# Foundations of Statistical Inference

Julien Berestycki

Department of Statistics
University of Oxford

MT 2016

# Course arrangements

- **Lectures** Mon. 3 pm and Tue.11 weeks 1-8
- **Classes** (1) & (2) Weeks 3,5,7,8 (J. Berestycki) Wed. 10am and 11:30am, LG04 (3)& (4) Weeks 3,4,7,8 (S. Filippi) Thur. 9am and 12 am.
- Hand in solutions by noon Monday of the week. Class Tutors : Julien Berestycki and Sarah Filippi
- 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

# 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)

# 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)

# 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)

# 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)

# 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)

# 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 )

# 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?

# 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.

# 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 $\theta$?

# 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 $\theta$? $\Rightarrow$ Sufficiency and the Factorization 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 $\theta$? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter $\theta$?

# 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 $\theta$? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter $\theta$? $\Rightarrow$Cramer-Rao inequality (and bound).

# 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 $\theta$? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter $\theta$? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter $\theta$?

# 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 $\theta$? $\Rightarrow$ Sufficiency and the Factorization Theorem.
- What are limits of how well we can estimate a parameter $\theta$? $\Rightarrow$Cramer-Rao inequality (and bound).
- How can we find good estimators of a parameter $\theta$? $\Rightarrow$Rao-Blackwell Theorem and Lehmann-Scheffé Theorem.

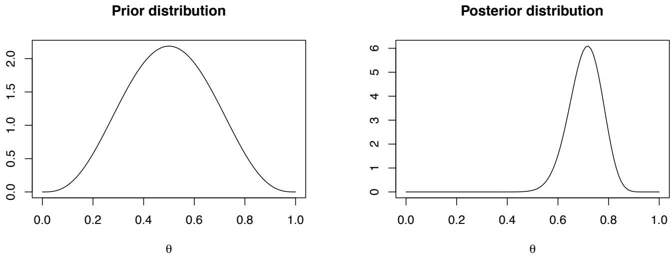
# Bayesian inference

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

# Bayesian inference

Parameters are treated as random variables. Inference starts by specifying a prior distribution on $\theta$ 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 $\theta$. If we observe 30 heads and 10 tails we can use probability theory to infer a posterior distribution for $\theta$ 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.

# 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)

# 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

# Parametric families

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

# Parametric families

$f(x;\theta)$, $\theta \in \Theta$, probability density of a random variable (rv) which could be discrete or continuous. $\theta$ 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 $\ell(\theta; x) = \log(L)$.

# Parametric families

$f(x;\theta)$, $\theta \in \Theta$, probability density of a random variable (rv) which could be discrete or continuous. $\theta$ 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 $N(\theta, 1)$ : $f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$ $x \in \mathbb{R},\ \theta \in \mathbb{R}$.

# Parametric families

$f(x;\theta)$, $\theta \in \Theta$, probability density of a random variable (rv) which could be discrete or continuous. $\theta$ 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 $N(\theta, 1)$ : $f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$ $x \in \mathbb{R},\ \theta \in \mathbb{R}$.
2. Poisson: $f(x;\theta) = \frac{\theta^x}{x!} e^{-\theta}, x = 0, 1, 2, \ldots, \theta > 0$.

# Parametric families

$f(x;\theta)$, $\theta \in \Theta$, probability density of a random variable (rv) which could be discrete or continuous. $\theta$ 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 $N(\theta,1)$ : $f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$ $x \in \mathbb{R},\ \theta \in \mathbb{R}$.
2. Poisson: $f(x;\theta) = \frac{\theta^x}{x!} e^{-\theta}, x = 0,1,2,\dots,\theta > 0$.
3. Regression:

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

# Parametric families

$f(x;\theta)$, $\theta \in \Theta$, probability density of a random variable (rv) which could be discrete or continuous. $\theta$ 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 $N(\theta,1)$ : $f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$ $x \in \mathbb{R}$, $\theta \in \mathbb{R}$.
2. Poisson: $f(x;\theta) = \frac{\theta^x}{x!} e^{-\theta}$, $x = 0,1,2,\dots,\theta > 0$.
3. 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

**Definition 1** GJJ 2.6, DRC 2.3
A rv $X$ belongs to a $k$-parameter exponential family if its probability density function (pdf) can be written as

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

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

# Exponential families of distributions

Definition 1 GJJ 2.6, DRC 2.3

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

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

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

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

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

# Example 1 : Poisson

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

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

# Example 1 : Poisson

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

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

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

# Example 1 : Poisson

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

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

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

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 | $f(x;\theta)$ | $A(\theta)$ | $B(x)$ | $C(x)$ | $D(\theta)$ |
|---|---|---|---|---|---|
| $\mathrm{Bin}(n,p)$ | $\binom{n}{x}p^x(1-p)^{n-x}$ | $\log\frac{p}{(1-p)}$ | $x$ | $\log\binom{n}{x}$ | $n\log(1-p)$ |

# Examples of 1-parameter Exponential families

Binomial, Poisson, Normal, Exponential.

| Distn | $f(x;\theta)$ | $A(\theta)$ | $B(x)$ | $C(x)$ | $D(\theta)$ |
|---|---|---|---|---|---|
| $\mathrm{Bin}(n,p)$ | $\binom{n}{x}p^x(1-p)^{n-x}$ | $\log\frac{p}{(1-p)}$ | $x$ | $\log\binom{n}{x}$ | $n\log(1-p)$ |
| $\mathrm{Pois}(\theta)$ | $e^{-\theta}\theta^x/x!$ | $\log\theta$ | $x$ | $-\log(x!)$ | $-\theta$ |

# Examples of 1-parameter Exponential families

Binomial, Poisson, Normal, Exponential.

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

# Examples of 1-parameter Exponential families

Binomial, Poisson, Normal, Exponential.

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

# 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$(\theta)$ | $e^{-\theta}\theta^x/x!$ | $\log\theta$ | $x$ | $-\log(x!)$ | $-\theta$ |
| $N(\mu,1)$ | $\frac{1}{\sqrt{2\pi}}\exp\{-\frac{(x-\mu)^2}{2}\}$ | $\mu$ | $x$ | $-x^2/2$ | $\frac{1}{2}(\mu^2-\log(2\pi))$ |
| Exp$(\theta)$ | $\theta e^{-\theta x}$ | $-\theta$ | $x$ | $0$ | $\log\theta$ |

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

# Example 2 : a 2-parameter family (Gamma)

If $X \sim \text{Gamma}(\alpha, \beta)$ then let $\theta = (\alpha, \beta)$ so

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

# Example 2 : a 2-parameter family (Gamma)

If $X \sim \text{Gamma}(\alpha, \beta)$ then let $\theta = (\alpha, \beta)$ so

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

# Example 2 : a 2-parameter family (Gamma)

If $X \sim \text{Gamma}(\alpha, \beta)$ then let $\theta = (\alpha, \beta)$ so

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

# Example 2 : a 2-parameter family (Gamma)

If $X \sim \text{Gamma}(\alpha, \beta)$ then let $\theta = (\alpha, \beta)$ so

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

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

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

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

$$= \omega(x)\psi(\theta) \exp\left\{\sum_{j=1}^{k} \phi_j B_j(x) + C(x)\right\}.$$

$\phi_j, j = 1, ..., k$ are the canonical *parameters*,
$B_j, j = 1, ..., k$ are the canonical *observations*.
These are sometimes called the natural parameters and observations.

# Exponential family canonical form

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

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

Because $\psi(\theta) = \left(\int \omega(x)\exp\left\{\sum_{j=1}^{k} \phi_j B_j(x)\right\} dx\right)^{-1}$ depends only on $\phi$
$\phi_j, j = 1, ..., k$ are the canonical *parameters*,
$B_j, j = 1, ..., k$ are the canonical *observations*.
These are sometimes called the natural parameters and observations.
$\Phi := \{\phi : \int \omega(x)\exp\left\{\sum_{j=1}^{k} \phi_j B_j(x)\right\} dx < \infty\}$ is natural parameter space.

# Exponential family canonical form

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

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

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

Since

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

Since

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

we have

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

Since

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

we have

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

Since

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

we have

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

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

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

Differentiate with respect to $\phi_i$.

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

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

Differentiate with respect to $\phi_i$.

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

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

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

Differentiate with respect to $\phi_i$.

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

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

Differentiate with respect to $\phi_i$.

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

Exercise $$\mathrm{Cov}[B_i(X), B_j(X)] = -\frac{\partial^2}{\partial \phi_i \partial \phi_j} D(\phi)$$

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

Differentiate with respect to $\phi_i$.

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

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

$$\mathbb{E}[B_i(X)] = -\frac{\partial}{\partial \phi_i} D(\phi)$$

Exercise $$\mathrm{Cov}[B_i(X), B_j(X)] = -\frac{\partial^2}{\partial \phi_i \partial \phi_j} D(\phi)$$

Exercise $$\mathrm{Var}[B_i(X)] = -\frac{\partial^2}{\partial \phi_i^2} D(\phi)$$

# Example 3 : Gamma

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

# Example 3 : Gamma

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

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

# Example 3 : Gamma

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

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

$\phi_1 = -\beta$, $\phi_2 = \alpha - 1$, $B_1(x) = x$, $B_2(x) = \log x$

# Example 3 : Gamma

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

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

$\phi_1 = -\beta$, $\phi_2 = \alpha - 1$, $B_1(x) = x$, $B_2(x) = \log x$

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

# Example 3 : Gamma

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

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

$\phi_1 = -\beta$, $\phi_2 = \alpha - 1$, $B_1(x) = x$, $B_2(x) = \log x$

$$\begin{aligned} D(\phi) &= \alpha \log \beta - \log \Gamma(\alpha) \\ &= (\phi_2 + 1)\log(-\phi_1) - \log \Gamma(\phi_2 + 1) \end{aligned}$$

# Example 3 : Gamma

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

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

$\phi_1 = -\beta$, $\phi_2 = \alpha - 1$, $B_1(x) = x$, $B_2(x) = \log x$

$$\begin{aligned} D(\phi) &= \alpha \log \beta - \log \Gamma(\alpha) \\ &= (\phi_2 + 1)\log(-\phi_1) - \log \Gamma(\phi_2 + 1) \end{aligned}$$

$$\mathbb{E}[X] = -\frac{\partial}{\partial \phi_1} D(\phi) = -\frac{(\phi_2 + 1)}{\phi_1} = \frac{\alpha}{\beta}$$

# Example 3 : Gamma

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

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

$\phi_1 = -\beta$, $\phi_2 = \alpha - 1$, $B_1(x) = x$, $B_2(x) = \log x$

$$\begin{aligned} D(\phi) &= \alpha \log \beta - \log \Gamma(\alpha) \\ &= (\phi_2 + 1)\log(-\phi_1) - \log \Gamma(\phi_2 + 1) \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X] &= -\frac{\partial}{\partial \phi_1} D(\phi) = -\frac{(\phi_2 + 1)}{\phi_1} = \frac{\alpha}{\beta} \\ \text{Var}[X] &= -\frac{\partial^2}{\partial \phi_1^2} D(\phi) = -\frac{(\phi_2 + 1)}{\phi_1^2} = \frac{\alpha}{\beta^2} \end{aligned}$$

# Example 3 : Gamma

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

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

$\phi_1 = -\beta$, $\phi_2 = \alpha - 1$, $B_1(x) = x$, $B_2(x) = \log x$

$$\begin{aligned} D(\phi) &= \alpha \log \beta - \log \Gamma(\alpha) \\ &= (\phi_2 + 1)\log(-\phi_1) - \log \Gamma(\phi_2 + 1) \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X] &= -\frac{\partial}{\partial \phi_1} D(\phi) = -\frac{(\phi_2 + 1)}{\phi_1} = \frac{\alpha}{\beta} \\ \text{Var}[X] &= -\frac{\partial^2}{\partial \phi_1^2} D(\phi) = -\frac{(\phi_2 + 1)}{\phi_1^2} = \frac{\alpha}{\beta^2} \end{aligned}$$

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

# Cumulant Generating Function

In a scalar canonical exponential family ($k = 1$)

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

# Cumulant Generating Function

In a scalar canonical exponential family ($k = 1$)

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

If $M_{B(X)}(s)$ is the Moment Generating Function (mgf) for $B(X)$, then

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

# Cumulant Generating Function

In a scalar canonical exponential family ($k = 1$)

$$\begin{aligned} \mathbb{E}_X[e^{sB(X)}] &= \int_{\mathbb{R}} \exp\{(\phi + s)B(x) + C(x) + D(\phi)\}\, dx \\ &= \exp\{D(\phi) - D(\phi + s)\} \end{aligned}$$

If $M_{B(X)}(s)$ is the Moment Generating Function (mgf) for $B(X)$, then

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

This is the cumulant generating function (defined as the log of the mgf) for the cumulants of $B(X)$ i.e.

$$\log(M_{B(X)}(s)) = \sum_{r=1}^{\infty} \kappa_r s^r / r!$$

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

# Example 4 : Binomial

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

# Example 4 : Binomial

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

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

# 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

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

# 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

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

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

# 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

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

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

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

$$f(x;\theta) = \frac{\theta e^{-x}}{(1+e^{-x})^{\theta+1}}$$

# Example 5 : Skew-logistic distribution

Consider the real valued random variable $X$ with pdf

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

# Example 5 : Skew-logistic distribution

Consider the real valued random variable $X$ with pdf

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

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

# 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 $D(\phi) = \log\theta = \log(-\phi)$

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

# Example 5 : Skew-logistic distribution

Consider the real valued random variable $X$ with pdf

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

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

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

$$\begin{aligned} \mathbb{E}(\log(1+e^{-x})) &= \left.\frac{-1}{\phi+s}\right|_{s=0} = \frac{1}{\theta} \\ \mathrm{Var}(\log(1+e^{-x})) &= \left.\frac{1}{(\phi+s)^2}\right|_{s=0} = \frac{1}{\theta^2} \end{aligned}$$

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

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

- The family is minimal if no linear constraint on the $B_i$ or the $\phi_i$.

- The family is minimal if no linear constraint on the $B_i$ or the $\phi_i$.
- 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.

- The family is minimal if no linear constraint on the $B_i$ or the $\phi_i$.
- 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.
- If the family is minimal and the parameter space contains a $d$-dimensional open rectangle the family is said to be full rank.
- 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

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

# Example

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

- $d = 1$

# Example

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

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

# Example

Normal family $N(\theta, \theta)$ (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?

# Example

Normal family $N(\theta, \theta)$ (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 (1,2).

# Exponential family conditions

1. The range of $X$, $\chi$ does not depend on $\theta$.

# Exponential family conditions

1. The range of $X$, $\chi$ does not depend on $\theta$.
2. $\big(A_1(\theta), A_2(\theta), \ldots, A_k(\theta)\big)$ have continuous second derivatives for $\theta \in \Theta$.

# Exponential family conditions

1. The range of $X$, $\chi$ does not depend on $\theta$.
2. $\big(A_1(\theta), A_2(\theta), \ldots, A_k(\theta)\big)$ 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 $\theta \in \Theta$.

# Exponential family conditions

1. The range of $X$, $\chi$ does not depend on $\theta$.
2. $\big(A_1(\theta), A_2(\theta), \ldots, A_k(\theta)\big)$ 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.]

# Linear relations among $A$'s do not generate curvature

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.

# Linear relations among $A$'s do not generate curvature

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.

**Theorem**

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

# Examples not in an exponential family

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

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

# 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

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

# 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

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

Other examples include the F-distribution, hypergeometric distribution and logistic distribution.