Biostatistics for Dummies

Biomedical Computing Cross-Training Seminar

October 18th, 2002
What is “Biostatistics”? 

- Techniques
  - Mathematics
  - Statistics
  - Computing

- Data
  - Medicine
  - Biology
What is “Biostatistics”? 

Knowledge of biological process

Biological data

prey

predators

population density

time
Common Applications
(Medical and otherwise)

- Clinical medicine
- Epidemiologic studies
- Biological laboratory research
- Biological field research
- Genetics
- Environmental health
- Health services
- Ecology
- Fisheries
- Wildlife biology
- Agriculture
- Forestry
Biostatisticians Work

- Develop study design
- Conduct analysis
- Oversee and regulate
- Determine policy
- Training researchers
- Development of new methods
Some Statistics on Biostatistics

- Internet search (Google)
  > 210,000 hits
- > 50 Graduate Programs in U.S.

Too much to cover in one hour!
Center Focus

- MSU strengths
  - Computational simulation in physical sciences
  - Environmental health sciences
- Bioinformatics is crowded

- Computational simulation in environmental health sciences
  - Build on appreciable MSU strength
  - Establish ourselves
    - Unique capability
    - Particular appeal to NIEHS
Focus of Seminar

- Statistical methodologies
  - Computational simulation in environmental health sciences
  - Can be classified as “biostatistics”
- Stochastic modeling
  - Time series
  - Spatial statistics*
The Application

Of interest
- Cancer incidence rate
- Pesticide exposure

Of concern
- Age
- Gender
- Race
- Socioeconomic status

Objectives
- Suitably adjust cancer incidence rate
- Determine if relationship exists
- Develop model
  - Explain relationship
  - Estimate cancer rate
  - Predict cancer rate
The Data

  - Number of acres harvested
  - Type of crop

  - Tumor type
  - Age
  - Gender
  - Race
  - County of residence
  - Cancer morbidity
    - Crude incidence/100,000
    - Age adjusted incidence/100,000
Why (Bio)statistics?

Statistics
- Science of uncertainty
- Model order from disorder

Disorder exists
- Large scale rational explanation
- Smaller scale residual uncertainty

Entropy
$$E(\lambda^{(1)}; \mu^{(1)}) \geq E(\lambda^{(0)}; \mu^{(0)})$$

Chaos
$$f^{(k)}(x) = x_k = f(x_{k-1}) = f(f(\cdots f(x)))$$

Deterministic equation
Randomness
(Bio)statistical Data

- Independent identically distributed
- Inhomogeneous data
- Dependent data
  - Time series
  - Spatial statistics
Time Series

- Identically distributed
- Time dependent
- Equally spaced

Randomness
Objectives in Time Series

- Graphical description
  - Time plots
  - Correlation plots
  - Spectral plots
- Modeling
- Inference
- Prediction
**Time Series Models**

**Linear Models**

\[ X(t) = \sum_{j=0}^{\infty} \phi_j \varepsilon(t-j) \]

- \( \varepsilon(t) \sim \text{i.i.d} \)
- Zero mean
- Finite variance
- \( \phi_j \) square summable

**Covariance stationary**

- Constant mean
- Constant variance
- Covariance function of distance in time

\[ \gamma(\nu) = \mathbb{E}[X(t)X(t+\nu)], \quad \nu = 0, 1, 2, \ldots \]
Nonlinear Time Series

- Amplitude-frequency dependence
- Jump phenomenon
- Harmonics
- Synchronization
- Limit cycles

Biomedical applications
- Respiration
- Lupus-erythematositis
- Urinary introgen excretion
- Neural science
- Human pupillary system
Some Nonlinear Models

- Nonlinear AR
  - Additive noise
- Threshold
  - AR
  - Smoothed TAR
  - Markov chain driven
  - Fractals
- Amplitude-dependent exponential AR
- Bilinear
- AR with conditional heteroscedasticity
- Functional coefficient AR
A Threshold Model

\[
\begin{cases}
\begin{bmatrix}
0 & 0 \\
0 & 0.25
\end{bmatrix}
\begin{bmatrix}
x_{t-1} \\
y_{t-1}
\end{bmatrix}
+ \begin{bmatrix}
0 \\
0
\end{bmatrix}
\quad \text{with probability 0.01} \\
\begin{bmatrix}
0.85 & 0.04 \\
-0.01 & 0.85
\end{bmatrix}
\begin{bmatrix}
x_{t-1} \\
y_{t-1}
\end{bmatrix}
+ \begin{bmatrix}
0 \\
1.6
\end{bmatrix}
\quad \text{with probability 0.85} \\
\begin{bmatrix}
0.20 & -0.26 \\
0.26 & 0.22
\end{bmatrix}
\begin{bmatrix}
x_{t-1} \\
y_{t-1}
\end{bmatrix}
+ \begin{bmatrix}
0 \\
0.8
\end{bmatrix}
\quad \text{with probability 0.07} \\
\begin{bmatrix}
-0.15 & 0.28 \\
0.26 & 0.24
\end{bmatrix}
\begin{bmatrix}
x_{t-1} \\
y_{t-1}
\end{bmatrix}
+ \begin{bmatrix}
0 \\
1
\end{bmatrix}
\quad \text{with probability 0.07}
\end{cases}
\]

with initial conditions \[
\begin{bmatrix}
x_0 \\
y_0
\end{bmatrix} = \begin{bmatrix}
1 \\
1
\end{bmatrix}
\]
A Threshold Model
Describing Correlation

- **Autocorrelation**
  - AR: exponential decay
  - MA: $0$ past $q$

- **Partial autocorrelation**
  - AR: $0$ past $p$
  - MA: exponential decay

- **Cross-correlation**

- Relationship to spectral density
Spatial Statistics*

- Data components
  - Spatial locations
    \[ S = \{s_1, s_2, \ldots, s_n\} \]
  - Observable variable
    \[ \{Z(s_1), Z(s_2), \ldots, Z(s_n)\} \]
  - \[ s \subseteq D \subseteq R^k \]
- Correlation

- Data structures
  - Geostatistical
  - Lattice
  - Point patterns or marked spatial point processes
  - Objects
- Assumptions on \( Z \) and \( D \)
Biological Applications

- Geostatistics
  - Soil science
  - Public health
- Lattice
  - Remote sensing
  - Medical imaging
- Point patterns
  - Tumor growth rate
  - In vitro cell growth
Spatial Temporal Models

Combine time series with spatial data

Application

- Time element
  - Pesticide exposure → time → develop cancer

- Spatial element
  - Proximity to pesticide use