Foundations of Statistical Inference BS2a MT 2011

Lecturer: Geoff Nicholls

University of Oxford

Lecture Notes Version 1.0

Notes and Problem sheets are available at

www.stats.ox.ac.uk\ ∼nicholls\bs2a

Classes Fridays 3-4pm and 4-5pm in SPR1 lecture room.
Sign up today.

Core Text Books

Garthwaite, P. H., Jolliffe, I. T. and Jones, B. (2002) Statistical Inference, Oxford Science Publications

Leonard, T., Hsu, J. S. (2005) Bayesian Methods, Cambridge University Press.

D. R. Cox (2006) Principals of Statistical Inference

Acknowledgements

These notes are based on Prof. Griffiths' course in 2010.

Parametric families

$f(x;\theta)$, $\theta \in \Theta$, probability density of a random variable which could be discrete or continuous. θ can be 1-dimensional or of higher dimension. Equivalent notation: $f_\theta(x)$, $f(x \mid \theta)$, $f(x,\theta)$.

Likelihood $L(\theta;x) = f(x;\theta)$ and log-likelihood $\ell(\theta;x) = \log(L)$.

Examples 1. Normal: $f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$ $x \in \mathbb{R}$, $\theta \in \mathbb{R}$.

2. Poisson: $f(x;\theta) = \frac{\theta^x}{x!} e^{-\theta}, x = 0, 1, 2, \ldots, \theta > 0$.

3. Regression: $f(y;\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(y_i - \sum_{j=1}^p x_{ij}\beta_j)^2}$,
$y \in \mathbb{R}^n, \sigma > 0$, $\beta \in \mathbb{R}^p$. $\theta = \{\beta, \sigma\}$.

Exponential families of distributions GJJ 2.6, DRC 2.3

X belongs to the k-parameter exponential family if its probability density function can be written as

$$f(x;\theta) = \exp\left\{\sum_{j=1}^{k} A_j(\theta)B_j(x) + C(x) + D(\theta)\right\},$$

$x \in \chi$, $\theta \in \Theta$, where $A_1(\theta), \ldots, A_k(\theta), D(\theta)$ are functions of θ alone,and $B_1(x), B_2(x), \ldots, B_k(x), C(x)$ are well behaved functions of x alone.

Example: Poisson

We want to put the Poisson distribution in the form (with $k = 1$)

$$f(x;\theta) = \exp\left\{A(\theta)B(x) + C(x) + D(\theta)\right\},$$

$$\begin{aligned} e^{-\theta}\theta^x/x! &= e^{-\theta + x\log\theta - \log x!} \\ &= \exp\left\{(\log\theta)x - \log x! - \theta\right\} \end{aligned}$$

Some one parameter Exponential families

1. Binomial, 2. Poisson, 3. $N(\mu, 1)$ 4. Exponential

Distn	$f(x;\theta)$	$A(\theta)$	$B(x)$	$C(x)$	$D(\theta)$
1.	$\binom{n}{x}p^x(1-p)^{n-x}$	$\log(p/(1-p)$	x	$\log\binom{n}{x}$	$n\log(1-p)$
2.	$e^{-\theta}\theta^x/x!$	$\log\theta$	x	$-\log(x!)$	$-\theta$
3.	$\sqrt{2\pi}\exp\{-(x-\mu)^2/2\}$	μ	x	$-x^2/2$	$-\frac{1}{2}\log(2\pi)-\mu^2/2$
4.	$\theta e^{-\theta x}$	$-\theta$	x	0	$\log\theta$

Example: two parameter family, Gamma distribution

$$\begin{aligned}\frac{\beta^{\alpha}x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} &= \exp\{-\beta x + (\alpha-1)\log x + \alpha\log\beta - \log\Gamma(\alpha)\} \\ &= \exp\left\{(\alpha-1)\log x - \beta x - \log\left[\Gamma(\alpha)\beta^{-\alpha}\right]\right\}\end{aligned}$$

$\theta = (\alpha, \beta)$, $A_1(\theta) = \alpha - 1$, $A_2(\theta) = -\beta$.

Some two-parameter Exponential families

Distribution	$f(x;\theta)$	$A(\theta)$	$B(x)$	$C(x)$	$D(\theta)$
$N(\mu,\sigma^2)$	$\frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$	$A_1(\theta) = -1/2\sigma^2$	$B_1(x) = x^2$	0	$-\frac{1}{2}\log(2\pi\sigma^2)$
		$A_2(\theta) = \mu/\sigma^2$	$B_2(x) = x$	0	$-\frac{1}{2}\mu^2/\sigma^2$
Gamma	$\frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}$	$A_1(\theta) = \alpha - 1$	$B_1(x) = \log x$	0	$-\log\left[\Gamma(\alpha)\beta^{-\alpha}\right]$
		$A_2(\theta) = -\beta$	$B_2(x) = x$		

Exponential family canonical form: $A_j(\theta) = \theta_j$, $j = 1, \ldots, k$

$$f(x;\theta) = \exp\left\{\sum_{j=1}^{k} \theta_j B_j(x) + C(x) + D(\theta)\right\}.$$

θ_j and B_j, $j = 1, ..., k$ are the natural or canonical parameters and observations, respectively. Since

$$\int_{\mathbb{R}} f(x;\theta)dx = 1$$

we have

$$\int_{\mathbb{R}} \exp\left\{\sum_{j=1}^{k} \theta_j B_j(x) + C(x)\right\} dx = \exp\{-D(\theta)\}$$

$$\int_{\mathbb{R}} \exp\left\{\sum_{j=1}^{k} \theta_j B_j(x) + C(x)\right\} dx = \exp\{-D(\theta)\}$$

Differentiate with respect to θ_i and θ_j.

$$\begin{aligned}
\mathbb{E}[B_i(X)] &= -\frac{\partial}{\partial \theta_i} D(\theta) \\
\text{Covar}[B_i(X), B_j(X)] &= -\frac{\partial^2}{\partial \theta_i \partial \theta_j} D(\theta) \\
\text{Var}[B_i(X)] &= -\frac{\partial^2}{\partial \theta_i^2} D(\theta)
\end{aligned}$$

Cumulant Generating Function DRC 2.3 In a scalar canonical exponential family

$$\begin{aligned}\mathbb{E}_\theta[e^{sB(x)}] &= \int_{\mathbb{R}} \exp\left\{(\theta+s)B(x)+C(x)+D(\theta)\right\}dx \\ &= \exp\{D(\theta)-D(\theta+s)\}\end{aligned}$$

and hence if $M_{B(X)}(s)$ is the Moment Generating Function for $B(X)$, then

$$\log(M_{B(x)}(s)) = D(\theta)-D(\theta+s)$$

is the cumulant generating function (defined as the log of the MGF, generates mean, variance etc) for the cumulants of $B(X)$.

Example: Gamma Check with already known mean α/β and variance (α/β^2).

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

$$\theta_1 = -\beta,\ \theta_2 = \alpha - 1,\ B_1(x) = x,\ B_2(x) = \log x$$

$$D(\theta) = \alpha\log\beta - \log\Gamma(\alpha)$$

$$\begin{aligned}
\mathbb{E}[X] &= -\frac{\partial}{\partial(-\beta)} D(\theta) = \frac{\alpha}{\beta} \\
\mathrm{Var}[X] &= -\frac{\partial^2}{\partial(-\beta)^2} D(\theta) = \frac{\alpha}{\beta^2}
\end{aligned}$$

Exercise: show that $\mathbb{E}[\log X] = \psi_0(\alpha) - \log(\beta)$ where ψ_0 is the digamma function, and $\Gamma'(\alpha) = \Gamma(\alpha)\psi_0(\alpha)$.

Family preserved under transformations

For continuous distributions the family is preserved under fixed smooth invertible transformations $X \to Y$, $Y = Y(X)$ of the random variable. The Jacobian depends only on y and so the natural observation $B(x(y))$, the natural parameter $A(\theta)$, and $D(\theta)$ do not change, while

$$C(X) \to C(X(Y)) + \det |\partial X / \partial Y|.$$

Exponential family conditions

1. The range of X, χ does not depend on θ.

2. $(A_1(\theta), A_2(\theta), \ldots, A_k(\theta))$ have continuous second derivatives for $\theta \in \Theta$.

3. If $\dim \Theta = d$, then the Jacobian

$$J(\theta) = \left[\frac{\partial A_i(\theta)}{\partial \theta_j}\right]$$

has full rank d for $\theta \in \Theta$.

[Part of the "regularity conditions" which we cite later.]

The family is minimal if no linear combination exists such that

$$\lambda_0 + \lambda_1 B_1(x) + \cdots + \lambda_k B_k(x) = 0$$

for all $x \in \chi$.

The dimension of the family is defined to be d, the rank of $J(\theta)$.

A minimal exponential family is said to be curved when $d < k$ and linear when $d = k$. We refer to a (k, d) curved exponential family.

Examples not in an exponential family

1. Uniform on $[0, \theta]$, $\theta > 0$.

$$f(x;\theta) = \frac{1}{\theta},\ x \in [0,\theta]\ \theta > 0$$

2. The Cauchy distribution

$$f(x;\theta) = [\pi(1 + (x-\theta)^2)]^{-1},\ x \in \mathbb{R}$$

Example of a curved exponential family

1. (X_1, X_2) independent, normal, unit variance, means $(\theta, c/\theta)$, c known.

$$\log f(x;\theta) = x_1\theta + cx_2/\theta - \theta^2/2 - c^2\theta^{-2}/2 + ...$$

is a $(2,1)$ curved exponential family.

Linear relations among A's do not generate curvature

2. Take a k-parameter exponential family and impose $A_1 = aA_2 + b$ for constants a and b.

$$\sum_{i=1}^{k} A_i B_i + C + D = \sum_{i=3}^{k} A_i B_i + A_2(aB_1 + B_2) + (C + bB_1) + D$$

This gives a $k-1$ parameter EF, not a $(k, k-1)$-CEF.
(old example ex Casella and Berger not curved!)

Sufficiency (GJJ 2.5) Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta)$. $T(X_1, \ldots, X_n)$ is a statistic. That is $t(x_1, \ldots, x_n)$ does not depend on θ. T is a sufficient statistic for θ if the conditional distribution of $X_1, \ldots, X_n$, given T, does not depend on θ.

T then contains all the information there is in the sample about θ.

A statistic is minimal sufficient if it can be expressed as a function of every other sufficient statistic.

Example: n independent trials where the probability of success is p. Let $X_1, \ldots, X_n$ be indicator variables which are 1 or 0 depending if the trial is a success or failure. Let $T = \sum_{i=1}^{n} X_i$ with distribution $h(t)$. The conditional distribution of $X_1, \ldots, X_n$ given $T = t$ is

$$\begin{aligned} g(x_1, \ldots, x_n \mid t) &= \frac{f(x_1, \ldots, x_n)}{h(t)} \\ &= \frac{\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} \\ &= \frac{p^t (1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} \\ &= \binom{n}{t}^{-1}, \end{aligned}$$

not depending on p, so T is sufficient for p.

Example illustrating minimal sufficiency: (GJJ Example 2.7)

Suppose T above is not minimal sufficient but another statistic U is MS. Then U can be given as a function of T (and not *vis versa* or T is MS) and there exist $t_1 \neq t_2$ values of T so that $U(t_1) = U(t_2)$ (ie $T \to U$ is many to one so $U \to T$ is not a function, and we assume for the moment no other t make $U(t) = U(t_1)$). The event $U = u$ is the event $T \in \{t_1, t_2\}$. Let $x_1, ...x_n$ contain t_1 successes. Then

$$\begin{aligned} g(x_1, ...x_n|u) &= g(x_1, ...x_n|t_1)P(T = t_1|T \in \{t_1, t_2\}) \\ &= \frac{p^{t_1}(1-p)^{n-t_1}}{\binom{n}{t_1}p^{t_1}(1-p)^{n-t_1} + \binom{n}{t_2}p^{t_2}(1-p)^{n-t_2}} \end{aligned}$$

which depends on p, so U is not sufficient, a contradiction, and hence T must be MS (similar reasong for multiple t_i).

Factorization Criterion $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ if and only if there exist two non-negative functions K_1, K_2 such that the likelihood function $L(\theta; \mathbf{x})$ can be written

$$L(\theta; x) = K_1[t(x_1, \ldots, x_n); \theta] K_2[x_1, \ldots, x_n],$$

where K_1 depends only on the sample through T, and K_2 does not depend on θ.

Proof for discrete random variables.

Let $h(t)$ be the distribution of T.

1. Assume that T is sufficient, then the distribution of the sample is

$$L(\theta; x) = g(x \mid t)h(t \mid \theta).$$

$g(x \mid t)$ does not depend on θ since T is sufficient, and $h(t \mid \theta)$ depends on x through $t(x)$ only. We set $L(\theta; x) = K_1 K_2$, where $K_1 \equiv h$, $K_2 \equiv g$.

2. Suppose $L(\theta; x) = K_1[t; \theta]K_2[x]$. Then

$$h(t \mid \theta) = \sum_{\{x:T(x)=t\}} L(\theta; x) = K_1[t; \theta] \sum_{\{x:T(x)=t\}} K_2(x).$$

Thus

$$g(x_1, \ldots, x_n \mid t) = \frac{L(\theta; x)}{h(t \mid \theta)} = \frac{K_2[x]}{\sum_{\{x:T(x)=t\}} K_2(x)},$$

not depending on θ. (K_1 cancels out in numerator and denominator.)

Goodness of Fit

Sufficiency is useful in model-checking. The model says data $Y \sim L(\theta; \cdot)$ with $L(\theta; x) = g(x \mid t)h(t \mid \theta)$, so the model asserts a distribution $g(x \mid t)$ which does not depend on the unknown θ. We can check the data respects this aspect of the model without needing to know the model parameter θ.

Example See also DRC example 3.2

Binomial n independent trials, success probability p, $X_1, ..., X_n$ Bernoulli rv. $T = \sum_i X_i$ is sufficient and

$$g(x \mid t) = (\#\text{ seq's with } t \text{ successes})^{-1}.$$

In a GOF test, g becomes the null. Test for +ve correlation between X_i and $X_i + 1$ (say). Count $v(x)$ pairs (X_i, X_{i+1}) with $(X_i = X_{i+1})$ and report p-value $P(v(X) > v(x))$.

Sufficiency in an exponential family I

Random Sample $X_1, \ldots, X_n$

Likelihood

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

Exponential family form again.

Sufficiency in an exponential family II

Suppose the family is in canonical form, and let $t_j = \sum_{i=1}^{n} B_j(x_i)$, $C(x) = \sum_{i=1}^{n} C(x_i)$.

$$L(\theta; x) = \exp\left\{ \sum_{j=1}^{k} \theta_j t_j + nD(\theta) + C(x) \right\}.$$

By the factorization criterion $t_1, \ldots, t_k$ are sufficient statistics for $\theta_1, \ldots, \theta_k$.

Estimators

Classical estimation of parameters.

$$\widehat{\theta} = \widehat{\theta}(x) = t(x_1, \ldots, x_n).$$

An interval estimate is a set valued function $C(X)$ such that $\theta \in C(X)$ with a specified probability.

Maximum likelihood estimation

If $L(\theta)$ is differentiable and there is a unique maximum in the interior of $\theta \in \Theta$, then $\widehat{\theta}$ is the solution of

$$\frac{\partial}{\partial\theta} L(\theta; x) = 0 \quad \text{or} \quad \frac{\partial}{\partial\theta} \ell(\theta) = 0,$$

where $l(\theta) = \log L(\theta; x)$.

$T = t(\mathbf{x})$ is unbiassed for a function $g(\theta)$ if

$$\mathbb{E}_\theta(T) = \int_\chi t(x) f(\mathbf{x};\theta) d\mathbf{x} = g(\theta), \quad \text{for all } \theta \in \Theta.$$

The Bias of an estimator T is

$$\text{bias}_\theta(T) = \mathbb{E}_\theta\left[T - g(\theta)\right]$$

and the Mean square error (MSE) of T is

$$\text{mse}_\theta(T) = \mathbb{E}_\theta\left[T - g(\theta)\right]^2 = V_\theta(T) + \left[\text{bias}_\theta(T)\right]^2$$

Example: $N(\mu, \sigma^2)$. $\widehat{\mu} = \bar{X}$ and $S^2 = (n-1)^{-1} \sum_{i=1}^n \left(X_i - \bar{X}\right)^2$ are unbiassed estimates of μ and σ^2.

Minimum variance estimator T.

If T and T' are unbiassed estimators of θ, then T is a MVE if

$$\text{var}_\theta(T) \leq \text{var}_\theta(T'), \text{ for all } \theta \in \Theta$$

Maximum likelihood estimation and exponential families

$$L(\theta; x) = \exp\left\{ \sum_{j=1}^{k} A_j(\theta) \left(\sum_{i=1}^{n} B_j(x_i) \right) + nD(\theta) + \sum_{i=1}^{n} C(x_i) \right\}.$$

Let $T_j(X) = \sum_{i=1}^{n} B_j(X_i)$, $j = 1, \ldots, k$. If the realized data are $X = x$, then the statistics evaluated on the data are $T_j(x) = t_j$.

The MLE of $\theta_1, \ldots, \theta_k$ are the solution of

$$t_j = \mathbb{E}_\theta(T_j),\ j = 1, \ldots, k.$$

Proof

$$\ell = \log L = \text{const} + \sum_{j=1}^{k} \theta_j t_j + nD(\theta)$$

so

$$\frac{\partial}{\partial \theta_j}\ell = t_j + n\frac{\partial}{\partial \theta_j}D(\theta)$$

However we know that

$$\mathbb{E}_\theta[T_j] = -n\frac{\partial}{\partial \theta_j}D(\theta), \text{ so}$$

$$\frac{\partial}{\partial \theta_j}\ell = t_j - \mathbb{E}_\theta(T_j) = 0$$

is equivalent to $t_j = \mathbb{E}(T_j)$.