WORKSHOP ON THE ANALYSIS OF NEURAL DATA 2001

MARINE BIOLOGICAL LABORATORY

WOODS HOLE, MASSACHUSETTS

A REVIEW OF STATISTICS

PART 2: STATISTICS

EMERY N. BROWN

NEUROSCIENCE STATISTICS RESEARCH LABORATORY
DEPARTMENT OF ANESTHESIA AND CRITICAL CARE
MASSACHUSETTS GENERAL HOSPITAL

DIVISION OF HEALTH SCIENCES AND TECHNOLOGY
HARVARD MEDICAL SCHOOL / MIT

AUGUST 20, 2001

STATISTICS

A. THE STATISTICAL PARADIGM

B. DATA REDUCTION PRINCIPLES

C. ESTIMATION THEORY

D. [HYPOTHESIS TESTING]

E. CONFIDENCE INTERVALS

Definition. A family of pdf 's and pmf 's is called an **exponential family** if it can be expressed as

$$f(x\mid\theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\},$$

where $h(x) > 0,\ \ t_1(x), \ldots, t_k(x)$ are real-valued functions of x, not depending on θ. $c(\theta) \geq 0$ and $w_1(\theta), \ldots, w_k(\theta)$ are real-valued functions of θ, not depending on x.

This family will play a central role in our discussions. The binomial, Poisson, exponential, gamma and Gaussian probability models are members of the exponential family.

Binomial Random Variable

$$f(x\mid p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \binom{n}{x}(1-p)^n \left(\frac{p}{1-p}\right)^x$$

$$\binom{n}{x}(1-p)^n \exp\left(\log\left(\frac{p}{1-p}\right)x\right).$$

Take $h(x) = \binom{n}{x},\ \ c(p) = (1-p)^n\, w_1(p) = \log\left(\frac{p}{1-p}\right),\ \ t_1(x) = x.$ Hence the binomial model belongs to the exponential family.

Gaussian Random Variable

$$f(x\mid\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right\}.$$

Take $h(x) = 1, \quad c(\mu,\sigma^2)(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}, \quad w_1(\mu,\sigma^2) = \frac{1}{\sigma^2} \quad w_2(\mu,\sigma^2) = \frac{\mu}{\sigma^2} \quad t_1(x) = -\frac{x^2}{2}, \quad t_2(x) = x.$

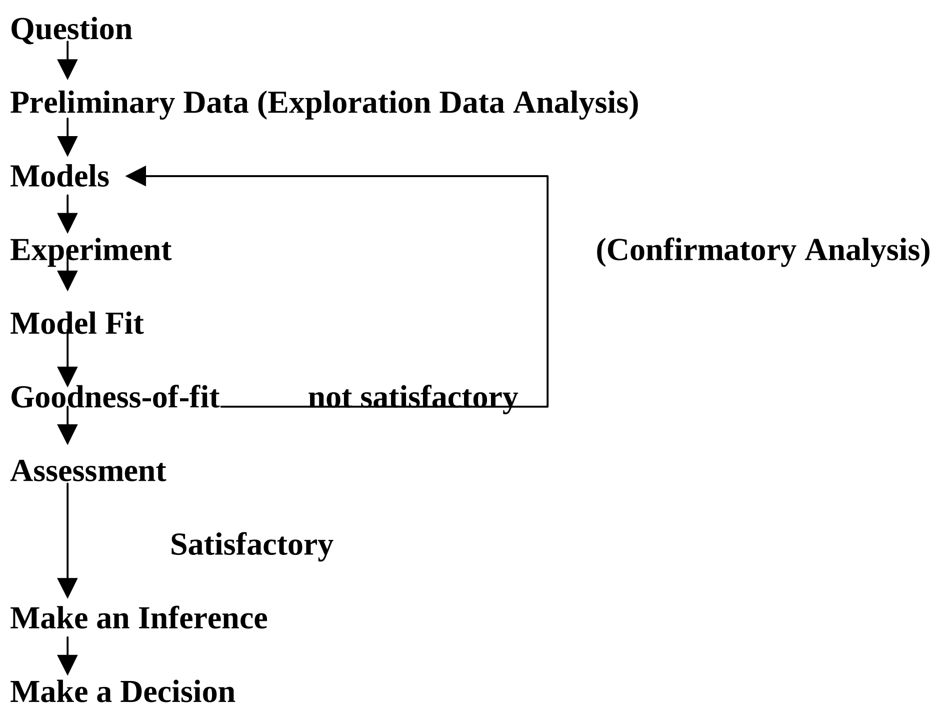
The Gaussian is in the exponential family.

Exercise: Is the inverse Gaussian probability model in the exponential family?

II. STATISTICS

The science of making decisions under uncertainty using mathematical models derived from probability theory.

A. THE STATISTICAL PARADIGM (Box, Tukey)



Example: **Neuroninhib**

Question: **Does neuroinhibin (a new GABA – a agonist, potential new general anesthetic) decrease neuronal spiking activity in neurons in constant conditions in isolated cultures?**

↓

Preliminary Data: **Recorded spiking activity from individual neurons in culture**

↓

Preliminary Model: **1. Poisson process with constant** λ
2. Gamma model with parameters α, λ

↓

Experiments: **Several days of recording with and without neuroinhibin**

↓

Model Fit: **1. Poor fit**
2. Better fit

↓

Goodness-of-Fit Assessment: Satisfactory **Not satisfactory** (→ back to Preliminary Model)

↓

Make an Inference: **When neuroinhibin was applied to neurons in culture, there was a 60% (statistically significant) decrease in mean spiking activity**

↓

Make a Decision: **Neuroinhibin is a potential inhibitor of neural spiking activity**

A. Data Reduction Principles

Notation

Observations: $x_1, \ldots, x_n = x$.

Probability Model: $f(x_k \mid \theta) \quad k = 1, \ldots, n \quad f(x \mid \theta) = \prod_{k=1}^{n} f(x_k \mid \theta)$. The parameters of the probability are denoted by θ. Let $T(x) =$ an arbitrary function of the data.

Definition: A statistic is any function of a set of data.

1. Sufficient Statistics

Definition: A statistic $T(x)$ is a sufficient statistic for θ if the conditional distribution of the sample x given the value of $T(x)$ does not depend on θ.

This statement says that once the statistic is computed, it summarizes all the information in the data sample about the parameter. To find a sufficient statistic we can use the Factorization Theorem.

Factorization Theorem: Let $f(x \mid \theta)$ be the joint *pdf* or *pmf* of a sample x. A statistic $T(x)$ is sufficient for if and only if these exist functions $g(t \mid \theta)$ and $h(x)$ such that for all sample points x and all parameter points θ,

$$f(x \mid \theta) = g(T(x) \mid \theta) h(x).$$

The dimension of the sufficient statistics equals the dimension of θ.

Example: Let $x_1, \ldots, x_n$ be sample from a Poisson distribution with parameter λ.

$$f(x \mid \lambda) = \prod_{k=1}^{n} f(x_k \mid \lambda) = \prod_{k=1}^{n} \frac{\lambda^{x_k} e^{-\lambda}}{x_k!} = \exp(\log \lambda \sum_{k=1}^{n} x_k - n\lambda) \prod_{k=1}^{n} (x_k!)^{-1}.$$

Take $g(T(x) \mid \lambda) = \exp(\log \lambda \sum_{k=1}^{n} x_k - n\lambda)$ and $h(x) = \prod_{k=1}^{n} (x_k!)^{-1}$ and we conclude that the sum of the observations (sample mean) is the sufficient statistic for estimating λ.

Example: $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$ with μ and σ^2 unknown

$$f(x \mid \mu, \sigma^2) = \prod_{k=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right\}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -(n(t_1 - \mu)^2 + (n-1)t_2) / 2\sigma^2 \right\}$$

$$= g(t_1, t_2 \mid \mu, \sigma^2) h(x),$$

where $t_1 = \overline{x}$ and $t_2 = \sum_{k=1}^{n} (x_k - \overline{x})^2 / (n-1),$ where $h(x) = 1.$

If $x_1, \ldots, x_n$ are *iid* observation from a *pdf* or *pmf*, $f(x \mid \theta)$. Suppose $f(x \mid \theta)$ belongs to an exponential family given by

$$f(x \mid \theta) = h(x) c(\theta) \exp\left(\sum_{i=1}^{k} w_i(\theta) t_1(x) \right).$$

Then $T(x) = \left(\sum_{j=1}^{n} t_1(x_j), \ldots, \sum_{j=1}^{n} t_k(x_j) \right).$

2. Likelihood Principle

Definition: Let $f(x \mid \theta)$ denote the joint *pdf* or *pmf* of the sample $x = (x_1, \ldots, x_n)$. Then given $X = x$ is observed, the function of θ defined as

$$L(\theta \mid x) = f(x \mid \theta),$$

is called the likelihood function.

Likelihood Principle. If x and y are two samples points such that $L(\theta \mid x)$ is proportional to $L(\theta \mid y)$, that is, there exists a constant $c(x, y)$ such that $L(\theta \mid x) = c(x, y) L(\theta \mid y)$ for all θ then the conclusions drawn from x and y should be identical.

Remarks: The Likelihood Principle states how the likelihood should be used as a data reduction device. Likelihoods that are proportional contain the same information. It depends critically on the specification of a parametric model. Hence it requires diagnostics. Information comes only from the current data sample and prior knowledge may not be "formally" used in the estimation and inference process.

ESTIMATION THEORY

Definition: An estimator is any function of the data sample used to determine a parameter. As estimate is the estimator evaluated for a given data sample.

1. Method of Moment

Given $(x_1, \ldots, x_n)$ a sample from a pdf or pmf $f(x \mid \theta_1, \ldots, \theta_k)$. The method of moments estimate is obtained by equation the first k moments to their sample values.

Example: Gaussian Random Sample

$x_1, \ldots, x_n \sim N(\mu, \sigma^2)$ and μ and σ^2 are unknown

$$m_1 = \bar{x} \qquad m_2 = n^{-1}\sum_{i=1}^{n} x_i^2$$

$$\mu_1 = \mu \qquad \mu_2 = \sigma^2 + \mu^2$$

The method of moments estimates are

$$\tilde{\mu} = \bar{x} \qquad \tilde{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 .$$

Example: Gamma Random Sample

$x_1, \ldots, x_n \sim \Gamma(\alpha, \lambda)$ α and λ are unknown

$$\mu_1 = \frac{\alpha}{\lambda} \qquad \mu_2 = \frac{\alpha}{\lambda^2}$$

$$\underset{\sim}{x} = \frac{\alpha}{\lambda} \qquad \tilde{\sigma}^2 = \frac{\alpha}{\lambda^2}$$

$$\tilde{\alpha} = \frac{\bar{x}^2}{\tilde{\sigma}^2} \qquad \tilde{\lambda} = \frac{\bar{x}}{\tilde{\sigma}^2} .$$

2. Maximum Likelihood Estimators

Given $x_1, \ldots, x_n$ iid sample from a pdf or pmf .

$f(x \mid \theta_1, \ldots, \theta_k)$, is the likelihood function

$$L(\theta \mid \underset{\sim}{x}) = \prod_{k=1}^{n} f(x_k \mid \theta) .$$

For each sample point $\underset{\sim}{x}$ let $\hat{\theta}(\underset{\sim}{x})$ be a parameter value of which $L(\theta \mid x)$ attains a maximum as a function of θ for fixed x . $\hat{\theta}(x)$ is a maximum likelihood estimator of the parameter θ .

Problems: Finding a global maximum

numerical sensitivity

If $L(\theta \mid \underset{\sim}{x})$ is differentiable we can consider $\frac{\partial L}{\partial \theta} = 0$ and check the conditions on $\partial^2 L / \partial^2 \theta$. Usually easier to work with log L instead of L .

Example: Gaussian Random Sample

$$x_1, \ldots, x_n \sim N(\mu, \sigma^2) ,$$

$\bar{x}$ is the ML estimate of μ .

$\tilde{\sigma}^2$ is the ML estimate of σ^2 .

This is straightforward to show by differentiating the Gaussian log likelihood equating the set of 1st partials to zero and solving for μ and σ^2. A check of second derivatives shows that this point is an interior maximum.

Example: Gamma Random Sample

$$x_1, \dots, x_n \sim \Gamma(\alpha, \lambda) \, .$$

If α is known then

$$f(x_1, \dots, x_n \mid \alpha, \lambda) = \prod_{k=1}^{n} \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x_k^{\alpha-1} e^{-\lambda x_k}$$

$$\log f(x \mid \alpha, \lambda) = n\Gamma(\alpha) + n\alpha \log \lambda + (\alpha - 1)\sum_{k=1}^{n} \log(x_k) - \lambda \sum_{k=1}^{n} x_k$$

$$\frac{\partial \log f(x \mid \alpha, \lambda)}{\partial \alpha} = \frac{n\alpha}{\lambda} - \sum_{k=1}^{n} x_k$$

$$0 = \frac{\alpha}{\lambda} - \bar{x}$$

$$\hat{\lambda} = \frac{\alpha}{\bar{x}}.$$

If $\alpha = 1$ we have $\hat{\lambda} = \bar{x}^{-1}$ is ML for exponential model. If α is unknown then there is no closed form solution for either α and λ. The estimates must be found numerically. Good starting values can be obtained from the method of moments estimates. Notice that the sufficient statistics for α and λ are $\sum_{k=1}^{n} \log(x_k)$ and $\sum_{k=1}^{n} x_k$. This shows that the simple method of moments estimates are not efficient.

Exercise: Inverse Gaussian Distribution. If $x_1, \ldots, x_n$ is a random sample from an inverse Gaussian distribution with parameters α and λ. Recall that the mean is α and the variance is α^3 / λ. Find the ML estimate is it the same as the method of moments estimate. The pdf is

$$f(x \mid \alpha, \lambda) = \left(\frac{\lambda}{2\pi x^3} \right)^{\frac{1}{2}} \exp\left\{ -\frac{\lambda (x-\alpha)^2}{2x\alpha^2} \right\}.$$

Answer: The ML estimate is $\hat{\alpha} = n^{-1} \sum_{i=1}^{n} x_i$, $\hat{\lambda}^{-1} = n^{-1} \sum_{i=1}^{n} (\frac{1}{x_i} - \frac{1}{\hat{\alpha}})$. What is the method of moments estimate.

Bayes' Estimator

$f(x \mid \theta)$ sample probability density

Assuming θ is a random variable the

$f(\theta)$ prior probability density

$$f(\theta \mid x) = \frac{f(\theta) f(x \mid \theta)}{\int f(\theta) f(x \mid \theta) d\theta}$$ posterior density.

θ has all its uncertainty characterized by its posterior density. We can take a summary statistic (function) from $f(\theta \mid x)$ to be a point estimate of θ. [Get the interval estimate first].

Example: $x_1, \ldots, x_n \sim B(n, p)$, $p \sim \beta eta(\alpha, \beta)$ Find the posterior distribution of p. Take $y = \sum_{k=1}^{n} x_k$

$$f(y \mid x) \propto f(p) f(x \mid p)$$

$$\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \binom{n}{y} p^{y}(1-p)^{n-y}$$

$$\propto p^{\alpha-1}(1-p)^{\beta-1} p^{y}(1-p)^{n-y}$$

$$\propto p^{\alpha+y-1}(1-p)^{n-y+\beta-1}.$$

Hence by the definition of a $\beta \quad pdf$

$$f(p \mid x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(n-y+\beta)} p^{\alpha+y-1}(1-p)^{n-y+\beta-1}.$$

The β distribution is a conjugate prior distribution for the binomial.

Example: Gaussian Likelihood and Gaussian Prior

$$x \sim N(\theta, \sigma^2)$$

$$\theta \sim N(\mu, \tau^2).$$

We want to find the posterior distribution of θ. The posterior is Gaussian (why?) and given as

$$E[\theta \mid x] = \frac{\tau^2}{\tau^2+\sigma^2} x + \frac{\sigma^2}{\sigma^2+\tau^2} \mu$$

$$v[\theta \mid x] = \frac{\sigma^2 \tau^2}{\sigma^2+\tau^2}.$$

Now let $\theta_t = E[\theta \mid x]$ and $\mu = \theta_{t-1}$

$$\theta_t = \theta_{t-1} + \frac{\tau^2}{\tau^2+\sigma^2}(x - \theta_{t-1}),$$

we obtain the simplest version of the Kalman filter. This is a part of departure for a recursive-decoding scheme for neural spike trains.

Evaluating Estimators

Let $w(x)$ be an estimator of θ then we can suggest several criteria for evaluating how well it performs.

Criteria for Evaluation

1. Mean-Squared Error $E_\theta[w(x)-\theta]^2$
2. Unbiasedness $E_\theta(w(x))=\theta$
3. Consistency $w(x)\to\theta$ as $n\to\infty$
4. Efficiency Achieves a minimum variance (Cramer-Rao Lower Bound)

Cramer-Rao Lower Bound. Given $x_1,\ldots,x_n$ be a sample from of *pdf* $f(x|\theta)$, $w(x)$ is an estimator and $E[\theta(w(x))]$ is a differentiable function of θ. Suppose also that

$$\frac{d}{d\theta}\int h(x)f(x|\theta)=\int h(x)\frac{df(x|\theta)}{d\theta}dx\,,$$

for $\forall h(x)$ with $E_\theta\,|(x)|<\infty$. Then

$$\operatorname{var}(w(x))\geq\frac{(\frac{dE_\theta w(x)}{d\theta})^2}{E_\theta((\frac{\partial\log f(x|\theta)}{\partial\theta})^2}\,.$$

CRLB give the lowest bound on the variance of an estimate. And if the estimate is unbiased, then the numerator is 1 and the denominator is the Fisher information. If θ is a p x 1 vector then the Fisher information is a pxp matrix given by

$$I(\theta)=E_\theta[(\frac{\partial\log f(x|\theta)}{\partial\theta})^T\frac{\partial\log f(x|\theta)}{\partial\theta}]=-E_\theta[\frac{\partial^2\log f(x|\theta)}{\partial\theta\partial\theta'}]$$

We will make extensive use of the Fisher information to derive confidence intervals for our estimates.

Factoids about Maximum Likelihood Estimates

1. ML Estimates are generally biased.
2. ML Estimates are consistent, they are hence asymptotically unbiased.
3. ML Estimates are asymptotically efficient.
4. The variance of ML estimate may be approximated by the inverse Fisher information matrix

$$\left[E\left(\frac{\partial \log f}{\partial \theta}\right)^2\right]^{-1} = -E\left[\frac{\partial^2 \log f}{\partial \theta^2}\right]^{-1}$$

5. If $\hat{\theta}$ is the ML estimate of θ then $h(\hat{\theta})$ is the ML Estimate of $h(\theta)$.

Exercise: Gaussian Random Sample

$x_1, \ldots, x_x \sim N(\mu, \sigma^2)$ with σ^2 unknown. The ML estimate of μ is $\bar{x}$. Use the definition of the Fisher information to show that $Var(\bar{x}) = \frac{\sigma^2}{n}$.

Exercise: Is a Bayes' estimator unbiased? How can Zhang et al. 1998 use the CRLB to evaluate the optimality of a Bayes' estimator?

Exercise: Gaussian Random Sample Revisited

$x_1, \ldots, x_n \sim N(\mu, \sigma^2)$ μ and σ^2 are unknown.

The ML estimate of σ^2 is $\hat{\sigma}^2 = \frac{\sum_{k=1}^{N}(x_i - \bar{x})^2}{N}$. Is $\hat{\sigma}^2$ an unbiased estimate?

D. HYPOTHESIS TESTING (To Appear)

E. CONFIDENCE INTERVALS

Definition (Classic): A $1-\alpha$ confidence interval for θ has probability $1-\alpha$ of covering the true parameter. There are several methods of construction.

1. Inverting a Test

2. Finding a Pivot

3. ML Approximation

4. CLT and Slutsky's Theorem

Example: Gaussian Random Sample (Pivot)

Definition: $Q(x,\theta)$ is a pivot if the distribution of $Q(x,\theta)$ is independent of all parameters, i.e. $x \sim F(x|\theta)$ has the same distribution for all θ. $x_1,\ldots,x_n \sim N(\mu,\sigma^2)$. We want a CI for μ given σ^2 is known $\bar{x}$ is the ML estimate of μ.

$$\Pr\left(\left|\frac{n^{\frac{1}{2}}(\bar{x}-\mu)}{\sigma}\right| < c\right) = 1-\alpha$$

$$\Pr(\bar{x} - n^{-\frac{1}{2}}c\sigma < \mu < \bar{x} + n^{-\frac{1}{2}}c\sigma = 1-\alpha.$$

$n^{\frac{1}{2}}\frac{(\bar{x}-\mu)}{\sigma}$ is a pivot. Pick $c = z_{\alpha/2}$ then we have a $1-\alpha$ CI since $\bar{x} \approx N(\mu, \frac{\sigma^2}{n})$

Example: Maximum Likelihood

$\hat{\theta}_{ML} \sim N(\hat{\theta},[-I_N(\hat{\theta})]^{-1})$ where $I_N(\theta)$ is the Fisher information.

By Taylor series approximation

$$h(\hat{\theta}_{ML}) \sim N(h(\theta),(h'(\theta))^2[-I_N(\theta)]^{-1}).$$

Therefore an approximate $1-\alpha$ CI is

$$h(\hat{\theta}) \pm z_{\alpha/2}[h'(\hat{\theta})^2[-I_N(\hat{\theta})]^{-1}].$$

Bayes' Credibility Interval. A Bayesian credibility interval evaluates the probable values of the parameter relative to the posterior density. The parameter is a random variable and not a fixed quantity.

REFERENCE

Casella G, Berger RL (1990). *Statistical Inference*. Duxbury Press: Belmont, CA .