# Statistical Data Mining and Medical Signal Detection

## Lecture Two: Medical Signal Detection and Bayesian Methodology

June 15, 2011

Motoya Machida (mmachida@tntech.edu)

# Data Mining with R

R is a language and environment for statistical computing and graphics, and is available as free software. The system runs on Windows, Linux, and Mac, and can be downloaded from

`http://cran.r-project.org`

Each "command" is executed in an interactive manner, known as "interpretor," and is requested in a form of "function," for example, it is a function `demo(graphics)` to show a demonstration of R graphics.

```
> demo("graphics")
```

R has a very strong data visualization capability along with flexible database interfaces, critical for data mining.

# Code Execution in R

The function "dgamma(x)" returns the gamma density at the value $x$, and the graph of the density is obtained by

```
> x = seq(0, 5, by=0.001)
> density = dgamma(x, shape=2, rate=2)
> plot(x, density, type="l", main="Gamma Density")
```

We can create a new function gmixture() which returns the density value of the mixture of two gamma density.

```
> gmixture = function(x,alpha1,beta1,alpha2,beta2,p){
+ p*dgamma(x,alpha1,beta1) + (1-p)*dgamma(x,alpha2,beta2)
+ }
> plot(x, gmixture(x, 0.2, 0.1, 2, 4, 1/3), type="l")
```

$\alpha$ and $\beta$ correspond respectively to the shape and the rate parameter of gamma density.

# Scripts and Working Directory in R

A script file (usually with extension ".r" or ".R") can be prepared as an external file, and executed in R with the command

```
> source("[script filename]")
```

Your external file must be found in the working directory to be recognized from R. You can always change the working directory from R via [File]→[Change dir...]. Alternatively you can set the working directory by

```
> setwd("[pathname]")
```

# Adverse Event Reporting System (AERS)

To improve drug safety it is important to develop methodologies detecting adverse drug events using postmarketing drug surveillance data. A strong association of drug and adverse reaction forms the basis for further epidemiological study and consequently for regulatory actions.

Adverse event reporting system (AERS) is created to monitor a possible causal relationship between drug and event. The database contains the information about the entire list $D$ of medical products and $R$ of medical terms of adverse reaction. Each event is reported exactly once alone with the list of medical products prescribed to a patient at the point of event, say "Rosinex & Ganclex," and the list of medical terms describing adverse events, say "Nausea."

# AERS Data Set

Each event is reported with the list $A$ of drug names and $B$ of adverse reactions, and the entire data are summarized in terms of the frequency of such events, denoted by $N_{A,B}$. Note that a pair $(A, B)$ is not necessarily labeled as a valid association of model. For example, an adverse event of "Rosinex & Ganclex" and "Nausea" is reported, but the drug combination of Rosinex and Ganclex may not be necessarily the cause of nausea.

`http://math.tntech.edu/machida/AERS.zip`

A data set contains a total of $1,090$ drugs and $1,072$ medical terms which were reported at least 50 individual incidents from January 2004 to March 2005.

# Event Frequencies and Report Counts

Let $D$ be the collection of drug names, and $R$ be the collection of medical terms for adverse reaction. A drug-adverse reaction relationship is formed as an edge of a bipartite graph $G$ between $D$ and $R$. Report counts can be obtained from the frequency $N_{A,B}$ of event. Here for a pair $(i, j)$ of individual drug and AE we can define the cell count

$$C_{ij} = \sum\{N_{A,B} : i \in A, j \in B\}$$

Note that the total number of reporting events is substantially smaller than the sum of all the cell counts of the contingency table.

```
> load("AERS.save")
> AERS[1:10,1:10]
> summary(as.numeric(AERS))
> hist(AERS[AERS < 50], breaks=seq(0,50,by=1), col="blue")
```

# Marginal Count and Baseline

- $C_{i\cdot} = \sum_j C_{ij}$ (marginal count for the $i$-th drug)
- $C_{\cdot j} = \sum_i C_{ij}$ (marginal count for the $j$-th AE)
- $C_{\cdot\cdot} = \sum_i C_{i\cdot} = \sum_j C_{\cdot j} = \sum_{(i,j)} C_{ij}$

where the summation $\sum_i$ indicates the sum over the index $i$. Then we can define the *baseline* by

$$E_{ij} = C_{i\cdot} C_{\cdot j} / C_{\cdot\cdot}$$

```
> load("DRUG.save")
> DRUG[1:10,]
> load("REAC.save")
> REAC[1:10,]
```

# Hierarchical Multinomial Model

By $\mathcal{L}(X|Y=y)$ we denote the law of probability of a random variable $X$ conditionally given $Y=y$ for another random variable $Y$, and by $B(n,p)$ the binomial distribution with parameter $(n,p)$. Then the hierarchical binomial model of report count is formed by a series of binomial distributions.

1. $\mathcal{L}(C_{.j}|C_{..}=n) \sim B(n,p_{.j})$ for the list $B$ of adverse reactions.
2. $\mathcal{L}(C_{ij}|C_{i.}=n_i) \sim B(n_i,p_{ij})$ for the pair $(i,j)$ of valid association

Then we can define the *relative report rate* by

$$\lambda_{ij} = p_{ij}/p_{.j}$$

# Poisson Distribution Model

The hierarchical model of binomial distribution is conditioned upon $C_{\cdot\cdot} = n$ and $C_{i\cdot} = n_{i\cdot}$, and related to the unconditional model $C_{ij} \sim \mathrm{Poisson}(\mu_{ij})$ via $p_{ij} = \mu_{ij}/\mu_{i\cdot}$ and $p_{\cdot j} = \mu_{\cdot j}/\mu_{\cdot\cdot}$ where

$$\mu_{i\cdot} = \sum_{j} \mu_{ij}; \quad \mu_{\cdot j} = \sum_{i} \mu_{ij}; \quad \mu_{\cdot\cdot} = \sum_{(i,j)} \mu_{ij}$$

It is also used to derive the model $\mathcal{L}(C_{i\cdot} | C_{\cdot\cdot} = n) \sim B(n, p_{i\cdot})$ of conditional distribution with $p_{i\cdot} = \mu_{i\cdot}/\mu_{\cdot\cdot}$

# Parameters of Interest

Hierarchical multinomial or Poisson distribution model can achieve the interpretability of relative report rates (RRR's). Assume that each report count $C$ is a draw from a Poisson distribution with unknown mean $\mu$. Here the values

$$\lambda = \mu / E$$

is treated as parameters, drawn from a common prior distribution.

```
> load("RRrank.save")
> RR.rank[1:10,]
> summary(RR.rank$LAMBDA)
> hist(RR.rank$LAMBDA[RR.rank$LAMBDA < 500], col=1)
> source("lambda.r")
> load("RR.save")
> plot.lambda(RR,grid.size=100,hue.size=64,hue.low=0.18)
```

# What is Bayes?

*Data*

$$X_1, \ldots, X_n$$

are regarded as independent and identically distributed (iid) random variables governed by an underlying probability density function $f(x;\theta)$. A value $\theta$ represents the characteristics of this underlying distribution, and is called a *parameter*. A *point estimate* is a "best guess" for the true value $\theta$. Bayesian uses the concept of prior belief about the parameter $\theta$ of interest. Then the uncertainty of $\theta$ changes according to the data

$$\mathbf{x} = (x_1, \ldots, x_n).$$

Here Bayesian interprets $\theta$ as a random variable, and the prior belief is given in the form of probability density $\pi(\theta)$ of $\theta$. The objective of Bayesian model is to investigate the posterior density $\pi(\theta \mid \mathbf{x})$ of $\theta$.

# Bayesian Model

Let $f(\mathbf{x};\theta)$ be a density function with parameter $\theta \in \Omega$. In a Bayesian model the parameter space $\Omega$ has a distribution $\pi(\theta)$, called a *prior distribution*. Furthermore, $f(\mathbf{x};\theta)$ is viewed as the conditional distribution of $\mathbf{X}$ given $\theta$. By the Bayes' rule the conditional density $\pi(\theta \mid \mathbf{x})$ can be derived from

$$\pi(\theta \mid \mathbf{x}) = \begin{cases} \pi(\theta) f(\mathbf{x};\theta) \Big/ \displaystyle\sum_{\theta \in \Omega} \pi(\theta) f(\mathbf{x};\theta) & \text{if } \Omega \text{ is discrete;} \\ \pi(\theta) f(\mathbf{x};\theta) \Big/ \displaystyle\int_{\Omega} \pi(\theta) f(\mathbf{x};\theta)\, d\theta & \text{if } \Omega \text{ is continuous.} \end{cases}$$

# Conjugate Family of Distributions

The distribution $\pi(\theta \mid \mathbf{x})$ is called the *posterior distribution*. Whether $\Omega$ is discrete or continuous, the posterior distribution $\pi(\theta \mid \mathbf{x})$ is "proportional" to $\pi(\theta)f(\mathbf{x};\theta)$ up to the constant. Thus, we write

$$\pi(\theta \mid \mathbf{x}) \propto \pi(\theta)f(\mathbf{x};\theta).$$

It is often the case that both the prior density function $\pi(\theta)$ and the posterior density function $\pi(\theta \mid \mathbf{x})$ belong to the same family of density function $\pi(\theta;\eta)$ with parameter $\eta$. Then $\pi(\theta;\eta)$ is called *conjugate* to $f(\mathbf{x};\theta)$.

# Exponential Conjugate Family

Suppose that the pdf has the form

$$f(\mathbf{x};\theta) = \exp\left[nc_0(\theta) + \sum_{j=1}^{m} c_j(\theta)k_j(\mathbf{x}) + h(\mathbf{x})\right],$$

and that a prior distribution is given by

$$\pi(\theta;\eta_0,\eta_1,\ldots,\eta_m) \propto \exp\left[c_0(\theta)\eta_0 + \sum_{j=1}^{m} c_j(\theta)\eta_j\right].$$

Then we obtain the posterior density

$$\pi(\theta \mid \mathbf{x}) = \pi(\theta;\eta_0 + n, \eta_1 + k_1(\mathbf{x}), \ldots, \eta_m + k_m(\mathbf{x})).$$

Thus, the family of $\pi(\theta;\eta_0,\eta_1,\ldots,\eta_m)$ is conjugate to $f(\mathbf{x};\theta)$, and the parameter $(\eta_0,\eta_1,\ldots,\eta_m)$ of prior distribution is called the *hyperparameter*.

# Prior Density for RRR's

The prior distribution of relative report rate (RRR) is assumed to be the mixture of two gamma distributions

$$\pi(\lambda) = pg(\lambda; \alpha_1, \beta_1) + (1 - p)g(\lambda; \alpha_2, \beta_2)$$

where $\alpha_1, \beta_1, \alpha_2, \beta_2, p$ are hyperparameters, and $g(\lambda; \alpha, \beta) = \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda} / \Gamma(\alpha)$ is a gamma density function. The determination of hyperparameters may not be so important; $\alpha_1 = 0.2, \beta_1 = 0.1, \alpha_2 = 2, \beta_2 = 4, p = 1/3$ can be a good choice, suggested by the fact that the majority of RRR's are well below one.

# Posterior Density

If the prior density $\pi(\lambda)$ and the baseline $E$ are known then the posterior density $\pi(\lambda \,|\, n)$ given the report count $C = n$ is proportional to $\phi(\lambda; n, E) = e^{-E\lambda+E}\lambda^n\pi(\lambda)$. Here we can observe that

$$\Phi(n, E) = \int_0^\infty \rho(\lambda; n)\, d\lambda = \pi(n) \Bigg/ \left( e^{-E}\frac{E^n}{n!} \right)$$

where $\pi(n) = p\, f(n; \alpha_1, \beta_1, E) + (1-p)\, f(n; \alpha_2, \beta_2, E)$ with

$$f(n; \alpha, \beta, E) = (1 + \beta/E)^{-n}(1 + E/\beta)^{-\alpha}\Gamma(\alpha + n)/\Gamma(\alpha)n!$$

Here $\pi(n)$ represents the marginal probability distribution of the report count $C = n$.

# Gamma-Poisson Shrinker

The posterior probability $q$ of the first component can be derived as

$$q = \frac{p\, f(n; \alpha_1, \beta_1, E)}{\pi(n)}$$

Then the posterior distribution of $\lambda$ given $C = n$ is expressed as the mixture

$$f(\lambda | n, E) = \pi(\lambda; \alpha_1 + n, \beta_1 + E, \alpha_2 + n, \beta_2 + E, q)$$

```
> load("EBGMrank.save")
> EBGM.rank[1:10,]
> source("lambda.r")
> load("EBGM.save")
> plot.lambda(EBGM,grid.size=100,hue.size=64,hue.low=0.18)
```