Reading

- M.J. van der Laan, S. Rose (2011), Targeted Learning: Causal Inference for Observational and Experimental Data, Springer, New York.
- We will post specific readings through Bcourse announcements. Lots of material on www.bepress.com/ucbbiostat.
- Popular read motivating targeted learning: Why we need a Statistical Revolution, http://www.stats.org/super-learning-and-the-revolution-in-knowledge/

Topics

Why we need targeted learning

- Problems with classical statistics, machine learning

Roadmap for targeted learning

- Causal models and identifying causal parameters

Estimating nuisance parameters

- Super learning, theory and practice
- Highly adaptive lasso*
- Online super learning*

Efficient estimation of statistical parameters

- Asymptotic linearity
- Offline one-step, TMLE estimation
- Online one-step, TMLE estimation*

Time permitting topics

- CV-TMLE* and data-adaptive target parameters*

Lab topics

Definite topics

- Simulating from causal models
- SuperLearner
- tmle
- h2O, h2Oensemble
- onlinesl*

Possible additional topics (suggestions welcome)

- Amazon computing
- ggplot
- Spark computing
- Shell programming
- big packages
- Rcpp

The role of statistics in science

Science is about discovering laws that govern the universe and understanding the impact/interplay of those laws. Progress in science if often achieved through analyzing experiments tailored to answer specific scientific questions about a stochastic system.

Statistics is about the analysis of real world experiments. Given a particular type of experiment on a stochastic system and a specific question, a statistical method aims to answer the question through a point estimate as well as through quantification of uncertainty (confidence interval).

Statistics quantifies our knowledge (or lack thereof) about scientific laws.

The role of statistics in science

The sorts of scientific laws we are interested in are those involving biomedical or public health interventions.

“If I perform *some intervention* on *some population*, what *effect* will it have?"

The actual observed data experiment is a crucial factor in determining

- what scientific questions can be answered;
- what scientific assumptions are needed to justify answers.

Examples of observed data experiments

We generally refer to the observed output/data of the experiment on n units as n independent and identically (i.i.d.) distributed copies of a random variable O. But "i.i.d." is already a real assumption!

Since O is a random variable it has a probability distribution which we will denote with P_0 (i.e. $_0$ standing for the true probability distribution).

In general, we will use W to denote baseline characteristics, A to denote some kind of treatment/intervention, Δ to denote some form of missingness, and Y to denote an outcome of interest.

We will generally be interested in asking the question: what happens to Y if we perform an intervention on A.

Examples of observed data experiments

One observes n i.i.d. copies of $O = (W, A, Y)$, W baseline covariates, A binary treatment, Y final outcome.

Example: Vaccine development – immunogenecity experiments

- Volunteers/lab animals receive a vaccine/placebo and have their immune response measured.
- Does the vaccine create an immune response?

Big Data Example: Electronic Medical Records – drug efficacy

- Patients are recorded as receiving e.g., a heart medication
- Does the medication protect against cardiovascular events?

Examples of observed data experiments

We observe $O = (W, A, \Delta, \Delta Y)$: the outcome is subject to missingness.

Example: Vaccine development – immunogenecity experiments

- In the lab, a fraction of immune response measures cannot be measured

Big Data Example: Electronic Medical Records – drug efficacy

- Some patients never appear again in EMR database.

Examples of observed data experiments

We observe $O = (W, \Delta, \Delta A, Y)$: the treatment assignment is subject to missingness.

Example: Biomarkers studies in epidemiology

- A fraction of biomarkers cannot be measured

Big Data Example: Electronic Medical Records – drug efficacy

- Patients prescriptions may not be recorded

Examples of observed data experiments

We observe $O = (W, A, (\Delta(t), \Delta(t)Y(t) : t = 1, \ldots, \tau))$, where $\Delta(t)$ is indicator of subject being monitored at time t and $Y(t)$ is indicator of event/failure at time t.

Example: Preventive HIV vaccine efficacy trial

- Individuals are tested for HIV at regular clinic visits (e.g., every six months). Some individuals miss clinic visits.

Big Data Example: Electronic health records – depression

- Patients depressive symptoms are assessed at regular clinic visits. Some individuals miss clinic visits.

Examples of observed data experiments

We observe n_1 observations of (W, A), from population with $Y = 1$, and n_0 observations of (W, A), from a population with $Y = 0$. This is called standard case-control sampling.

Example: Rare forms of cancer

- Cohort studies are expensive and for rare diseases it is more cost-efficient to sample (e.g., from cancer registry) based on outcome.

Example: Immune responses in vaccine trials

- Measuring immune responses is expensive, so in practice we measure immune response on infected participants and a subset of uninfected participants.

Simple Randomized Trials

Not all observed data are created equal. We must also consider what we know about how the data were generated.

In a randomized trial the assignment of A is controlled by the experiment. For example, we flip a coin and assign $A = 1$ if it is head.

- Examples: Clinical trials, lab experiments

However, one can also assign $A = 1$ with a probability depending on patient characteristics W.

- Examples: Clinical trials where randomization is stratified by site

However, even though a randomized trial controls assignment of A, it does usually not completely control missingness Δ or censoring/monitoring $\Delta(t)$.

Sequential RCT

One observes n i.i.d. observations of $O = (L_0, A_0, L_1, A_1, L_2 = Y) \sim P_0$. In a sequential RCT, one controls the assignment of both initial treatment $A(0)$ and subsequent treatment $A(1)$.

Example: Cancer treatment trial with multiple possible courses of treatment

- We assign a drug among 4 drugs, see how well the patient responds, and possibly assign a different treatment if the patient is not responding.

Adaptive (Group) Sequential RCT

Patients enroll over time. Treatment assignment is controlled for each patient, but can be based on the observed past among all previously enrolled patients.

Example: Drug development

- If, e.g., women appear to be responding well to the drug, we may want to increase the probability of assigning treatment to newly enrolled women.

Simple Observational Studies

In an observational study the data generating experiment does not control the assignment of A.

Example: Epidemiological cohort studies on drug use

- The treatment decision is in the hands of the individual's medical doctor.
- We try to collect all patient characteristics the doctor might take into account when making a treatment decision.

Complex Observational Studies

One observes n i.i.d. observations of

$$(L_0, A_0, \Delta_0, \Delta_0 L_1, A_1, \Delta_1, \Delta_1 L_2, \ldots, A_K, \Delta_K, \Delta_K Y),$$

where A_t represents treatment assignment at time t, Δ_t is an indicator of being monitored at time $t+1$, L_t are time dependent covariates measured at time t, and $Y = L_{K+1}$ is the final outcome.

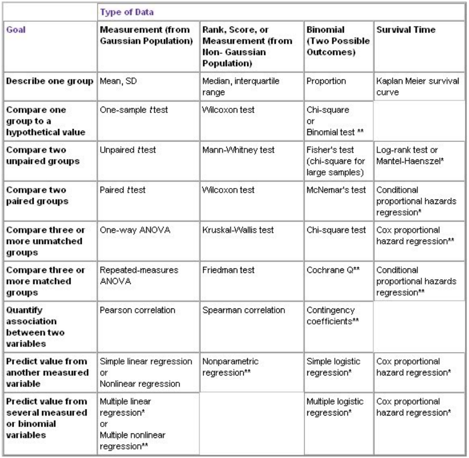
Example: Drug safety study, also called Phase IV-study, are of this type.

Key Points

- The scientific question and data generating experiment determines what statistical methods are needed.
- Better: The scientific question determines the data generating experiment and statistical method needed.
- The design of a study has a large impact on what we know about the underlying distribution of the observed data.

Traditional Approach in Epidemiology and Clinical Medicine

In general, conventional statistical practice lets the type of data at hand dictate the scientific question of interest:



	Type of Data			
Goal	**Measurement (from Gaussian Population)**	**Rank, Score, or Measurement (from Non- Gaussian Population)**	**Binomial (Two Possible Outcomes)**	**Survival Time**
Describe one group	Mean, SD	Median, interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample *t*test	Wilcoxon test	Chi-square or Binomial test **	
Compare two unpaired groups	Unpaired *t*test	Mann-Whitney test	Fisher's test (chi-square for large samples)	Log-rank test or Mantel-Haenszel*
Compare two paired groups	Paired *t*test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression*
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox proportional hazard regression**
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane Q**	Conditional proportional hazards regression**
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients**	
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression**	Simple logistic regression*	Cox proportional hazard regression*
Predict value from several measured or binomial variables	Multiple linear regression* or Multiple nonlinear regression**		Multiple logistic regression*	Cox proportional hazard regression*

Traditional Approach in Epidemiology and Clinical Medicine

The internet abounds with prescriptive statistical recommendations.

- continuous outcome + covariates → linear regression
- continuous and positive outcome + covariates → log-linear regression
- survival outcome and covariates → proportional hazards regression

The “parameter of interest" is defined as the regression coefficient for A in the chosen parametric model.

These parametric models are often too simplistic to reflect the underlying complexity, but remain popular because they are easy to implement.

Traditional Approach

In this approach, statistical convenience is allowed to determine the scientific question of interest.

Several critical questions naturally arise:

1. Is the “parameter of interest" truly of scientific interest?
2. If the model is misspecified, what are you actually estimating?

Why not begin by defining the estimand of interest?

- Dialogue with collaborators to determine the true question of interest.
- The definition should not rely on a parametric model, i.e., the parameter should be defined nonparametrically.
- Nevertheless, we could use parametric or semiparametric models to inspire interesting parameters.

Traditional Approach in Epidemiology and Clinical Medicine

Suppose we observe $O = (A, Y)$ for some continuous exposure A and some continuous outcome Y.

Our "Statistics for Dummies" book recommends we use linear regression, i.e., assume something like

$$Y = \alpha + \beta A + \epsilon \ ,$$

where ϵ is a mean zero, Normal random variable and (α, β) are unknown real numbers.

Traditional Approach in Epidemiology and Clinical Medicine

What if the relationship between A and Y is not linear?

- Example: threshold effect

It turns out that the least-squares estimator β_n of β is estimating

$$\tilde{\beta}_0 = \frac{\mathrm{Cov}(A, Y)}{\mathrm{Var}(A)} \ .$$

Is this a relevant parameter for assessing treatment efficacy?

Traditional Approach in Epidemiology and Clinical Medicine

An example from survival analysis: Suppose the outcome of interest is T, a survival time we are interested in the effect of A on Y.

Over the course of the study, some subjects may be lost to followup, say at time C.

The observed data consist of n i.i.d. copies of (A, Δ, Y) where $Y = \min(T, C)$ and $\Delta = I(T \leq C)$.

Our “Statistics for Dummies" book recommends that we should use a Cox model:

$$\lambda(t|1) = \exp(\gamma^*)\lambda(t|0) \ ,$$

where $\lambda(t|a)$ is the conditional hazard of T at time t given $A = a$.

Traditional Approach in Epidemiology and Clinical Medicine

What if the hazards between the treatment groups are not actually proportional over time?

- Example: waning vaccine efficacy

It turns out that the estimator γ_n from a Cox model is converging to the solution in γ of the equation

$$0 = \int w(t) \left\{ \frac{\lambda(t|1)}{\lambda(t|0)} - \exp(\gamma) \right\} dt \ ,$$

where the weights are given by

$$\frac{\lambda(t|0) P(T \geq t|A=1) {\color{red} P(C \geq t|A=1)}}{1 + \exp(\gamma) \frac{P(A=1)}{P(A=0)} \frac{P(T \geq t|A=1) {\color{red} P(C \geq t|A=1)}}{P(T \geq t|A=0) {\color{red} P(C \geq t|A=0)}}} \ .$$

In other words, the “parameter of interest" depends on the censoring distribution.

Traditional Approach in Epidemiology and Clinical Medicine

What is the alternative to this approach? We could define the parameter of interest as

$$\gamma_0 = \underset{\gamma}{\text{argmin}} \int \left\{ \frac{\lambda(t|1)}{\lambda(t|0)} - \exp(\gamma) \right\}^2 d\mu(t) \ ,$$

where μ is some weight function that we choose such that $\int d\mu(t) = 1$.

We can easily check that

$$\gamma_0 = \log \int \frac{\lambda(t|1)}{\lambda(t|0)} d\mu(t) \ .$$

If the Cox model holds, then the two parameters correspond. If not (as will almost always be the case), γ_0 defines a user-specified weighted average of the log-hazard ratio over time.

Traditional Approach in Epidemiology and Clinical Medicine

In both examples, the “recommended" statistical procedures are estimating some contrast across different levels of A, but probably not the contrast that your medical collaborators actually care about.

The difference between what you think you’re estimating and what you’re actually estimating could be drastic and amplified by a large sample size.

Inference, even for the true estimand, can also be misleading without modification. This has spawned an entire branch of statistics – so-called “robust" statistics.

Complications of Human Art in Statistics

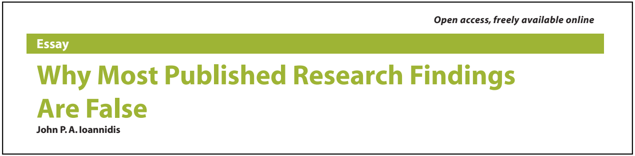
Returning to the linear regression example, classical statistical practice encourages users to "check" models after they have been fit.

If one of these checks fails, choose a new model: add quadratic term, remove a covariate, choose a different link function, choose a different error distribution.

Problems with post-hoc procedures:

1. Murky definition of target parameter; no correspondence with a scientific question.
2. Even if model checks are performed honestly (a big if), inference does not account for ad-hocery!

Complications of Human Art in Statistics

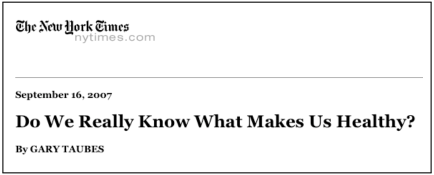


Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

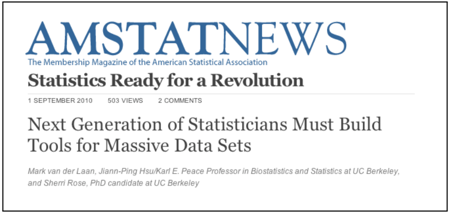


The New York Times
nytimes.com

September 16, 2007

Do We Really Know What Makes Us Healthy?

By GARY TAUBES



AMSTATNEWS
The Membership Magazine of the American Statistical Association

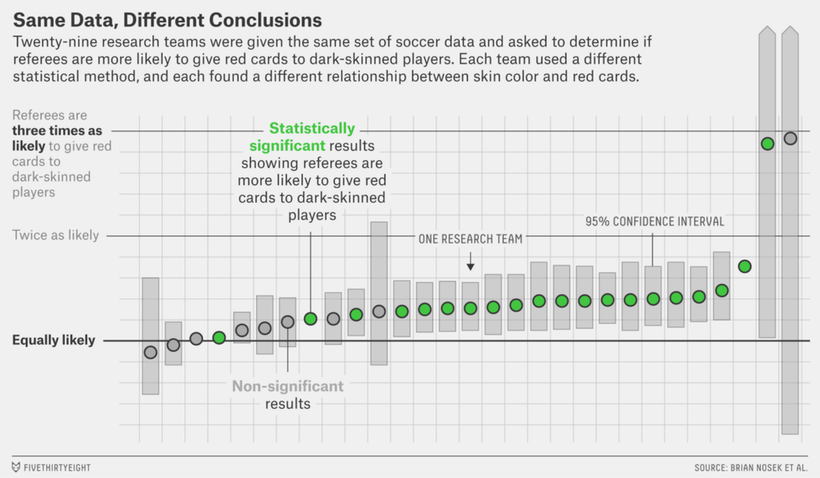
Statistics Ready for a Revolution

1 SEPTEMBER 2010 503 VIEWS 2 COMMENTS

Next Generation of Statisticians Must Build Tools for Massive Data Sets

Mark van der Laan, Jiann-Ping Hsu/Karl E. Peace Professor in Biostatistics and Statistics at UC Berkeley, and Sherri Rose, PhD candidate at UC Berkeley

Complications of Human Art in Statistics



Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

SOURCE: BRIAN NOSEK ET AL.

http://fivethirtyeight.com/features/science-isnt-broken/# part1

Complications of Human Art in Statistics

“Even the most skilled researchers must make subjective choices that have a huge impact on the result they find."

- This should cause extreme discomfort!

“On the one hand, our study shows that results are heavily reliant on analytic choices. On the other hand, it also suggests theres a ***there*** *there. Its hard to look at that data and say theres no bias against dark-skinned players."*

- Why is this statement false?

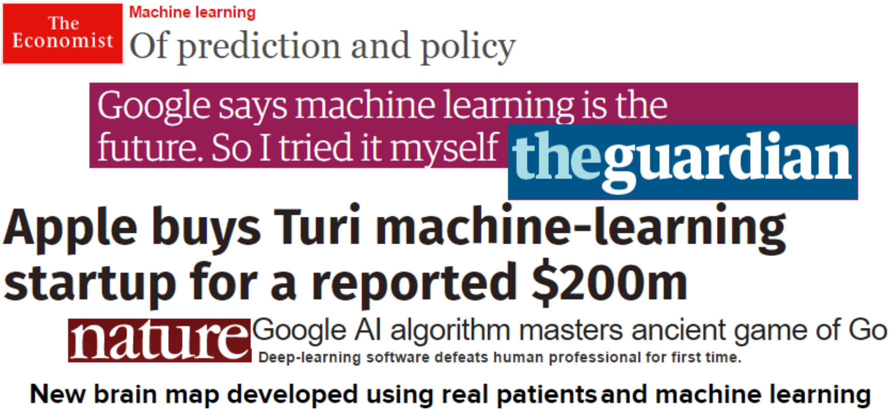
Summary

Problems with classical statistics:

1. Parametric models are misspecified.
2. Researchers often interpret the target parameter as if the parametric model is correct.
3. The parametric model is often data-adaptively (or worse!) selected, and this part of the estimation procedure is not accounted for in the variance.
4. Due to guaranteed bias in coefficient, if the null hypothesis of no effect is true, with probability tending to 1 as sample size converges to infinity, **one will falsely reject the null**.

Machine learning

Machine learning is garnering huge attention in the press.



The Economist

Machine learning

Of prediction and policy

Google says machine learning is the future. So I tried it myself

theguardian

Apple buys Turi machine-learning startup for a reported $200m

nature

Google AI algorithm masters ancient game of Go

Deep-learning software defeats human professional for first time.

New brain map developed using real patients and machine learning

Machine learning

Machine learning primarily deals with training algorithms that can predict an outcome based on input features.

- Examples: self-driving cars, text recognition, image recognition, Alpha-Go

These algorithms are able to predict much better than the tools developed in classical statistics (e.g., prediction based on parametric regression models) owing to their flexibility and ability to adapt to underlying data.

However, this does not mean that we no longer have a need for statistics!

Machine learning

Consider the (possibly apocryphal) story of the US Army using neural networks to train an algorithm to recognized camouflaged tanks.

- Training data consisted of pictures of tanks camouflaged in trees and trees without tanks.

The classifier was found to have incredible predictive accuracy on the training data, but essentially no accuracy when used in practice.

It turns out, all the tank pictures were taken on a sunny day and all the control pictures were taken on a cloudy day and this is what the neural net had learned.

- Understanding sampling and experimental design is important!

Machine learning

Prediction is not always of scientific interest. Even in settings where prediction is of interest, we still need statistics to assess the performance of our predictions (and corresponding uncertainty!).

Nevertheless, these methods for learning from data are exciting and have been shown time and again to perform far better than the tools in our classical statistics tool box.

The challenge facing modern statisticians (and you, as students in this class) is how to develop statistical methods aimed at answering specific scientific questions of interest, while utilizing state-of-the-art algorithms.