HST 190: Introduction to Biostatistics

Lecture 1:
Basic principles of statistical data analysis
Welcome!

• Statistical reasoning is the process of drawing scientific conclusions from data in a rational, consistent way

• Goals for the course:
  ▪ develop an intuition for the key concepts that underpin the statistical analysis of data
  ▪ read the “Methods” section of an article, and understand/critique the approach taken
  ▪ learn to analyze and draw scientific conclusions from your own data
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Basic principles of statistical data analysis</td>
</tr>
<tr>
<td>2</td>
<td>Principles of probability &amp; Estimation of parameters</td>
</tr>
<tr>
<td>3</td>
<td>Two-sample comparisons, hypothesis testing and power/sample size calculations</td>
</tr>
<tr>
<td>4</td>
<td>Clinical trials &amp; Simple linear regression</td>
</tr>
<tr>
<td>5</td>
<td>Multiple linear regression</td>
</tr>
<tr>
<td>6</td>
<td>Methods for binary outcomes</td>
</tr>
<tr>
<td>7</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>8</td>
<td>Analysis of time-to-event data</td>
</tr>
<tr>
<td>9</td>
<td>Project presentations</td>
</tr>
<tr>
<td>10</td>
<td>Review before the exam</td>
</tr>
</tbody>
</table>
Course Logistics

- Eight lectures
  - each 2-2.5 hours long

- Reading will be assigned prior to each lecture
  - given the pace of the course, this is strongly encouraged

- Problem sets following each lecture
  - include Matlab exercises
  - due at 9am on the day of the following lecture (unless specified otherwise)
• During breaks in the middle we will:
  ▪ complete group exercises
  ▪ learn Matlab
  ▪ discuss course projects

• You will also work on a **group project** and present results during one of the class meetings

• In-class **exam** will take place during last meeting
  ▪ 28\(^{th}\) August
  ▪ open-book
Suggestions

• Ask questions during the lecture as well as on Piazza
  ▪ take notes!

• Material presented in different sequence from Rosner
  ▪ consult Rosner for a different approach

• Lots of material in a short time
  ▪ feel free to ask for help!

• There will be many formulae
  ▪ goal is not to memorize them
  ▪ even though we have access to software, hand calculations can help cultivate intuition
How to Prioritize

• The course is pass/fail.

• Exam is open-book, so don’t spend time memorizing formulas. Learn when and why to use each procedure; you can always refer to your notes to see how.

• To get the most out of this course, you should:
  ▪ attend lectures
  ▪ submit solutions to all the problem sets
  ▪ participate in class discussions, group exercises, and Piazza
  ▪ complete a project
  ▪ take the final exam
Resources

- Lecture Notes (*Canvas* -> *Files*)
  - *Get bonus points for finding typos!*
- Introduction to Matlab (*Canvas* -> *Files*)
- Rosner textbook, 7th ed. (required; a lifelong reference)
- Piazza
- Pagano & Gavreau textbook
- See Syllabus for additional references.
Basic steps of data analysis

• To set the stage, let’s consider two motivating questions:
  
  1) is there an association between time spent in the operating room and post-surgical outcomes for lung cancer resection?
  
  2) can we develop an enhanced breast cancer risk model?

• The questions have been left deliberately vague! it’s often the case that scientific questions are initially imprecisely posed

• Integral to the process of research is translating science into statistics, and back again:
  
  ▪ as you read papers, it is important to consider how the authors thought through this process
There are many (possibly infinite!) ways in which one could characterize ‘basic steps’ but a reasonable outline might be:

I. Understand the context of the analysis
II. Establish the scientific goals
III. Translate the scientific goals into statistical language
IV. Choose statistical methods to employ
V. Implementation and running the analysis
VI. Interpretation
• Sometimes, the way forward is clear and, in that sense, the process is prescriptive
  ▪ features/issues that are common to all analyses

• In many instances, however, the way forward isn’t clear
  ▪ aspects of the analysis don’t fit in with what you currently know
  ▪ these may relate to the science, data and/or statistics aspects

• Solutions include:
  ▪ appealing to the published literature (scientific and statistical)
  ▪ adopting or adapting existing methods
  ▪ developing new methods

• Regardless, dealing with these issues will require some creativity, and there is seldom, if ever, one ‘correct’ data analysis
  ▪ different data analyses correspond to different scientific questions
  ▪ which scientific question is ‘right’?
I. Understanding the Context

- From the perspective of a biostatistician, the purpose of data analysis is to learn about some population using information in a sample.
- Learn about covariates in terms of association with or prediction of an outcome:
  - notationally we often think in terms of $X$ and $Y$
  - possibly within or across certain sub-populations denoted, say, by $Z$
- Context usually involves three things:
  1) the background science
  2) the nature of the available data
  3) the population of interest, often called the ‘target population’
Q: Is there an association between time spent in the operating room and post-surgical outcomes?

• **Background science:**
  - longer operating time $\rightarrow$ greater exposure to anesthesia
  - shortening operating time might reduce adverse post-surgical outcomes
    - complications during the hospital stay
    - recurrence of lung cancer
    - mortality
  - may also lead to decreased costs/increased efficiency
    - increased capacity for the operating room
    - shorter post-surgical hospital stay
• Available data:
  ▪ ≈400 surgeries at Brigham and Women’s Hospital
  ▪ performed between 1997-2008
  ▪ demographic, clinical, tumor and follow-up information

• Target population:
  ▪ patients who undergo elective surgery for early stage non-small cell lung cancer
  ▪ need to be aware of different surgery sub-types
    o lobectomy, segmentectomy, wedge resection
    o thorachotomy, video assisted thoracic surgery
  ▪ what do we think about the (relatively) long time frame?
  ▪ generalizability beyond BWH?
Q: Can we develop an enhanced breast cancer risk model?

• Background science:
  ▪ the ‘Gail model’ for breast cancer risk was developed in the late 1980s
    o age, race,
    o age at menarche, age at birth of first child
    o family history, number of prior biopsy examinations and atypical hyperplasia
  ▪ the model was validated in a number of subsequent studies
  ▪ subsequent research identified a number of additional risk factors for breast cancer
    o breast density, use of hormone replacement therapy and body mass index
• **Available data:**
  - 2,392,998 screening mammograms from the Breast Cancer Surveillance Consortium
    - NCI-funded nationwide network of mammography registries
  - mammograms performed between 1996-2002
  - outcomes are ascertained via linkages with cancer registries

• **Target population:**
  - screening mammograms performed on women aged 35-84 years
    - unit of analysis is the mammogram, not the woman
  - who undergoes screening? who doesn’t?
    - how might this impact the interpretation of the study?
Nature of the available data

• What were the data collection procedures?
  ▪ convenience sample or part of a designed study?
  ▪ what was the setting/timeframe?
  ▪ observational study or randomized design?
  ▪ cross-sectional, prospective, or retrospective?
  ▪ stratification and/or matching?

• How were the procedures followed?
  ▪ any systematic deviations from the ‘ideal’ data collection process?
  ▪ may be due to patients?
    o refusal to participate/respond
    o inaccurate responses
- may be due to researchers?
  - were uniform procedures applied to all (potential) participants?
  - are we actually measuring what we think we are measuring?

- Have there been any interim data cleaning/manipulation efforts?
  - cleaning of ‘strange’ values
    - set to some threshold value or to missing
    - exclusion from the dataset
  - construction of derived variables
Populations

• In practice, the ‘population’ can be
  ▪ an actual, potentially observable population
  ▪ a hypothetical (sometimes infinite) population

• Might refer to the ‘target population’ to emphasize that there is a specific population in mind

• Defining the target population is crucial in that it provides the context the scientific question of interest
  ▪ who would we like our results to generalize too?

• Narrow vs. broad definitions of the target population
  ▪ heterogeneity vs. homogeneity
  ▪ what are the trade-offs?
• What comes first ... the data or the population?
  ▪ depends on when you get involved

• If the data has already been collected:
  ▪ for which population could we consider the sample as being ‘representative’?
  ▪ may need to focus the dataset by excluding certain folks
    o implicitly changes the population to which one can generalize
    o sample size vs. mixing of effects
  ▪ is there scope for additional data collection efforts?

• If the data has not been collected:
  ▪ much greater flexibility for choosing/defining the population of interest
• Recall, the goal is to learn about the relationships between a subset of covariates

• Achieved by collecting and analyzing a sample from the population
  ▪ an important aspect of ‘context’ is that this is indeed what we are doing
    o or, at least, hoping to do!

• Suppose we could enumerate the entire population
  ▪ that is, the sample is the population

• In this case observed data characterizes relationships completely
• Note when we have a complete enumeration, there is no sampling variability
  ▪ we don’t have to worry about making statements about the population on the basis of information in the sample
  ▪ the sample is the population

• We don’t have to consider or quantify uncertainty associated with only observing a sub-sample
  ▪ no need for standard errors, confidence intervals or p-values
  ▪ may be no need for statistical methods!

• Most of the time we can’t enumerate the entire population
  ▪ typically, this isn’t logistically and/or financially feasible

• So...
II. Establish the scientific goals

• Broadly speaking one can classify scientific goals as:
  - description or exploration of a population
  - evaluation of some hypothesis
  - prediction of future outcomes

• A single analysis may have several goals
  - depends on scientific setting and background
Q: Is there an association between time spent in the operating room and post-surgical outcomes?

- **Description/exploration:**
  - what is the nature of the association?
  - does the association differ across surgery types?

- **Hypothesis testing:**
  - a priori hypothesis among the collaborators that shorter times are associated with better post-surgical outcomes
Q: Can we develop an enhanced breast cancer risk model?

• Prediction:
  ▪ use all the available information in the best possible way to predict the risk of breast cancer
  ▪ build prediction models that cater to specific settings with varying amounts/type of information?
    ○ at home/online
    ○ in the physicians office

• Why might description/exploration and hypothesis testing be of less interest?
Goal is to characterize the relationships among a set of covariates in the population of interest.

An important issue is whether or not the goal is to establish causation:
- Typically requires a greater understanding of the science.

Typically, although not always, viewed as hypothesis generating:
- We have a cool dataset, let’s see what we can find ...
- There is a fine, often blurry line between exploration and hypothesis testing.
  - What came first ... the data or the question?
Hypothesis testing

• Goal is to make some confirmatory statement

• Typically framed in the context of making a ‘decision’ between two competing hypotheses

  \[ H_0: \text{null hypothesis} \]
  \[ H_1: \text{alternative hypothesis} \]

• Assume the null hypothesis holds and look for evidence to the contrary

• Standard hypothesis testing reduces the potential decisions to:
  1. fail to reject \( H_0 \)
  2. reject \( H_0 \) (implicitly in favor of \( H_1 \))

    ▪ decision should be accompanied by some measure of uncertainty
Prediction

• Goal is to estimate future outcomes or risk
  ▪ Typically framed in terms of building the best possible model

• What do we mean by ‘best’?
  ▪ need some means of judging accuracy and penalizing poor predictions
  ▪ ideally based on real world consequences
    ○ e.g. false-positive vs. false-negative for breast cancer

• Sometimes a single best model is inappropriate
  ▪ a model may work well in one population and not others
  ▪ inputs may not always be available (e.g. genetic information)

• To what extent do we need to care about causation?
  ▪ do we need to understand the ‘true’ underlying mechanisms?
The real world

• Unfortunately, the scientific goals are not always clear at the outset

• Typically, it is the case that:
  ▪ there are many scientific goals that are of interest, and/or
  ▪ the goal can be interpreted in a number of ways

• Primarily a problem because investigators need precise statements to be able to proceed
  ▪ to translate the scientific goals into statistical ones

• Towards refining study goals, a couple of useful questions are:
  1) who is the intended (primary) audience?
  2) what will be actionable from the results?
• Consider the question: What is Mrs. Jones’ risk of breast cancer?

• How one proceeds depends, at least in part, on how this information will be used:

  **Researchers**
  o determine eligibility for a randomized study of some novel preventative agent

  **Patients**
  o decision as to whether or not she should get in touch with her physician

  **Physicians**
  o planning for future screening schedule

  **Policy-makers**
  o monitor the public health burden of breast cancer
• Related questions include:
  ▪ is interest in all breast cancers or some specific tumor type?
  ▪ risk over which timeframe?
    o 1 year?
    o 5 years?
    o lifetime?
  ▪ how much information will the interested ‘user’ have access to?
    o will detailed family history information be available?
    o will genetic information be available?

• Different answers to all these questions define different scientific goals
III. Translating scientific goals into statistical terms/tasks

- Once the scientific goals are ‘established’ we need to translate them into the language of statistics.

<table>
<thead>
<tr>
<th>Scientific goal</th>
<th>Statistical task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description/exploration</td>
<td>Estimation</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>Inference</td>
</tr>
<tr>
<td>Prediction</td>
<td>Estimation</td>
</tr>
</tbody>
</table>

- Moving forward requires:
  - precise and clear definitions of all relevant covariates
  - specification of key relationships of interest
Precisely defining covariates

- Each of the potential goals is trying to say something about the relationships among a set of covariates
- Prior to any analysis we need clear definitions for all relevant covariates:
  - response variables
  - exposure(s) of interest
  - interaction terms and/or effect modifiers
  - predictors of the response
  - predictors of the exposure(s) of interest
- There will be overlap across these various types of variables
  - e.g., a covariate may be a predictor of both the response and of the exposure of interest
• Often not as straightforward as one might think, mainly because there is often choice involved

• Suppose the response of interest is ‘diagnosis of breast cancer’
  ▪ over which time frame?
  ▪ for which sub-types?

• Suppose the exposure of interest is ‘operating time’
  ▪ when does time start?
  ▪ when does time stop?

• Define (and perhaps re-define) until everything is clear!
Q: Is there an association between time spent in the operating room and post-surgical outcomes?

- **Responses:**
  - hospital stay of > 7 days (binary)
  - number of major complications during hospital stay (count)
    - need a list of ‘major’ complications
  - time to death (continuous, right-censored)

- **Exposure of interest:**
  - operating time defined as the time from the first incision to the time of the first stitch to close up (continuous)
Q: Can we develop an enhanced breast cancer risk model?

- Response:
  - diagnosis of breast cancer within 1 year of the screening mammogram (binary)

- Exposure of interest:
  - age, race, education, breast density, HRT use ...
  - a total of 13 potential predictors
  - all categorical
    - at least in the available dataset
IV. Choosing statistical methods

• One way of viewing all the statistical methods available is as a collection of tools
  ▪ different statistical tools for different statistical tasks
  ▪ develop understanding of a collection of tools over the course of your career

• A toolbox of statistical tools/methods
  ▪ basic methods, that everyone should be able to use
  ▪ specialized methods
    ○ sophisticated tools that require ‘training’
    ○ constantly being developed and published in the literature
  ▪ sometimes new questions require new methods
• For the most part, the tools that researchers employ are determined by the issues we’ve considered so far
  - scientific goals
  - nature of the available data
  - population of interest

• Even given all this information, there are often several choices of statistical tools/methods

• How to choose between all the available approaches?
  - interpretation (to be discussed later)
  - operating characteristics
    - e.g. bias and statistical efficiency
V. Implementation and running the analysis

- Seemingly the most ‘prescriptive’ of the steps
  - in a perfect world, turn the handle ... and you’re done!

- Unfortunately, actually performing the analysis is not always straightforward

- Many choices for statistical software
  - R, Matlab, SAS, Stata, WinBUGS, ...
  - each has numerous resources, including already-written code available online
  - not all methods have been implemented in all software packages
• Performing the analyses can also highlight all sorts of problems
  - EDA might highlight data issues
    - missing data
    - unusual values
    - unusual observed relationships

• Issues like this may require a re-think of the scientific goals
  - if you can’t answer this question, which question can you answer?
VI. Interpretation

• It’s important to distinguish interpretation of the *model* from interpretation of the *results*

**Interpretation of a model**

• Specification of the model is something that we have control over
  - it should be straightforward to provide a precise interpretation of its’ components
    - you cannot be pedantic enough on this point
  - should be able to do this before you even see that data

• Consider the linear regression model:

\[ E[Y|X] = \beta_0 + \beta_1 X \]

  - How do we interpret $\beta_1$?
• Here are some results ... what does it all mean?!?
  ▪ translation of statistics back to science

• Interpreting the results requires a detailed understanding both the scientific and statistical context
  ▪ usually requires discussion with collaborators

• Sometimes the results don’t support the initial hypotheses!
  ▪ e.g., Breitner et al (2008) Neurology
  ▪ Risk of dementia and AD with prior exposure to NSAIDs in an elderly community-based cohort
  ▪ see the next slide
ABSTRACT

Background: Nonsteroidal anti-inflammatory drugs (NSAIDs) may prevent Alzheimer dementia (AD).

Methods: We analyzed the association of prior NSAID exposure with incident dementia and AD in the Adult Changes in Thought population-based cohort aged ≥65 years (median 74.8) at enrollment. Participants were members of Group Health, which provided computerized pharmacy dispensing records from 1977 onward. We studied 2,736 dementia-free enrollees with extensive prior pharmacy data, following them biennially for up to 12 years to identify dementia and AD. Cox proportional hazards regression assessed association of dementia or AD with NSAID use graded in standard daily doses (SDD) dispensed over 2 years (e.g., heavy use = 500+ SDD), with some analyses also adding consecutive biennial self-reports of NSAID use.

Results: Pharmacy records identified 351 participants (12.8%) with history of heavy NSAID use at enrollment. Another 107 became heavy users during follow-up. Some 476 individuals developed incident dementia, 356 with AD (median onset ages 83.5 and 83.8 years). Contrary to the hypothesis that NSAIDs protect against AD, pharmacy-defined heavy NSAID users showed increased incidence of dementia and AD, with adjusted hazard ratios of 1.66 (95% confidence interval, 1.24–2.24) and 1.57 (95% confidence interval, 1.10–2.23). Addition of self-reported exposure data did not alter these results.

Conclusions: These findings differ from those of other studies with younger cohorts. The results observed elsewhere may reflect delayed onset of Alzheimer dementia (AD) in nonsteroidal anti-inflammatory drug (NSAID) users. Conceivably, such delay could result in increased AD incidence in late old age. The relation of NSAID use and AD pathogenesis needs further investigation.

Neurology® 2009;72:1899-1905
• These can be particularly challenging situations

• Are these results ‘right’?
  ▪ are we misinterpreting our assumptions/models?
  ▪ are there data issues that we aren’t aware of?
  ▪ is the code wrong?
  ▪ are the results sensitive to particular analysis choices?

• It may be that the results are ‘right’
  ▪ perhaps a new understanding of the mechanism of interest
  ▪ perhaps the results pertain to a population that hasn’t been studied before
Learning about populations

• It is seldom possible to specify one, single target population
  ▪ often the case there are many interesting target populations

• Flexibility to consider different populations depends on whether or not the sample has been collected

• If the sample has not been collected, one might consider
  ▪ a range of scientific questions
  ▪ the feasibility of collecting data across different populations

• If the sample has been collected, flexibility depends on the nature and scope of the available data
Breast cancer screening

- Broad goal of screening is to detect cancer as early as possible
  - balance between public health goals and costs
  - cannot screen everyone all of the time
  - there are also ‘harms’ associated with screening
  - mammography is not perfect
  - real consequences associated with false-positives

- Current recommendations are (broadly):
  - all women aged 50 or older get screened every two years
  - also, women in their 40’s who are at ‘high risk’

Q: How good is mammography as a screening modality?
  - answer depends, in part, on the population of interest
Performance Benchmarks for Screening Mammography

Purpose:
To retrospectively evaluate the range of performance outcomes of the radiologist in an audit of screening mammography by using a representative sample of U.S. radiologists to allow development of performance benchmarks for screening mammography.

Materials and Methods:
Institutional review board approval was obtained, and study was HIPAA compliant. Informed consent was or was not obtained according to institutional review board guidelines. Data from 188 mammographic facilities and 807 radiologists obtained between 1996 and 2002 were analyzed from six registries from the Breast Cancer Surveillance Consortium (BCSC). Contributed data included demographic information, clinical findings, mammographic

  - all women who undergo screening mammography
Performance of First Mammography Examination in Women Younger Than 40 Years

Bonnie C. Yankaskas, Sebastien Haneuse, Julie M. Kapp, Karla Kerlikowske, Berta Geller, Diana S. M. Buist; for the Breast Cancer Surveillance Consortium

Manuscript received May 26, 2009; revised February 15, 2010; accepted February 25, 2010.

Correspondence to: Bonnie C. Yankaskas, PhD, Department of Radiology, University of North Carolina at Chapel Hill, 204 Wing F, CB#7510, Chapel Hill, NC 27599-7510 (e-mail: bcy@med.unc.edu).

Background Few data have been published on mammography performance in women who are younger than 40 years.

Methods We pooled data from six mammography registries across the United States from the Breast Cancer Surveillance Consortium. We included 117,738 women who were aged 18–39 years when they had their first screening or diagnostic mammogram during 1995–2005 and followed them for 1 year to determine accuracy of mammography assessment. We measured the recall rate for screening examinations and the sensitivity, specificity, positive predictive value, and cancer detection rate for all mammograms.

Results For screening mammograms, no cancers were detected in 637 mammograms for women aged 18–24 years. For women aged 35–39 years who had the largest number of screening mammograms (n = 73,335) in this study, the recall rate was 12.7% (95% confidence interval [CI] = 12.4% to 12.9%), sensitivity was 76.1% (95% CI = 69.2% to

Effect of Breast Augmentation on the Accuracy of Mammography and Cancer Characteristics

Diana L. Miglioretti, PhD
Carolyn M. Rutter, PhD
Berta M. Geller, EdD
Gary Cutter, PhD
William E. Barlow, PhD
Robert Rosenberg, MD
Donald L. Weaver, MD
Stephen H. Taplin, MD
Rachel Ballard-Barbash, MD, MPH
Patricia A. Carney, PhD
Bonnie C. Yankaskas, PhD
Karla Kerlikowske, MD

**Context**  Breast augmentation is not associated with an increased risk of breast cancer; however, implants may interfere with the detection of breast cancer thereby delaying cancer diagnosis in women with augmentation.

**Objective**  To determine whether mammography accuracy and tumor characteristics are different for women with and without augmentation.

**Design, Setting, and Participants**  A prospective cohort of 137 women with augmentation and 685 women without augmentation diagnosed with breast cancer between January 1, 1995, and October 15, 2002, matched (1:5) by age, race/ethnicity, previous mammography screening, and mammography registry, and 10533 women with augmentation and 974915 women without augmentation and without breast cancer among 7 mammography registries in Denver, Colo; Lebanon, NH; Albuquerque, NM; Chapel Hill, NC; San Francisco, Calif; Seattle, Wash; and Burlington, Vt.

**Main Outcome Measures**  Comparison between women with and without augmentation of mammography performance measures and cancer characteristics, including invasive carcinoma or ductal carcinoma in situ, tumor stage, nodal status, size, grade, and estrogen-receptor status.

An Assessment of the Quality of Mammography Care at Facilities Treating Medically Vulnerable Populations

L. Elizabeth Goldman, MD, MCR, * Sebastien J.-P. A. Haneuse, PhD, ‡‡ Diana L. Miglioretti, PhD, ‡‡ Karla Kerlikowske, MD, §‖ Diana S. M. Buist, PhD, † Bonnie Yankaskas, PhD, †‖ and Rebecca Smith-Bindman, MD, **§ for the National Cancer Institute Sponsored Breast Cancer Surveillance Consortium

**Background:** Women in medically vulnerable populations, including racial and ethnic minorities, socioeconomically disadvantaged, and residents of rural areas, experience higher breast cancer mortality than do others. Whether mammography facilities that treat vulnerable women demonstrate lower quality of care than other facilities is unknown.

**Objectives:** To assess the quality of mammography women receive at facilities characterized as serving a high proportion of medically vulnerable populations.

**Conclusions:** Facilities serving high proportions of vulnerable populations provide screening mammography with equal or better quality (as reflected in higher specificity with no corresponding decrease in sensitivity) than other facilities. Further research is needed to understand the mechanisms underlying these findings.

**Key Words:** mammography, screening, quality of care, vulnerable populations

*(Med Care 2008;46: 701–708)*

Remarks

• Except in the most trivial of settings, the data analysis process is collaborative and iterative

• How you proceed will depend on many things:
  ▪ the nature of the data
  ▪ your philosophy
  ▪ the philosophy of your collaborators

• Getting the science ‘right’ is often the hardest part
  ▪ goals are seldom precise at the outset
  ▪ going back-and-forth between the science and statistics is typically a very instructive process
  ▪ to do a good job usually requires knowledge of the science
• More often than not, there is scope for prescription as well as for creativity
  ▪ sometimes there is an obvious way forward
  ▪ other times there isn’t

• What came first ... the question or the data?

• There is seldom one ‘right’ scientific question or data analysis
  ▪ Box and Draper (1987):

  *Essentially, all models are wrong but some are useful.*