Lecture Outlines for Prob-Stats

Maxwell B. Stinchcombe

Fall Semester, 2018

Contents

Preface

This course will start with an introduction to measure theoretic probability. We will first use this for a set of limit theorems, and for decision problems, both static and dynamic, in the face of uncertainty. We will then spend time on parametric classes of probability distributions and the properties of estimators, including: bias and expected squared error; sufficiency, completeness, minimality and the Cramer-Rao lower bound. This leads directly to hypothesis testing, the main language used in communicating statistical results both within and across fields. We will then give some of the introductory results for non-random samples — especially those arising from selection biases — that are now the focus much of current econometric and applied research. We will end the course with the beginnings of decision theory for such problems.

We will use the following sources. There is overlap between them, and you may use whichever you want when there are multiple coverages.

C&B is

> G. Casella and R. L. Berger. *Statistical Inference.* The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, second edition, 2002.

HB is

> H. J. Bierens. *Topics in Advanced Econometrics.* Cambridge University Press, Cambridge, 1994. Estimation, testing, and specification of cross-section and time series models.

CSZ is

> D. Corbae, M. B. Stinchcombe, and J. Zeman. *An Introduction to Mathematical Analysis for Economic Theory and Econometrics.* Princeton University Press, Princeton, NJ, 2009.

We will also use a variety of papers through the semester, they are listed below and I will e-mail them as they become relevant.

CHAPTER I

Why We're Doing This

A. The Basic Problem

A random number (vector, object) X will happen in the future, say at time $t = n + 1$. Before it happens, now at time n, we must choose an action $a \in A$. Preferences are such that we wish to maximize $E\,u(a, X)$, that is, to solve the problem in

$$V(\mu) := \max_{a\in A} \int u(a, x)\, d\mu(x) \tag{I.1}$$

where μ is the distribution of X. We have seen previous realizations of X, aka the **data**, $X_1, \ldots, X_n$. For now assume that X_t, $t = 1, \ldots, n+1$, take values in $\mathbb{R}$. The distribution of X is a point μ in $\Delta(\mathbb{R})$, which denotes the set of distributions on $\mathbb{R}$.

B. The Two Approaches

There are two main approaches. Both are variants of the idea that the past is predictive of the future.

B.1. Parametric.

Assume/guess/intuit/have a vision telling you that $X_1, \ldots, X_n$ are independent and identically distributed (iid) with distribution P_θ for some $\theta \in \Theta$ where $\Theta \subset \mathbb{R}^k$ is a set of parameters. For the best estimate $\widehat{\theta}_n = \widehat{\theta}(X_1, \ldots, X_n)$ and then choose the a that solves

$$V(P_{\widehat{\theta}_n}) = \max_{a\in A} \int u(a, x)\, dP_{\widehat{\theta}_n}(x). \tag{I.2}$$

The **statistical model** is $\{P_\theta : \theta \in \Theta\}$. It is **correctly specified** if there exists a "true" $\theta^\circ \in \Theta$ such that the X_t, $t = 1, \ldots, n+1$ are iid P_{θ° and $\widehat{\theta}_n \to \theta^\circ$, then for n large enough that $\widehat{\theta}_n$ is close to θ°, we can hope that $V(P_{\theta^\circ})$ is close to $V(P_{\widehat{\theta}_n})$.

To study: properties of $\{P_\theta : \theta \in \Theta\}$; $\widehat{\theta}_n \to \theta^\circ$; $V(P_{\theta^\circ}) \to V(P_{\widehat{\theta}_n})$; and, most crucially for economics, what can we say if the statistical model is wrong, that is, **mis-specified**?

One of the common ways that the model can be wrong is that for large n (aka the long-run distribution of the data), $\frac{1}{n}\sum_{t=1}^n 1_E(X_n) \simeq P_{\theta^\circ}(E)$, but X_t is not independent of X_{t+1}. Since the average correlation of times series of US data is more than 0.9, this is a frequent occurence.

B.2. Empirical.

Assume/guess/intuit/have a vision telling you that $X_1, \ldots, X_n$ are independent and identically distributed (iid) with distribution $\mu \in \Delta(\mathbb{R})$. Form the **empirical estimator** $\widehat{\mu}_n(E) := \frac{1}{n}\sum_{t=1}^n 1_E(X_n)$ and then chose the a that solves

$$V(\widehat{\mu}_n) = \max_{a\in A} \int u(a, x)\, d\mu_n(x) = \frac{1}{n}\sum_{t=1}^n u(a, X_n). \tag{I.3}$$

To study: $\widehat{\mu}_n \to \mu$; $V(\widehat{\mu}_n) \to V(\mu)$; comparison with the parametric approach; time series issues.

Weighted rolling averages are a common solution to the time series issues, but that means that the questions now include, at a minimum, both "What set of weights?" and "How do we detect if the time series structure has changed?"

C. The Conditional Problem

Now suppose Y_{n+1} is the random variable that will enter in the problem

$$\max_{a \in A} E(u, Y_{n+1}) \tag{I.4}$$

that the data is $(X_0, Y_1), \ldots, (X_{n-1}, Y_n), X_n$, and that X_{t-1} has predictive value for Y_t.

C.1. Parametric. Assume that (X_{t-1}, Y_t) is iid P_θ, $\theta \in \Theta$, estimate $\widehat{\theta}_n$, calculate the conditional distribution of Y_{n+1} given X_n and $P_{\widehat{\theta}_n}(\cdot|X_n)$, then solve

$$V(P_{\widehat{\theta}_n}(\cdot|X_n) = \max_{a \in A} \int (u, y)\, d\, P_{\widehat{\theta}_n}(y|X_n). \tag{I.5}$$

We have the same issues here, but in a more complicated format: convergence of the estimator; correctness of specification; times series problems.

C.2. Empirical. Assume enough to guarantee that the function $x \mapsto a^*(x)$ that solves

$$\max_{a \in A} E\left(E\left(u(a(x), Y)|X = x\right)\right) \tag{I.6}$$

is well-enough behaved to be approximated by some numerical procedure that you know about.

Typical: find a sequence of classes of functions $\mathbb{A}_n$, possibly data determined, with the property that $d(a^*, \mathbb{A}_n) \to 0$; let $\widehat{\mu}_n$ be the empirical distribution of (X_{t-1}, Y_t), $t = 1, \ldots, n)$, solve the problem

$$\max_{a(\cdot) \in \mathbb{A}_n} \int u(a(x), y)\, d\widehat{\mu}_n(x, y), \tag{I.7}$$

and use the solution $\widehat{a}^*_n$ at the observation X_n to calculate the optimal action $\widehat{a}^*_n(X_n)$.

D. Partial Information

In the following examples, it seems likely that observational data will not, indeed cannot, reveal the information we care about. The essential problem is that people with different and *unobservable* characteristics make different choices, and their characteristics are correlated with, or even causal, for the outcomes we care about.

- Evaluating the effectiveness of a job training program that people must volunteer themselves into is confounded by the observation that the characteristics of the people who choose to enter the program are different than those who choose not to enter the program.
- Evaluating school or teacher effectiveness is confounded by the observation that better schools and teachers are chosen by families that put more emphasis on their childrens' academic performance. And the other resources that these families provide to their children also have long-run benefits.

- Evaluating the value of a college degree is confounded by the observation that the people choosing to go to college are different than those who do not choose. Some of these differences involve some notion of 'raw ability,' but in the US, other differences involve the cultural habits of the upper middle class.
- Evaluating the value of mosquito netting in malarial regions has found the paradoxical result that exposure to the disease (as measured by a quick and cheap blood test) is correlated with a lower willingness to pay for mosquito netting. The initally unobserved variable was history of family income, which was lower, often dangerously lower, for families stricken by the disease.
- Evaluating religion-based rehabilitation programs in prisons is confounded by the observation that the people choosing to enter them have unobservably different motivations and previous life experiences.
- Evaluating the effects of lead exposure on academic performance in elementary school children has to deal with an *errors in variables* problem, the blood tests for lead exposure were sporadic and lead leaves the blood stream (a half life of 2 or 3 months), to settle into fat, especially the fat in the nervous system.

This class of problems, those with unobservables confounding the estimation, is central to much of modern econometrics and applied economic research.

E. Causality and Useful Information

Sources

D. Deutsch. *The Beginning of Infinity: Explanations that Transform the World.* Penguin Books, New York, 2011.

E. Durkheim. *Suicide: A Study in Sociology.* Routledge, 2005.

Deutsch provides a lovely coverage of what goes into a "good" explanation, a causal explanation. There are two basic parts. Minimality — one cannot change any part of the explanation and have it be true, and reach — the explanation gives solutions beyond those that it was invented to solve.

Durkheim believed that social forces, such as *anomie*, had as much reality as the entities that physicists study, e.g. gravity and light. One key difference that tells you that the social sciences are very different is that theories in the social sciences do not give functional forms, e.g. the inverse square laws for gravity and light.

There are no explanations of why things happen in economics that are "good" in Deutsch's sense. There is a big difference between failing to be "good" and failing to be useful. Consider logistic regressions and the study of e.g. bus ridership/opposition to the death penalty/suicide.

Consider the **logistic** cdf $\Phi(x) = \frac{e^x}{1+e^x}$. Logistic regression supposes that the probability that something happens, e.g. that someone takes the bus to work, or is against the death penality, or that someone commits suicide, is a function of their characteristics, given in a vector $\boldsymbol{x} \in \mathbb{R}^k$. The logistic enters in that the function is assumed to take the form

$$f(\boldsymbol{x};\theta) = \Phi(\beta_0 + \beta'\boldsymbol{x}), \tag{I.8}$$

$\theta = (\beta_0, \beta) \in \mathbb{R}^{1+k}$.

After observing many people, $i = 1, \ldots, n$, with measured characteristics $\boldsymbol{x}_i$, and behavior $Y_i = 1$ if bus/opposition to death penalty/suicied, $Y_i = 0$ otherwise, one estimates $\widehat{\theta}$ by solving

$$\min_{\theta \in \mathbb{R}^{1+k}} \textstyle\sum_{i=1}^{n} (Y_i - f(\boldsymbol{x}_i; \theta))^2. \tag{I.9}$$

Even though we are certain that the causality does not take the form given in (I.8), the resulting $\widehat{\theta}$ is useful if $f(\cdot; \widehat{\theta})$ is a good predictor of Y_i. If it is a good predictor, then: for bus systems, it can help design routes and schedules; for political advertising it can help avoid sending policy platform points that would cause voters to vote against the candidate; for suicide prevention efforts, it can help identify who should receive help.

One easy extension of this idea comes from thinking about e.g. opposition to the death penalty. One might suppose that there is a positive fraction, ρ, of the population that finds the death penalty so morally objectionable that they will not support it in any circumstances. To capture this, change the functional form to

$$f(\boldsymbol{x}; \theta, \rho) = (1 - \rho)\Phi(\beta_0 + \beta' \boldsymbol{x}), \tag{I.10}$$

and estimate both $\widehat{\rho}$ and $\widehat{\theta}$ as before.

F. Summary

What we need is a systematic way to talk about random variables, about distributions, conditional probabilities, and sequences of optimal choices. The essential model of randomness in all of the models/theories in economics is due to Kolmogorov. It goes by the name of **measure theoretic probability**.

CHAPTER II

Measure Theoretic Probability

Possible sources for this material are C&B, Ch. 1, HB, Ch. 1 and 2, and CSZ, Ch. 7.

A. Basics

A.1. Probability spaces. $(\Omega, \mathcal{F}, P)$, $\Omega \neq \emptyset$, $\mathcal{F}$ a σ-field (read "sigma-field"), $P : \mathcal{F} \to [0, 1]$ a countably additive probability.

A.2. Random variables. $X : \Omega \to \mathbb{R}$ with $X^{-1}((a, b]) \in \mathcal{F}$, $-\infty \leq a < b < \infty$.

A.3. Distributions. CDF for X defined by $F_X(x) = P(X \in (-\infty, x])$.

A.4. Interpretation. Three interpretations of probability: long-run frequency; subjective; fair bets.

A.5. Limit Constructions. Sums and products, almost always, infinitely often, convergence sets for sequences of random variables.

A.6. The Integral. For simple functions, then for limits.

B. Some Limit Theorems

B.1. Weak Law. Weak Law of Large Numbers

B.2. Borel-Cantelli. The Borel-Cantelli Lemma.

B.3. Dominated Convergence. Lebesgue's dominated convergence theorem.

B.4. Fatou's Lemma. Fatou's Lemma.

B.5. The Strong Law. The strong law of large numbers (SLLN).

B.6. Two 0-1 Laws. The Kolmogorov 0-1 law and the Hewitt-Savage 0-1 law.

CHAPTER III

Conditional Expectations

Sources, HBCh. 3, C&BCh. 1, CSZCh. 8.1.

A. Basics

A.1. Predictors. Mean squared loss best predictors.

A.2. Beliefs.

A.3. Exchangeability. Urns and hysteresis. Exchangeable sequences and de Finetti's exchangeability Theorem.

B. Choice Theory Under Risk

B.1. Expected Utility. Prefer rv X to Y if $E\,u(X) > E\,u(Y)$.

B.2. Curvature of $u(\cdot)$. Source

L. Eeckhoudt and H. Schlesinger. Putting risk in its proper place. *American Economic Review*, 96(1):280–289, 2006.

Risk aversion, prudence, temperance.

B.3. Optimal Action. $a^*(\mu) = \arg\max_{a \in A} \int u(a,x)\,d\mu(x)$.
$V_u(\mu) := \max_{a \in A} \int u(a,x)\,d\mu(x)$.

B.4. The Value of Information. Source CSZCh. 8.12 and

D. Blackwell. Equivalent comparisons of experiments. *Ann. Math. Statistics*, 24:265–272, 1953.

D. Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 93–102, Berkeley and Los Angeles, 1951. University of California Press.

The Blackwell ordering on information structures.

C. Basic Dynamic Programming

C.1. Search Models. Some comparative statics of search.

CHAPTER IV

Classes of Distributions

Source, C&BCh. 2 and 3

A. Transformations

X has cdf $F_X(\cdot)$, the cdf of $Y := g(X)$ is given by $F_Y(r) = P(X \in g^{-1}((-\infty, r])$. When g is monotonic increasing or decreasing, or has a small number of regions in which it is monotonic, can go directly to $F_Y(\cdot)$.

A.1. Moment Generating Functions. Stone-Weierstrass theorem.

A.2. Characteristic Functions. Stone-Weierstrass theorem and inversion.

B. The Two Basic Limit Distributions

The Gaussian and the Poisson parts of the Central Limit Theorem (CLT).

C. Infinite Divisibility

Cauchy, Poisson sums.

C.1. Exponential Classes. A tour of some of your worst memories from integral calculus.

D. Statistics and Estimators

Bias.
Mean squared error.
Cramer-Rao lower bound, see especially

R. R. Bahadur. On fisher's bound for asymptotic variances. *The Annals of Mathematical Statistics*, 35(4):1545–1552, 1964.

Sufficiency, minimality, completeness.

E. Decisions with Estimation Uncertainty

Source

R. W. Klein, L. C. Rafsky, D. S. Sibley, and R. D. Willig. Decisions with estimation uncertainty. *Econometrica: Journal of the Econometric Society*, pages 1363–1387, 1978.

CHAPTER V

Non-Random Samples

A. Duration Estimators

Source,

S. W. Salant. Search theory and duration data: a theory of sorts. *The Quarterly Journal of Economics*, 91(1):39–57, 1977.

Memoryless.
Hazard rates.
Population diversity.

B. Selection Bias

Source, Chapter 1 in

C. F. Manski. *Identification for prediction and decision.* Harvard University Press, 2009.

Treatment effects.
Manski bounds.
Tighter bounds.

C. Set Identification

C.1. Definition. $\theta \in A_n(X_1, \ldots, X_n) \subset \Theta$, $A_n \to A$, and A is not a singleton set.

C.2. Decision Theory. Decision theory for set-identified distributions.

Bibliography

R. R. Bahadur. On fisher's bound for asymptotic variances. *The Annals of Mathematical Statistics*, 35(4):1545–1552, 1964.

H. J. Bierens. *Topics in Advanced Econometrics*. Cambridge University Press, Cambridge, 1994. Estimation, testing, and specification of cross-section and time series models.

D. Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 93–102, Berkeley and Los Angeles, 1951. University of California Press.

D. Blackwell. Equivalent comparisons of experiments. *Ann. Math. Statistics*, 24: 265–272, 1953.

G. Casella and R. L. Berger. *Statistical Inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, second edition, 2002.

D. Corbae, M. B. Stinchcombe, and J. Zeman. *An Introduction to Mathematical Analysis for Economic Theory and Econometrics*. Princeton University Press, Princeton, NJ, 2009.

D. Deutsch. *The Beginning of Infinity: Explanations that Transform the World*. Penguin Books, New York, 2011.

E. Durkheim. *Suicide: A Study in Sociology*. Routledge, 2005.

L. Eeckhoudt and H. Schlesinger. Putting risk in its proper place. *American Economic Review*, 96(1):280–289, 2006.

R. W. Klein, L. C. Rafsky, D. S. Sibley, and R. D. Willig. Decisions with estimation uncertainty. *Econometrica: Journal of the Econometric Society*, pages 1363–1387, 1978.

C. F. Manski. *Identification for prediction and decision*. Harvard University Press, 2009.

S. W. Salant. Search theory and duration data: a theory of sorts. *The Quarterly Journal of Economics*, 91(1):39–57, 1977.