Biostatistics in Public Health

Abdelmonem A. Afifi, Ph.D.
Dean Emeritus and Professor of Biostatistics
UCLA School of Public Health
afifi@ucla.edu

What Will I Talk About?

- The role(s) of biostatistics in P.H.
- Tools available to the biostatistician.
- Example: bioinformatics.

Introduction

- The press frequently quotes scientific articles about:
  - Diet
  - The Environment
  - Medical care, etc.
- Effects are often small and vary greatly from person to person
- We need to be familiar with statistics to understand and evaluate conflicting claims

Public Health
What Is Public Health?

“Public Health is the science and art of protecting and improving the health of communities through education, promotion of healthy lifestyles, and research for disease and injury prevention.”

- Association of Schools of Public Health, 2010


- The mission of public health is defined as:
  “Assuring the conditions in which people can be healthy.”

The Functions of Public Health

- Assessment: Identify problems related to the public’s health, and measure their extent
- Policy Setting: Prioritize problems, find possible solutions, set regulations to achieve change, and predict effect on the population
- Assurance: Provide services as determined by policy, and monitor compliance

Evaluation is a theme that cuts across all these functions, i.e., how well are they performed?
Approach and Rationale

- In 1988 report: public health refers to the efforts of society, both government and others, to assure the population’s health.
- The 2002 report elaborates on the efforts of the other potential public health system actors.

Areas of Action and Change

- Adopt a population-level approach, including multiple determinants of health
- Strengthen the governmental public health infrastructure
- Build partnerships
- Develop systems of accountability
- Base policy and practice on evidence
- Enhance communication

Determinants of Population Health

- Employment and occupational
- Biology of disease
- Education
- Socioeconomic status
- Psychosocial factors
- Environment: natural and built
- Public health services
- Health care services
What is Biostatistics?

- **Statistics** is the art and science of making decisions in the face of uncertainty
- **Biostatistics** is statistics as applied to the life and health sciences

The Functions of Public Health

- **Assessment**: Identify problems related to the public’s health, and measure their extent
- **Policy Setting**: Prioritize problems, find possible solutions, set regulations to achieve change, and predict effect on the population
- **Assurance**: Provide services as determined by policy, and monitor compliance

*Evaluation is a theme that cuts across all these functions, i.e., how well are they performed?*

Role of the Biostatistician in Assessment

- decide which information to gather,
- find patterns in collected data, and
- make the best summary description of the population and associated problems

It may be necessary to

- design general surveys of the population needs,
- plan experiments to supplement these surveys, and
- assist scientists in estimating the extent of health problems and associated risk factors.
The Functions of Public Health

- **Assessment**: Identify problems related to the public’s health, and measure their extent
- **Policy Setting**: Prioritize problems, find possible solutions, set regulations to achieve change, and predict effect on the population
- **Assurance**: Provide services as determined by policy, and monitor compliance

*Evaluation is a theme that cuts across all these functions, i.e., how well are they performed?*

Role of the Biostatistician in Policy Setting

- develop mathematical tools to:
  - measure the problems,
  - prioritize the problems,
  - quantify associations of risk factors with disease,
  - predict the effect of policy changes, and
  - estimate costs, including monetary and undesirable side effects of preventive and curative measures.

Role of the Biostatistician in Assurance and Evaluation

- use sampling and estimation methods to study the factors related to compliance and outcome.
- decide if improvement is due to compliance or something else, how best to measure compliance, and how to increase the compliance level in the target population.
- take into account possible inaccuracy in responses and measurements, both intentional and unintentional.

*Survey instruments should be designed to make it possible to check for inaccuracies, and to correct for nonresponse and missing values*
Examples of Community Engagement

Infant Mortality: Genesee County, Michigan, USA (Pestronk, 2003)

Infant mortality rate*, Michigan, 1999

- All of Michigan 8.0
- Whites, Genesee County 7.4
- African Americans, Genesee County 23.5

*Deaths per 1,000 live births

Infant Mortality: Genesee County, Contd.

- Programs to Reduce Infant Deaths Effectively (PRIDE)
- Community-based public health used to identify contributing factors
- Three areas emerged as important underlying determinants of infant mortality

Statisticians help gather, analyze and interpret the data necessary for convincing the public and the policy makers.

Addressing Infant Mortality in Genesee County

MATERNAL-INFANT HEALTH

- Fostering Community Mobilization
- Enhancing the Perinatal System
- Reducing Racism
Tobacco Control Initiatives in the United States

- Government regulations to ban television advertising of tobacco in the 1970s.
- Public Health campaigns for smoking cessation increased.
- Stricter enforcement of under-age sales with expensive fines
- Smoking banned in most public places

Youth engagement in Boston: Sociedad Latina (CDC, 2010)

- Empower Latino youth and their families.
- Tobacco use a major community problem.
- In 2004, petitioned Boston City Council to pass laws that would:
  - increase merchant fees,
  - strengthen enforcement of laws, and
  - regulate the advertising of tobacco products.

Health Research Center: Northern Ghana (Tindana, 2011)

- Navrongo Health Research Center (NHRC)
  - Began as field site for a Vitamin A trial in 1989.
  - After successful trial, Ghana Ministry of Health commissioned site as a regional health research center.
  - Mandated to investigate health problems affecting northern Ghana to inform policy.
  - Runs demographic surveillance system.

Northern Ghana Health Research Center: Strategies

- Strategies used by research center for successful community engagement:
  1. Identifying and understanding structure of local authorities and decision-making.
  2. Using strategies familiar to community.
  3. Ensuring that tangible benefits flow from research to the community.
  4. Planning, flexibility, and funding.
- Results: Vitamin A supplementation, bed net use for malaria, delivery of family planning services.
Motor Cycle Helmets

- Since 1975, states started passing laws requiring helmet use
- 1992: a California state law required safety helmets meeting US Department of Transportation standards

Evaluation of Law

- The Southern California Injury Prevention Research Center conducted study to determine:
  - Change in helmet use with the 1992 helmet law, and
  - Impact of the law on crash fatalities and injuries

Results of Center Study

- Helmet use increased from about 50% in 1991 to more than 99% throughout 1992
- Statewide motorcycle crash fatalities decreased by 37.5%
- An estimated 92 to 122 fatalities were prevented
- The proportion of riders likely to sustain head-injury related impairments decreased by 34.1%

Statisticians work with public health professionals to gather, analyze and interpret the data.

Back to Biostatistics and Biostatisticians
Understanding Variation in Data

- Variation from person to person is ubiquitous, making it difficult to identify the effect of a given factor or intervention on one's health.
- For example, a habitual smoker may live to be 90, while someone who never smoked may die at age 30.
- The key to sorting out such seeming contradictions is to study properly chosen groups of people (samples).

Next steps

- Look for the aggregate effect of something on one group as compared to another.
- Identify a relationship, say between lung cancer and smoking.
- This does not mean that every smoker will die from lung cancer, nor that if you stop smoking you will not die from it.
- It does mean that the group of people who smoke are more likely than those who do not smoke to die from lung cancer.

Probability

- How can we make statements about groups of people, but cannot do so about any given individual in the group?
- Statisticians do this through the ideas of probability.
- For example, we can say that the probability that an adult American male dies from lung cancer during one year is 9 in 100,000 for a non-smoker, but is 190 in 100,000 for a smoker.

Events and their Probabilities

- We call dying from lung cancer during a particular year an “event”.
- Probability is the science that describes the occurrence of such events.
- For a large group of people, we can make quite accurate statements about the occurrence of events, even though for specific individuals the occurrence is uncertain and unpredictable.
Statistical Model

- A model for the event “dying from lung cancer” relies on two assumptions:
  - the probability that an event occurs is the same for all members of the group (common distribution); and
  - a given person experiencing the event does not affect whether others do (independence).
- This simple model can apply to all sorts of Public Health issues.
- Its wide applicability lies in the freedom it affords us in defining events and population groups to suit the situation being studied.

Example

Brain Injury of Bicycle Riders

- Groups: rider used helmet? Yes/no
- Events: crash resulted in severe brain injury? Yes/no.

<table>
<thead>
<tr>
<th></th>
<th>Wearing Helmet</th>
<th>Not Wearing Helmet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe Head Injury</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Not Severe Head Injury</td>
<td>19</td>
<td>8</td>
</tr>
</tbody>
</table>

Analysis of Evidence

- We see that:
  - 20% (2 out of 10) of those not wearing a helmet sustained severe head injury,
  - But only 5% (1 out of 20) among those wearing a helmet.
- **Relative risk is 4 to 1.**
- Is this convincing evidence?
  - Probability tells us that it is not, and the reason is that, with such a small number of cases, this difference in rates is just not that unusual. Let’s see why.
Probability Model: the Binomial Distribution

- Suppose that the chance of severe head injury following a bicycle crash is 1 in 10.
- Use a child's spinner with numbers “1” through “10”. The dial points to a number from “1” to “10”; every number is equally likely; and the spins are independent.
- Let the spin indicate severe head injury if a "1" shows up, and no severe head injury for "2" through "10".

This model is known as the Binomial Distribution.

Probability of Observed Data

- We spin the pointer ten times to see what could happen among ten people not wearing a helmet.
- The Binomial distribution says the probability:
  - That we do not see a "1" in ten spins is .349,
  - That we will see exactly one "1" in ten spins is .387,
  - Exactly two "1’s" is .194, Exactly three is .057, exactly four is .011, with negligible probability for five or more.

So if this is a good model for head injury, the probability of 2 or more people experiencing severe head injury in ten crashes is 0.264.

Hypothesis Testing

- We hypothesize that no difference exists between two groups (called the "null" hypothesis), then use the theory of probability to determine how tenable such an hypothesis is.
- In the bicycle crash example, the null hypothesis is that the risk of injury is the same whether or not you wear a helmet.
- Probability calculations tell how likely it is under the null hypothesis to observe a risk ratio of four or more in samples of 20 people wearing helmets and ten people not wearing helmets.

Results of the Test

- With such a small sample, one will observe a risk ratio ≥ four about 16% of the time, far too large to give us confidence in asserting that wearing helmets prevents head injury.
- If the probability were small, say < 5%, we would conclude that there is an effect.
- To thoroughly test whether helmet use does reduce the risk of head injury, we need to observe a larger sample.
2x2 Tables

- This type of data presentation is called a 2x2 table
- The test we used is called the Chi-square test.

Relationships Among variables

Studying Relationships among Variables

- A major contribution to our knowledge of Public Health comes from understanding:
  - trends in disease rates and
  - relationships among different predictors of health.
- Biostatisticians accomplish these analyses by fitting mathematical models to data.

Example: Blood Lead

- Blood lead levels in children are known to cause serious brain and neurologic damage
  - at levels as low as ten micrograms per deciliter.
- Since the removal of lead from gasoline, blood levels of lead in children in the United States have been steadily declining,
  - but there is still a residual risk from environmental pollution.
Blood Lead versus Soil Lead

- In a survey, we relate blood lead levels of children to lead levels from a sample of soil near their residences.
- A plot of the blood levels and soil concentrations shows some curvature.
- So we use the logarithms to produce an approximately linear relationship.
- When plotted, the data show a cloud of points as in the following example for 200 children.

Analysis of Lead Data

- The plot was produced by a statistical software program called Stata.
- We fitted a straight line to the data, called the regression equation of $y$ on $x$.
- The software also printed out the fitted regression equation: $y = 0.29x + 0.01$.
- It says that an increase of 1 in log(soil-lead) concentration will correspond, on average, to an increase in log(blood-lead) of 0.29.

Data on Blood Lead versus Soil Lead (in log scale)

- For example, a soil-lead level of 100 milligrams per kilogram, whose log is two, predicts an average log blood-lead level of $0.29 \times 2 + 0.01 = 0.59$,
  - corresponding to a measured blood level of 3.8 micrograms per deciliter.
- For 1000 mg per kg soil-lead level, the blood lead level is computed to be 7.6 mcg per dL.

Interpretation
Public Health Conclusion

- From the public health viewpoint, there is a positive relationship between the level of lead in the soil and blood-lead levels in the population,
  - i.e., soil-lead and blood-lead levels are positively correlated.

Correlation

- To study the degree of the relationship between two variables, we:
  - Estimate a quantity called the correlation coefficient, or “r”
  - This “r” must lie between -1 and +1, and is interpreted as a measure of how close to a straight line the data lie.

Correlation Analysis

- Values near ±1: nearly perfect line,
- Values near 0: no linear relationship,
  - but there may be a non-linear relationship.
- For the lead data, r = 0.42
  - It can be used to test for the statistical significance of the regression.

Significance Analysis

- Test of correlation r = .42 declares that the regression is significant at the 5% level.
  - This means that the chance of such a correlation happening by chance alone is less than 1 in 20.
  - We conclude that the observed association must be real.
Another Analysis

- We can use the 2x2 table analysis discussed earlier.
- For each child, we measure whether the soil lead was high or low, and classify a child’s blood lead levels as high and low, choosing appropriate definitions.

2x2 Table Analysis of Lead Data

- Choosing a median cutoff value for low and high produces the following table:

<table>
<thead>
<tr>
<th></th>
<th>low blood lead level</th>
<th>high blood lead level</th>
</tr>
</thead>
<tbody>
<tr>
<td>low soil lead level</td>
<td>63</td>
<td>37</td>
</tr>
<tr>
<td>high soil lead level</td>
<td>37</td>
<td>63</td>
</tr>
</tbody>
</table>

Interpretation of 2x2 Table Analysis

- The chi square statistic for this table also indicates a significant association between blood lead levels and soil lead levels in children.
- The conclusion is not as compelling as in the linear regression analysis, and
  - we have lost a lot of information in the data by simplifying them in this way.
- One benefit, however, of this simpler analysis is that we do not have to take logarithms of our data, or worry about the appropriate choice of a regression model.

Common Biostatistical Methods
Multiple Regression Analysis

- Outcome, $Y$, is continuous.
- Predictors, or covariates, the $X$’s, can be on any scale.
- Relationship between $Y$ and the $X$’s is assumed linear.
- Objective is to examine and quantify the relationship between $Y$ and the $X$’s, and
- Derive an equation to predict $Y$ from the $X$’s.

Example of Multiple Regression Analysis

- $Y =$ reduction in SBP
- $X_1 =$ treatment (1=new, 0=standard)
- $X_2 =$ gender (1=female, 0=male)
- $X_3 =$ age in years
- $X_4 =$ ethnicity (coded)
- Question: after accounting for all the covariates, is the new treatment effective?

Logistic Regression Analysis

- Outcome, $Y$, is binary (1 = yes, 0 = no).
- Predictors, or covariates, the $X$’s, can be on any scale.
- For given $X$’s, we denote the probability that $Y = 1$ by $p$. The odds are $p/(1-p)$.
- We assume that the relationship between the logarithm of the odds and the $X$’s is linear.
- Objective is to examine and quantify the relationship between $Y$ and the $X$’s, and
- Derive an equation to predict $Y$ from the $X$’s.

Example of Multiple Logistic Regression Analysis

- $Y =$ patient cured? 1=yes, 0=no.
- $X_1 =$ treatment (1=drug, 0=placebo)
- $X_2 =$ gender (1=female, 0=male)
- $X_3 =$ age in years
- $X_4 =$ ethnicity (coded)
- Question: after accounting for all the covariates, is the drug effective?
Survival Analysis

- The outcome $Y$ is the time till a specific event occurs (survival time).
- Other measurements can include covariates and treatment.
- We wish to study the survival distribution, either by itself or as it relates to the covariates.
- Several models exist.

Example of survival Analysis

- $Y =$ survival in years since onset of cancer
- $X_1 =$ treatment (1=new, 0=standard)
- $X_2 =$ gender (1=female, 0=male)
- $X_3 =$ age in years, $X_4 =$ ethnicity (coded)
- $X_5 =$ size of tumor

Question: after accounting for all the covariates, is the new treatment effective?

New Frontiers: Bioinformatics

“Bioinformatics is the study of the inherent structure of biological information and biological systems. It brings together the avalanche of systematic biological data (e.g. genomes) with the analytic theory and practical tools of mathematics and computer science.” (UCLA Bioinformatics Interdisciplinary Program)
What Do Physicians Understand by Medical Informatics?

- Practitioners will look up Best Practices on-line
- Hospital Infosystems will be available 24x7 through the Internet
- Clinicians will receive new research information directly relevant to their practice
- Physicians will routinely use Computer facilitated diagnostic & therapeutic algorithms
- Physicians will manage similar patient problems using computer facilitated tools.

The Focus of Public Health Informatics

- Prevention
- The health of populations
  - Example: NHLBI guidelines regarding cholesterol.
    - It’s an algorithm based on LDL, HDL and other risk factors,
    - followed by a recommendation to the patient regarding whether or not taking a cholesterol-reducing medication is advisable.

Uses of Bioinformatics and Medical Informatics

- It is within our grasp to be able to generalize this example many-fold.
- Based on the individual’s profile, it will be possible to formulate individual tailor-made guidelines for a healthier life.
Challenges in Data Analysis: Adjustments Needed

- The flood of information from genomics, proteomics, and microarrays can overwhelm the current methodology of biostatistics.
- Example: microarrays.

Example: DNA Microarrays

- Plate smaller than a microscope slide
- Can be used to measure thousands of gene expression levels simultaneously
- Microarrays can detect specific genes or measure collective gene activity in tissue samples.
- 2 basic types:
  - cDNA arrays
  - oligonucleotide arrays
Uses of Microarrays

- Gene expression patterns are compared between different tissue samples
- Question: Can the gene expression profile predict cancer tissue? (Diagnosis).
- Question: Can a gene expression predict survival outcomes? (Prognosis).
- Question: can we tailor the drug to the patient’s profile? (Treatment)

Ethical Issues of Bioinformatics and Medical Informatics

- Some discrimination based on whether a person smokes or is overweight takes place right now.
- The eligibility of individuals for health and life insurance can become threatened by whether they fit certain criteria based on genetic profiles.
- Employment opportunities may also be jeopardized.

References


Summary

- It is indeed an exciting time for biostatistics and public health.

- Thank you very much.
  Abdelmonem A. Afifi
  afifi@ucla.edu