Foundations of Statistical Inference

J. Berestycki & D. Sejdinovic

Department of Statistics
University of Oxford

MT 2019

Course arrangements

- **Lectures** Tue. 3 pm and Thu.10 am weeks 1-4. Then Mon. 11 am and Tue 1pm weeks 5-8.
- **UG Classes** Three sets of classes. Times, dates and enrolment via Minerva
- **MSc classes**
- Notes and Problem sheets will be available at `www.stats.ox.ac.uk\~berestyc\SB2a.html`
- **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
- This course builds on notes from Bob Griffiths, Geoff Nicholls and Jonathan Marchini

Part A Statistics

The majority of the statistics that you have learned up to now falls under the philosophy of classical (or Frequentist) statistics. This theory makes the assumption that we can randomly take repeated samples of data from the same population.

You learned about three types of statistical inference

- Point estimation (Maximum likelihood, bias, consistency, efficiency, information)
- Interval estimation(exact and approximate intervals using CLT)
- Hypothesis testing (Neyman-Pearson lemma, uniformly most powerful tests, generalised likelihood ratio tests)

You also had an introduction to Bayesian Statistics

- Posterior inference (Posterior $\propto$ Likelihood $\times$ Prior)
- Interval estimation(credible intervals, HPD intervals)
- Priors(conjugate priors, improper priors, Jeffreys' prior)
- Hypothesis testing (marginal likelihoods, Bayes Factors)

Part A Statistics

The majority of the statistics that you have learned up to now falls under the philosophy of classical (or Frequentist) statistics. This theory makes the assumption that we can randomly take repeated samples of data from the same population.

You learned about three types of statistical inference

- Point estimation (Maximum likelihood, bias, consistency, efficiency, information)
- Interval estimation(exact and approximate intervals using CLT)
- Hypothesis testing (Neyman-Pearson lemma, uniformly most powerful tests, generalised likelihood ratio tests)

You also had an introduction to Bayesian Statistics

- Posterior inference (Posterior $\propto$ Likelihood $\times$ Prior)
- Interval estimation(credible intervals, HPD intervals)
- Priors(conjugate priors, improper priors, Jeffreys' prior)
- Hypothesis testing (marginal likelihoods, Bayes Factors)

Part A Statistics

The majority of the statistics that you have learned up to now falls under the philosophy of classical (or Frequentist) statistics. This theory makes the assumption that we can randomly take repeated samples of data from the same population.

You learned about three types of statistical inference

- Point estimation (Maximum likelihood, bias, consistency, efficiency, information)
- Interval estimation(exact and approximate intervals using CLT)
- Hypothesis testing (Neyman-Pearson lemma, uniformly most powerful tests, generalised likelihood ratio tests)

You also had an introduction to Bayesian Statistics

- Posterior inference (Posterior $\propto$ Likelihood $\times$ Prior)
- Interval estimation(credible intervals, HPD intervals)
- Priors(conjugate priors, improper priors, Jeffreys' prior)
- Hypothesis testing (marginal likelihoods, Bayes Factors)

Part A Statistics

The majority of the statistics that you have learned up to now falls under the philosophy of classical (or Frequentist) statistics. This theory makes the assumption that we can randomly take repeated samples of data from the same population.

You learned about three types of statistical inference

- Point estimation (Maximum likelihood, bias, consistency, efficiency, information)
- Interval estimation(exact and approximate intervals using CLT)
- Hypothesis testing (Neyman-Pearson lemma, uniformly most powerful tests, generalised likelihood ratio tests)

You also had an introduction to Bayesian Statistics

- Posterior inference (Posterior $\propto$ Likelihood $\times$ Prior)
- Interval estimation(credible intervals, HPD intervals)
- Priors(conjugate priors, improper priors, Jeffreys' prior)
- Hypothesis testing (marginal likelihoods, Bayes Factors)

Part A Statistics

The majority of the statistics that you have learned up to now falls under the philosophy of classical (or Frequentist) statistics. This theory makes the assumption that we can randomly take repeated samples of data from the same population.

You learned about three types of statistical inference

- Point estimation (Maximum likelihood, bias, consistency, efficiency, information)
- Interval estimation(exact and approximate intervals using CLT)
- Hypothesis testing (Neyman-Pearson lemma, uniformly most powerful tests, generalised likelihood ratio tests)

You also had an introduction to Bayesian Statistics

- Posterior inference (Posterior $\propto$ Likelihood $\times$ Prior)
- Interval estimation(credible intervals, HPD intervals)
- Priors(conjugate priors, improper priors, Jeffreys' prior)
- Hypothesis testing (marginal likelihoods, Bayes Factors)

Part A Statistics

The majority of the statistics that you have learned up to now falls under the philosophy of classical (or Frequentist) statistics. This theory makes the assumption that we can randomly take repeated samples of data from the same population.

You learned about three types of statistical inference

- Point estimation (Maximum likelihood, bias, consistency, efficiency, information)
- Interval estimation(exact and approximate intervals using CLT)
- Hypothesis testing (Neyman-Pearson lemma, uniformly most powerful tests, generalised likelihood ratio tests)

You also had an introduction to Bayesian Statistics

- Posterior inference (Posterior $\propto$ Likelihood $\times$ Prior)
- Interval estimation(credible intervals, HPD intervals)
- Priors(conjugate priors, improper priors, Jeffreys' prior)
- Hypothesis testing (marginal likelihoods, Bayes Factors)

Part A Statistics

The majority of the statistics that you have learned up to now falls under the philosophy of classical (or Frequentist) statistics. This theory makes the assumption that we can randomly take repeated samples of data from the same population.

You learned about three types of statistical inference

- Point estimation (Maximum likelihood, bias, consistency, efficiency, information)
- Interval estimation(exact and approximate intervals using CLT)
- Hypothesis testing (Neyman-Pearson lemma, uniformly most powerful tests, generalised likelihood ratio tests)

You also had an introduction to Bayesian Statistics

- Posterior inference (Posterior $\propto$ Likelihood $\times$ Prior)
- Interval estimation(credible intervals, HPD intervals)
- Priors(conjugate priors, improper priors, Jeffreys' prior)
- Hypothesis testing (marginal likelihoods, Bayes Factors)

Part A Statistics

The majority of the statistics that you have learned up to now falls under the philosophy of classical (or Frequentist) statistics. This theory makes the assumption that we can randomly take repeated samples of data from the same population.

You learned about three types of statistical inference

- Point estimation (Maximum likelihood, bias, consistency, efficiency, information)
- Interval estimation(exact and approximate intervals using CLT)
- Hypothesis testing (Neyman-Pearson lemma, uniformly most powerful tests, generalised likelihood ratio tests)

You also had an introduction to Bayesian Statistics

- Posterior inference (Posterior $\propto$ Likelihood $\times$ Prior)
- Interval estimation(credible intervals, HPD intervals)
- Priors(conjugate priors, improper priors, Jeffreys' prior)
- Hypothesis testing (marginal likelihoods, Bayes Factors)

Frequentist inference

In BS2a we develop the theory of point estimation further.

- Are there families of distributions about which we can make general statements? $\Rightarrow$ Exponential families.
- How can we summarise all the information in a dataset about a parameter θ? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter θ? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter θ? $\Rightarrow$Rao-Blackwell Theorem and Lehmann-Scheffé Theorem.

Frequentist inference

In BS2a we develop the theory of point estimation further.

- Are there families of distributions about which we can make general statements? $\Rightarrow$ Exponential families.
- How can we summarise all the information in a dataset about a parameter θ? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter θ? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter θ? $\Rightarrow$Rao-Blackwell Theorem and Lehmann-Scheffé Theorem.

Frequentist inference

In BS2a we develop the theory of point estimation further.

- Are there families of distributions about which we can make general statements? $\Rightarrow$ Exponential families.
- How can we summarise all the information in a dataset about a parameter θ? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter θ? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter θ? $\Rightarrow$Rao-Blackwell Theorem and Lehmann-Scheffé Theorem.

Frequentist inference

In BS2a we develop the theory of point estimation further.

- Are there families of distributions about which we can make general statements? $\Rightarrow$ Exponential families.
- How can we summarise all the information in a dataset about a parameter θ? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter θ? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter θ? $\Rightarrow$Rao-Blackwell Theorem and Lehmann-Scheffé Theorem.

Frequentist inference

In BS2a we develop the theory of point estimation further.

- Are there families of distributions about which we can make general statements? $\Rightarrow$ Exponential families.
- How can we summarise all the information in a dataset about a parameter θ? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter θ? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter θ? $\Rightarrow$Rao-Blackwell Theorem and Lehmann-Scheffé Theorem.

Frequentist inference

In BS2a we develop the theory of point estimation further.

- Are there families of distributions about which we can make general statements? $\Rightarrow$ Exponential families.
- How can we summarise all the information in a dataset about a parameter θ? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter θ? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter θ? $\Rightarrow$Rao-Blackwell Theorem and Lehmann-Scheffé Theorem.

Frequentist inference

In BS2a we develop the theory of point estimation further.

- Are there families of distributions about which we can make general statements? $\Rightarrow$ Exponential families.
- How can we summarise all the information in a dataset about a parameter θ? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter θ? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter θ? $\Rightarrow$Rao-Blackwell Theorem and Lehmann-Scheffé Theorem.

Frequentist inference

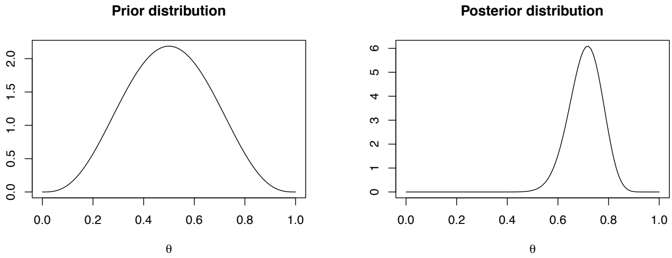
In BS2a we develop the theory of point estimation further.

- Are there families of distributions about which we can make general statements? $\Rightarrow$ Exponential families.
- How can we summarise all the information in a dataset about a parameter θ? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter θ? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter θ? $\Rightarrow$Rao-Blackwell Theorem and Lehmann-Scheffé Theorem.

Bayesian inference

Parameters are treated as random variables. Inference starts by specifying a prior distribution on θ based on prior beliefs. Having collected some data we use Bayes' Theorem to update our beliefs to obtain a posterior distribution.

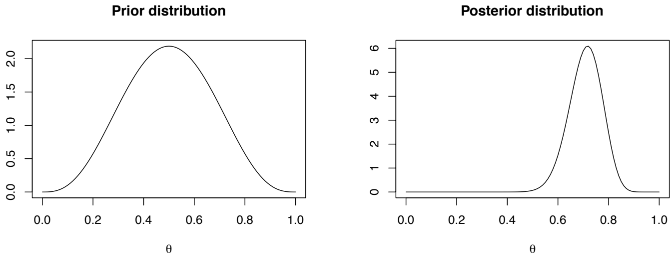
Quick Example Suppose I give a coin and tell you that it is bit biased. We might use a Beta(4,4) distribution to represent our beliefs about the θ. If we observe 30 heads and 10 tails we can use probability theory to infer a posterior distribution for θ of Beta(34, 14).



Bayesian inference

Parameters are treated as random variables. Inference starts by specifying a prior distribution on θ based on prior beliefs. Having collected some data we use Bayes' Theorem to update our beliefs to obtain a posterior distribution.

Quick Example Suppose I give a coin and tell you that it is bit biased. We might use a Beta(4,4) distribution to represent our beliefs about the θ. If we observe 30 heads and 10 tails we can use probability theory to infer a posterior distribution for θ of Beta(34, 14).



Computational techniques for Bayesian inference

It is not always possible to obtain an analytic solution when doing Bayesian Inference, so we study approximate computational techniques in this course.

These include

- Approximations to marginal likelihoods NEW
  - Variational Approximations
  - Laplace approximations
  - Bayesian Information Criterion (BIC)
- The EM algorithm NEW
  - useful in Frequentist and Bayesian inference of missing data problems

Computational techniques for Bayesian inference

It is not always possible to obtain an analytic solution when doing Bayesian Inference, so we study approximate computational techniques in this course.

These include

- Approximations to marginal likelihoods NEW
  - Variational Approximations
  - Laplace approximations
  - Bayesian Information Criterion (BIC)
- The EM algorithm NEW
  - useful in Frequentist and Bayesian inference of missing data problems

Computational techniques for Bayesian inference

It is not always possible to obtain an analytic solution when doing Bayesian Inference, so we study approximate computational techniques in this course.

These include

- Approximations to marginal likelihoods NEW
  - Variational Approximations
  - Laplace approximations
  - Bayesian Information Criterion (BIC)
- The EM algorithm NEW
  - useful in Frequentist and Bayesian inference of missing data problems

Decision theory

Quick Example

Zed and Adrian and run a small bicycle shop called "Z to A Bicycles". They must order bicycles for the coming season. Orders for the bicycles must be placed in quantities of twenty (20). The cost per bicycle is 70 GBP if they order 20, 67 GBP if they order 40, 65 GBP if they order 60, and 64 GBP if they order 80. The bicycles will be sold for 100 GBP each. Any bicycles left over at the end of the season can be sold (for certain) at 45 GBP each. If Zed and Adrian run out of bicycles during the season, then they will suffer a loss of "goodwill" among their customers. They estimate this goodwill loss to be 5 GBP per customer who was unable to buy a bicycle. Zed and Adrian estimate that the demand for bicycles this season will be 10, 30, 50, or 70 bicycles with probabilities of 0.2, 0.4, 0.3, and 0.1 respectively.

Notation

X, Y, Z Capital letters for random variables.

x, y, z Lower case letters for realisations of random variables.

$\mathbb{E}_X(\cdot)$ Expectation with respect to the random variable X.

$\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_k\}$ Sometimes we will use bold symbols to denote a vector of parameters.

Lecture 1 - Exponential families

A class of (very) regular models – common building block of (more) complex models

Parametric families

$f(x;\theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, probability density of a random variable (rv) which could be discrete or continuous.

Parametric $1 \leq d < +\infty$

Likelihood $L(\theta;x) = f(x;\theta)$: think of $\theta \mapsto L(\theta;x)$a s function of θ while $x \mapsto f(x:\theta)$ is a pdf/pmf for each θ.

Notation : log-likelihood $\ell(\theta;x) = \log(L(\theta;x))$.

Examples

1. Normal $N(\theta,1)$: $f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$ $x \in \mathbb{R}$, $\theta \in \mathbb{R}$.
2. Poisson: $f(x;\theta) = \frac{\theta^x}{x!} e^{-\theta}, x = 0,1,2,\dots,\theta > 0$.
3. Gaussian regression:

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

Exponential families of distributions

Definition (1-parameter Exponential family)

A rv X belongs to a 1-*parameter exponential family* if its probability density function (pdf) or probability mass function (pmf) can be written as

$$\begin{aligned} f(x;\theta) &= \exp\left\{A(\theta)B(x) + C(x) - D(\theta)\right\} \\ &= h(x)\exp\left\{A(\theta)B(x)\right\}\psi(\theta), \end{aligned}$$

where $\theta \in \Theta$ and $B(x), C(x)$ are well behaved (measurable) functions of x alone.

$\psi(\theta)$ is a normalising factor

$$\psi(\theta) = \left[\int h(x)\exp\left\{A(\theta)B(x)\right\}dx\right]^{-1}.$$

Exponential families of distributions

Definition (1-parameter Exponential family)

A rv X belongs to a 1-*parameter exponential family* if its probability density function (pdf) or probability mass function (pmf) can be written as

$$\begin{aligned} f(x;\theta) &= \exp\left\{A(\theta)B(x) + C(x) - D(\theta)\right\} \\ &= h(x)\exp\left\{A(\theta)B(x)\right\}\psi(\theta), \end{aligned}$$

where $\theta \in \Theta$ and $B(x), C(x)$ are well behaved (measurable) functions of x alone.

$\psi(\theta)$ is a normalising factor

$$\psi(\theta) = \left[\int h(x)\exp\left\{A(\theta)B(x)\right\} dx\right]^{-1}.$$

Example 1 : Poisson

We want to put the Poisson distribution in the form

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

$$\begin{aligned} e^{-\theta}\theta^x/x!\,\mathbb{I}_{x\in\mathbb{N}} &= e^{-\theta+x\log\theta-\log x!}\,\mathbb{I}_{x\in\mathbb{N}} \\ &= \exp\left\{(\log\theta)x - \log x! - \theta\right\}\mathbb{I}_{x\in\mathbb{N}} \end{aligned}$$

So $A(\theta) = \log\theta, B(x) = x, C(x) = -\log x!, D(\theta) = \theta$.

Examples of 1-parameter Exponential families

Binomial, Poisson, Normal, Exponential.

Distn	$f(x;\theta)$	$A(\theta)$	$B(x)$	$C(x)$	$D(\theta)$
Bin(n,p)	$\binom{n}{x}p^x(1-p)^{n-x}$	$\log\frac{p}{(1-p)}$	x	$\log\binom{n}{x}$	$-n\log(1-p)$
Pois(θ)	$e^{-\theta}\theta^x/x!$	$\log\theta$	x	$-\log(x!)$	θ
$N(\mu,1)$	$\frac{1}{\sqrt{2\pi}}\exp\{-\frac{(x-\mu)^2}{2}\}$	μ	x	$-x^2/2$	$\frac{-1}{2}(\mu^2-\log(2\pi)$
Exp(θ)	$\theta e^{-\theta x}$	$-\theta$	x	0	$-\log\theta$

Others : negative binomial, Pareto (with known minimum), Weibull (with known shape), Laplace (with known mean), Log-normal, inverse Gaussian, beta, Dirichlet, Wishart. Exercise: check these distributions

Examples not in the Exponential family

- To be in an exponential family, it is a necessary condition that the support of the pdf/pmf $f(x;\theta)$ does not depend on θ.
  Exemple: The shifted exponential $f(x;\theta) = e^{\theta - x}\mathbf{1}_{x>\theta}$ cannot be in the Exponential family.
- There are other reasons for which a parametric distribution might not be in the Exponential family.
  Exemple: Cauchy distributions with given location parameter

$$f(x;\mu) = \frac{1}{\pi(1+(x-\mu))^2}\mathbf{1}_{x\in\mathbb{R}}$$

  cannot be in the Exponential family. (Prove it).

Exponential families of distributions

Definition (k-parameters Exponential family)

A rv X belongs to a k*-parameter exponential family* if its probability density function (pdf) or probability mass function (pmf) can be written as

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

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

Exponential families are widely used in practice - for example in generalised linear models (see BS1a).

Example 2 : a 2-parameter family (Gamma)

If $X \sim$ Gamma(α, β) then let $\theta = (\alpha, \beta)$ so

$$\begin{aligned} f(x;\theta) &= \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \mathbf{1}_{x \geq 0} \\ &= \exp\left\{\alpha \log \beta + (\alpha - 1)\log x - \beta x - \log \Gamma(\alpha)\right\} \mathbf{1}_{x \geq 0} \\ &= \exp\left\{(\alpha - 1)\log x - \beta x - \log\left[\Gamma(\alpha)\beta^{-\alpha}\right]\right\} \mathbf{1}_{x \geq 0} \end{aligned}$$

And we have
$A_1(\theta) = \alpha - 1$, $B_1(x) = \log x$,
$A_2(\theta) = -\beta$, $B_2(x) = x$.

Some other 2-parameter Exponential families

Distribution	$f(x;\theta)$	$A(\theta)$	$B(x)$	$C(x)$	$D(\theta)$
$N(\mu,\sigma^2)$	$\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$	$A_1(\theta) = -1/2\sigma^2$	$B_1(x) = x^2$	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

Definition (Canonical form)

Let $\phi_j = A_j(\theta)$, $j = 1, \ldots, k$ then

$$f(x;\boldsymbol{\phi}) = \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x) + C(x) - D(\theta)\right\}$$
$$= h(x)\psi(\theta)\exp\left\{\sum_{j=1}^{k} \phi_j B_j(x)\right\}.$$

is called the canonical form of the density.
$\phi_j, j = 1, ..., k$ are the canonical *parameters*,
$B_j, j = 1, ..., k$ are the canonical *observations*.
(sometimes called the natural parameters and observations)
$\Phi := \{\phi : \int h(x)\exp\left\{\sum_{j=1}^{k} \phi_j B_j(x)\right\} dx < \infty\}$ is the natural parameter space.

Exponential family canonical form

Definition (Canonical form)

Let $\phi_j = A_j(\theta)$, $j = 1, \dots, k$ then

$$f(x; \boldsymbol{\phi}) = \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x) + C(x) - D(\phi)\right\}$$
$$= h(x)\psi(\phi)\exp\left\{\sum_{j=1}^{k} \phi_j B_j(x)\right\}.$$

Can always do this since

$$\psi(\theta)^{-1} = \int h(x) \exp\{\sum_j \phi_j B_j(x)\} dx$$

is called the canonical form of the density.

$\phi_j, j = 1, ..., k$ are the canonical *parameters*,

$B_j, j = 1, ..., k$ are the canonical *observations*

Exponential family canonical form

Definition (Canonical form)

Let $\phi_j = A_j(\theta)$, $j = 1, \ldots, k$ then

$$\begin{aligned} f(x;\boldsymbol{\phi}) &= \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x) + C(x) - D(\phi)\right\} \\ &= h(x)\psi(\phi)\exp\left\{\sum_{j=1}^{k} \phi_j B_j(x)\right\}. \end{aligned}$$

is called the canonical form of the density.

$\phi_j, j = 1, ..., k$ are the canonical *parameters*,

$B_j, j = 1, ..., k$ are the canonical *observations*.

(sometimes called the natural parameters and observations)

$\Phi := \{\phi : \int h(x) \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x)\right\} dx < \infty\}$ is the natural parameter space.

Exemple: Binomial pmf in canonical form

Let $X \sim Bin(n,p)$ with pmf

$$f(x;p) = \binom{n}{x} p^x (1-p)^{n-x} \mathbf{1}_{x \in \{0,1,\dots,n\}}$$
$$= e^{x \log \frac{p}{1-p} - n \log(1-p)} \binom{n}{x} \mathbf{1}_{x \in \{0,1,\dots,n\}}$$

Writing $\phi = \log \frac{p}{1-p}$ this can be rewritten

$$f(x;\phi) = e^{\phi x - n \log(1+e^{\phi})} \binom{n}{x} \mathbf{1}_{x \in \{0,1,\dots,n\}}$$

and so the natural parameter space is $\Phi = \mathbb{R}$.

Assumption

There is an implicit assumption that the number of freely varying θ_i's is the same as the number of free ϕ_i's

Assumption 1

$\text{Dim}(\Theta) = \text{Dim}(\phi(\Theta))$.

If this is not satsified we say that X belongs to a curved exponential family.

Assumption 2

The family is minimal meaning that there are no linear constraints on the B_i.

Definition

A family of distribution $\{f(\cdot;\phi), \phi \in \Phi\}$ belonging to the canonical k-parameter exponential family is called full-rank if for every $\phi \in \text{Int}(\Phi)$ the $k \times k$ covarience matrix $\left(\frac{\partial^2}{\partial\phi_i\partial\phi_j} D(\phi)\right)$ is nonsingular or equivalently if Φ contains a k-dimensional rectangle.

Convexity

Theorem

The natural/canonical parameter space of a full rank linear k-dimensional exponential family is convex and contains a k-dimensional open interval.

Definition (Regular Exponential family)

We say that the (canonical) exponential family is regular if Φ is an open set.

Theorem (Computation of moments)

If $\phi \in \mathit{Int}(\Phi)$ then $L(\cdot; x)$ is infinitely differentiable at ϕ and for all $k \geq 0$ $\mathbb{E}[\|B(X)\|^k] < \infty$ and the cumulant generating function (log of moment generating function) exists and verifies

$$\log \mathbb{E}_\phi[e^{s' \cdot B(X)}] = D(\phi + s) - D(\phi)$$

(s' is the transpose of $s \in \mathbb{R}^k$ and $B(X) = (B_1(X), \ldots, B_k(X))$)
The function $D(\Phi)$ is infinitely differentiable at every $\phi \in \mathit{Int}(\Phi)$ and

$$E[B_i(X)] = \frac{\partial}{\partial \phi_i} D(\phi), \; \mathit{Cov}(B_i(X), B_j(X)) = \frac{\partial^2}{\partial \phi_i \partial \phi_j} D(\phi).$$

Proof

$$\underbrace{\int_{\mathbb{R}} \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x) + C(x) - D(\boldsymbol{\phi})\right\} dx}_{\int_{\mathbb{R}} f(x;\boldsymbol{\phi})dx} \quad = \quad 1$$

$$\int_{\mathbb{R}} \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x) + C(x)\right\} dx \quad = \quad \exp\{D(\boldsymbol{\phi})\}$$

$$M_{B(X)}(s) = \mathbb{E}_X[e^{s' \cdot B(X)}] = \int_{\mathbb{R}} \exp\left\{(\phi + s)' \cdot B(x) + C(x) - D(\phi)\right\} dx$$
$$= \exp\{-D(\phi) + D(\phi + s)\}$$

Then $\quad \log(M_{B(X)}(s)) = D(\phi + s) - D(\phi)$

This is the cumulant generating function

Proof - 2nd part

What am I missing in the proof ? Existence of $M_{B(X)}(s)$
$\phi \in \text{Int}(\Phi)$, hence $\exists \delta > 0$ s. t. $\forall |s| < \delta$

$$\int_{\mathcal{X}} e^{(\phi+s)' \cdot B(x) + C(x)} dx < +\infty \quad \Rightarrow$$

$$\forall |s| \leq \delta, \quad M_{B(X)}(s) < +\infty, \quad \Rightarrow \forall k \geq 0, \quad \mathbb{E}_X[\|B(X)\|^k] < +\infty$$

Proof - part 3 : Differentiability

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

Differentiate with respect to ϕ_i. Why is it differentiable?

$$\begin{aligned}
\int_{\mathbb{R}} B_i(x) \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x) + C(x)\right\} dx &= \frac{\partial}{\partial \phi_i} D(\boldsymbol{\phi}) \exp\{D(\boldsymbol{\phi})\} \\
\int_{\mathbb{R}} B_i(x) \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x) + C(x) - D(\boldsymbol{\phi})\right\} dx &= \frac{\partial}{\partial \phi_i} D(\boldsymbol{\phi}) \\
\mathbb{E}[B_i(X)] &= \frac{\partial}{\partial \phi_i} D(\boldsymbol{\phi})
\end{aligned}$$

By induction $\partial^k D(\phi)/\partial \phi^k$ exists for all k.

About the cumulant generating function

We have seen that

$$\begin{aligned} \mathbb{E}[B_i(X)] &= \frac{\partial}{\partial \phi_i} D(\boldsymbol{\phi}) \\ \text{Exercise} \qquad \mathsf{Cov}[B_i(X), B_j(X)] &= \frac{\partial^2}{\partial \phi_i \partial \phi_j} D(\boldsymbol{\phi}) \\ \text{Exercise} \qquad \mathsf{Var}[B_i(X)] &= \frac{\partial^2}{\partial \phi_i^2} D(\boldsymbol{\phi}) \end{aligned}$$

and more generally $\log(M_{B(X)}(s))$: power series expansion ($k = 1$)

$$\log(M_{B(X)}(s)) = \sum_{r=1}^{\infty} \kappa_r s^r / r! \quad \kappa_r : \text{cumulants of } B(x)$$

where $\kappa_1 = \mathbb{E}(B(X))$ and $\kappa_2 = V(B(X))$ Exercise : prove this

Example 3 : Gamma

We already know that if $X \sim \text{Gamma}(\alpha, \beta)$ the $\mathbb{E}(X) = \frac{\alpha}{\beta}$ and $Var(X) = \frac{\alpha}{\beta^2}$.

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

$\phi_1 = -\beta$, $\phi_2 = \alpha$, $B_1(x) = x$, $B_2(x) = \log x$

$$\begin{aligned} D(\boldsymbol{\phi}) &= -\alpha \log \beta + \log \Gamma(\alpha) \\ &= -\phi_2 \log(-\phi_1) + \log \Gamma(\phi_2) \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X] &= \frac{\partial}{\partial \phi_1} D(\boldsymbol{\phi}) = -\frac{\phi_2}{\phi_1} = \frac{\alpha}{\beta} \\ \text{Var}[X] &= \frac{\partial^2}{\partial \phi_1^2} D(\boldsymbol{\phi}) = \frac{\phi_2}{\phi_1^2} = \frac{\alpha}{\beta^2} \end{aligned}$$

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

Example 4 : Binomial

We already know that if $X \sim \text{Binom}(n,p)$ then $\mathbb{E}(X) = np$.

$$\phi = \log \frac{p}{(1-p)} \Rightarrow p = \frac{e^{\phi}}{1+e^{\phi}}$$

and

$$B_1(x) = x, D(\boldsymbol{\phi}) = -n\log(1-p) = n\log(1+e^{\phi})$$

$$\begin{aligned} \log(M_X(s)) &= -D(\boldsymbol{\phi}) + D(\boldsymbol{\phi}+s) \\ &= -n\log(1+e^{\phi}) + n\log(1+e^{\phi+s}) \end{aligned}$$

Therefore

$$\kappa_1 = \frac{\partial}{\partial s}\log(M_X(s))\Big|_{s=0} = n\frac{e^{\phi}}{1+e^{\phi}} = np$$

Example 5 : Skew-logistic distribution

Consider the real valued random variable X with pdf

$$\begin{aligned} f(x;\theta) &= \frac{\theta e^{-x}}{(1+e^{-x})^{\theta+1}} \\ &= \exp\Big\{ -\theta\log(1+e^{-x}) + \log\Big(\frac{e^{-x}}{1+e^{-x}}\Big) + \log\theta\Big\} \end{aligned}$$

and $\phi = \theta$, $B_1(x) = -\log(1+e^{-x})$ and $D(\phi) = -\log\theta = -\log(\phi)$

$$\Rightarrow \log(M_X(s)) = -D(\boldsymbol{\phi}) + D(\boldsymbol{\phi}+s) = \log(\phi) - \log(\phi+s)$$

$$\begin{aligned} \mathbb{E}(\log(1+e^{-x})) &= \tfrac{-1}{\phi} = \frac{-1}{\theta} \\ \text{Var}(\log(1+e^{-x})) &= \frac{1}{\phi^2} = \frac{1}{\theta^2} \end{aligned}$$

These results are harder to derive directly.

Family preserved under transformations

A smooth invertible transformation of a rv from the Exponential family is also within the Exponential family. If $X \to Y$, $Y = Y(X)$ then

$$\begin{aligned} f_Y(y;\theta) &= f_X(x(y);\theta)|\partial X/\partial Y| \\ &= \exp\left\{\sum_{j=1}^{k} A_j(\theta)B_j(x(y)) + C(x(y)) + D(\theta)\right\}|\partial X/\partial Y|, \end{aligned}$$

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

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

Curved families

If $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ and $d < k$ the family is said to be curved and linear when $d = k$. We refer to a (k, d) curved exponential family.
Example 6 (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.

Curved families

If $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ and $d < k$ the family is said to be curved and linear when $d = k$. We refer to a (k, d) curved exponential family.

Example 6 (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.

Example 7

Normal family $N(\theta, \theta^2)$ (mean =variance).

- $d = 1$
- $$\log f(x, \theta) = \frac{1}{\theta} x - \frac{1}{2\theta^2} x^2 - \frac{1}{2} + D(\theta)$$
- Minimal?
- Curved (2,1).

Exponential family are regular models (in the sense of asymptotic regularity)

- Canonical exponential families with Φ open set are regular parametric families (see asymptotic normality of mle)
  [if the range of $\mathcal{X}$ does not depend on ϕ]
- For non canonical exponential families :
  If $\phi(\theta) = \big(A_1(\theta), A_2(\theta), \ldots, A_k(\theta)\big)$ have continuous second derivatives for $\theta \in \Theta \subset \mathbb{R}^d$ and $d \leq k$. and the Jacobian

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

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

Then $f_\theta(x)$ is a regular parametric family.