^T \ell_{\theta_n}(x) h \, p_{\theta_0}(x) \, dx + o(1) \\ &= \frac{1}{8} \mathbb{E}_{\theta_0}[h^T I(\theta_0) h] \end{aligned}$$

where I is the *information matrix* $I(\theta) := \mathbb{E}_\theta[\dot{\ell}_\theta^T \dot{\ell}_\theta]$, to be discussed at length in the next section.. In particular, if $\ell(x) = x$ and $\nu(p_\theta) = c^T\theta$ for some vector c, then

$$\begin{aligned} \inf_{T_n}\{|T_n - c^T\theta_n|, |T_n - c^T\theta_0|\} &\geq \frac{1}{4}|c^T\theta_n - c^T\theta_0| \exp\left(-\frac{1}{4}h^T I(\theta_0) h\right) \\ &= \frac{1}{4\sqrt{n}}|h^T c| \exp\left(-\frac{1}{4}h^T I(\theta_0) h\right) \end{aligned}$$

where the exponential comes from estimating $(1 - H^2(p_{\theta_n}, p_{\theta_0})^2)^{2n}$, noting that $\left(1 + \frac{x}{n}\right)^n \geq e^x$ for all x and all n (I think).

4.2 Fisher's information

In the univariate case, with $\{p_\theta, \theta \in \Theta \subseteq \mathbb{R}\}$, *Fisher's information* is

$$I(\theta) = \int_{\mathbb{R}} \left(\frac{p_\theta'(x)}{p_\theta(x)}\right)^2 p_\theta(x) dx = \mathbb{E}_\theta\left[\left(\frac{d}{d\theta}\log(p_\theta(X))\right)^2\right]$$

The name may come from information theory. In regions where the density is changing quickly, we can estimate θ more accurately. Correspondingly, the information is large because the derivative is large. See *Elements of information theory* by Thomas Cover.

4.2.1 Example. For $X \sim N(\theta, \sigma^2)$, with σ^2 known,

$$\begin{aligned} \frac{p_\theta'(x)}{p_\theta(x)} &= \frac{1}{\sigma^2}(x - \theta) \\ I(\theta) &= \mathbb{E}_\theta\left[\frac{1}{\sigma^4}(X - \theta)^2\right] = \frac{1}{\sigma^2} \end{aligned}$$

Unfortunately, $I(\theta)$ depends on the parameterization of θ, so reparameterizing can change the information. Suppose that $\xi = h(\theta)$, where h is differentiable.

Then

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta\left[\left(\frac{d}{d\theta}\log p(x;h(\theta))\right)^2\right] \\ &= \mathbb{E}_\theta\left[\left(\frac{d\xi}{d\theta}\frac{d}{d\xi}\log p(x;\xi)\right)^2\right] \\ &= (h'(\theta))^2 I(\xi). \end{aligned}$$

Therefore $I(\xi) = I(\theta)/(h'(\theta))^2$.

In the multivariate case, where $\Theta \subseteq \mathbb{R}^k$, $I(\theta)$ is a matrix with

$$I_{ij}(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i}\log p_\theta(x)\frac{\partial}{\partial\theta_j}\log p_\theta(x)\right]$$

for $1 \le i,j \le k$. If $\xi_i = h(\theta_1,\dots,\theta_k)$ for $i = 1,\dots,k$ then we have a similar relationship between the information matrices, $I(\xi) = JI(\theta)J^T$, where J is the Jacobian of the inverse transformation: $J_{ij} = \frac{\partial\theta_j}{\partial\xi_i}$. Note that J is usually symmetric in applications.

4.2.2 Example. Let X be distributed as an element of an exponential family

$$p_\theta(x) = \exp\left(\sum_{i=1}^k \eta_i(\theta)T_i(x) - A(\eta(\theta))\right)h(x).$$

Let $\tau_i(\theta) := \mathbb{E}_\theta[T_i(X)]$ for $i = 1,\dots,k$. Then $I(\tau) = C^{-1}$, where C is the covariance matrix of the random vector $(T_1(X),\dots,T_k(X))$. Indeed,

$$\begin{aligned} 1 &= \int_{\mathbb{R}} \exp\left(\sum_{i=1}^k \eta_i T_i(x) - A(\eta)\right)dx \\ 0 &= \frac{\partial}{\partial\eta_i}\int_{\mathbb{R}} \exp\left(\sum_{i=1}^k \eta_i T_i(x) - A(\eta)\right)dx \\ &= \int_{\mathbb{R}}\left(T_i(x) - \frac{\partial}{\partial\eta_i}A(\eta)\right)p_\theta(x)dx \end{aligned}$$

so $\tau_i = \mathbb{E}_\theta[T_i(X)] = \frac{\partial}{\partial\eta_i}A(\eta)$. Taking the derivative again yields

$$0 = \int_{\mathbb{R}}\left(\left(T_i(x) - \frac{\partial}{\partial\eta_i}A(\eta)\right)\left(T_j(x) - \frac{\partial}{\partial\eta_j}A(\eta)\right) - \frac{\partial^2}{\partial\eta_i\partial\eta_j}A(\eta)\right)p_\theta(x)dx$$

so $\mathrm{Cov}_\theta(T_i(X),T_j(X)) = \frac{\partial^2}{\partial\eta_i\partial\eta_j}A(\eta)$. From the definition,

$$I_{ij}(\eta) = \mathbb{E}_\theta\left[\left(T_i(x) - \frac{\partial}{\partial\eta_i}A(\eta)\right)\left(T_j(x) - \frac{\partial}{\partial\eta_j}A(\eta)\right)\right] = \mathrm{Cov}_\theta(T_i(X),T_j(X)),$$

so $I(\eta) = C$. The Jacobian of the transformation taking η to τ is again C, so $I(\tau) = C^{-1}CC^{-1} = C^{-1}$.

4.2.3 Example. Let $X \sim N(\mu, \sigma^2)$.

$$I(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} \quad \text{and} \quad I(\mu,\sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

4.2.4 Example. Consider X distributed as a member of an exponential family with $k = 1$, so that

$$p_\theta(x) = \exp(\eta(\theta)T(x) - B(\theta))h(x).$$

We have seen that $I(g(\theta)) = I(\theta)/(g'(\theta))^2$. By differentiating under the integral sign it can be shown that $\eta'(\theta)\mathbb{E}_\theta[T(X)] = B'(\theta)$. Therefore

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{p'_\theta(X)}{p_\theta(X)}\right)^2\right] = \mathbb{E}_\theta[(\eta'(\theta)T(X) - B'(\theta))^2] = (\eta'(\theta))^2 \operatorname{Var}_\theta(T(X))$$

Combining these formulae,

$$I(g(\theta)) = \left(\frac{\eta'(\theta)}{g'(\theta)}\right)^2 \operatorname{Var}_\theta(T(X)).$$

For $\tau(\theta) := \mathbb{E}[T(X)]$ we have seen that $I(\tau) = 1/\operatorname{Var}_\theta(T)$, so we get the formulae

$$\operatorname{Var}_\theta(T(X)) = \left|\frac{\tau'(\theta)}{\eta'(\theta)}\right| \quad \text{and} \quad \tau(\theta) = \frac{B'(\theta)}{\eta'(\theta)}.$$

(i) For $X \sim \Gamma(\alpha, \beta)$, with α known,

$$p_\beta(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

So $\eta(\beta) = -\frac{1}{\beta}$, $T(x) = x$, and $B(\beta) = \alpha \log \beta$. Therefore

$$\tau(\beta) := \mathbb{E}_\beta[X] = \frac{\alpha/\beta}{1/\beta^2} = \alpha\beta,$$

$\operatorname{Var}_\beta(X) = \alpha\beta^2$, and $I(\alpha\beta) = 1/(\alpha\beta^2)$.

(ii) For $X \sim N(\mu, \sigma^2)$ with σ^2 known, $B(\mu) = \mu^2/(2\sigma^2)$, $\eta(\mu) = \mu/\sigma^2$, and $T(x) = x$. Therefore $I(\mu^2) = (\frac{1/\sigma^2}{2\mu})^2 \operatorname{Var}(X) = 1/(4\mu^2\sigma^2)$. These formulae can be used to do many calculations for exponential families.

4.3 The Cramér-Rao lower bound

One-dimensional case

Let $\Theta \subseteq \mathbb{R}$ be a one-dimensional parameter space, and let $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures with densities p_θ with respect to a σ-finite measure μ on $\mathbb{R}$, and let T be a measurable function.

4.3.1 Conditions.

(C1) *Θ is open in $\mathbb{R}$.*

(C2) *(a) There is a set $B \subseteq \mathbb{R}$ such that $\mu(B^c) = 0$ and $\frac{d}{d\theta}p_\theta(x)$ exists for all $\theta \in \Theta$ and all $x \in B$.*

(b) $A := \{x : p_\theta(x) = 0\}$ does not depend on θ

(C3) *$I(\theta) := \mathbb{E}_\theta[(\dot{\ell}_\theta(X))^2] < \infty$, where $\dot{\ell}_\theta(x) = \frac{d}{d\theta}\log p_\theta(x)$ is the score.*

(C4) *$\int_{\mathbb{R}} p_\theta(x)\mu(dx)$ and $\int_{\mathbb{R}} T(x)p_\theta(x)\mu(dx)$ can both be differentiated with respect to θ under the integral sign. In this case $\mathbb{E}_\theta[\dot{\ell}_\theta(X)] = 0$.*

(C5) *$\int_{\mathbb{R}} p_\theta(x)\mu(dx)$ can be differentiated twice with respect to θ under the integral sign. In this case we have the formula*

$$I(\theta) = -\mathbb{E}[\ddot{\ell}_\theta(X)] = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log p_\theta(x)\right].$$

4.3.2 Theorem. *Let $T(X)$ be an estimator for $q(\theta)$, and let $b(\theta)$ be the bias. Then, given that (C1)–(C4) hold,*

$$\mathrm{Var}_\theta(T(X)) \geq \frac{(q'(\theta) + b'(\theta))^2}{I(\theta)}$$

for all $\theta \in \Theta$. When T is unbiased this reduces to $(q'(\theta))^2/I(\theta)$.

Proof: Assuming (C1)–(C4) all hold,

$$\begin{aligned}
\mathbb{E}_\theta[T(X)] = q(\theta) + b(\theta) = \int_{\mathbb{R}} T(x)p_\theta(x)\mu(dx) &= \int_{B^c \cap A^c} T(x)p_\theta(x)\mu(dx) \\
q'(\theta) + b'(\theta) &= \int_{B^c \cap A^c} T(x)\frac{d}{d\theta}p_\theta(x)\mu(dx) \\
&= \int_{B^c \cap A^c} T(x)\dot{\ell}_\theta(x)p_\theta(x)\mu(dx) \\
&= \mathrm{Cov}_\theta(T(X), \dot{\ell}_\theta(X)) \\
(q'(\theta) + b'(\theta))^2 &= (\mathrm{Cov}_\theta(T(X), \dot{\ell}_\theta(X)))^2 \leq \mathrm{Var}_\theta(T(X))I(\theta)
\end{aligned}$$

since the score has mean zero and by the Cauchy-Schwarz inequality. □

If the lower bound is achieved then, from the Cauchy-Schwarz inequality, it must be the case that $\dot{\ell}_\theta(X) = A(\theta)(T(X) - \mathbb{E}_\theta[T(X)])$ for some constant $A(\theta)$. This is why the MLE often achieves the lower bound, while other statistics fail to do so. It can be shown that the lower bound can only be achieved for parameters of an exponential family.

This may be false.

Multivariate case

Let $\Theta \subseteq \mathbb{R}^k$ be a k-dimensional parameter space, and let $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures with densities p_θ with respect to a σ-finite measure μ on $\mathbb{R}^k$, Suppose $X \sim P_\theta$ for some $\theta \in \Theta$ and $T(X)$ is an estimator for $q(\theta)$, where $q : \mathbb{R}^k \to \mathbb{R}$ is differentiable. As usual, $b(\theta) := \mathbb{E}_\theta[T(X)] - q(\theta)$ is the bias.

4.3.3 Theorem. *Suppose that (M1)–(M4) below hold. Then*

$$\mathrm{Var}_\theta(T(X)) \geq \alpha^T I(\theta)^{-1}\alpha$$

for all $\theta \in \Theta$, where $\alpha := \nabla(q(\theta) + b(\theta))$. If $T(X)$ is unbiased this reduces to $\mathrm{Var}_\theta(T(X)) \geq \dot{q}(\theta)^T I^{-1}(\theta)\dot{q}(\theta)$.

4.3.4 Conditions.
(M1) Θ is open in $\mathbb{R}^k$.
(M2) (a) There is a set $B \subseteq \mathbb{R}$ such that $\mu(B^c) = 0$ and $\frac{\partial}{\partial\theta_i}p_\theta(x)$ exists for all $\theta \in \Theta$, $1 \leq i \leq k$, and all $x \in B$.
(b) $A := \{x : p_\theta(x) = 0\}$ does not depend on θ
(M3) The $k \times k$ matrix $I_{ij}(\theta) := \mathbb{E}_\theta[\dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^T]$ is positive definite, where $\dot{\ell}_\theta(x) = \nabla_\theta \log p_\theta(x)$ is the score vector.
(M4) $\int_{\mathbb{R}^k} p_\theta(x)\mu(dx)$ and $\int_{\mathbb{R}^k} T(x)p_\theta(x)\mu(dx)$ can both be differentiated with respect to θ under the integral sign. In this case $\mathbb{E}_\theta[\dot{\ell}_\theta(X)] = 0$.
(M5) $\int_{\mathbb{R}^k} p_\theta(x)\mu(dx)$ can be differentiated twice with respect to θ under the integral sign. In this case we have the following formula.

$$I(\theta) = -\mathbb{E}_\theta\, \ddot{\ell}_\theta(X) = \left(-\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\right]\right)_{ij}$$

5 Decision Theory

Let $\theta \in \Theta$, the *parameter space*, let $a \in A$, the *action space*, and let $x \in X$, the *sample space*. Let $L : \Theta \times A \to \mathbb{R}_+$ be a loss function and $d : A \times X \to [0,1]$ be defined by $d(a,x) = d(a|x) =$ "probability of taking action a given x is observed." This is a *randomized decision rule*. Given a decision rule $d(x) := d(\cdot|x)$ (actually a probability distribution on A), the *decision loss* is $L(\theta, d(x))$ and the *decision risk* is

$$R(\theta, d) = \int_X \int_A L(\theta, a)d(da|x)P_\theta(dx).$$

If the decision rule is non-randomized then the risk is $\int_X L(\theta, d(x))P_\theta(dx)$. Let D denote the collection of all decision rules.

5.0.5 Definition. A decision rule d is an *inadmissible decision rule* if there is another rule d' such that $R(\theta, d') \leq R(\theta, d)$ for all θ and $R(\theta, d') < R(\theta, d)$ for some

θ. An *admissible decision rule* is one that is not inadmissible. A decision rule d^* is *minimax* if

$$\inf_{d \in D} \sup_{\theta \in \Theta} R(\theta, d) = \sup_{\theta \in \Theta} R(\theta, d^*).$$

A probability distribution Λ on Θ is called a *prior distribution*. Given a prior Λ and $d \in D$, the *Bayes risk* is

$$R(\Lambda, d) = \int_\Theta R(\theta, d) \Lambda(d\theta).$$

A *Bayes rule* is any decision rule d_Λ satisfying $R(\Lambda, d_\Lambda) = \inf_{d \in D} R(\Lambda, d)$.

5.1 Game theory: the finite case

Suppose Θ, X, and A are all finite sets, of sizes ℓ, n, and k, respectively.

5.1.1 Lemma. *D is a bounded convex set.*

Proof: In this setting, D is the collection of $k \times n$ matrices with non-negative entries, all of whose columns sum to one. □

5.1.2 Theorem. $R := \{(R(\theta_1, d), \dots, R(\theta_\ell, d)) \in \mathbb{R}^\ell_+ : d \in D\}$ *is convex.*

Proof: Let $\alpha \in [0,1]$ and $d_1, d_2 \in D$ be given.

$$\begin{aligned} &\alpha R(\theta, d_1) + (1-\alpha) R(\theta, d_2) \\ &= \alpha \sum_{j=1}^n \sum_{i=1}^k d_{ij}^{(1)} L(\theta, a_i) p_\theta(x_j) + (1-\alpha) \sum_{j=1}^n \sum_{i=1}^k d_{ij}^{(2)} L(\theta, a_i) p_\theta(x_j) \\ &= \sum_{j=1}^n \sum_{i=1}^k (\alpha d_1 + (1-\alpha) d_2)_{ij} L(\theta, a_i) p_\theta(x_j) \\ &= R(\theta, \alpha d_1 + (1-\alpha) d_2) \end{aligned}$$

□

5.1.3 Theorem. *Every $d \in D$ can be expressed as a convex combination of non-randomized decision rules.*

This formula for λ_r is unverified.

Proof: The extreme points of D are the non-randomized decision rules. In fact $d = \sum_{r=1}^{k^n} \lambda_r \delta^{(r)}$, where

$$\lambda_r := \prod_{j=1}^n \prod_{i=1}^k d_{ij}^{\delta_{ij}^{(r)}}$$

and $\{\delta^{(r)} : 1 \le r \le k^n\}$ enumerate the non-randomized rules. □

5.1.4 Example. Let $\Theta = \Theta_0 \amalg \Theta_1$, $A = \{0,1\}$, $d(1|x) = \phi(x) = 1 - d(0|x)$, $L(\theta,0) = \ell_0 \mathbf{1}_{\theta\in\Theta_1}$, and $L(\theta,1) = \ell_1 \mathbf{1}_{\theta\in\Theta_0}$. Then

$$R(\theta,d) = \begin{cases} \ell_1 \mathbb{E}_\theta[\phi(X)] & \theta \in \Theta_0 \\ \ell_0 \mathbb{E}_\theta[1-\phi(X)] & \theta \in \Theta_1. \end{cases}$$

Thus decision theory is clearly a generalization of hypothesis testing, where the risk corresponds to the power.

5.1.5 Example (Urns). Let $\Theta = \{1,2\}$, the urns. In urn 1 there are 10 red, 20 blue, and 70 green balls, and in urn 2 there are 40 red, 40 blue, and 20 green balls. One ball is draw, and the problem is to decide which urn it came from. Whence $X = \{R,G,B\}$ and $A = \Theta$. Arbitrarily, let

$$L(\theta,a) = \begin{bmatrix} 0 & 10 \\ 6 & 0 \end{bmatrix}$$

which can be regarded as Hamming loss $\mathbf{1}_{a\neq\theta}$, θ-by-θ. (In particular it is not convex.) Let $d = (d_R, d_B, d_G)$ be the decision rule, specified as the probability of choosing urn 1 given that the colour R, B, or G was shown.

The risk is computed as follows.

$$\begin{aligned} R(1,d) &= \sum_{i=1}^{2}\sum_{j=1}^{3} L(1,j)d(j,x_j)p_1(x_j) \\ &= 10\sum_{j=1}^{3} d(2,x_j)p_1(x_j) \\ &= 10(0.1(1-d_R) + 0.2(1-d_B) + 0.7(1-d_G)) \\ &= 10 - d_R - 2d_B - 7d_G \\ \text{and } R(2,d) &= 2.4d_R + 2.4d_B + 1.2d_G \end{aligned}$$

The Bayes risk is simply

$$\begin{aligned} R(\Lambda,d) &= \lambda R(1,d) + (1-\lambda)R(2,d) \\ &= \lambda(10 - 3.4d_R - 4.4d_B - 8.2d_G) + (2.4d_R + 2.4d_B + 1.2d_G) \\ &= 10\lambda + (2.4 - 3.4\lambda)d_R + (2.4 - 4.4\lambda)d_B + (1.2 - 8.2\lambda)d_G \end{aligned}$$

The Bayes rule for $\lambda = 0.5$ is $d_R = d_B = 0$ (since they have positive coefficients) and $d_G = 1$ (because it has a negative coefficient). In fact, the Bayes rule will always be a non-randomized rule. For the case $\lambda = 12/17$ the Bayes rule is not unique, because d_A may be any number and the Bayes risk is still minimized.

5.2 Minimax rules, Bayes rules, and admissibility

Remark (Non-randomized rules vs. randomized rules). Corollary 1.7.9 of the textbook says that if the loss function is convex then we can focus on non-randomized rules. More precisely, if $L(\theta, a)$ is convex in a for all θ then, given any randomized estimator, there is a non-randomized estimator that is at least as good, and uniformly better if strictly convex.

Remark (Admissibility and Bayes). Under mild conditions it can be shown that every admissible rule is a Bayes rule for some prior distribution. Theorem 4.2.4 in the textbook says that if a Bayes rule is unique (i.e. if there is a unique d minimizing $\int_\Theta R(\theta, d)\Lambda(d\theta)$) then it must be admissible.

5.2.1 Definition. A prior Λ is a *least favourable prior* if $\gamma_\Lambda \geq \gamma_{\Lambda'}$ for all other priors Λ', where $\gamma_\Lambda := \int R(\theta, d_\Lambda)\Lambda(d\theta)$ is the Bayes risk of the Bayes rule.

5.2.2 Theorem (Least favourable Bayes rules are minimax).
If $\gamma_\Lambda = \sup_\theta R(\theta, d_\Lambda)$ then Λ is a least favourable prior and d_Λ is minimax. Further, if d_Λ is the unique Bayes rule with respect to Λ then it is the unique minimax rule.

Proof: Let d be any other decision procedure.

$$\begin{aligned}\sup_\theta R(\theta, d) &\geq \int R(\theta, d)\Lambda(d\theta) \\ &\geq \int R(\theta, d_\Lambda)\Lambda(d\theta) && \text{by definition of } d_\Lambda \\ &= \sup_\theta R(\theta, d_\Lambda) && \text{by assumption}\end{aligned}$$

If d_Λ is the unique Bayes rule for Λ then the second inequality is strict. □

5.2.3 Corollary. *If a Bayes rule has constant risk (i.e. $R(\theta, d_\Lambda)$ does not depend on θ) then it is minimax.*

5.2.4 Corollary. *Let $\Theta_\Lambda := \{\theta \in \Theta : R(\theta, d_\Lambda) = \sup_\theta R(\theta, d_\Lambda)\}$. d_Λ is minimax if $\Lambda(\Theta_\Lambda) = 1$*

5.2.5 Example (Urns, revisited). The minimax rule is found by solving

$$\max\{10 - d_R - 2d_B - 7d_G, 2.4d_R + 2.4d_B + 1.2d_G\}.$$

The solution is $d^* = (d_R, d_B, d_G) = (0, 9/22, 1)$. This is a Bayes rule with $\lambda = 6/11$, but there is no unique Bayes rule for this Λ. This d^* has constant risk $R(1, d^*) = R(2, d^*) = 24/11$.

5.2.6 Theorem. *Let $L(\theta, a) = K(\theta)(\theta - a)^2$ be weighted squared error loss (K is a positive function). Then the Bayes estimator for $g(\theta)$ is*

$$d_\Lambda(x) = \frac{\mathbb{E}[K(\theta)g(\theta)|x]}{\mathbb{E}[K(\theta)|x]} = \frac{\int_\Theta K(\theta)g(\theta)\Lambda(d\theta|x)}{\int_\Theta K(\theta)\Lambda(d\theta|x)},$$

where $\Lambda(\theta|x)$ is the posterior distribution. When $K(\theta) \equiv 1$, $d_\Lambda(x) = \mathbb{E}[g(\theta)|x]$.

PROOF: Let d be any other estimator and let $h(t)$ be the Bayes risk of $d_\Lambda + td$ conditional on x.

$$h(t) := \int_\Theta K(\theta)(g(\theta) - d_\Lambda(x) - td(x))^2 \Lambda(d\theta|x).$$

For each x, h is a differentiable function of t, and $h'(0) = 0$ since d_Λ minimizes the Bayes risk.

$$0 = \frac{d}{dt}h(t)\bigg|_{t=0} = -2d(x)\int_\Theta K(\theta)(g(\theta) - d_\Lambda(x))\Lambda(d\theta|x)$$

For each x there is d such that $d(x) \neq 0$, so we can solve for $d_\Lambda(x)$ and see that it has the desired expression. □

5.3 Conjugate Priors

5.3.1 Definition. Suppose that x has density $p(x|\theta)$ and Λ is a prior distribution. If the posterior $\Lambda(\cdot|x)$ has the same form (i.e. lies in the same parametric family of probability distributions) as the prior then Λ is a *conjugate prior*.

5.3.2 Example. Let $X \sim \text{binomial}(n,\theta)$ and $\theta \sim \text{Beta}(\alpha,\beta) = \Lambda$. It can be shown that $(\theta|x) \sim \text{Beta}(\alpha+x, \beta+n-x)$. If $L(\theta,a) = (\theta - a)^2$ is squared error loss then

$$d_\Lambda = \mathbb{E}[\theta|X] = \frac{\alpha+X}{\alpha+\beta+n} = \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\frac{X}{n}$$

The risk in this case is

$$\begin{aligned} R(\theta, d_\Lambda) &= \mathbb{E}[(d_\Lambda) - \theta)^2] \\ &= \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)^2 \left(\frac{\alpha}{\alpha+\beta}\right)^2 + \left(\frac{n}{\alpha+\beta+n}\right)^2 \frac{\theta(1-\theta)}{n} \\ &= \frac{\alpha^2 + (n - 2\alpha(\alpha+\beta))\theta + ((\alpha+\beta)^2 - n)\theta^2}{(\alpha+\beta+n)^2} \end{aligned}$$

If we can choose α and β so that is does not depend on θ then we will have the minimax rule by 5.2.3. We need $(\alpha+\beta)^2 = 2\alpha(\alpha+\beta) = n$, so solving, $\alpha = \beta = \sqrt{n}/2$ and the minimax rule is obtained by taking $\Lambda = \text{Beta}(\sqrt{n}/2, \sqrt{n}/2)$. In this case

$$d^* = \frac{1}{1+\sqrt{n}}\left(\frac{1}{2} + \frac{x}{\sqrt{n}}\right) = \overline{x}_n + o\left(\frac{1}{\sqrt{n}}\right)$$

and the minimax risk is $1/(4(1+\sqrt{n})^2)$.

5.3.3 Example (Normal with known variance).
Let $X \sim N(\theta, \sigma^2)$ and $\theta \sim N(\mu, \tau^2)$. It can be shown that

$$(\theta|x) \sim N\left(\frac{1/\tau^2}{1/\tau^2 + 1/\sigma^2}\mu + \frac{1/\sigma^2}{1/\tau^2+1/\sigma^2}X, \frac{1}{1/\tau^2+1/\sigma^2}\right)$$

and so, for squared error loss,

$$d_\Lambda = \mathbb{E}[\theta|X] = \frac{1/\tau^2}{1/\tau^2 + 1/\sigma^2}\mu + \frac{1/\sigma^2}{1/\tau^2 + 1/\sigma^2}X.$$

The Bayes risk is

$$\begin{aligned} R(\theta, d_\Lambda) &= \mathbb{E}[(d_\Lambda - \theta)^2] \\ &= \left(\frac{1/\tau^2}{1/\tau^2 + 1/\sigma^2}\right)^2 (\mu - \theta)^2 + \left(\frac{1/\sigma^2}{1/\tau^2 + 1/\sigma^2}\right)^2 \sigma^2. \end{aligned}$$

Until now we have not used the fact that σ^2 is known. Even if it is know, there is no prior that can be chosen so that the Bayes risk does not depend on θ, so we cannot apply 5.2.3. But by taking τ larger and larger, the term with θ can be made as small as desired. In the limit with $\tau = \infty$, there is no dependence on θ. This motivates the definition of *improper prior*, a measure Λ on Θ with $\Lambda(\Theta) = \infty$. This example is continued below.

5.3.4 Theorem. *Suppose that $\theta \sim \Lambda$ and $(X|\theta) \sim P_\theta$. Assume there is a rule d_0 that has finite risk. If, for a.e. x, $d_\Lambda(x)$ minimizes*

$$\mathbb{E}[L(\theta, d_\Lambda(\cdot|x))|X = x] = \int_\Theta L(\theta, d(\cdot|x))\Lambda(d\theta|x).$$

then d_Λ is the Bayes rule.

5.3.5 Definition. A *generalized Bayes rule* is a rule satisfying the condition of 5.3.4, but with Λ an improper prior instead. A rule d is a *limiting Bayes rules* if there is a sequence of proper priors Λ_k such that $d(x) = \lim_{k\to\infty} d_{\Lambda_k}(x)$ a.s. Such a sequence $\{\Lambda_k\}_{k=1}^\infty$ is said to be a *least favourable sequence of priors* if $\gamma_{\Lambda_k} \to \gamma$ and $\gamma_\Lambda \le \gamma$ for all proper priors Λ.

5.3.6 Theorem. *Suppose that $\{\Lambda_k\}_{k=1}^\infty$ is a sequence of priors with $\gamma_{\Lambda_k} \to \gamma$ and d^* is an estimator with $\sup_\theta R(\theta, d^*) = \gamma$. Then d^* is minimax and the sequence is least favourable.*

PROOF: Let d be any other decision procedure. For all $k \ge 1$,

$$\sup_\theta R(\theta, d) \ge \int_\Theta R(\theta, d)\Lambda_k(d\theta) \ge \gamma_{\Lambda_k}.$$

Taking $k \to \infty$, and noting that the right hand side converges to γ,

$$\sup_\theta R(\theta, d) \ge \gamma = \sup_\theta R(\theta, d^*),$$

so d^* is minimax. Further, for any prior Λ,

$$\gamma_\Lambda = \int_\Theta R(\theta, d_\Lambda) d\theta \le \int_\Theta R(\theta, d^*)\Lambda(d\theta) \le \sup_\theta R(\theta, d^*) = \gamma,$$

so the sequence is least favourable. □

In 5.3.3, take $\Lambda_k = N(\mu, k)$, where μ is arbitrary because it is irrelevant. Then $\gamma_{\Lambda_k} \to \sigma^2$ and the rules d_{Λ_k} converge to the rule $d(x) = x$. We have seen that $R(\theta, X) = \sigma^2$, so the sample mean is minimax by 5.3.6.

5.3.7 Lemma. *Let X be a random quantity with distribution F and let $g(F)$ be a functional defined over a set $\mathscr{F}_1$ of distributions. Suppose that δ is a minimax estimator of $g(F)$ when F is restricted to a subset $\mathscr{F}_0 \subseteq \mathscr{F}_1$. Then δ is a minimax estimator of $g(F)$ for $F \in \mathscr{F}_1$ if $\sup_{F\in\mathscr{F}_0} R(F, \delta) = \sup_{F\in\mathscr{F}_1} R(F, \delta)$.*

5.3.8 Example (Normal with unknown variance).
Let $X \sim N(\theta, \sigma^2)$, where $\theta \in \mathbb{R}$ and $\sigma^2 \le M$. For squared error loss, from the calculations in 5.3.3, $\sup_{\theta,\sigma^2} R(\theta, x) = M$. (Note that there is a big problem if we consider $M = \infty$.) In 5.3.3 we saw that the minimax estimator is the sample mean when σ^2 is known. Applying the lemma, the sample mean remains minimax when σ^2 is unknown but known to be bounded.

5.3.9 Example. Let $X_1, \dots, X_n \sim F$, where $F \in \mathscr{F}_\mu := \{F : \mathbb{E}_F |X| < \infty\}$.

(i) Let $\mathscr{F}_B := \{F \in \mathscr{F}_\mu : \mathbb{P}_F[0 \le X \le 1] = 1\}$. What is the minimax estimator for $\theta = \mathbb{E}_F[X]$ under squared error loss? Consider the subclass $\mathscr{F}_0 := \{\text{Bernoulli}(p) : 0 \le p \le 1\}$. The sum of n Bernoulli r.v.'s is binomial, so the estimator $d(x) = (1/2 + \sqrt{n}\overline{x})/(1 + \sqrt{n})$ is minimax over the class $\mathscr{F}_0$. For any $F \in \mathscr{F}_B$, write

$$\begin{aligned} d &= \frac{1}{1+\sqrt{n}}\frac{1}{2} + \frac{\sqrt{n}}{1+\sqrt{n}}\overline{x} \\ \mathbb{E}[(d-\theta)^2] &= \frac{1}{(1+\sqrt{n})^2}\left(\frac{1}{2} - \theta\right)^2 + \left(\frac{\sqrt{n}}{1+\sqrt{n}}\right)^2 \text{Var}(\overline{X}) \\ &= \frac{1}{(1+\sqrt{n})^2}\left(\mathbb{E}_F[X^2] - \theta + \frac{1}{4}\right) \end{aligned}$$

Now, $\mathbb{E}_F[X^2] \le \mathbb{E}_F[X] = \theta$, so $\mathbb{E}[(d-\theta)^2] \le 1/(4(1+\sqrt{n})^2)$, which is the minimax risk in the class $\mathscr{F}_0$. Therefore the minimax estimator is d.

(ii) Let $\mathscr{F}_V := \{F \in \mathscr{F}_\mu : \text{Var}_F(X) \le M\}$ and again $\theta = \mathbb{E}_F[X]$. Consider the subclass $\mathscr{F}_0 := \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \le M\}$. It can be shown that $\sup_{F\in F_V} R(F, \overline{X}) = M/n$, which is the minimax risk in the subclass, so the sample mean is minimax for θ.

5.4 Inadmissibility

Recall that any unique Bayes rule is admissible by Theorem 4.2.4 in the textbook.

5.4.1 Example. Let $X_1, \dots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. For $\theta \sim N(\mu, \tau^2)$ and squared error loss, we derived that $d_\Lambda(x) = p\mu + (1-p)\overline{x}$, where

$$p := \frac{1/\tau^2}{1/\tau^2 + 1/\sigma^2}.$$

This is the unique Bayes rule, so it is admissible. It is interesting to ask for which (a, b) is $d(x) := a\overline{x} + b$ admissible? This example shows that any such rule with $0 < a < 1$ and b arbitrary is admissible.

5.4.2 Theorem. *Let X have mean θ and variance σ^2 (but not necessarily normally distributed). Then $d_{a,b}(x) = ax + b$ is inadmissible for θ in the above example under squared error loss if*

(i) $a > 1$*;*
(ii) $a < 0$*; or*
(iii) $a = 1$*,* $b \neq 0$*.*

Proof: We must demonstrate a rule d such that $R(\theta, d_{a,b}) \geq R(\theta, d)$ for all θ and strictly greater for some θ.

$$\begin{aligned} R(\theta, d_{a,b}) &= \mathbb{E}_\theta[(aX + b - \theta)^2] \\ &= \mathbb{E}_\theta[(a(X-\theta) + b + (a-1)\theta)^2] \\ &= a^2\,\mathbb{E}_\theta[(X-\theta)^2] + (b + (a-1)\theta)^2 \\ &= a^2\sigma^2 + (b + (a-1)\theta)^2 \end{aligned}$$

Recall that the risk of $d(x) = x$ is σ^2. If $a > 1$ or $a = 1$ and $b \neq 0$ then $R(\theta, d_{a,b}) > \sigma^2 = R(\theta, d)$, so this takes care of (i) and (ii). For the third case, if $a < 0$ then

$$R(\theta, d_{a,b}) > (b + (a-1)\theta)^2 > \left(\theta - \frac{b}{1-a}\right)^2 = R\left(\theta, \frac{b}{1-a}\right). \qquad \square$$

5.4.3 Theorem (Stein, 1956). *Let $\mu \in \mathbb{R}^k$ and $X_1, \dots, X_n \sim N(\mu, \sigma^2 I_k)$, so that $\overline{X} = \frac{1}{n}\sum_{i=1i}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n} I_k)$.*

(i) If $k = 1$ then $\overline{X}$ is admissible.
(ii) If $k = 2$ then $\overline{X}$ is admissible.
(iii) If $k \geq 3$ then $\overline{X}$ is not admissible.

Remark. This is discussed in Stein (1959) and discussed at length in James and Stein (1962). They showed that the estimator

$$\hat{\mu} := \overline{X}\left(1 - \frac{\sigma^2(k-2)}{n\|\overline{X}\|^2}\right)_+$$

(or something similar) has strictly smaller risk than $\overline{X}$ when $k \geq 3$. When k is very large then apparently this is very close to $\overline{X}$. In many practical cases μ will have most coordinates equal to zero, so in these cases the threshold estimator is better still.

$$\hat{\mu}_i := \begin{cases} X_i & |X_i| > t \\ 0 & |X_i| \leq t \end{cases} \quad \text{or} \quad \hat{\mu}_i := \begin{cases} X_i - t & |X_i| > t \\ X_i + t & |X_i| < -t \\ 0 & |X_i| \leq t \end{cases}$$

Proof ($k = 1$ using the Cramér-Rao bound):
We prove that $\overline{X}$ is admissible when $k = 1$ and $\sigma^2 = 1$. For squared error loss, $R(\theta, \overline{X}) = 1/n$. Assume that d is an estimator such that $R(\theta, d) \leq 1/n$ for all θ. Note that

$$R(\theta, d) = \mathbb{E}_\theta[(d(X) - \theta)^2] = \text{Var}_\theta(d(X)) + (b(\theta))^2,$$

where $b(\theta) := \mathbb{E}_\theta[d(X)] - \theta$ is the bias. By the Cramér-Rao lower bound,

$$\text{Var}_\theta(d(X)) \geq \frac{(1 + b'(\theta))^2}{nI(\theta)} = \frac{(1 + b'(\theta))^2}{n}.$$

Hence, for all θ,

$$(b(\theta))^2 + \frac{(1 + b'(\theta))^2}{n} \leq \frac{1}{n}.$$

In particular, $|b(\theta)| \leq 1/\sqrt{n}$ for all θ, so b is a bounded function, and $b'(\theta) \leq 0$ for all θ, so b is a decreasing function. Since b is bounded, there are sequences $\theta_i^+ \to \infty$ and $\theta_i^- \to -\infty$ such that $b'(\theta_i^+) \to 0$ and $b'(\theta_i^-) \to 0$. From the inequality, it follows that $b(\theta_i^+) \to 0$ and $b(\theta_i^-) \to 0$. Since b is decreasing, it must be the zero function. Since $\overline{X}$ is the UMVUE, $d(X) = \overline{X}$. □

Proof ($k = 1$ using Bayes estimation):
Here is another proof for the case $k = 1$ and $\sigma^2 = 1$. Assume for contradiction that $\overline{X}$ is inadmissible then there is d^* such that $R(\theta, d^*) \leq 1/n$ for all θ and strictly less for some θ. The risk is a continuous function of θ, so there is an interval $[\theta_0, \theta_1]$ and an $\varepsilon > 0$ such that $R(\theta, d^*) < 1/n - \varepsilon$ for all $\theta_0 \leq \theta \leq \theta_1$. Let Λ_τ denote the prior $N(0, \tau^2)$, and let d_τ be the associated Bayes estimator. Then

$$\gamma_\tau^* := \int R(\theta, d^*)\Lambda_\tau(\theta) d\theta \geq \int R(\theta, d_\tau)\Lambda(\theta) d\theta = \gamma_\tau$$

by the definition of Bayes estimator. We have seen that, since $\sigma^2 = 1$,

$$\gamma_\tau = \frac{1}{1/\tau^2 + n/\sigma^2} = \frac{\tau^2}{1 + n\tau^2}.$$

Now,

$$\frac{1/n - \gamma_\tau^*}{1/n - \gamma_\tau} = \frac{\frac{1}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty}(1/n - R(\theta, d^*))e^{-\theta^2/2\tau^2} d\theta}{1/n - \tau^2/(1 + n\tau^2)}$$

$$
\begin{aligned}
&= \frac{n(1+n\tau^2)}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty}(1/n - R(\theta,d^*))e^{-\theta^2/2\tau^2}d\theta \\
&\geq \frac{n(1+n\tau^2)}{\sqrt{2\pi}\tau}\int_{\theta_0}^{\theta_1}(1/n - R(\theta,d^*))e^{-\theta^2/2\tau^2}d\theta \\
&> \frac{n(1+n\tau^2)}{\sqrt{2\pi}\tau}\varepsilon(\Phi(\theta_1)-\Phi(\theta_0))
\end{aligned}
$$

This last quantity goes to ∞ as $\tau \to \infty$. Therefore, for sufficiently large τ, $1/n - \gamma_\tau^* > 1/n - \gamma_\tau$. But this contradictions that $\gamma_\tau \leq \gamma_\tau^*$. □

Recall *Stein's Lemma*: When $k = 1$, if $X \sim N(\theta, \sigma^2)$ and g is a function with $\mathbb{E}[|g'(X)|] < \infty$ then

$$\sigma^2\,\mathbb{E}[g'(X)] = \mathbb{E}[(X-\theta)g(X)].$$

In k dimensions, if $X \sim N(\theta, \sigma^2 I_k)$ and $g : \mathbb{R}^k \to \mathbb{R}$ is with $\mathbb{E}[|\frac{\partial}{\partial x_i}g(X)|] < \infty$ then

$$\sigma^2\,\mathbb{E}\left[\frac{\partial}{\partial x_i}g(X)\right] = \mathbb{E}[(X_i-\theta_i)g(X)].$$

Proof ($k \geq 3$, inadmissibility): For simplicity, take $\sigma^2 = 1$. Then we have seen that $\mathbb{E}[|\overline{X}-\theta|^2] = 1/n$. We look for an estimator $\hat{\theta}_n$ of the form

$$\hat{\theta}_n = \overline{X} + \frac{1}{n}g(\overline{X})$$

where $g : \mathbb{R}^k \to \mathbb{R}^k$ is to be determined, such that for all θ,

$$
\begin{aligned}
0 &< \mathbb{E}_\theta[|\overline{X}-\theta|^2] - \mathbb{E}_\theta\left[\left|\overline{X}+\frac{1}{n}g(\overline{X})-\theta\right|^2\right] \\
&= -\frac{1}{n^2}\,\mathbb{E}[|g(\overline{X})|^2] - \frac{2}{n}\sum_{i=1}^{k}\mathbb{E}[(\overline{X}_i-\theta_i)g_i(\overline{X})] \\
&= -\frac{1}{n^2}\,\mathbb{E}[|g(\overline{X})|^2] - \frac{2}{n^2}\sum_{i=1}^{k}\mathbb{E}\left[\frac{\partial}{\partial x_i}g_i(\overline{X})\right].
\end{aligned}
$$

Furthermore, we would like $g = \nabla \log \psi$ for some $\psi : \mathbb{R}^k \to (0,\infty)$. Then

$$\frac{\partial}{\partial x_i}g_i(x) = \frac{\partial}{\partial x_i}\left(\frac{1}{\psi(x)}\frac{\partial\psi}{\partial x_i}\right) = -\frac{1}{\psi(x)^2}\left(\frac{\partial\psi}{\partial x_i}\right)^2 + \frac{1}{\psi(x)}\frac{\partial^2\psi}{\partial x_i^2}.$$

With regards to the desired expression for g, x-by-x,

$$-\frac{1}{n^2}\sum_{i=1}^{k}\frac{1}{\psi(x)^2}\left(\frac{\partial\psi}{\partial x_i}\right)^2 - \frac{2}{n^2}\left(\sum_{i=1}^{k}-\frac{1}{\psi(x)^2}\left(\frac{\partial\psi}{\partial x_i}\right)^2 + \frac{1}{\psi(x)}\frac{\partial^2\psi}{\partial x_i^2}\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{k}\frac{1}{\psi(x)^2}\left(\frac{\partial\psi}{\partial x_i}\right)^2 - \frac{2}{n^2}\frac{\Delta\psi(x)}{\psi(x)}.$$

So we want

$$0 < \frac{1}{n^2}\mathbb{E}[|g(\overline{X})|^2] - \frac{2}{n^2}\mathbb{E}\left[\frac{\Delta\psi(\overline{X})}{\psi(\overline{X})}\right].$$

If we take ψ to be harmonic then we get what we wanted. Take $\psi(x) = |x|^{2-k}$, which is harmonic when $k \geq 3$. Then $g(x) = \nabla \log\psi(x) = \frac{2-k}{|x|^2}x$. Plugging this in,

$$\hat{\theta}_n = \left(1 - \frac{k-2}{n|\overline{X}|^2}\right)\overline{X}$$

is a strictly better estimator, with a difference in squared error risk of

$$\frac{(k-2)^2}{n^2}\mathbb{E}\left[\frac{1}{|\overline{X}|^2}\right].$$

□

The problem of estimating the mean of n normals has been studied in depth for a long time. This problem is important in non-parametric estimation, in that such problems can often be reduced to estimating the mean of n normals, where n is a large number. Often Stein's shrinkage is not enough of a correction.

Herbert, Robbins, Bradley, Efron, Larry, and Brown introduced and studied the *empirical Bayes method* to solve the same problem. If $X \sim N(\theta, \sigma^2 I_k)$, with σ^2 known, and $\theta \sim N(0, \tau^2 I_k)$ then $d_\tau = \frac{\tau^2}{\sigma^2+\tau^2}X$ is the Bayes estimator. The idea is to use the data to pick the "best" τ, say $\tau = \hat{\tau}$, and then use $d_{\hat{\tau}}$ as the estimator.

Some of these are first names, but I don't know which are which.

5.4.4 Example. Let X be as above, so that $X \sim N(0, (\sigma^2 + \tau^2)I_k)$, where σ^2 is known and τ^2 is to be estimated. The MLE for τ^2 is $\hat{\tau}^2 = (\|X\|^2/k - \sigma^2)_+$, so

$$d_{\hat{\tau}^2}(X) = \frac{\hat{\tau}^2}{\hat{\tau}^2+\sigma^2} = \left(1 - \frac{k\sigma^2}{\|X\|^2}\right)_+ X,$$

which is every close to Stein's estimator for large k and $n = 1$.

6 Hypothesis Testing

6.1 Simple tests: the Neyman-Pearson lemma

Consider the problem of testing $H_0 : X \sim P_0$ vs. $H_1 : X \sim P_1$, where P_0 and P_1 have densities p_0 and p_1 with respect to a σ-finite measure μ. Recall that a (generalized) decision rule is

$$d(a|x) = \text{probability of taking action } a \text{ when we observe outcome } x.$$

In this case the action space is $\{0,1\}$, corresponding to the hypothesis that we accept.

Let $\psi(x) := d(1|x)$, the probability of accepting the alternative hypothesis H_1 given datum x. The *level of a test* or *size of a test* is $\mathbb{E}_0[\psi(X)]$, the probability of a *type I error* (rejecting the null hypothesis despite it being true). The *power of a test* is $\mathbb{E}_1[\psi(X)]$, the probability of accepting the alternative when it is true. The probability of a *type II error* (accepting the null hypothesis despite the alternative being true) is $\mathbb{E}_1[1-\psi(X)]$.

Recall that (type I error + type II error) $\geq 1 - \frac{1}{2}\|p_0 - p_1\|_1$. In proving this note that the function

$$\phi_1(x) = \begin{cases} 1 & p_1(x) > p_0(x) \\ 0 & p_1(x) < p_0(x) \end{cases}$$

played an important role.

6.1.1 Definition. A level α *most powerful test*, or *MP test*, is a test ϕ with the property that $\mathbb{E}_0[\phi] \leq \alpha$ (i.e. it is level α) and, for every other test ψ of level α, $\mathbb{E}_1[\psi] \leq \mathbb{E}_1[\phi]$.

In this Neyman-Pearson setting, the objective is to fix the level of a test and then maximize the power. Equivalently, fix an acceptable level of type I error and then find the test with the minimum probability of type II error. It turns out that in this simple setup, the most powerful test is of the form

$$\phi_k(x) := \begin{cases} 1 & p_1(x) > kp_0(x) \\ \gamma & p_1(x) = kp_0(x) \\ 0 & p_1(x) < kp_0(x). \end{cases}$$

6.1.2 Theorem (Neyman-Pearson). *Let $0 \leq \alpha \leq 1$.*

(i) There are constants k and γ such that $\mathbb{E}_0[\phi_k(X)] = \alpha$.

(ii) The test ϕ_k from (i) is most powerful for testing P_0 vs. P_1.

(iii) If ψ is a level α most powerful test then it must be equal to ϕ_k μ-a.e., unless there is a test of size α with power 1.

Proof: Let $Y = p_1(X)/p_0(X)$, and let F_Y be the CDF of Y, so that

$$\mathbb{P}\left[\frac{p_1(X)}{p_0(X)} > k\right] = 1 - F_Y(k)$$

Let $k := \inf\{c : 1 - F(c) < \alpha\}$, a non-negative number since $Y \geq 0$, and let

$$\gamma := \frac{\alpha - P_0[Y > k]}{P_0[Y = k]}$$

if $P_0[Y = k] \neq 0$, otherwise take $\gamma = 1$. Then

$$\mathbb{E}_0[\phi_k(X)] = P_0[Y > k] + \gamma P_0[Y = k]$$

$$= P_0[Y > k] + \frac{\alpha - P_0[Y > k]}{P_0[Y = k]} P_0[Y = k] = \alpha$$

Therefore ϕ_k is a level α test. Now we check that is it most powerful. Assume for simplicity that $\mu[p_1 = kp_0] = 0$. Let ψ be another test with $\mathbb{E}_0[\psi] \le \alpha$.

$$\begin{aligned}
(\mathbb{E}_1[\phi_k] - \mathbb{E}_1[\psi]) - k(\mathbb{E}_0[\phi_k] - \mathbb{E}_0[\psi]) \\
&= \int (\phi_k - \psi)(p_1(x) - kp_0(x))\mu(dx) \\
&= \int_{p_1 > kp_0} + \int_{p_1 < kp_0} (\phi_k - \psi)(p_1(x) - kp_0(x))\mu(dx) \\
&= \int_{p_1 > kp_0} (1 - \psi(x))(p_1(x) - kp_0(x))\mu(dx) \\
&\qquad - \int_{p_1 < kp_0} \psi(x)(p_1(x) - kp_0(x))\mu(dx) \ge 0
\end{aligned}$$

because $\phi_k = 1$ on the first set and $\phi_k = 0$ on the second set and $0 \le \psi \le 1$ everywhere. Now, $k \ge 0$ and $\mathbb{E}_0[\phi_k] - \mathbb{E}_0[\psi] \ge \alpha - \alpha = 0$, so $\mathbb{E}_1[\phi_k] - \mathbb{E}_1[\psi] \ge 0$. Therefore ϕ_k is most powerful.

For uniqueness, assume now that ψ is also most powerful at level α. Then $\mathbb{E}_0[\psi] \le \alpha = \mathbb{E}_0[\phi_k]$ and $\mathbb{E}_1[\psi] = \mathbb{E}_1[\phi_k]$. From the latter,

$$-k(\mathbb{E}_0[\phi_k] - \mathbb{E}_0[\psi]) \ge 0,$$

so either $\mathbb{E}_0[\psi] = \alpha$ or $\mathbb{E}_0[\psi] < 0$ and $k = 0$. In the first case

This proof was particularly confusingly presented in class. Try to fill in the case $\mu[p_1 = kp_0] > 0$.

$$0 = \int (\phi_k - \psi)(p_1(x) - kp_0(x))dx,$$

so it follows that $\psi = \phi_k = 1$ on $\{p_1/p_0 > k\}$ and $\psi = \phi_k = 0$ on $\{p_1/p_0 < k\}$, μ-a.e. If $\mathbb{E}_0[\psi] < \alpha$ then $k = 0$, so there is a test of power 1. □

6.1.3 Corollary. *If $0 < \alpha < 1$ and β is the power of the most powerful level α test then $\alpha < \beta$, unless $p_0 = p_1$ μ-a.e.*

Proof: Consider the trivial test of level α, namely the constant α. Then

$$\beta = \mathbb{E}_1[\phi_k] \ge \mathbb{E}_1[\alpha] = \alpha.$$

Moreover, if $\alpha = \beta$ then α is most powerful by the Neyman-Pearson lemma, so $p_1/p_0 = k$ has probability 1, so $k = 1$ and the densities are equal. □

6.2 Complex tests

Let $\{p_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ be a family of densities. The last section was concerned with testing $\theta = \theta_0$ vs. $\theta = \theta_1$, a *simple test*. In this case we know the most powerful tests. A *complex test* involves testing whether θ lies in some infinite subsets of Θ. We will consider the following four complex tests in detail.

$$\begin{aligned}
&H_1 : \theta \le \theta_0 \text{ vs. } K_1 : \theta > \theta_0\\
&H_2 : \theta \le \theta_1 \text{ or } \theta \ge \theta_2 \text{ vs. } K_2 : \theta_1 < \theta < \theta_2\\
&H_3 : \theta_1 \le \theta \le \theta_2 \text{ vs. } K_3 : \theta < \theta_1 \text{ or } \theta > \theta_2\\
&H_4 : \theta = \theta_0 \text{ vs. } K_4 : \theta \ne \theta_0
\end{aligned}$$

What are the most powerful tests in each case?

6.2.1 Definition. For testing $\theta \in \Theta_0$ vs. $\theta \in \Theta_1$, a test ϕ is *uniformly most powerful*, or *UMP*, at level α if $\sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\phi] \le \alpha$ (i.e. it has level α) and if ψ is any other level α test then $\mathbb{E}_\theta[\psi] \le \mathbb{E}_\theta[\phi]$ for all $\theta \in \Theta_1$.

6.2.2 Definition. The family of densities $\{p_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ has the *monotone likelihood ratio property*, or *MLR*, on the interval $[\theta_0, \theta_1] \subseteq \Theta$ if there is a statistic T and a nondecreasing function g such that

$$\frac{p_{\theta'}(x)}{p_\theta(x)} = g(T(x))$$

for each pair $\theta < \theta'$ in $[\theta_0, \theta_1]$,

6.2.3 Example. Consider a one-parameter exponential family,

$$p_\theta(x) = c(\theta)e^{Q(\theta)T(x)}h(x).$$

If $Q(\theta)$ is nondecreasing then this family has MLR. Indeed, if $\theta < \theta'$ then

$$\frac{p_{\theta'}(x)}{p_\theta(x)} = \frac{c(\theta')}{c(\theta)}e^{(Q(\theta')-Q(\theta))T(x)},$$

which is a nondecreasing function of $T(x)$. The binomial and Poisson families have this property.

6.2.4 Example. The Cauchy location family $p_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}$ does not have MLR since $\frac{1+(x-\theta')^2}{1+(x-\theta)^2} \to 1$ as $x \to \infty$ and also as $x \to -\infty$, but is zero at $x = \theta'$, so the ratio cannot be nondecreasing.

6.2.5 Example. Some important MLR families are

(i) $N(\theta, 1)$, $\theta \in \mathbb{R}$
(ii) non-central $\chi_k^2(\delta)$, $\delta \ge 0$, with k fixed
(iii) non-central $t_r(\delta)$, $\delta \ge 0$, with r fixed, where

$$t_r(\delta) = \frac{N(\delta, 1)}{\sqrt{\chi_r^2/r}}$$

(iv) non-central $F_{m,n}(\delta)$, $\delta \geq 0$, with m and n fixed, where

$$F_{m,n}(\delta) = \frac{\chi^2_m(\delta)}{\chi^2_n(0)}.$$

These families show up often, and we often wish to test $\delta = 0$ vs. $\delta > 0$.

6.2.6 Theorem (Karlin-Rubin). *Let X be distributed according to a density p_θ in a family with MLR involving the statistic $T(X)$.*

(i) There is a UMP level α test ϕ_c of $H_1 : \theta \leq \theta_0$ vs. $K_1 : \theta > \theta_1$ with the property that $\mathbb{E}_{\theta_0}[\phi_c(X)] = \alpha$, namely

$$\phi_c(x) = \begin{cases} 1 & T(x) > c \\ \gamma & T(x) = c \\ 0 & T(x) < c. \end{cases}$$

(ii) The power function $\beta(\theta) := \mathbb{E}_\theta[\phi_c(X)]$ is increasing in θ when $\beta(\theta) < 1$.

(iii) For all $\theta < \theta_0$, if ϕ is a test with $\mathbb{E}_\theta[\phi(X)] = \alpha$ then ϕ_c has the smaller power function.

Proof (Sketch): The idea is to construct a test $H : \theta = \theta_0$ vs. $K : \theta = \theta_1$, where θ_1 is fixed for now. By the Neyman-Pearson lemma we get a test $\phi_0(x)$, which we hope to write as $\psi(T(x))$. From the MLR property, $\beta(\theta) := \mathbb{E}_\theta[\phi(X)]$ increasing. As a result, $\mathbb{E}_{\theta_0}[\phi_0(X)] = \alpha$, and $\mathbb{E}_\theta[\phi_0(X)] \leq \alpha$ for all $\theta \leq \theta_0$, so it is level α in the current setting. The second key fact that follows from the MLR property is that ϕ_0 is uniquely determined by (θ_0, α), and in particular does not depend on θ_1. □

There is a geometric representation of the simple test $H : \theta = \theta_0$ vs. $K : \theta = \theta_1$. Let $\mathscr{C} := \{0 \leq \phi \leq 1 \text{ measurable}\}$ be the set of *critical functions* and

$$\Omega := \{(\alpha, \beta) = (\mathbb{E}_0[\phi(X)], \mathbb{E}_1[\phi(X)]) : \phi \in \mathscr{C}\} \subseteq [0,1]^2.$$

This set has the following properties.

(i) $(\alpha, \alpha) \in \Omega$ for all $0 \leq \alpha \leq 1$.

(ii) If $(\alpha, \beta) \in \Omega$ then $(1-\alpha, 1-\beta) \in \Omega$.

(iii) Ω is convex because integration is linear and the convex combination of tests is a test.

(iv) Ω is closed because $\{0 \leq \phi \leq 1\}$ is a weakly compact subset of L^∞(?)

6.2.7 Theorem (Generalized Neyman-Pearson lemma).
Let $f_1, \dots, f_{m+1}$ be real-valued functions defined on a Euclidean space $\mathscr{X}$ and integrable with respect to a σ-finite measure μ (e.g. density functions). Suppose that, for given constants $\alpha_1, \dots, \alpha_m$, there exists a critical function $\phi \in \mathscr{C}$ such that

$$\int_{\mathscr{X}} \phi(x) f_i(x) \mu(dx) = \alpha_i$$

for $i = 1, \dots, m$. Let $\mathscr{C}_0$ be the class of all critical functions such that the above equations hold. Then

(i) Over all $\phi \in \mathscr{C}_0$ there is one that maximizes $\int_{\mathscr{X}} \phi(x) f_{m+1}(x)\mu(dx)$. (When $m = 1$ this quantity is the power.)

(ii) A sufficient condition for $\phi \in \mathscr{C}_0$ to maximize $\int_{\mathscr{X}} \phi(x) f_{m+1}(x)\mu(dx)$ is that there are constants $k_1, \dots, k_m \in \mathbb{R}$ such that

$$\phi(x) = \begin{cases} 1 & f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x) \\ 0 & f_{m+1}(x) < \sum_{i=1}^m k_i f_i(x). \end{cases}$$

(iii) Further, if the k_i may be chosen to be non-negative then ϕ maximizes $\int_{\mathscr{X}} \phi(x) f_{m+1}(x)\mu(dx)$ over the larger class

$$\left\{ \phi : \int_{\mathscr{X}} \phi(x) f_i(x)\mu(dx) \le \alpha_i \text{ for all } i = 1, \dots, m \right\}.$$

(iv) Define $\mathscr{M} := \{(\int_{\mathscr{X}} \phi(x) f_1(x)\mu(dx), \dots, \int_{\mathscr{X}} \phi(x) f_m(x)\mu(dx)) : \phi \in \mathscr{C}\} \subseteq \mathbb{R}^m$. If $(\alpha_1, \dots, \alpha_m)$ is an interior point of $\mathscr{M}$ then there are constants $k_1, \dots, k_m$ and a test function $\phi \in \mathscr{C}_0$ of the form above that maximizes $\int_{\mathscr{X}} \phi(x) f_{m+1}(x)\mu(dx)$. Moreover, a necessary condition for an element of $\mathscr{C}_0$ to maximize $\int_{\mathscr{X}} \phi(x) f_{m+1}(x)\mu(dx)$ is that it is of the form above μ-a.e.

6.2.8 Theorem. *Consider a one parameter exponential family*

$$p_\theta(x) = c(\theta) e^{Q(\theta)T(x)} h(x)$$

with Q strictly increasing. For the problem of testing

$$H_2 : \theta \le \theta_1 \text{ or } \theta \ge \theta_2 \text{ vs. } K_2 : \theta_1 < \theta < \theta_2.$$

there is a test

$$\phi(x) = \begin{cases} 1 & c_1 < T(x) < c_2 \\ \gamma_i & T(x) = c_i, i = 1, 2 \\ 0 & T(x) < c_1 \text{ or } T(x) > c_2. \end{cases}$$

where $c_1, c_2, \gamma_1, \gamma_2$ are determined by $\alpha = \mathbb{E}_{\theta_1}[\phi(X)] = \mathbb{E}_{\theta_2}[\phi(X)]$, such that

(i) ϕ is UMP, i.e. ϕ maximizes $\mathbb{E}_\theta[\phi(X)]$ over all tests of the same level for every $\theta_1 < \theta < \theta_2$.

(ii) ϕ minimizes $\mathbb{E}_\theta[\phi(X)]$ over all tests of the same level for every $\theta < \theta_1$ and $\theta > \theta_2$.

(iii) The associated power function is unimodal, unless $T(X)$ is concentrated on the same set $\{t_1, t_2\}$ for all θ.

We will refer to Lemma 3.4.2 from the textbook: If $\phi = \mathbf{1}_{[a,b]}$ and satisfies the constraints $\alpha = \mathbb{E}_{\theta_1}[\phi(X)] = \mathbb{E}_{\theta_2}[\phi(X)]$ and $\theta_1 \neq \theta_2$ then ϕ is uniquely determined μ-a.e.

PROOF: We claim that (α,α) is an interior point of the region Ω. Indeed, it can be shown that $(\alpha,\alpha+\Delta)$ is in the region for Δ sufficiently small. Take ϕ to be a level α MP test for $\theta=\theta_1$ vs. $\theta=\theta_2$, so that $\beta=\mathbb{E}_{\theta_1}[\phi(X)]>\alpha$. The symmetry of Ω completes the proof of the claim.

For part (i), note that T is a sufficient statistic for θ, so we may focus on functions of the form $\phi(x)=\psi(T(x))=\psi(t)$. Let $p_\theta^T(t)dt=c(\theta)e^{Q(\theta)t}\nu(dt)$ denote the density of $T=T(X)$.

First fix θ' such that $\theta_1<\theta'<\theta_2$ and choose a test $\psi(T)$ such that $\mathbb{E}_{\theta_1}[\psi(T)]=\mathbb{E}_{\theta_2}[\psi(T)]=\alpha$ and $\mathbb{E}_{\theta'}[\psi(T)]$ is as large as possible. Note that for the region $\mathscr{M}=\{(\mathbb{E}_{\theta_1}[\psi(T)],\mathbb{E}_{\theta_2}[\psi(T)]):\psi\}$, (α,α) is an interior point. Applying Theorem 3.6.1 with $f_1=p_{\theta_1}(t)$, $f_2=p_{\theta_2}(t)$, and $f_3=p_{\theta'}(t)$ (so $m=2$) and $\int\psi f_1=\alpha=\int\psi f_2$, to maximize $\int\psi f_3$. There is ψ_0 that does this, and there are numbers k_1 and k_2 such that

$$\phi_0(t)=\begin{cases}1 & f_3(t)>k_1f_1+k_2f_2\\ 0 & f_3(t)<k_1f_1+k_2f_2.\end{cases}$$

The desired test will be $\phi_0(x)=\psi_0(T(x))$. We have expressions for all these densities, so $\phi_0(t)=1$ only if

$$k_1c(\theta_1)e^{Q(\theta_1)t}+k_2c(\theta_2)e^{Q(\theta_2)t}<c(\theta')e^{Q(\theta')t}.$$

Write this as $a_1e^{b_1t}+a_2e^{b_2t}<1$, where $a_1=k_1c(\theta_1)/c(\theta')$, $a_2=k_2c(\theta_2)/c(\theta')$, $b_1=Q(\theta_1)-Q(\theta')<0$, and $b_2=Q(\theta_2)-Q(\theta')>0$. There are three cases:

(a) $a_1\le 0$ and $a_2\le 0$
(b) $a_1\le 0$ and $a_2>0$
(c) $a_1>0$ and $a_2\le 0$
(d) $a_1>0$ and $a_2>0$.

We claim that only (d) can hold. If (a) holds then $\psi_0\equiv 1$, which contradicts that $\mathbb{E}_{\theta_1}[\psi(T)]=\alpha<1$. If (b) or (c) holds then the left hand side of the inequality is a monotone function of t, implying that the rejection region is a half-infinite interval. By Lemma 3.4.2 and Lemma 3.7.1, $\mathbb{E}_\theta[\phi(T(X))]$ is a strictly monotone function of θ when ψ is of this form, so this contradicts that $\mathbb{E}_{\theta_1}[\psi(T)]=\mathbb{E}_{\theta_2}[\psi(T)]=\alpha$.

So we know we are in case (d). The inequality holds then for $c_1<t<c_2$ for some constants c_1 and c_2, and $\psi=\mathbf{1}_{[c_1,c_2]}$, more or less. It remains to determine $\gamma_1:=\psi(c_1)$ and $\gamma_2:=\psi(c_2)$. But by Lemma 3.7.1 there is only one choice μ-a.e. Further, these depend on $(\theta_1,\theta_2,\alpha)$ and not on θ'. It follows that ϕ_0 is MP for testing $\theta\in\{\theta_1,\theta_2\}$ vs. $\theta=\theta'$, and is independent of θ'. Therefore it is MP for testing $\theta\in\{\theta_1,\theta_2\}$ vs. $K_3:\theta_1<\theta<\theta_2$. To finish the proof we need to show that this test works for $H_3:\theta\le\theta_1$ or $\theta\ge\theta_2$, which involves showing that $\mathbb{E}_\theta[\psi_0(T)]\le\alpha$ for all θ with $\theta\le\theta_1$ or $\theta\ge\theta_2$.

Now fix, without loss of generality, $\theta'<\theta_1$. Find another test ϕ^* for which $\mathbb{E}_{\theta_1}[\psi(T)]=\mathbb{E}_{\theta_2}[\psi(T)]=\alpha$ and $\mathbb{E}_{\theta'}[\psi(T)]$ is as small as possible. We will be able to show that $\psi^*=\psi_0$, which is prove in particular that $\mathbb{E}_{\theta'}[\psi_0(T)]\le\alpha$ (by comparing with the constant test α). Similar things can be done for $\theta'>\theta_2$.

To find ψ^*, instead consider $1-\psi^*$ and trying to maximize $\mathbb{E}_{\theta'}[\psi(T)]$ subject to $\mathbb{E}_{\theta_1}[\psi(T)] = \mathbb{E}_{\theta_2}[\psi(T)] = 1-\alpha$ As before, there are k_1 and k_2 such that the optimum is

$$1-\phi^*(t) = \begin{cases} 1 & f_3(t) > k_1 f_1 + k_2 f_2 \\ 0 & f_3(t) < k_1 f_1 + k_2 f_2. \end{cases}$$

The same analysis shows that ψ^* is 1 only if $a_1 + a_2 e^{b_2 t} > e^{b_1 t}$, after dividing by $c(\theta')e^{Q(\theta_1)t}$. Since Q is increasing, $b_1 > 0$ and $b_2 < 0$. We can show that $a_2 < 0$, so again the test looks like $\psi = \mathbf{1}_{[c_1,c_2]}$ for some c_1 and c_2 depending only on $(\theta_1, \theta_2, \alpha)$, and it must be the same as ψ_0.

For the third part... (omitted, I guess). □

6.3 UMPU tests

6.3.1 Definition. For testing $\theta \in \Theta_0$ vs. $\theta \in \Theta_1$, a test ϕ is an *unbiased test* at level α if $\mathbb{E}_\theta[\phi] \le \alpha$ for all $\theta \in \Omega_0$ and $\mathbb{E}_\theta[\phi] \ge \alpha$ for all $\theta \in \Theta_1$. A test is *UMPU* at level α if it is UMP over the class of unbiased tests at level α.

There is no UMP for the testing H_3 vs. K_3. Indeed, such a test would have to satisfy $\sup_{\theta_1 \le \theta \le \theta_2} \mathbb{E}_\theta[\phi(X)] \le \alpha$. Consider the problem of testing H_3 vs. $K : \theta = \theta'$ where $\theta' > \theta_2$ (this is easier than K_3). When the power function is strictly monotone, there is a MP test. But in this case the MP test cannot be unbiased for testing H_3 vs. K_3 (see homework, I guess?). If we restrict the class of tests to those have $\mathbb{E}_{\theta_1}[\phi(X)] = \alpha = \mathbb{E}_{\theta_1}[\phi(X)]$ then a UMP exists. Such tests will be unbiased, so the most powerful test over this class is the UMPU.

This paragraph is quite confusing. See the homework solutions for a slightly clearer explanation for why there is no UMP test for H_3.

6.3.2 Definition. Let ω be the common boundary of two sets Ω_H and Ω_K. We say that ϕ is *similar on the boundary*, or *SB*, if $\beta_\phi(\theta) = \mathbb{E}_\theta[\phi] = \alpha$ for all $\theta \in \omega$.

6.3.3 Lemma. *If $\beta_\phi(\theta)$ is continuous and ϕ is unbiased then ϕ is SB.*

PROOF: Intermediate value theorem. □

6.3.4 Lemma. *If the distribution p_θ is such that the power function of every test is continuous and ϕ_0 is UMP over the class of SB tests then ϕ_0 is UMPU (i.e. UMP over the class of unbiased tests).*

PROOF: If the power function is always continuous then the class of SB tests contains the class of unbiased tests. □

6.3.5 Theorem. *Consider a one parameter exponential family*

$$p_\theta(x) = c(\theta)e^{Q(\theta)T(x)}h(x)$$

with Q strictly increasing. For the problem of testing

$$H_3 : \theta_1 \le \theta \le \theta_2 \text{ vs. } K_3 : \theta < \theta_1 \text{ or } \theta > \theta_2$$

there is a unique level α UMPU test, namely

$$\phi = \begin{cases} 1 & T(x) < c_1 \text{ or } T(x) > c_2 \\ \gamma_i & T(x) = c_i, i = 1, 2 \\ 0 & c_1 < T(x) < c_2. \end{cases}$$

PROOF: The power functions are all continuous for this family. Consider the problem of maximizing $\mathbb{E}_{\theta'}[\psi(T)]$ over ψ unbiased, for some fixed $\theta' \in [\theta_1, \theta_2]$. This is equivalent to the minimization problem for $1 - \psi$, which brings us back to the setting of the proof of the last theorem. □

Last week we talked about the tests:

$$\begin{aligned} &H_1 : \theta \leq \theta_0 \text{ vs. } K_1 : \theta > \theta_0 \text{ (UMP)} \\ &H_2 : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2 \text{ vs. } K_2 : \theta_1 < \theta < \theta_2 \text{ (UMP)} \\ &H_3 : \theta_1 \leq \theta \leq \theta_2 \text{ vs. } K_3 : \theta < \theta_1 \text{ or } \theta > \theta_2 \text{ (UMPU)} \end{aligned}$$

We expect the test for H_4 to be a special case of the test for H_3, and so we expect there to be a UMPU test. The problem is to determine c_1, c_2, γ_1, and γ_2 with some constraints. In the third case, the constraints $\mathbb{E}_{\theta_1}[\phi(X)] = \alpha$ and $\mathbb{E}_{\theta_2}[\phi(X)] = \alpha$ determine the four parameters. In case four we only have the one constraint $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$, which does not seem like it will be enough. Recall that ϕ is unbiased for this case if $\mathbb{E}_\theta[\phi(X)] \leq \alpha$ when $\theta = \theta_0$ and $\mathbb{E}_\theta[\phi(X)] \geq \alpha$ when $\theta \neq \theta_0$. Therefore β_ϕ attains its minimum at θ_0. If it is differentiable then $\beta'_\phi(\theta_0) = 0$. Furthermore, if $P_\theta(dt) = c(\theta)e^{\theta t}\nu(dt)$ is the distribution of $T(X)$ and we can write $\phi(x) = \psi(T(x))$ then we can use Stein's lemma to compute

$$\frac{d}{dt}\mathbb{E}_\theta[\psi(T(X))] = \mathbb{E}_\theta[T\psi(T)] + \frac{c'(\theta)}{c(\theta)}\mathbb{E}_\theta[\psi(T)].$$

This holds for all unbiased tests, so in particular it holds for $\phi(x) \equiv \alpha$. Then

$$0 = \alpha\,\mathbb{E}_\theta[T(X)] + \frac{c'(\theta)}{c(\theta)}\alpha,$$

giving $\mathbb{E}_\theta[T] = -c'(\theta)/c(\theta)$. Plugging this into the original expression yields

$$\frac{d}{dt}\mathbb{E}_\theta[\psi(T(X))] = \mathbb{E}_\theta[T\psi(T)] - \mathbb{E}_\theta[T]\,\mathbb{E}_\theta[\psi(T)] = \mathrm{Cov}(T, \psi(T)).$$

Finally, if ψ is unbiased then $\mathbb{E}_{\theta_0}[\psi(T)] = \alpha$ and the derivative is zero at $\theta = \theta_0$, so

$$\mathbb{E}_{\theta_0}[T\psi(T)] = \alpha\,\mathbb{E}_{\theta_0}[T] = -\alpha\frac{c'(\theta_0)}{c(\theta_0)}.$$

This is the second constraint that can be used to find c_1, c_2, γ_1, and γ_2.

6.4 Nuisance parameters

Difference, in addition, we have a nuisance parameter ϑ, e.g. in $X_i \sim N(\theta, \sigma^2)$, testing $\theta = \theta_0$ vs. $\theta \neq \theta_0$ has the nuisance parameter σ^2.

Review: unbiasedness of test + continuity of power function = SB property

6.4.1 Definition. Let T be sufficient for $P = \{p_\theta(x), \theta \in \omega = \partial\Theta\}$ and let $P^T = \{p_\theta^T = p_\theta(T(x)) : \theta \in \omega\}$. A test function is said to have *Neyman structure* with respect to T if $\mathbb{E}_\theta[\psi(X)|T] = \alpha$ a.e.-P_θ^T, for $\theta \in \omega$.

6.4.2 Lemma. *If ϕ has Neyman structure with respect to T then ϕ is SB.*

Proof: For all $\theta \in \omega$,

$$\mathbb{E}_\theta[\phi(X)] = \mathbb{E}_\theta[\mathbb{E}[\phi(X)|T]] = \mathbb{E}_\theta[\alpha] = \alpha. \qquad \square$$

Neyman structure allows us to use conditioning (on complete sufficient statistics) to reduce a high dimensional problem to fewer dimensions. The question that remains is, when does SB imply Neyman structure? Note that if all unbiased tests are SB and all SB tests have Neyman structure, then to find the UMPU test it suffices to find the UMP test over the class of tests of Neyman structure. Why does this make the problem any easier?

$$\mathbb{E}_\theta\,\phi_0 = \alpha \qquad\qquad \alpha \in \omega$$

$$\mathbb{E}_{\theta,\vartheta}\,\phi_0 = \iint \phi(u,t) P_\theta^{U|t}(du) P_{\theta,\vartheta}^T(dt)$$

For every fixed t, if $\phi(u,t)$ is level α UMP, then the power is greater than or equal to the power of any other level α Neyman structured test.

6.4.3 Theorem. *Let X be a random variable with distribution $p_\theta \in P = \{p_\theta : \theta \in \Theta\}$ and let T be sufficient for $P_B = \{p_\theta : \theta \in \omega = \partial\Theta\}$. Then all SB tests have Neyman structure if and only if the family of distributions $P^T = \{p_\theta^T : \theta \in \omega\}$ is bounded complete, i.e. "if ψ is bounded and $\mathbb{E}_\theta[\psi(T)] = 0$ for all $\theta \in \omega$ then $\psi = 0$ a.e."*

Proof: If it is bounded complete then if ϕ is a level α SB test then let $\psi = \mathbb{E}[\phi|T]$. $\mathbb{E}_\theta\,\phi = \alpha$ for all $\theta \in \omega$ so $\mathbb{E}[\psi(T) - \alpha] = 0$ for all $\theta \in \omega$, so $\mathbb{E}[\phi|T] = \alpha$ a.e.-P^T.

Suppose that all SB tests have Neyman structure, but there is a statistic such that P^T is not bounded complete. Then there is $h \neq 0$ bounded such that $\mathbb{E}_\theta\,h(T) = 0$ for all $\theta \in \omega$. Define $\psi(T) = ch(T) + \alpha$ where c is chosen sufficiently small that ψ is a test. Then $\mathbb{E}_\theta[\psi] = \alpha$ and $\mathbb{E}[\psi(T)|T] = \psi(T) = \alpha + ch(T)$, which is not α with positive p_{θ_0}-probability for some $\theta_0 \in \omega$. $\square$

The setting is now as follows. We have a $(k+1)$-parameter exponential family with density functions

$$p_{\theta,\vartheta} = c(\theta,\vartheta)\exp\left(\theta U(x) + \sum_{i=1}^k \xi_i T_i(x)\right)$$

with respect to a σ-finite dominating measure μ. Suppose that $\Theta \subseteq \mathbb{R}^{k+1}$ is convex with non-empty interior. Then $T := (T_1, \dots, T_k)$ is complete and sufficient for the nuisance parameters ϑ when θ is fixed. In this setting the hypotheses of the above theorem hold, so we need merely to find the UMP test over the class of tests with Neyman structure.

1. $H : \theta \le \theta_0$ vs. $K : \theta > \theta_0$
2. $H : \theta \le \theta_1$ or $\theta \ge \theta_2$ vs. $K : \theta_1 < \theta < \theta_2$
3. $H : \theta_1 \le \theta \le \theta_2$ vs. $K : \theta < \theta_1$ or $\theta > \theta_2$
4. $H : \theta = \theta_0$ vs. $K : \theta \ne \theta_0$

The following are UMPU tests.

$$\phi_1(u,t) = \begin{cases} 1 & u > c(t) \\ \gamma(t) & u = c(t) \\ 0 & u < c(t) \end{cases}$$

$$\phi_2(u,t) = \begin{cases} 1 & c_1(t) < u < c_2(t) \\ \gamma_i(t) & u = c_i(t) \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_3(u,t) = \begin{cases} 1 & u > c_1(t) \text{ or } u < c_2(t) \\ \gamma_i(t) & u = c_i(t) \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_4(u,t) = \begin{cases} 1 & u > c_1(t) \text{ or } u < c_2(t) \\ \gamma_i(t) & u = c_i(t) \\ 0 & \text{otherwise} \end{cases}$$

where $t = T(x)$. For ϕ_1 the conditions that $\mathbb{E}_{\theta_0}[\phi_1(U,T)|T = t] = \alpha$ for all t determine the coefficients. For ϕ_2 and ϕ_3 the conditions are $\mathbb{E}_{\theta_i}[\phi(U,T)|T = t] = \alpha$ for $i = 1,2$ and all t, and for ϕ_4 the conditions are $\mathbb{E}_{\theta_0}[\phi_4(U,T)|T = t] = \alpha$ and $\mathbb{E}_{\theta_0}[U\phi_4(U,T)|T = t] = \alpha\,\mathbb{E}_{\theta_0}[U|T = t]$ for all t.

Proof: T is sufficient for ϑ if θ if fixed. The associated family of distributions of T is

$$dP^T_{\theta_j,\theta}(t) = c(\theta_j,\theta)\,\mathbb{E}\left(\sum_{i=1}^{k} \vartheta_i T_i(x)\right) d\nu_{\theta_j}(t)$$

where $\theta = \theta_0$, θ_1, or θ_2 and $j = 0$, 1, or 2. (ν is like a push-forward or pull-back of μ wrt $t = T(x)$.) By the assumptions on the domain Θ, $\omega_j := \{\theta = \theta_j : (\theta, \vartheta) \in \Theta\}$ contains a k-dimensional rectangle. Therefore T is sufficient and complete. If ϕ is SB then

$$\mathbb{E}_{\theta_j}[\phi(U,T)|T = t] = \alpha$$

for all t. We now show that ϕ_1 is UMP over all SB tests (in order to conclude that it is UMPU).

Consider case 1 for now. For any point (θ,ϑ) in the alternative K_1,

$$\begin{aligned}\mathbb{E}_{\theta,\vartheta}[\phi(U,T)] &= \iint \phi(u,t)P_\theta^{U|t}(du)P_{\theta,\vartheta}^T(dt)\\ &= \mathbb{E}_{\theta,\vartheta}[\mathbb{E}_{\theta,t}[\phi(U,T)|T=t]]\end{aligned}$$

(The distribution $(U|T=t)$ does not depend on ϑ because T is sufficient.) For all SB tests, for all t,

$$\mathbb{E}_{\theta_0}[\phi(U,T)|T=t]=\alpha.$$

We claim that ϕ_1 maximizes the inner integration over all t subject to this constraint. The rest of the proof is an exercise? □

6.4.4 Example. Suppose that $X\sim$ Poisson(λ) and $Y\sim$ Poisson(μ). We are interested in comparing λ and μ (say, testing $\lambda\le\mu$ or $\lambda=\mu$).

$$\mathbb{P}[X=x,Y=y]=\frac{e^{-(\lambda+\mu)}}{x!y!}\exp\big(y\log(\mu/\lambda)+(x+y)\log\lambda\big).$$

Let $\theta=\log(\mu/\lambda)$, $\vartheta=\log\lambda$, $U=Y$ and $T=X+Y$. For the test $H:\mu\le a\lambda$ (where a is fixed) $H:\log(\mu/\lambda)\le\log a$, so $H:\theta\le\log(a)$ vs. $K:\theta>\log a$ with nuisance parameter ϑ. There is a UMPU test with form

$$\phi(u,t)=\begin{cases}1 & u>c(t)\\ \gamma(t) & u=c(t)\\ 0 & u<c(t)\end{cases}$$

where c and γ are determined by $\mathbb{E}_{\theta_0}[\phi(U,T)|T=t]=\alpha$ for all t. Here $\theta_0=\log a$ so given t,

$$\mathbb{P}_{\theta_0}[U>c(T)|T=t]+\gamma(t)\mathbb{P}_{\theta_0}[U=c(T)|T=t]=\alpha$$

What is $(U|T=t)$? It is Binomial($t,\mu/(\lambda+\mu)$). Under $\mathbb{P}_{\theta_0}$, $\mu/(\lambda+\mu)=a/(a+1)$. This (in principle) determines c and γ as functions of t.

$$\sum_{k=c(t)+1}^{t}\binom{t}{k}\left(\frac{a}{a+1}\right)^k\left(\frac{1}{a+1}\right)^{t-k}+\gamma(t)\binom{t}{c(t)}\left(\frac{a}{a+1}\right)^{c(t)}\left(\frac{1}{a+1}\right)^{t-c(t)}=\alpha.$$

6.4.5 Example. Let $X_i\sim N(\xi,\sigma^2)$, $1\le i\le m$. We could test $H_1:\sigma\le\sigma_0$ (UMP exists), or $H_2:\sigma\ge\sigma_0$ (UMP does not, but UMPU does), or $H_3:\xi\le\xi_0$ (UMP does not, but UMPU does), or $H_4:\xi\ge\xi_0$ (UMP does not, but UMPU does).

Suppose $Y_j\sim N(\eta,\tau^2)$, $1\le j\le n$. We could test $H:\xi\le\eta$ or $H':\sigma^2\le a\tau^2$.

For H_1, the joint density is

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left(-\frac{n\xi^2}{2\sigma^2}\right)\exp\left(-\frac{1}{2\sigma^2}\sum x_i^2+\frac{n\xi}{\sigma^2}\overline{X}\right)$$

Take $\theta = -1/(2\sigma^2)$, $\vartheta = n\xi/\sigma^2$, $U = \sum X_i^2$ and $T = \overline{X}$. Then H_1 is equivalent to testing $\theta \geq \theta_0$. There is a UMPU test of the form

$$\phi(u,t) = \begin{cases} 1 & u < c(t) \\ 0 & u > c(t) \end{cases}$$

and we have the constraint

$$\begin{aligned} \alpha &= \mathbb{P}_{\theta_0}[U < c(T) | T = t] \\ &= \mathbb{P}_{\theta_0}[U - nT^2 < c_0(T) | T = t] \\ &= \mathbb{P}_{\theta_0}[U - nt^2 < c_0(t)] \end{aligned}$$

(Note that $U - nT^2 = S^2$ is independent of $\overline{X} = T$.) It follows that $c_0(t) = c_0$ is a constant independent of t. Now $U - nT^2$ is $\sigma_0^2 \chi^2_{n-1}$ under $\mathbb{P}_{\theta_0}$, so we obtain the usual χ^2-test. Taking c_0/σ_0^2 to be the α-percentile of χ^2_{n-1} determines the test. We reject when $\sum (X_i - \overline{X})^2 \leq c_0$.

Suppose that $V = h(U,T)$ is monotonically increasing in u for any fixed t. Then

$$\phi(u,t) = \begin{cases} 1 & u < c(t) \\ 0 & u > c(t) \end{cases} = \begin{cases} 1 & v < c_0(t) \\ 0 & v > c_0(t) \end{cases} =: \phi(v)$$

are the same test. If V is independent of T when $\theta = \theta_0$ then $\phi(v)$ is UMPU for H_1. If V is independent of T when $\theta = \theta_1$ and $\theta = \theta_2$ then we get corresponding statements for H_2 and H_3. If V is independent of T for $\theta = \theta_0$ and $h(u,t) = a(t)u + b(t)$ then $\phi_4(v)$ is UMPU for H_4. (See the theorem on the bottom of page 151.)

6.4.6 Example (Two sample normal with equal variances).
Let $X_1, \dots, X_m \sim N(\xi, \sigma^2)$ and $Y_1, \dots, Y_n \sim N(\eta, \sigma^2)$. Consider testing the hypotheses $H : \eta = \xi$ or $H' : \eta \leq \xi$. The joint density of the data is

$$c(\xi, \eta, \sigma^2) \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^m X_i^2 + \sum_{j=1}^n Y_j^2 \right) + \frac{m\xi}{\sigma^2}\overline{X} + \frac{n\eta}{\sigma^2}\overline{Y} \right),$$

so the natural parameters are $-\frac{1}{2\sigma^2}$, $\frac{m\xi}{\sigma^2}$, and $\frac{n\eta}{\sigma^2}$. Rearrange the parameters to get something useful.

$$\frac{m\xi}{\sigma^2}\overline{X} + \frac{n\eta}{\sigma^2}\overline{Y} = \frac{\eta - \xi}{(\frac{1}{m} + \frac{1}{n})\sigma^2}(\overline{Y} - \overline{X}) + \frac{m\xi + n\eta}{(m+n)\sigma^2}(m\overline{X} + n\overline{Y})$$

so we may write the density as

$$c \exp\left(\frac{\eta - \xi}{(\frac{1}{m} + \frac{1}{n})\sigma^2}(\overline{Y} - \overline{X}) + \frac{m\xi + n\eta}{(m+n)\sigma^2}(m\overline{X} + n\overline{Y}) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^m X_i^2 + \sum_{j=1}^n Y_j^2 \right) \right)$$

Now write $\theta = \frac{\eta-\xi}{(\frac{1}{m}+\frac{1}{n})\sigma^2}$, with nuisance parameters $\vartheta_1 = \frac{m\xi+n\eta}{(m+n)\sigma^2}$ and $\vartheta_2 = \frac{1}{2\sigma^2}$. We look for $V = h(U,T)$ such that h is increasing in u and $h(U,T)$ is independent of T when $\theta = 0$. Note that when $\theta = 0$ then $\overline{Y} - \overline{X} \sim N(0, (\frac{1}{m} + \frac{1}{m})\sigma^2)$, so with this in mind take

$$h(U,T) = \frac{U}{\sqrt{T_2 - \frac{1}{m+n}T_1^2 - \frac{mn}{m+n}U}}$$

and note that

$$V' := \frac{U/\sqrt{\frac{1}{m}+\frac{1}{n}}}{\sqrt{(T_2 - \frac{1}{m+n}T_1^2 - \frac{mn}{m+n}U)/(m+n-2)}} \sim t_{m+n-2}.$$

Therefore we recover the classical t-test.
If the variances are not the same then there is no good answer for the UMPU.

6.4.7 Example (Two sample normal with equal means). From the exam.

Review for second midterm: How to find a Bayes/minimax rule? To find a Bayes rule calculate the posterior and then find the posterior mean (for squared error loss). For minimax, if you can find $\delta(X)$ such that it is Bayes and it has constant risk then it is minimax (and the prior is least favourable). Less rigidly, if δ has constant risk γ and $\gamma = \lim_{n\to\infty} \gamma_{\Lambda_n}$ then δ is minimax (think improper priors, normal mean with variance known). A third approach is to find a minimax estimator for a restricted family and then show that the risk over the smaller set is the same as the risk over the larger set (think normal mean with variance unknown by bounded).

How to find/determine admissible estimators? Recall that $a\overline{X} + b$ is inadmissible when $a < 0$, $a > 1$, or $a = 1$ and $b \neq 0$, otherwise it is admissible. When the dimension is three or greater, $\overline{X}$ is inadmissible (Stein).

Testing: Neyman-Pearson lemma for the simple test setting shows MP test is of the form

$$\phi(x) = \begin{cases} 1 & f_K/f_H > k \\ \gamma & f_K/f_H = k \\ 0 & f_K/f_H < k \end{cases}$$

and the proof is not very complicated (read it!). For complex tests, we defined a monotone likelihood ratio family to be one for which $f_{\theta'}/f_\theta$ is monotonically increasing in x (or $T(x)$) when $\theta' > \theta$. E.g. one parameter exponential family $c(\theta)e^{Q(\theta)U(x)}h(x)$ with increasing Q, location family for some distributions. There is a UMP for $\theta \leq \theta_0$, found by showing the MP for $\theta = \theta_0$ vs. $\theta = \theta_1$ does not depend on θ_1 for all $\theta_1 > \theta_0$, and then that the power is increasing in θ for $\theta < \theta_0$. See Karlin-Rubin theorem. For the one parameter exponential family in particular, we considered four testing problems. There are UMP tests for the first two but not the last two, and if we restrict to the family of unbiased tests then there are UMPU tests. For the multi-parameter exponential family we know the UMPU tests. See theorem on page 151.

7 Maximum Likelihood Estimators

This section follows TPE §1.8.

7.0.8 Theorem (Delta method). *Let g be a function that is continuously differentiable at θ_0, with $g'(\theta_0) \neq 0$. If $\sqrt{n}(T_n - \theta_0) \xrightarrow{(d)} N(0, \tau^2)$ then*

$$\sqrt{n}(g(T_n) - g(\theta_0)) \xrightarrow{(d)} N(0, \tau^2(g'(\theta_0))^2).$$

Proof: Note that $T_n - \theta_0 \xrightarrow{\text{(d)}} 0$, so $T_n - \theta_0 \xrightarrow{\text{(p)}} 0$. By continuity, $g'(\xi_n) \xrightarrow{\text{(p)}} g'(\theta_0)$ whenever $\xi_n \xrightarrow{\text{(p)}} \theta_0$. By Taylor's theorem, for some ξ_n between T_n and θ_0,

$$\begin{aligned}\sqrt{n}(g(T_n) - g(\theta_0)) &= g'(\xi_n)\sqrt{n}(T_n - \theta_0) \\ &\xrightarrow{\text{(d)}} g'(\theta_0)N(0, \tau^2) \overset{\text{(d)}}{=} N(0, \tau^2(g'(\theta_0))^2).\end{aligned}$$ □

7.0.9 Assumptions. *Let $\{p_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ be a family of densities with respect to a σ-finite measure μ on $\mathbb{R}$ and satisfying the following assumptions.*

A0: (Identifiable) $\theta \neq \theta'$ implies $p_\theta \neq p_{\theta'}$.

A1: (Common support) $A = \{x : p_\theta > 0\}$ does not depend on θ.

A2: $X_1, \dots, X_n \sim P_{\theta_0} =: P_0$.

A3: (Smooth around θ_0) Θ contains a neighbourhood Θ_0 of θ_0 such that, for all $\theta \in \Theta_0$,

- *a) For μ-a.e. x, $\ell(\theta|x) = \log p_\theta(x)$ is twice continuously differentiable.*
- *b) For μ-a.e. x, $\dddot{\ell}_{jkl}(\theta|x)$ exists and satisfies $|\dddot{\ell}_{jkl}(\theta|x)| \leq M_{jkl}(x)$ for all $1 \leq j, k, l \leq d$, where M is family of constants independent of θ.*

A4: (Regular at θ_0)

- *a) $\mathbb{E}_0[\dot{\ell}(\theta_0|X)] = 0$.*
- *b) $\mathbb{E}_0[|\dot{\ell}(\theta_0|X)|^2] < \infty$.*
- *c) $I(\theta_0)_{ij} := -\mathbb{E}_0[\ddot{\ell}_{ij}(\theta_0|X)]$ is a positive definite matrix.*

We fix the following notation. $L_n(\theta) := \prod_{i=1}^n p_\theta(x_i)$ is the *likelihood function*, and $\ell_n(\theta) := \log L_n(\theta)$ is the *log-likelihood function*. For $B \subseteq \Theta$, $L_n(B) := \sup_{\theta \in B} L_n(\theta)$ and $\ell_n(B) := \sup_{\theta \in B} \ell_n(\theta)$.

7.0.10 Theorem. *Assume that A0–A4 are satisfied.*

(i) (Consistency of MLE) With probability one, there is a sequence $\tilde{\theta}_n$ of solutions to the sequence of equations $\dot{\ell}_n(\theta|x_1, \dots, x_n) = 0$ such that $\tilde{\theta}_n \xrightarrow{(p)} \theta_0$.

(ii) (Asymptotic efficiency of MLE) $\tilde{\theta}_n$ is "asymptotically linear" with respect to $\ell_n(\theta_0|x_1, \dots, x_n)$, i.e.

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = I^{-1}(\theta_0)Z_n + o_P(1)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}(\theta_0|x_i) + o_P(1)$$

$$\xrightarrow{(d)} I^{-1}(\theta_0)Z \sim N(0, I^{-1}(\theta_0))$$

where $Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}(\theta_0|x_i)$ *and* $\tilde{\ell}(\theta_0|x) := I^{-1}(\theta_0)\dot{\ell}(\theta_0|x)$

(iii) (Asymptotic efficiency of LRT) $2\log\tilde{\lambda}_n \xrightarrow{(d)} Z^T I^{-1}(\theta_0) Z \sim \chi^2_d$ *where*

$$\tilde{\lambda}_n := \frac{L_n(\tilde{\theta}_n|x_1,\dots,x_n)}{L_n(\theta_0|x_1,\dots,x_n)}$$

Remark. It suffices to remember that "identifiability + common support + smooth around θ_0 + regular at θ_0 = MLE is consistent and efficient and $\log(LRT) \to \frac{1}{2}\chi^2_d$." Note that the *general likelihood ratio test statistic*

$$\hat{\lambda}_n := \frac{\max_{\theta\in\Theta} L_n(\theta|x_1,\dots,x_n)}{L_n(\theta_0|x_1,\dots,x_n)} = \frac{L_n(\hat{\theta}_n^{MLE}|x_1,\dots,x_n)}{L_n(\theta_0|x_1,\dots,x_n)}$$

is slightly different from the *likelihood ratio test statistic* $\tilde{\lambda}_n$.

PROOF: Fix $a > 0$ small and let $Q_a := \{\theta \in \Theta : |\theta - \theta_0| = a\}$. We will show that

$$P_0[\ell_n(\theta) < \ell_n(\theta_0) \text{ for all } \theta \in Q_a] \to 1 \text{ as } n \to \infty.$$

It will follow from this that, for n large enough, $\dot{\ell}_n(\theta) = 0$ will have one or more solutions inside Q_a with high probability. By Taylor's theorem, for $\theta \in Q_a$,

$$\begin{aligned}\frac{1}{n}(\ell_n(\theta) - \ell_n(\theta_0)) &= (\theta-\theta_0)^T \frac{1}{n}\dot{\ell}_n(\theta_0) \\ &\quad - \frac{1}{2}(\theta-\theta_0)^T \left(-\frac{1}{n}\ddot{\ell}_n(\theta_0)\right)(\theta-\theta_0) + O(|\theta-\theta_0|^3),\end{aligned}$$

where we have used the regularity and the boundedness of the third derivative. By the SLLN,

$$\frac{1}{n}\dot{\ell}_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}(\theta_0|x_i) \to \mathbb{E}_0[\dot{\ell}(\theta_0|X)] = 0$$

and

$$-\frac{1}{n}\ddot{\ell}_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} -\ddot{\ell}(\theta_0|x_i) \to \mathbb{E}_0[-\ddot{\ell}(\theta_0|X)] = I(\theta_0).$$

Therefore, since $|\theta - \theta_0| = a$,

$$\begin{aligned}\frac{1}{n}(\ell(\theta) - \ell(\theta_0)) &\le a o_P(1) - \frac{1}{2}(\theta-\theta_0)^T I(\theta_0)(\theta-\theta_0) + Ba^3 \\ &\le a o_P(1) - \frac{1}{2}\lambda_d a^2 + Ba^3\end{aligned}$$

where B is some constant and $\lambda_d > 0$ is the smallest eigenvalue of the information matrix. In particular, for small enough a and big enough n, $\ell(\theta) - \ell(\theta_0) < 0$.

For the next part,

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}}\dot{\ell}_n(\tilde{\theta}_n) \\
&= \frac{1}{\sqrt{n}}\dot{\ell}_n(\tilde{\theta}_n) - \left(-\frac{1}{n}\ddot{\ell}_n(\theta_n^*)\right)\sqrt{n}(\tilde{\theta}_n - \theta_0) \\
\sqrt{n}(\tilde{\theta}_n - \theta_0) &= -\left(\frac{1}{n}\ddot{\ell}_n(\theta_n^*)\right)^{-1}\frac{1}{\sqrt{n}}\dot{\ell}_n(\tilde{\theta}_n)
\end{aligned}$$

By the CLT, the "numerator" converges in distribution to a normal random variable with mean zero and covariance matrix $\mathbb{E}[\dot{\ell}_n(\theta_0)\dot{\ell}_n(\theta_0)^T] = I(\theta_0)$. For the "denominator," recalling that θ_n^* is a number between $\tilde{\theta}_n$ and θ_0,

$$\frac{1}{n}\ddot{\ell}_n(\theta_n^*) = \frac{1}{n}\ddot{\ell}_n(\theta_0) + \frac{1}{n}\dddot{\ell}_n(\theta_n^*)(\theta_n^* - \theta_0) \xrightarrow{(p)} -I(\theta_0),$$

by the SLLN and because the third derivative is bounded. Therefore

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{(d)} (I(\theta_0))^{-1}N(0, I(\theta_0)) \overset{(d)}{=} N(0, I(\theta_0)^{-1}).$$

The final part is more of the same. □

7.0.11 Definition. If $\delta(X)$ is an estimator of θ and $\sqrt{n}(\delta(X) - \theta) \xrightarrow{(d)} N(0, \tau^2)$ then we say that δ is an *asymptotically unbiased estimator* with *asymptotic variance* τ^2. It can be shown that the asymptotic variance of an asymptotically unbiased estimator is bounded below by $I(\theta)^{-1}$, and that the MLE (under certain conditions) attains this lower bound. A sequence of estimators that attains the lower bound is said to be *asymptotically efficient*.

7.0.12 Example (Newton-Ralphson). If $\sqrt{n}(\tilde{\theta}_n - \theta) = O_P(1)$ then $\tilde{\theta}_n$ is said to be a $\sqrt{n}$*-consistent* sequence of estimators for θ. For the MLE, $\hat{\theta}_n$,

$$\begin{aligned}
0 &= \dot{\ell}_n(\hat{\theta}_n) \\
&\approx \dot{\ell}_n(\tilde{\theta}_n) + \ddot{\ell}_n(\tilde{\theta}_n)(\hat{\theta}_n - \tilde{\theta}_n) \\
\hat{\theta}_n &\approx \tilde{\theta}_n - \frac{\dot{\ell}_n(\tilde{\theta}_n)}{\ddot{\ell}_n(\tilde{\theta}_n)}
\end{aligned}$$

This gives a method of obtaining a sequence of asymptotically efficient estimators given a $\sqrt{n}$-consistent sequence.

7.1 Kullback-Leibler information

7.1.1 Definition. Let P and Q be two probability measures with densities p and q with respect to a σ-finite measure μ. The *Kullback-Leibler information* is

$$K(P, Q) := d_{KL}(P, Q) := \mathbb{E}_P\left[\log\left(\frac{p(X)}{q(X)}\right)\right] = -\mathbb{E}_P\left[\log\left(\frac{q(X)}{p(X)}\right)\right]$$

7.1.2 Lemma. *$K(P,Q) \geq 0$ and is zero if and only if $P = Q$.*

Proof: Since $x \mapsto -\log x$ is a convex function, by Jensen's inequality,

$$K(P,Q) = -\mathbb{E}_P[\log(q/p)] \geq -\log \mathbb{E}_P[q/p] = -\log 1 = 0. \qquad \square$$

Note that, by the SLLN,

$$\frac{1}{n}\log\left(\frac{L_n(\theta_0)}{L_n(\theta)}\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}\right) \to \mathbb{E}_{\theta_0}\left[\log\left(\frac{p_{\theta_0}(X)}{p_\theta(X)}\right)\right] = K(P_{\theta_0}, P_\theta)$$

7.2 Examples

Note that the condition of common support excludes the uniform family $U(0,\theta)$ and the one-sided double parameter exponential family, and the smoothness requirement excludes the double exponential family $\frac{1}{2}\exp(-|x-\theta|)$.

7.2.1 Exercise. Show that if $X_1, \dots, X_n \sim U(0,\theta)$ then

$$\sqrt{n}(X_{(n)} - \theta) \xrightarrow{\text{(d)}} \text{Exponential}(1).$$

Also show that if $X_1, \dots, X_n \sim N(0,1)$ then

$$\frac{X_{(n)}}{\sqrt{2\log n}} \xrightarrow{\text{(p)}} 1.$$

It can be show that there are $b_n \sim \sqrt{2\log\log n}$ and $a_n \sim \sqrt{2\log n}$ such that

$$b_n(X_{(n)} - a_n) \xrightarrow{\text{(d)}} \log(\text{Exponential}(1)) =: \text{Gumbel}(1).$$

These limits come up in the theory of empirical processes and extreme value theory.

Clearly the equations $\dot{\ell}_n(\theta) = 0$ each have only one root then the sequence of roots converges to the true parameter. If those equations all have multiple roots then it is not necessarily the case that every sequence of roots converges to the true parameter.

7.2.2 Example (One parameter exponential family). In this case the conditions always all hold.

$$\begin{aligned}
p_\theta(x) &= \exp(\theta T(x) - A(\theta))h(x) \\
\ell(\theta) &= \theta T(x) - A(\theta) + \log(h(x)) \\
\dot{\ell}(\theta) &= T(x) - A'(\theta)
\end{aligned}$$

Then $\dot{\ell}(\theta) = 0$ if and only if $T(x) = A'(\theta)$. Note that $A''(\theta) = \text{Var}_\theta(T(X)) > 0$, so A' is an increasing function. Hence there is at most one root, and the MLE is

$\tilde{\theta} = (A')^{-1}(T(x))$. Here $I(\theta) = \mathrm{Var}_\theta(T(X))$ and $\tilde{\theta}$ satisfies $\frac{1}{n}\sum_{i=1}^n T(x_i) = A'(\theta)$. The binomial$(n, \theta)$, Poisson$(\theta)$, and normal (if one parameter is known) fall under this example.

7.2.3 Example (Multi-parameter exponential family). In this case the conditions always all hold.

$$p_\eta(x) = \exp(\sum_{i=1}^d \eta_i T_i(x) - A(\eta))h(x)$$
$$\dot{\ell}_{\eta_j} = T_j(x) - \frac{\partial}{\partial \eta_j} A(\eta)$$

The second derivative matrix is $\mathrm{Cov}(T, T)$, which is positive definite.

7.2.4 Example (Double exponential). Also referred to as the *Laplace distribution*. The joint density is $f(x_1, \dots, x_n; \theta) = \frac{1}{2^n}\exp(-\sum_{i=1}^n |x_i - \theta|)$. The MLE is found by making $\sum_{i=1}^n |x_i - \theta|$ as small as possible. It turns out that $\hat{\theta}$ = sample median. Indeed, if the sample median is t, then

$$\sum_{i=1}^n |x_i - t| = \sum_{x_i \ge t}^n (x_i - t) + \sum_{x_i < t}^n (t - x_i)$$

(Fill this in.) We cannot use the theorem to conclude that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\text{(d)}} N(0, I(\theta)^{-1}),$$

even though it is true. Let $U_1, \dots, U_n \sim U(0, 1)$. Following pages 89-93 of Ferguson's book, let $0 < q < 1$ and $k := \lfloor nq \rfloor$. The q-quantile is $U_{(k)}$. Then

$$\sqrt{n}(U_{(k)} - q) \xrightarrow{\text{(d)}} N(0, q(1 - q)).$$

Now if $Y_1, \dots, Y_n$ are distributed as some distribution function F, then $F(Y_i) \sim U(0, 1)$, and the order statistics are preserved. It follows that

$$\sqrt{n}(Y_{([nq])} - y_q) \xrightarrow{\text{(d)}} N\left(0, \frac{q(1-q)}{f(y_q)^2}\right)$$

where y_q is the q-quantile of F and f is the density. In the case of the double exponential, the theoretical median is θ and the value of the density at the median is $1/2$, so

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\text{(d)}} N(0, 1).$$

7.2.5 Example (Cauchy location family). Here $f(x) = 1/(\pi(1+x^2))$. The derivative of the log-likelihood function is hence

$$\dot{\ell}(\theta|x) = \frac{2(x - \theta)}{1 + (x - \theta)^2}.$$

Setting $\dot{\ell}(\theta|x_1,\dots,x_n) = 0$ and rearranging,

$$0 = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}$$
$$0 = \sum_{i=1}^{n} \left((x_i - \theta) \prod_{i \neq j} (1 + (x_j - \theta)^2) \right)$$

This is a polynomial equation for θ, and it may have as many as $2n - 1$ different roots. The method we use is as follows. Find a $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$ and then iterate one step to get

$$\hat{\theta}_n = \tilde{\theta}_n - \frac{\dot{\ell}_n(\tilde{\theta}_n)}{\ddot{\ell}(\tilde{\theta}_n)}.$$

The initial estimator we choose is $\tilde{\theta}_n$ = sample median (because the theoretical mean is infinite). Via the same analysis as the last example, since the theoretical median is θ,

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{\text{(d)}} N\left(0, \frac{\pi^2}{4}\right).$$

7.3 Method of Moments

The *method of moments* is very useful for constructing $\sqrt{n}$-estimators. Assume that f is a density function with sufficiently thin tails that a few moments exists. Let $X_1, \dots, X_n \sim f(x - \theta)$. Set

$$\mathbb{E}[\overline{X}] = \int_{\mathbb{R}} x f(x - \theta) dx = \int_{\mathbb{R}} (\theta + x) f(x) dx = \theta + m$$

where $m = \int_{\mathbb{R}} x f(x) dx$ is the theoretical mean. Then

$$\sqrt{n}((\overline{X}_n - m) - \theta) \to N(0, \sigma^2),$$

where σ^2 is the theoretical variance. Applying Newton-Ralphson, a good estimator is hence

$$\hat{\theta} = \overline{X}_n - a - \frac{\dot{\ell}(\overline{X}_n - a)}{\ddot{\ell}(\overline{X}_n - a)}.$$

7.3.1 Example (Mixture distributions). Assume $X_1, \dots, X_n \sim \theta F + (1 - \theta) G$. This can come up in Bayesian situations, e.g. X_i is class 1 with probability θ and class 2 with probability $1 - \theta$, and we wish to estimate θ, the proportion of the population in class 1.

$$\mathbb{E}[\overline{X}_n] = \theta m_F + (1 - \theta) m_G = \theta(m_F - m_G) + m_G$$

Assuming $m_F \neq m_G$ but both are known, then $(\overline{X}_n - m_G)/(m_F - m_G)$ is a $\sqrt{n}$-consistent estimator. In practical applications it is often the case that m_G is not known, but similar estimators can be found.

7.3.2 Example (Joe Hodge's super-efficiency). Suppose that $\sqrt{n}(\delta_n - \theta) \xrightarrow{(d)} N(0, v(\theta))$. One would expect that $v(\theta) \geq I(\theta)^{-1}$, but this is not always the case. Suppose $X_1, \dots, X_n \sim N(\theta, 1)$ and let

$$\delta_n = \begin{cases} \overline{X}_n & |\overline{X}_n| > n^{-1/4} \\ a\overline{X}_n & |\overline{X}_n| \leq n^{-1/4} \end{cases}$$

when $\theta \neq 0$. Then since $\sqrt{n}(\overline{X}_n - \theta) \overset{(d)}{=} N(0,1)$, so $\overline{X}_n = \theta + \frac{1}{\sqrt{n}} N(0,1)$, it follows that $\mathbb{P}[\overline{X}_n \neq \delta_n] = \mathbb{P}[|\overline{X}_n| \leq n^{-1/4}] \to 0$ as $n \to \infty$. It will follow that $\sqrt{n}(\delta_n - \theta) \xrightarrow{(d)} N(0,1)$ when $\theta \neq 0$. On the other hand, when $\theta = 0$, $\delta_n = aX_n$ with high probability, so $\sqrt{n}\delta_n \xrightarrow{(d)} N(0, a^2)$. In this case

$$v(\theta) = \begin{cases} 1 & \theta \neq 0 \\ a^2 & \theta = 0. \end{cases}$$

It turns out that this can only happen on a set of parameters of Lebesgue measure zero.

7.3.3 Example. Let $P_{\mu,\sigma}(x) = \frac{1}{2}\phi(x) + \frac{1}{2\sigma}\phi(\frac{x-\mu}{\sigma})$ be a mixture of normals, a standard normal half of the time, and a $N(\mu, \sigma^2)$ half the time. Then trying to maximize the likelihood yields an infinite value, so there is no MLE.

$$\sup_{\mu,\sigma} L(x|\mu, \sigma) \geq \left\{ \sup_{\sigma, \mu = x} \frac{1}{2}\phi(x) + \frac{1}{2\sigma}\phi(0) \right\} = \infty.$$

Index