Econ 240A (1st Half)
Section 1: Fall 2018
Friday, Aug 31

Fengshi Niu*

Contents

[*] Thanks to Mikkel Sølvsten and Matias D. Cattaneo for letting me use their old notes as an inspiration.

1 Big Picture

In this course we will be mainly working with statistics, which in turn is concerned with both the collection of data and their analysis and interpretation (Lehmann and Casella 1998). We will not deal with data collection issues in this class but rather take the data as given, and ask what they have to tell us. The answer depends not only on the data, on what is being observed, but also on background knowledge of the situation; the latter is formalized in the assumptions with which the analysis is entered. There are different ways to approach the analysis in order to extract useful information. In general, we can distinguish between two types of analysis, which in turn nest different assumptions:

1. *Exploratory Data Analysis (EDA)*: in this approach the data is taken as given and almost no assumptions are imposed. The main goal of this approach is to extract regularities, summarize and organize the data. The methods and implementation of this approach are discussed in STAT-215A in some years, as well as in many applied courses in economics. In this class, however, we will work directly with more structure by imposing assumptions on the data generating process. This in turn implies that no EDA will be discussed in this course. (Notice that part of this approach preceeded what is now called *data mining*, something that we will not discus here either.) A historical reference for this approach is Tukey (1977).

2. *Confirmatory Data Analysis (CDA)*: after an EDA approach has been carried out, it is imperative to confirm and validate the findings. Moreover, in general (especially in economics) many hypotheses cannot be tested directly without imposing more structure (assumptions) to the problem at hand. Within the CDA framework, there are two (complementary) approaches:

   (a) *Classical Inference and Decision Theory*: here the data is now assumed to be realizations of some underlying random variables which follow a joint distribution, P, belonging to some known class $\mathcal{P}$ of distributions. In general, the distributions are indexed by a parameter $\theta \in \Theta$, not necessarily real-valued or scalar, where Θ is some index set (usually known as *parameter space*). So we have:

   $$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

   The aim of this analysis is to specify a plausible value for θ. This is the problem of point estimation. Alternatively, we could be less ambitious and look for a subset of Θ that may contain the true θ. This alternative strategy relies on the estimation of *confidence intervals*. Using either approach, we may perform *hypothesis testing*.

   (b) *Bayesian Analysis*: here not only the data but also the parameter, θ, is assumed to be random. Even though it is not observable, its distribution is assumed to be known. In particular, this distribution, known as a *prior distribution*, is used in combination with the distribution of the data to obtain the *posterior distribution* (the distribution of θ conditional on the data). The posterior distribution summarizes what can be said about θ on the basis of the assumptions made and the data.

In this class we will deal almost exclusively with Classical Inference Theory and we will present most results briefly with the main goal of providing the necessary background to undertake a more formal and technical discussion of econometrics. In terms of where this part of knowledge fits in a full data analysis pipeline, it's helpful to look at Figure 6 on the last page of this notes.

Before we begin the discussion of the main results in statistics we need to review the theory of probability. This is carried out in the next following sections.

2 Probability Space

A *probability space* consists of three core parts $(\Omega, \mathcal{F}, P)$.

1. $\Omega =$ the set of all possible outcomes or the universal domain.
2. $\mathcal{F} =$ collection of sets/events A for which $P(A)$ is defined. It is assumed that $\mathcal{F}$ contains the whole set Ω and is closed under finite and countable set operations (referring to $\cup$, $\cap$, ^{c}), and that $\Omega \in \mathcal{F}$. Technically, such a collection of sets $\mathcal{F}$ is known as a σ-algebra.
3. $P =$ probability function with domain $\mathcal{F}$ satisfying the following axioms (due to Kolmogorov):
   - $0 \le P(A) \le 1$
   - $P(\Omega) = 1$
   - $P(A \cup B) = P(A) + P(B)$ if A and B are are disjoint (mutually exclusive)
   - P also satisfies the countable additivity axiom.

     $$P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) \text{ if the } A_n \text{ are mutually exclusive}$$

     meaning $A_i \cap A_j = \emptyset$ for all $i \neq j$.
   - P may also be called a probability distribution on Ω, with $\mathcal{F}$ understood from context, usually as the smallest σ-algebra containing all events of interest, e.g. if $\Omega = \mathbb{R}$, then $\mathcal{F}$ is the *Borel σ-algebra* generated by intervals (or open sets, or ...), if $\Omega = \mathbb{R}^k$, then $\mathcal{F}$ is the Borel σ-algebra generated by intervals (or open sets, or ...). Though these measure theory niceties will not play a major role in this course, it's useful to have a basic understanding of them. Several advantages of using measure theory includes:
     (a) The language of measure theory is necessary for stating many results correctly, e.g. expectation, conditional expectation, notion of convergence, etc.
     (b) The notation of measure theory allows one to merge results for discrete, continuous, and mixed random variables.

3 Conditional Probability and Independence

In this section we explore some important results related to conditional probability and independence. Many of these results will be very useful later in the second half of the class. We begin with the basic definition of conditional probability. In this section we assume that there exists a background probability space $(\Omega, \mathcal{F}, P)$.

Definition 3.1. *Given $A, B \in \mathcal{F}$, and assuming that $P(B) > 0$, then the **conditional probability of** A **given** B is defined as*

$$P\left(A \middle| B\right) = \frac{P(A \cap B)}{P(B)}.$$

The following exercise shows that the set function $P\left(\cdot \middle| B\right)$ is a well-defined probability measure. Using this concept we have the following useful theorem.

Theorem 3.1. *(Total Probability) Let $\{B_n\}_{n=1}^{\infty} \in \mathcal{F}$ be a partition (i.e.,$\{B_n\}_{n=1}^{\infty}$ are pairwise disjoint and $\cup_{n=1}^{\infty} B_n = \mathcal{F}$), and let $A \in \mathcal{F}$ with $P(A) > 0$, then*

$$P(A) = \sum\nolimits_{n=1}^{\infty} P\left(A \middle| B_n\right) \cdot P(B_n).$$

Proof. By assumption we have $\cup_{n=1}^{\infty} B_n = \mathcal{F}$. Notice that $A = A \cap \mathcal{F} = A \cap (\cup_{n=1}^{\infty} B_n) = \cup_{n=1}^{\infty}(A \cap B_n)$, and as a direct consequence we have (why?)

$$(A \cap B_i) \cap (A \cap B_j) = \varnothing, \text{ for all } i \neq j.$$

Using countable additivity, we have

$$P(A) = P\left(\cup_{n=1}^{\infty}(A \cap B_n)\right) = \sum\nolimits_{n=1}^{\infty} P(A \cap B_n) = \sum\nolimits_{n=1}^{\infty} P\left(A \middle| B_n\right) \cdot P(B_n),$$

where the last equality follows by the definition of conditional probability. □

The next theorem is a well-known theorem in probability and statistics. We will not use it much but it is extremely easy to prove using the last result.

Theorem 3.2. *(Bayes' rule) Let $\{B_n\}_{n=1}^{\infty} \in \mathcal{F}$ be a partition, and let $A \in \mathcal{F}$ with $P(A) > 0$, then for each $i = 1, 2, \cdots$ we have*

$$P\left(B_i \middle| A\right) = \frac{P(B_i \cap A)}{\sum_{n=1}^{\infty} P(B_n \cap A)} = \frac{P\left(\left(A \middle| B_i\right)\right) \cdot P(B_i)}{\sum_{n=1}^{\infty} P\left(A \middle| B_n\right) \cdot P(B_n)}.$$

The last definition will be one of the most useful definition for the rest of this course.

Definition 3.2. *Two events $A, B \in \mathcal{F}$ are statistically independent if*

$$P(A \cap B) = P(A) \cdot P(B).$$

4 Random Variables

A *random variable* is technically defined as a measurable function $\omega \mapsto X(\omega)$ from $\omega \in \Omega$ to the real space $\mathbb{R}$ (or any other measurable space). A random variable X by default maps to $\mathbb{R}$. However, you can have a random vector X with values in $\mathbb{R}^k, k = 1, 2, 3, \ldots$. This goes on to random sequences or random functions, all that we're doing is modifying the target space.

The Ω space is a background space of all possibilities. We want to map that to a target space through a function. However we need to be able to measure the probability of an event in our new space. Therefore our random variable X must be *measurable.*

Measurability

Measurability for real random variables means that we need to be able to compute the probability $P(X \leq x) = P(\{\omega : X(\omega) \leq x\})$ for every $x \in \mathbb{R}$. This requires that $\{X \leq x\} := \{\omega : X(\omega) \leq x\}$ is in the σ-algebra $\mathcal{F}$ for very $x \in \mathbb{R}$.

Discrete random variables

X is *discrete* if the range of X is a finite or countably infinite set of possible values X. The *distribution of* X on the target space is then defined by the *probability mass function* $P(X = x)$ as x ranges over the finite or countably infinite set of possible values X. By the axioms of the probability function P, this function of x is non-negative, and it must sum to 1 over all x.

(Absolutely) continuous random variables

Such variables X can take any value in $\mathbb{R}$. Technically, it is not X that is continuous, but the distribution of X, in a sense made precise later.

5 Distributions

The *distribution of* X is created by projecting the P from the Ω space on to another space via the function X. For a real random variable X, the *cumulative distribution function of* X (often abbreviated to *cdf* or *distribution function* is the function usually denoted F or F_X which is defined as:

$$F(x) := F_X(x) := P(X \leq x)$$

This function F of a real variable x is defined for all real random variables X, whether continuous or discrete or neither. The subscript X is just used to indicate that F is derived from X. General properties of a cdf F, which follow from the axioms of probability, include

1. $F(\cdot)$ is not decreasing.

2. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
3. $F(\cdot)$ is right continuous, that is, $\lim_{z\downarrow x} F(z) = F(x)$, for all $x \in \mathbb{R}$.
4. Denote $F_X(x-) = \lim_{z\uparrow x} F_X(z)$. Then $F_X(x-) = P(X < x)$.
5. $P(X = x) = F_X(x) - F_X(x-)$.

Note that X is called (absolutely) continuous if F_X is continuous. It is important to know whether you're dealing with a discrete or continuous random variable. If you have X with a density f_X, meaning

$$f_X(x) = \frac{d}{dx}F_X(x) \qquad F_X(x) = \int_{-\infty}^{x} f_X(y)dy$$

then $F = F_X$ is continous, even differentiable with derivative $F'(x) = f(x)$. Then all point probabilities $P(X = x)$ are 0. In general $P(X = x)$ is the size of the jump of F_X at x.

Example: Uniform$[0, 1]$ If X falls uniformly between [0,1], with the probability of an interval defined by its length, the distribution of X is called uniform$[0, 1]$. The cdf is

$$F_X(x) = x \text{ if } 0 \le x \le 1$$

with $F_X(x) = 0$ for $x < 0$ and $F_X(x) = 1$ for $x > 1$.

You can differentiate the cumulative distribution function in order to get the probability density function: which shows the probability density of X at x as a function of x. Note there are two exceptional points 0 and 1 where the definition of the density is somewhat arbitrary, either 0 or 1 depending on which side you take the derivative on. But the value of the density at any finite number of points is of no importance, because probabilities are defined by integration of the density. It is also possible to have a continuous cdf that does not have a density in this sense, but that will not concern us here.

It is often that we need to compute the distribution of a function or transformation of a random variable. For example, in economics we usually encounter latent variable models. The following discussion reviews the tools presented in lectures.

Univariate Transformations

We begin by considering the easy, but insightful, case of scalar random variables. As usual, we assume a background probability space $(\Omega, \mathcal{F}, P)$ and we assume that X is a random variable with cdf $F_X(x)$, which induces the probability space $(\mathbb{R}, \mathcal{B}, F_X)$. First recall that most of the interesting functions of X are random variable themselves (technically, the function g has to be a measurable map itself), and letting $Y = g(X)$ we have a third induced probability space $(\mathbb{R}, \mathcal{B}, F_Y)$. To save on notation we drop the subscripts on each probability measure.

Observe that $P(Y \in A) = P(g(X) \in A)$ and assuming that $g(\cdot)$ is invertible we have

$$P(Y \in A) = P(g(X) \in A) = P(\{x : g(x) \in A\}) = P\left(X \in g^{-1}(A)\right).$$

Notice that this result is, in fact, always true. However, what we just did does not give us a clear way of computing this new random variable. Sometimes we can use particular results such as the following theorem for the case of continuous random variables and strictly monotone transformations.

Theorem 5.1. *(*MONOTONE UNIVARIATE TRANSFORMATIONS*) Let X be a random variable with pdf $f_X(x)$ and let $Y = g(X)$. Assume the following holds:*

1. *g is a strictly monotone function,*
2. *$f_X(x)$ is continuous on the set $\mathbb{X} \equiv \{x : f_X(x) > 0\}$, and*
3. *$g^{-1}(y)$ has a continuous derivative on the set $\mathbb{Y} \equiv \{y : y = g(x) \text{ for some } x \in \mathbb{X}\}$.*

Then the pdf of $Y = g(X)$ is given by

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \left|\frac{d}{dy} g^{-1}(y)\right| \cdot f_X\left(g^{-1}(y)\right) \cdot \mathbb{I}\{y \in \mathbb{Y}\}.$$

We discussed in Lecture the intuition behind the derivation of this formula. Observe that this derivation is done using the cdf of Y and then we reexpress everything in terms of X. A similar expression can be found for the case of a discrete random variable.

6 Expectation

In the language of measure theory, *the expectation of a random variable X* is

$$\mathbb{E}(X) = \int X dP = \int X(\omega) dP(\omega) = \int_{\mathbb{R}} x dF_X(x).$$

More specifically, if X is a discrete random variable then for a function g from the range of X to $\mathbb{R}$ we define

$$\mathbb{E}(g(X)) := \sum_x g(x) P(X = x)$$

where the $\sum_x$ is over all possible values x of X. In particular, if X itself is real valued

$$\mathbb{E}(X) := \sum_x x P(X = x)$$

$$\mathbb{E}(X^2) := \sum_x x^2 P(X = x)$$

$$\mathbb{E}(|X - 3|) := \sum_x |x - 3| P(X = x)$$

and so on. In all these equations, $P(X = x)$ is the probability mass function of X.

Remark 6.1. *There is one catch if there are an infinite number of values for X. To define $Eg(X)$ we have to assume that the $\sum_x$ is* absolutely convergent*, meaning that*

$$\mathbb{E}(|g(X)|) := \sum_x |g(x)| P(X = x) < \infty$$

.

This definition for discrete X can be extended by approximation to random variables X that are not discrete. Details of this approximation are not treated here. In particular, if X has a density $f_X(x)$ then for g as above,

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

provided the integral is finite with $g(x)$ replaced by $|g(x)|$. Similarly if we have a pair of random variables (X, Y) with a joint density $f_{X,Y}(x, y)$ we can compute the expectation of a generic function $g(X, Y)$ as

$$\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy$$

Here the subscript X, Y is just used as a label to indicate the source of the joint probability density function $f_{X,Y}(x, y)$ whose arguments for integration are x and y. You should think of $f_{X,Y}(x, y)$ as the probability per unit area for values of (X, Y) near (x, y). This meaning is made precise by integration as in the above formula. The formula for g the indicator function of a region B in the plane gives the probability that $(X, Y) \in B$. So, basically you can keep adding variables and you just keep performing that same integration.

Some Properties of Expectations

If X, Y are defined on a common probability space, and both $E(X)$ and $E(Y)$ are well defined (by absolutely convergent sums or integrals) then $E(X + Y)$ is defined, and

$$E(X + Y) = E(X) + E(Y).$$

It is important that this *addition rule for expectations* holds whether or not X and Y are *independent* meaning

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$$

for all choices of intervals A and B. If X and Y are independent, then also $E(XY) = E(X)E(Y)$.

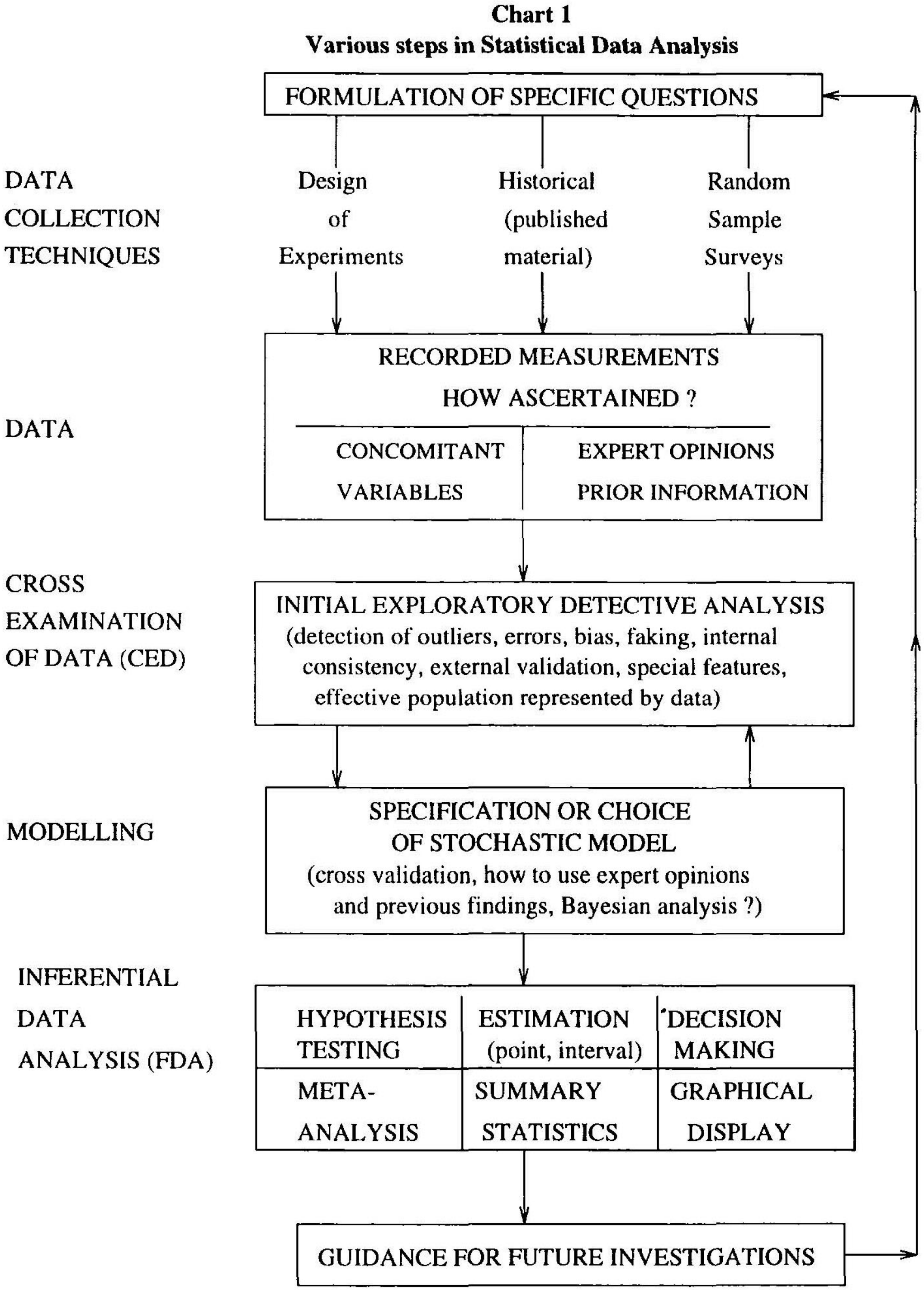
Chart 1
Various steps in Statistical Data Analysis
FORMULATION OF SPECIFIC QUESTIONS
DATA COLLECTION TECHNIQUES
Design of Experiments
Historical (published material)
Random Sample Surveys
DATA
RECORDED MEASUREMENTS
HOW ASCERTAINED ?
CONCOMITANT VARIABLES
EXPERT OPINIONS PRIOR INFORMATION
CROSS EXAMINATION OF DATA (CED)
INITIAL EXPLORATORY DETECTIVE ANALYSIS
(detection of outliers, errors, bias, faking, internal consistency, external validation, special features, effective population represented by data)
MODELLING
SPECIFICATION OR CHOICE OF STOCHASTIC MODEL
(cross validation, how to use expert opinions and previous findings, Bayesian analysis ?)
INFERENTIAL DATA ANALYSIS (FDA)
HYPOTHESIS TESTING
ESTIMATION (point, interval)
DECISION MAKING
META-ANALYSIS
SUMMARY STATISTICS
GRAPHICAL DISPLAY
GUIDANCE FOR FUTURE INVESTIGATIONS

Figure 1: Various steps in statistical data analysis from Rao 1997.