Econometrics Math Camp
Day Two: Even More Foundations of Statistics

Michael Droste

August 24, 2020

Introduction

- Welcome back!
- We have three main objectives for the today's half of econometrics math camp:
    1. Review the most important take-aways from Friday
    2. Review some additional material
    3. Tackle two important simulation-based exercises

Today's Outline

- Review from Day 1
    - Properties of Distributions
    - Properties of Expectations
    - Really Important Rules
- New topics
    - Variance and higher-order moments
    - Correlation and covariance
    - Useful Distributions
    - Estimators
    - Asymptotics

Continuous Random Variables

- Let X be a random variable. We say that X is a continuous random variable if (and only if) F_X can be written as:

  $$F_X(x) = \int_{-\infty}^{\infty} f_X(t)dt$$

  where f_X satisfies $f_X(x) \geq 0$ and $\int_{-\infty}^{\infty} f_X(t)dt = 1$.

- At the points where F_X is continuous, we have:

  $$f_X(x) = \frac{dF_x(x)}{dx}$$

- We call $f_X(x)$ the probability density function (or pdf) of X.

- The support of X is $S_X = \{x : f_X(x) > 0\}$

Cumulative Distribution Function

- Let X be a random variable. The cumulative distribution function (or cdf) or X, $F : \mathbb{R} \to [0, 1]$, is defined as:

  $$F_X(x) = P(X^{-1}(x)) = P(\{\omega \in \Omega : X(\omega) \leq x\})$$

  We often write the cdf of X as:

  $$F_x(x) = P(X \leq x)$$

- Loosely, the cdf of X is a function that tells you, for any given value of x, the probability that the random variable X takes on a value less than or equal to x.

Cumulative Distribution Functions: Quantiles

- The quantiles of a random variable X are given by the inverse of its cumulative distribution function.

- The quantile function is:
$$Q(u) = \inf\{x : F_X(x) \geq u\}$$
If F_X is invertible, then:
$$Q(u) = F_X^{-1}(u)$$

- Cool fact: for any function F that satisfies the handy properties on the previous slide, we can construct a random variable whose cumulative distribution function is F.

Expectations of Continuous Random Variables

- Let X be a continuous random variable. Its expectation (or expected value) is defined as:

  $$E[X] = \int_{S_x} x f_X(x) dx$$

  if $\int_{S_x} |x| f_X(x) dx < \infty$. Otherwise, the expectation does not exist.

- Note that expectations (still) play nicely with transformations of (continuous) random variables. For instance, let $g : \mathbb{R} \to \mathbb{R}$. Then:

  $$E[g(X)] = \int_{S_x} g(x) f_X(x) dx$$

Expectations as Linear Operators

- Expectations are a linear operator. What does this mean? Let X be a random variable, $a \in \mathbb{R}$ a constant, and $g_1(\cdot)$, $g_2(\cdot)$ be real-valued functions. Then:
    1. $E[a] = a$
    2. $E[ag_1(X)] = aE[g_1(X)]$
    3. $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$

Conditional Expectations

- Let X and Y be random variables with a joint density $f_{X,Y}(x,y)$. The conditional expectation of Y given X=x is:

$$E[Y|X=x] = \int_{S_Y} y f_{Y|X}(y|x) dy$$

- Note that this is a function of x, and is sometimes called the conditional expectation function or regression function. It is sometimes useful to denote the CEF of a variable Y as a function of x as $\mu_Y(x)$.

Conditional Expectations: Properties

- Conditional expectations are going to show up again and again in this course. They are intimately related to regression analysis.

- One incredibly useful property of the CEF is called the CEF decomposition. Let Y be a random variable. We can write:

  $$Y = E[Y|X] + \epsilon$$

  where ϵ is mean independent of X and is therefore uncorrelated with *any* function of X.

- You will try to prove this property in the second breakout session.

Law of Iterated Expectations

- The law of iterated expectations is a really, really useful law for manipulating conditional expectations. It will show up all the time in your homework in a variety of settings.

- One form of the law of iterated expectations can be stated as:

$$E_Y[Y] = E_X E_{Y|X}[Y]$$

  where E_X denotes the expectation taken with respect to the marginal density of X and $E_{Y|X}$ denotes the expectation taken with respect to the conditional density of Y given X.

- The way I think about this rule is as follows. Suppose I have the expectation of the conditional density of Y given X. If I take this expression and apply an expectation operator with respect to X, then I am left with the expectation of the marginal density of Y.

Conditional Probability: Law of Total Probability

- Consider K disjoint events C_k that partition the sample space Ω; that is, $C_i \cap C_j = \emptyset$ for all $i \neq j$ and $\cup_{i=1}^{K} C_i = \Omega$. Let A be some event.

- The law of total probability states that we can write $P(A)$ in terms of $P(A|C_i)$ and $P(C_i)$ in a way that 'adds up'.

$$P(A) = \sum_{i=1}^{K} P(A|C_i)P(C_i)$$

Conditional Probability: Bayes Rule

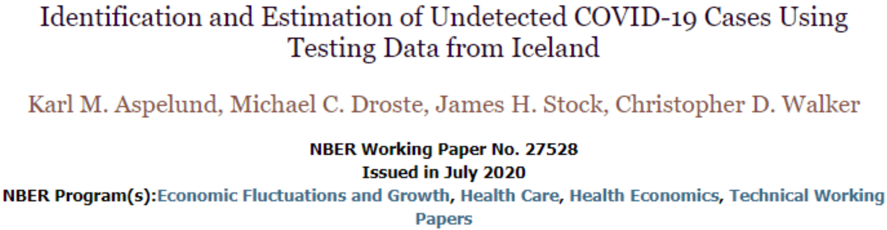
- Given two events A, B, Bayes' rule (sometimes seen as Bayes' law) relates the conditional probabilities $P(A|B)$, $P(B|A)$ and the marginal probabilities $P(A)$, $P(B)$.

- One simple formulation of Bayes law can be expressed as:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- Proof? (hint: use the multiplication rule)

- You can write entire papers that are basically just Bayes' law...



Identification and Estimation of Undetected COVID-19 Cases Using Testing Data from Iceland

Karl M. Aspelund, Michael C. Droste, James H. Stock, Christopher D. Walker

NBER Working Paper No. 27528
Issued in July 2020
NBER Program(s): Economic Fluctuations and Growth, Health Care, Health Economics, Technical Working Papers

Moments of a Distribution

- Our first piece of new content today will be to describe higher-order moments of a distribution.

- The k-th moment of a random variable X is $E[X^k]$. The first moment, $E[X]$, is just the mean.

- The k-th centered moment of a random variable X is $E[(X - E[X])^k]$

- The 2nd centered moment is called the variance, and is represented as $Var[X] = E[(X - E[X])^2]$.

Covariance

- Let X and Y two random variables with joint density $f_{X,Y}(x, y)$.

- The covariance of X and Y is:

$$\begin{aligned} Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

- The covariance operator $Cov(X, Y)$ is a linear operator, which basically means it distributes and we can pull out constants. That is, for any constants a, b and any three random variables X, Y, Z, the covariance operator distributes like so:

$$Cov(X, aY + bW) = aCov(X, Y) + bCov(X, W)$$

- Note that the covariance of a variable with itself is simply the variance.

Conditional Variance

- Because the variance is defined in terms of expectations, we can think about the idea of conditional variance, which is defined in a way similar to the conditional expectation.

- Let X and Y be random variables.

$$Var[Y|X] = E[(Y - E[Y|X])^2|X]$$

- Intuitively, the conditional variance tells us how the variance of Y changes with X.

Moments for Vectors

- It will be useful (and easy) to generalize the idea of a moment to a vector of random variables. So let $X = (X_1, ..., X_n)$ be an n-dimensional vector of random variables.

- Its mean vector is:

$$E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}$$

- Its covariance matrix is:

$$Var(X) = \Sigma$$

  where Σ is an $n \times n$ matrix whose ij-th entry is $\Sigma_{ij} = Cov(X_i, X_j)$.

- You will see a lot of covariance matrices this year. They're important objects! Elie will tell you more.

Useful Distributions

- Next, it will be helpful to briefly walk through some common distributions you will see this year.

- This list isn't exhaustive, and some distributions we won't cover (e.g. extreme value distributions) are really important in some applications.

- We will focus on the univariate case, but all of these generalize to multivariate (i.e. joint) distributions.

Useful Distributions: Uniform Distribution

- The first distribution we'll think about is a uniform distribution.

- The key property of a uniform distribution is that any value in its support is equally likely to be realized.

- Let X be a continuous random variable drawn from a uniform distribution over the interval $[a, b]$, with $b > a$. Then the pdf of x is $f_X(x) = \frac{1}{b-a}$ for $x \in [a, b]$ and 0 otherwise.

- We write $X \sim U[a, b]$.

- $E[X] = \frac{1}{2}(a+b)$, $Var[X] = \frac{1}{12}(b-a)^2$

Useful Distributions: Normal Distribution

- Next is everyone's favorite, the normal distribution. We say that a random variable X is normally distributed if the pdf of X looks like:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- We write $X \sim N(\mu, \sigma^2)$. The special case where $X \sim N(0, 1)$, with mean 0 and variance 1, is called a standard normal distribution. Note that if $X \sim N(\mu, \sigma^2)$, we can always write $X = \mu + \sigma Z$ for $Z \sim N(0, 1)$.

- $E[X] = \mu$, $Var[X] = 1$.

- Many estimators exhibit a property called "asymptotic normality" - which we will precisely define shortly - that makes your life a whole lot easier. This is fundamentally why normality is going to be so important for you this year.

Useful Distributions: Chi-Squared Distribution

- The chi-squared distribution arises whenever you take the sum of i.i.d. (independent and identically distributed) squared realizations of a standard normal distribution. Since the standard normal is common, squaring things is common, and summing things is common, that means the chi-squared distribution is also pretty common.

- Chi-squared distributions tend to show up when we construct test statistics in hypothesis testing, a topic you'll see a lot this year.

- Let $Z_i \sim N(0, 1)$ for $i = 1, ..., n$. Let $X = \sum_{i=1}^{n} Z_i^2$.

- We say that X is a chi-squared random variable with n degrees of freedom. We write $X \sim \chi_n^2$.

- $E[X] = n$, $Var[X] = 2n$

Useful Distributions: Others

- There are at least a half-dozen more common distributions you'll see this year. However, there is other material that I think is more important to cover in today's lecture, so I will leave the details for you.

- If you are interested and haven't seen many distributions before, I encourage you to read the econometrics math camp notes for a readable description of more distributions.

Useful Distributions: Others

- There are at least a half-dozen more common distributions you'll see this year. However, there is other material that I think is more important to cover in today's lecture, so I will leave the details for you.

- If you are interested and haven't seen many distributions before, I encourage you to read the econometrics math camp notes for a readable description of more distributions.

Break-Out Session #3

- We will now split out into breakout rooms to work through a useful problem
- Please see the metrics math camp worksheet 2, available on GitHub: github.com/mdroste/metrics-mathcamp-2020
- Take 15 minutes to work through this problem with your group, and we will re-convene afterwards.

Break

- Let’s take a short (10 minute) break.

- I will be here to chat and answer any questions!

Statistical Models, Data

- You might notice that any discussion of a statistical model, an estimator, or even a dataset has been conspicuously absent for metrics math camp. Let's fix that.

- Let our data be $D = (D_1, ..., D_n)$. Each D_i consists of a $1 \times K$ vector of variables.

- A statistical model is a set of assumptions regarding the joint distribution of the data. Most generally, a model can be thought of as saying that $D \sim F(\theta)$, for some distribution F and parameter θ. Note that θ can be a single parameter, a function, a distribution, or all of these. We sometimes write the model as $\{F(\theta) : \theta \in \Theta\}$, where Θ is the "parameter space".

- If θ is finite-dimensional, our model is parametric. Otherwise, it is non-parametric.

- Statistical models can be motivated by economic theory or not (ask Ed Glaeser what he thinks about atheoretical regressions).

Identification

- You have probably heard “identification” before, probably in the context of causal inference. What do you think it means?

- Isaiah will teach you that identification tells us what we can learn from our data under a given model.

- To be precise, we say that the parameter θ in a statistical model is point-identified under the model $\{D(\theta) : \theta \in \Theta\}$ if the mapping $\theta \rightarrow F(\theta)$ is one-to-one.

- Identification asks, ”If I knew F, could I always recover θ?”

- You will spend a lot of the year pondering identification in many settings (causal inference or not). Identification is always relevant.

Example: Linear Regression

- Consider a linear regression model in matrix form:

  $$Y_i = X_i' \beta + \epsilon_i$$

  where $E[\epsilon_i | X_i] = 0$.

- We can solve this with the ordinary least squares (OLS) estimator, which chooses β to minimize the sum of squared residuals ϵ_i:

  $$\hat{\beta} = (X'X)^{-1} X'Y$$

- Linear regression models like the above are quite generally the best linear approximation to an underlying CEF of Y given X. They can always be done if $(X'X)$ is invertible.

- Question: How can we accommodate a constant term in this regression model by modifying the set of regressors X?

Estimators

- OLS is one particular example of an estimator, which is simply a function from the data to the parameters of interest in your statistical model.

- When the parameter is θ, we often represent an estimator as $\hat{\theta}$.

- It is useful to characterize estimators by their properties. Estimators can be categorized in many ways, but you will spend a lot of the first year talking about consistency, efficiency, asymptotic normality, and unbiasedness.

Properties of Estimators

- An estimator $\hat{\theta}$ is consistent if it converges in probability to the true parameter θ as the sample size N (or some other relevant index) grows large. Consistent estimators are sometimes thought of as “asymptotically unbiased”.

- An estimator is asymptotically normal if the distribution of $\hat{\theta}$ converges in distribution to a normal distribution with standard deviation proportional to $1/\sqrt{n}$ as the sample grows large. What does it mean for an estimator to have a distribution? We’ll explore this in the second half of today’s class!

- Much more on this to come in 2120 and 2140!

Asymptotics

- How do we say something about the behavior of our estimators without strong parametric assumptions (e.g. assuming a distribution for the errors)?

- Answer: Use limiting behavior of estimators in large samples, where the behavior of most estimators becomes much simpler, thanks to a set of really powerful results (like the law of large numbers and the central limit theorem).

- Of course, the asymptotic (limiting) behavior of an estimator as the sample grows infinitely large is only an approximation for the estimator's behavior in finite samples - and potentially a poor approximation.

Stochastic Convergence

- The idea behind a lot of asymptotics is to consider “sequences” of estimators, usually indexed by the population size (N). We need to extend the idea of converging sequences of real numbers to random variables.

- It will be helpful to start by remembering the idea of convergence for a non-stochastic sequence of real numbers. Let $\{x_n\}$ be a sequence of real numbers. We say:

  $$\lim_{n \to \infty} x_n = x$$

  if for all $\epsilon > 0$, there exists some N such that for all $n > N$, $|x_n - x| < \epsilon$.

- We will now think about several different ways to generalize this definition to sequences of random variables.

Almost Sure Convergence

- Let $\{X_n\}$ denote a sequence of random variables. We say that the sequence $\{X_n\}$ converges to the random variable X almost surely if:

$$P(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$$

- This definition is a little bit unwieldy, so to make life simpler, we often just write:

$$X_n \xrightarrow{\text{a.s.}} X$$

Almost Sure Convergence: In English

- For a given outcome ω in the sample space Ω, we can ask whether:

$$\lim_{n \to \infty} X_n(\omega) = X(\omega)$$

  holds using the definition of non-stochastic convergence.

- If the set of outcomes for which this holds has probability = 1, then:

$$X_n \xrightarrow{\text{a.s.}} X$$

Convergence in Probability

- Let $\{X_n\}$ be a sequence of random variables and X be a random variable. The sequence of random variables $\{X_n\}$ converges in probability to the random variable X if for all $\epsilon > 0$,
$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) \to 0$$
- As with almost sure convergence, it is helpful and very common to express this in the shorthand:
$$X_n \xrightarrow{\text{p}} X$$

Convergence in Probability: In English

- Fix some $\epsilon > 0$. We can then compute:

$$P_n(\epsilon) = P(|X_n - X| > \epsilon)$$

- This is just a number, o we can check whether $P_n(\epsilon) \to 0$ using the standard definition of non-stochastic convergence.

- If $P_n(\epsilon) \to 0$ for all values $\epsilon > 0$, then we write $X_n \xrightarrow{p} X$

Convergence in Distribution

- Let $\{X_n\}$ be a sequence of random variables and $F_n(\cdot)$ be the cdf of X_n. Let X be a random variable with cdf $F(\cdot)$. The sequence $\{X_n\}$ converges in distribution (or weakly converges, or converges in law) to X if:

  $$\lim_{n\to\infty} F_n(x) = F(x)$$

  for all points x at which $F(x)$ is continuous.

- We often write $X_n \xrightarrow{\mathrm{d}} X$.

Convergence in Distribution: In English

- Convergence in distribution describes the convergence of cdfs of variables. It does not mean that individual realizations of variables get close to each other.
- Recall that the cdf of a variable is defined as:

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$$

- As a result, $F_n(x) \to F(x)$ does not inform us about whether $X_n(\omega)$ is getting close to $X(\omega)$ for any $\omega \in \Omega$

Multivariate Convergence

- We can extend each of these definitions to random vectors very easily.
- We say that a sequence of random vectors $\{X_n\}$ converges almost surely to X if each element of X_n converges almost surely to X. Likewise, the same element-wise extension applies for convergence in probability.
- A sequence of random vectors converges in distribution to a random vector if we apply the definition above to the joint cdf.

Relationship Between Types of Convergence

- We discussed three types of convergence for sequences of random variables: almost sure convergence, convergence in probability, and convergence in distribution.

- It turns out that the ordering of these topics was not accidental, and we moved from ’strong’ notions of convergence to ’weak’ notions.

- In particular, almost sure convergence implies convergence in probability, and convergence in probability implies convergence in distribution.

- The reverse direction does not hold: convergence in distribution does not imply convergence in probability.

Break

- Let’s take a short (10 minute) break.

- I will be here to chat and answer any questions!

Asymptotic Tools

- Now that we have defined three different types of convergence, we can think about four distinct theorems that are going to come up repeatedly when we talk about asymptotics.

- These theorems are statements involving the types of convergence we just discussed, and in practice, using them means that we (often) don't have to explicitly show that a sequence converges using the definition of convergence.

- These tools are Slutsky's Theorem; the Continuous Mapping Theorem; the Law of Large Numbers; and the Central Limit Theorem. All four of these are worth memorizing.

Slutsky's Theorem

- Let c be a constant, X_n and Y_n denote sequences of random variables indexed by n, and X and Y denote random variables. Suppose that $X_n \xrightarrow{\text{d}} X$ and $Y_n \xrightarrow{\text{p}} Y$. Then:

  1. $X_n + Y_n \xrightarrow{\text{d}} X + c$
  2. $X_n Y_n \xrightarrow{\text{d}} Xc$
  3. $X_n / Y_n \xrightarrow{\text{d}} X / c$ if $c \neq 0$.

  If $c = 0$, then $X_n Y_n \xrightarrow{\text{p}} 0$.

Continuous Mapping Theorem

- Let g be a continuous function. Then:
  1. If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.
  2. If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$

- The continuous mapping theorem is easy to remember. Intuitively, it tells us that convergence in probability and distribution are preserved over a continuous transformation g.

The Law(s) of Large Numbers

- This is a big one! The law of large numbers (there are actually several different flavors) provides conditions under which sample averages converge to expectations.

- One form is the weak law of large numbers (weak LLN or WLLN for short). Let $X_1, ..., X_n$ be a sequence of random variables with $E[X_i] = \mu$, $Var[X_i] = \sigma^2 < \infty$, $Cov(X_i, X_j) = 0$ for all $i \neq j$. Then:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{p}} \mu$$

- Another form is the strong law of large numbers (strong LLN or SLLN for short). Let $X_1, X_2, ...$ be i.i.d with $E[X_i] = \mu < \infty$. Then:

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu$$

The Central Limit Theorem(s)

- Another important class of theorems for asymptotics are central limit theorems, which provide conditions under which "properly-centered" sample averages converge in distribution to normal random variables.

- There are several different flavors, but I'll discuss the most useful one for you.

- Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with mean μ and variance σ^2. Then:
$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\text{d}} N(0, \sigma^2)$$

- This generalizes to random vectors, where you will often see this applied. If $X_1, X_2, \ldots$ are random vectors with mean vector μ and covariance matrix Σ, then we have:
$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\text{d}} N(0, \Sigma)$$

Break-Out Session #4

- This concludes our lecture-style material in math camp.

- We'll now move to break-out rooms once again to consider two problems that you should solve with computers.

Break

- Let’s take a short (10 minute) break.

- I will be here to chat and answer any questions!

Unsolicited Advice for 2120 and 2140

- The first-year econometrics sequence consists of two courses that feel very different, even though the structure of the courses is the same.
    - The first course, Principles of Econometrics (2120), is taught by Prof. Tamer and relies on a revised version of the late Gary Chamberlain's notes. Gary's notes are great and echo his (idiosyncratic) view of econometrics. Gary has particular notation (i.e. for the regression function) that you will likely not see anywhere else.
    - The second course, Econometric Methods (2140), is taught by Prof. Andrews and relies on Isaiah's own lecture slides. The material is less abstract; Isaiah motivates each topic with recent applications and empirical work, and the course has a nice unifying theme of asymptotic analysis.
- My best advice is to immerse yourself in the readings for the course. The math in econometrics is not terribly difficult, but there is quite a bit of it, so go slowly and read each week's materials (Gary's notes, Isaiah's slides) more than once.

Unsolicited Advice for 2120 and 2140

- It is an undisputed fact that you will learn more from your classmates during your first year than from your professors.

- This is why it's so important to work in groups - though of course you have your work cut out for you this year.

- The problem sets are intended to enhance your understanding of the material. A well-written problem will teach you something important when you solve it. Please do your problem sets with others. Seeing how your classmates interpret the material and tackle problems will teach you an enormous amount that is hard to gain from doing it yourself.

- Even if you're confident you have the hang of things, it will be really useful to talk through each problem and your attempt to solve them with others.

Wrapping Up

- Good luck in your first year! I hope you enjoyed this portion of math camp, and I hope you will find the material we covered useful this year.

- Feel free to reach out if you have any questions or want to chat about anything - I would love to chat with you at any time.