Lecture 4:

Parametric Inference – continued

Outline:

1. MLE Optimality
2. The Delta Method
3. Multiparameter models
4. Parametric Bootstrap
5. Sufficient Statistics
6. Exponential Families
7. Computing MLE *

* *covered in detail on the seminar*

MLE Optimality

- We already saw that MLE is asymptotically normal. But how does it compare (asymptotically) to other reasonable estimators?

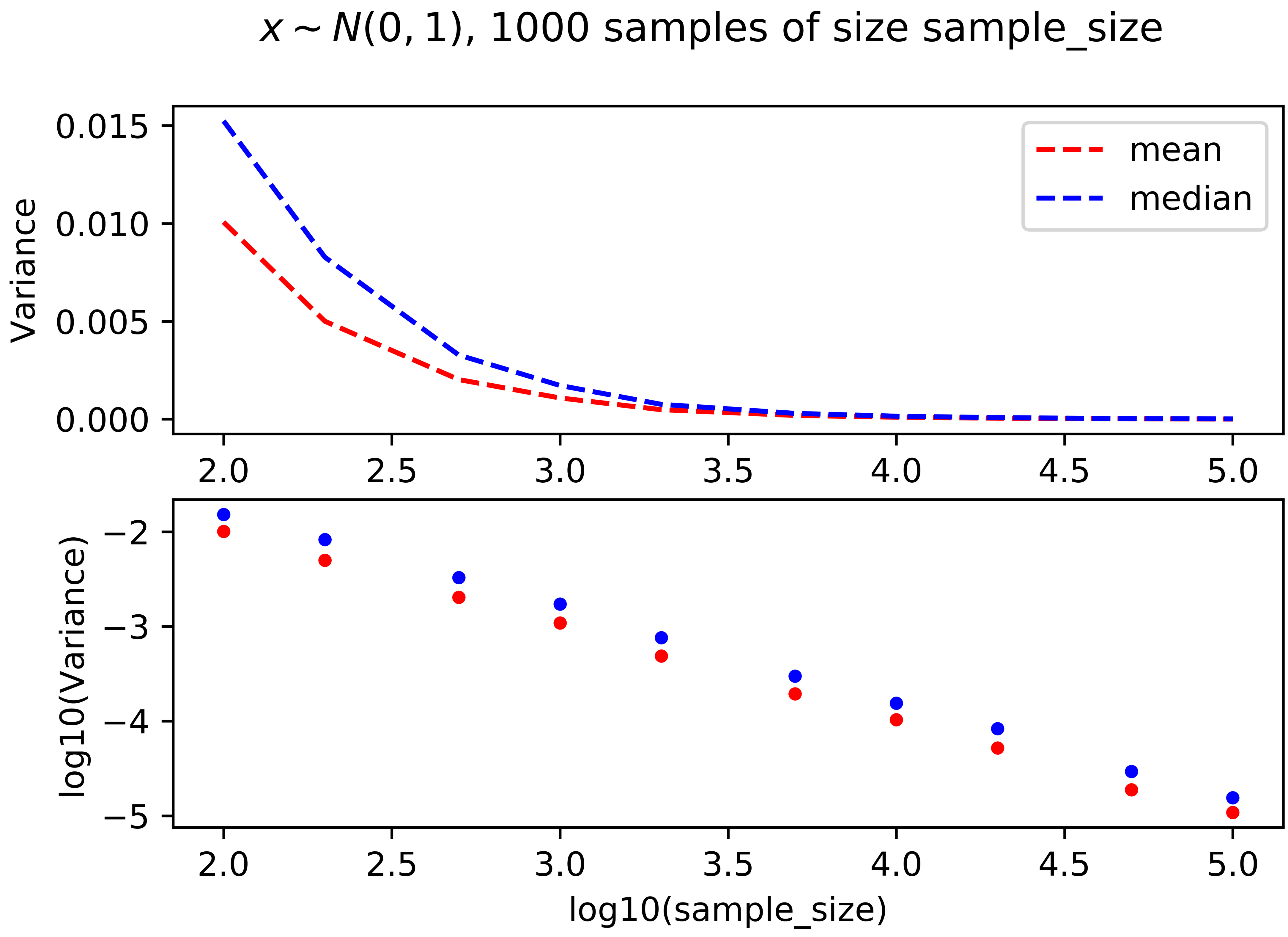
- **Example**: $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma)$. The MLE for θ is $\widehat{\theta}_n = \overline{X}_n$ the sample mean. For it,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{P} \mathcal{N}(0, \sigma)$$

  Another reasonable estimator for θ is the sample median $\tilde{\theta}_n$. For it we have

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{P} \mathcal{N}(0, \sigma\sqrt{\pi/2})$$

MLE Optimality



MLE Optimality

- More generally, for two estimators T_n and U_n such that

$$\sqrt{n}(T_n - \theta) \xrightarrow{P} \mathcal{N}(0, \sigma = t) \text{ and } \sqrt{n}(U_n - \theta) \xrightarrow{P} \mathcal{N}(0, \sigma = u)$$

define **asymptotic relative efficiency** as ARE$(U, T) = t^2/u^2$ – so in our normal example, ARE$(\widetilde{\theta}_n, \widehat{\theta}_n) = 2/\pi = 0.63$ – median is 0.63 times as effective as mean (we are effectively using a fraction of the data).

- **Theorem**: If $\widehat{\theta}_n$ is MLE and $\tilde{\theta}_n$ is any other estimator, then

$$\text{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1$$

so MLE is **efficient** or **asymptotically optimal!**

The Delta Method

The Delta method

- Let $\tau = g(\theta)$, where g is a smooth function. The MLE of τ is $\widehat{\tau} = g(\widehat{\theta})$ (equivariance). But what is **the distribution** of $\widehat{\tau}$?

- **Theorem** (The Delta method): If $\tau = g(\theta)$ where g is diff. and $g'(\theta) \neq 0$, then

$$(\widehat{\tau}_n - \tau)/\hat{\text{se}}(\widehat{\tau}) \xrightarrow{P} \mathcal{N}(0,1) \quad \text{where } \widehat{\tau}_n = g(\widehat{\theta}_n) \text{ and}$$

$$\hat{\text{se}}(\widehat{\tau}_n) = |\, g'(\widehat{\theta})\,|\, \hat{\text{se}}(\widehat{\theta}_n)$$

This allows to build a $(1 - \alpha)$-confidence interval:

$$C_n = \left(\widehat{\tau}_n - z_{\alpha/2}\hat{\text{se}}(\widehat{\tau}_n), \widehat{\tau}_n + z_{\alpha/2}\hat{\text{se}}(\widehat{\tau}_n)\right)$$

The Delta method

- **Example**: Let $X_1, \ldots, X_n \sim$ Bernoulli(p) and let $\psi = g(p) = \log\big(p/(1-p)\big)$. The **Fisher information** is $I(p) = 1/\big(p(1-p)\big)$, so the estimated s.e. of the MLE $\widehat{p}_n$ is

$$\hat{\text{se}} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$ The MLE of ψ is $\widehat{\psi} = \log \widehat{p}/(1-\widehat{p})$.

Since $g'(p) = 1/\big(p(1-p)\big)$, so

$$\hat{\text{se}}(\widehat{\psi}_n) = |\, g'(\widehat{p}_n) \,|\, \hat{\text{se}}(\widehat{p}_n) = \frac{1}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}$$

The Delta method

- **Example**: Recall that $\hat{\text{se}}(\hat{\theta}_n) = 1/\sqrt{I_n(\theta)}$ – **Fisher information.**

  Now let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$ – μ is known and we want to estimate $\psi = \log \sigma$. Fisher info is **minus** the expectation of:

$$\frac{\partial^2 \log f(X; \sigma)}{\partial \sigma^2} = \frac{\partial^2}{\partial \sigma^2}\left(\log \sigma - \frac{(X-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sigma^2} - \frac{3(X-\mu)^2}{\sigma^4}$$

so $I(\sigma) = 2/\sigma^2$, $I_n(\sigma) = n\, I(\sigma)$ and $\hat{\text{se}}(\widehat{\sigma}_n) = \widehat{\sigma}_n/\sqrt{2n}$ and so for $\psi = g(\sigma) = \log \sigma$ with $g' = 1/\sigma$ we have:

$$\hat{\text{se}}(\widehat{\psi}_n) = \frac{1}{\widehat{\sigma}_n}\frac{\widehat{\sigma}_n}{\sqrt{2n}} = \frac{1}{\sqrt{2n}}$$

Multiparameter models

Multiparameter models

- Our ideas can be directly extended to models with several parameters. $\theta = (\theta_1, \ldots, \theta_k)$ and MLE is $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$. The matrix of minus expectations of second derivatives:

$$H_{ij} = \frac{\partial^2 \ell_n}{\partial \theta_i \, \partial \theta_j} \quad \text{is} \quad \Big(I_n(\theta)\Big)_{ij} = -\,\mathbb{E}_\theta(H_{ij})$$

– called the **Fisher information matrix**.

It is handy to denote the **inverse of it** $J_n(\theta) = I_n^{-1}(\theta)$.

Multiparameter models

- **Theorem**: Under appropriate regularity conditions,

$$(\hat{\theta} - \theta) \approx \mathcal{N}(0, \Sigma = J_n)$$

and for θ_j – j-th component of θ

$$\frac{\hat{\theta}_j - \theta_j}{\hat{\text{se}}_j} \xrightarrow{P} \mathcal{N}(0,1)$$ where $\hat{\text{se}}_j^2 = J_n(j, j)$ is the j-th diagonal element of J_n.

And the approximate covariance $\text{Cov}(\hat{\theta}_j, \hat{\theta}_k) \approx J_n(j, k)$ is the (j,k)-th non-diagonal element

Multiparameter models

- **Theorem** (Multiparameter Delta method): For $\tau = g(\theta_1, \ldots, \theta_k)$, denote $\nabla g = \left(\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_k}\right)^T$ the **gradient** of g. If ∇g evaluated at $\widehat{\theta}$, $\widehat{\nabla g}$, is not zero, for $\widehat{\tau} = g(\widehat{\theta})$ we have

$$\frac{\widehat{\tau} - \tau}{\hat{\text{se}}(\widehat{\tau})} \xrightarrow{P} \mathcal{N}(0,1) \quad \text{where}$$

$$\hat{\text{se}}(\widehat{\tau}) = \sqrt{(\widehat{\nabla g})^T \widehat{J}_n (\widehat{\nabla g})} \ \text{ where } \widehat{J}_n = J_n(\widehat{\theta}_n).$$

Multiparameter models

- **Example**. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$ and $\tau = g(\mu, \sigma) = \sigma/\mu$. We have

$$I_n(\mu, \sigma) = \frac{n}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \text{ hence } J_n = I_n^{-1} = \frac{\sigma^2}{2n}\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

the gradient of g is $\nabla g = \left(-\sigma/\mu^2,\ 1/\mu\right)^T$

thus $\widehat{\text{se}}(\widehat{\tau}) = \sqrt{(\widehat{\nabla g})^T \widehat{J}_n (\widehat{\nabla g})} = \sqrt{\frac{1}{n\hat{\mu}^4} + \frac{\hat{\sigma}^2}{2\hat{\mu}^2}}$

Parametric Bootstrap

Parametric Bootstrap

- We've introduced bootstrap in the **non-parametric** setting – we sampled $X_1^*, \dots, X_n^*$ from the empirical CDF $\widehat{F}_n$. How does one extend it for parametric models? (To estimate standard errors and thus confidence intervals)

- Quite straightforward! We should sample $X \sim f(x; \hat{\theta}_n)$ – where $\hat{\theta}_n$ is either MLE or MOM-estimator.

- See an **example** on the next slide

Parametric Bootstrap

- **Example**: Recall estimating $\tau = \sigma/\mu$ for the normal distribution. To do parametric bootstrap:

1. Simulate (sample) $X_1^*, \ldots, X_n^* \sim \mathcal{N}(\widehat{\mu}, \widehat{\sigma})$
2. Compute $\widehat{\mu}^* = n^{-1} \sum_i X_i^*$ and $(\widehat{\sigma}^*)^2 = n^{-1} \sum_i (X_i^* - \hat{\mu}^*)^2$
3. Compute $\widehat{\tau}^* = g(\widehat{\mu}^*, \widehat{\sigma}^*) = \widehat{\sigma}^*/\widehat{\mu}^*$
4. Repeating this B times gives $\widehat{\tau}_1^*, \ldots, \widehat{\tau}_B^*$

From that bootstrap sample, get the std! This is simpler than the Delta method, but the latter gives an analytic expression.

Sufficient Statistics

Sufficient Statistics

- A **statistic** is a function $T(X^n)$ of the data.
- **Definition**: Write $x^n \leftrightarrow y^n$ if $f(x^n;\ \theta) = cf(y^n;\ \theta)$ for some constant c that might depend on x^n and y^n but not θ. A statistic $T(x^n)$ is **sufficient** if $T(x^n) \leftrightarrow T(y^n)$ implies $x^n \leftrightarrow y^n$.
- So if $x^n \leftrightarrow y^n$, then the likelihood function based on x^n has the same shape as the likelihood function based on y^n. Roughly, a statistic $T(X^n)$ is sufficient if we can calculate the likelihood knowing only $T(X^n)$.
- **Example**: $X_1, \ldots, X_n \sim$ Bernoulli(p). Then $\mathscr{L}(p) = p^S(1-p)^{n-S}$ where $S = \sum_i X_i$. So S is sufficient

Sufficient Statistics

- **Example**: Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$ and let $T = (\overline{X}, S)$. Then

$$f\left(X^n; \mu, \sigma\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\bar{X}-\mu)^2}{2\sigma^2}\right\}$$

where S^2 is sample variance. So $T = (\overline{X}, S)$ is a sufficient statistic, as well as $U = (2\overline{X}, S-1)$ is. These are sufficient as well: $T_1(X^n) = (X_1, \ldots, X_n)$, $T_2(X^n) = (\overline{X}, S)$. $T_3(X^n) = \overline{X}$ is not. $T_4(X^n) = (\overline{X}, S, X_3)$ is. Notice that T_2 is a function of T_1, T_2 is a function of T_4.

- **Definition**: A statistic is **minimal sufficient** if 1) it is sufficient, and 2) it is a function of every other sufficient statistic

Sufficient Statistics

- So for $\mathcal{N}(\mu, \sigma)$, $T = (\overline{X}, S)$ is a minimal sufficient statistic. For Bernoulli, $T = \sum X_i$ is. For Poisson, $T = \sum X_i$ as well.

- **Exercise**: Let $(X_1, X_2) \sim$ Bernoulli(p). Figure out that $T = X_1 + X_2$ is sufficient statistic.

- **Theorem** (Factorization): T is sufficient if and only if there are functions $g(t, \theta)$ and $h(x)$ such that $f(x^n;\ \theta) = g(t(x^n), \theta)\, h(x^n)$

- **Example**: For the above exercise, $t = x_1 + x_2$

$$f(x_1, x_2; \theta) = f(x_1; \theta) f(x_2; \theta) = \theta^{x_1}(1-\theta)^{1-x_1}\, \theta^{x_2}(1-\theta)^{1-x_2} = g(t, \theta) h(x_1, x_2)$$

where $g(t, \theta) = \theta^t (1-\theta)^{2-t}$ and $h(x_1, x_2) = 1$.

Sufficient Statistics

- Why bother with sufficient statistics? Let $\widehat{\theta}$ be an estimator of θ. The Rao-Blackwell theorem says that **an estimator should only depend on the sufficient statistic**, otherwise it can be improved! Denote $R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(\theta - \widehat{\theta})^2$ – the MSE of the estimator.

- **Theorem** (Rao-Blackwell): Let $\widehat{\theta}$ be an estimator and let T be a sufficient statistic. Define a new estimator by

$$\tilde{\theta} = \mathbb{E}_\theta(\widehat{\theta} \,|\, T)$$

then, for any θ, it holds that $R(\theta, \tilde{\theta}) \leq R(\theta, \widehat{\theta})$!

Exponential Families

Exponential Families

- Most of parametric models we studied so far are special cases of a general class of models called **exponential families.** We say that $\left\{f(x;\theta) :\ \theta \in \Theta\right\}$ is a one-parameter exponential family if there are functions $\eta(\theta),\ B(\theta),\ T(x)$ and $h(x)$ such that

$$f(x;\theta) = h(x)\, e^{\eta(\theta)T(x) - B(\theta)}$$

such $T(x)$ is called the **natural sufficient statistic**.

- **Example**: $X \sim$ Poisson(θ). Then $f(x;\theta) = \dfrac{\theta^x e^{-\theta}}{x!} = \dfrac{1}{x!} e^{x \log \theta - \theta}$

so $\eta(\theta) = \log \theta,\ B(\theta) = \theta,\ T(x) = x$ and $h(x) = 1/(x!)$

Exponential Families

- **Example**: $X \sim$ Binomial(n, θ). Then
$$f(x;\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{n}{x} \exp\left\{ x \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta) \right\}$$
so in this case $\eta(\theta) = \log(\theta/(1-\theta))$, $B(\theta) = -n\log\theta$ and
$T(x) = x,\ h(x) = \binom{n}{x}$.

- An exponential family can be rewritten as
$f(x;\eta) = h(x)\, e^{\eta T(x) - A(\eta)}$ where $\eta = \eta(\theta)$ is called the **natural parameter** and $A(\eta) = \log \int h(x)\, e^{\eta T(x)} dx$

- For example, Poisson $f(x;\eta) = e^{\eta x - e^{\eta}}/x!$ and $\eta = \log\theta$.

Exponential Families

- **Theorem**: For X with density from exp. family,

$$\mathbb{E}(T(X)) = A'(\eta) \quad \text{and} \quad \mathbb{V}(T(X)) = A''(\eta)$$

- If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector, then

$$f(x;\theta) = h(x)\exp\left\{\sum_{j=1}^{k} \eta_j(\theta)T_j(x) - B(\theta)\right\} \text{ where}$$

$T = (T_1, \ldots, T_k)$ is the sufficient statistic. IID sample of size n also has sufficient statistic $(\sum_i T_1(X_i), \ldots, \sum_i T_k(X_i))$. Then also

$$\mathbb{E}(T(X)) = A'(\eta) \text{ (a vector)} \quad \text{and} \quad \mathbb{V}(T(X)) = A''(\eta) \text{ (a matrix)}$$

Computing MLE

Computing MLE

- This is all good, but we only have analytic formulae for MLE for the simplest models – for the whole exponential family, but still – if a model is more complicated than that, we need **numerical methods** already.

- On the seminar, we'll cover:

  1. The Newton-Raphson method

  2. The **Expectation Maximization (EM) Algoritm**

Computing MLE

- **Motivation:** Mixture of 2 normal distributions: