Econometrics Math Camp

Day 1: Probability and Random Variables

Michael Droste

August 2021

Introduction

- The goal of the next three days is to get you acquainted with some of the tools you will be using throughout the first-year econometrics sequence.

- I have placed all of the material (notes, slides, and worksheets) in a GitHub repository: https://github.com/mdroste/metrics-mathcamp-2021

- These notes and slides were developed by myself, Ashesh Rambachan, and Frank pinter. All errors are mine - let me know if you find any!

- Please have both a computer and a paper/tablet ready! Today will be pencil and paper; tomorrow and Monday's content will require a computer in group work.

Today's Outline

- Probability
  - Random Experiments
  - Probabilities and Conditional Probabilities
- Random Variables
  - Continuous and Discrete Random Variables
  - Conditioning and Independence
  - Transformations of Random Variables
- Expectations
  - Definition and Properties
  - Conditional Expectations
- Moments

Probability

Motivation

- We will develop a theory of probability by using some tools from a branch of mathematics called measure theory, which you may not have seen before.

- Measure theory provides a unifying framework upon which to develop a theory of probability. It is elegant and requires little prerequisite knowledge.

- After we have defined probability in measure-theoretic terms, we can largely take this machinery for granted, in the sense that virtually all of your first-year homework will consist of basic algebra and applying laws of probability.

Random Experiments

- We are interested in a very general concept that we will call a random experiment: an experiment or process whose outcome is not known to us beforehand.

- Let Ω denote the sample space of a random experiment. The sample space Ω is the set of all possible outcomes of the experiment we are studying.

- A subset $A \subseteq \Omega$ is called an event of the random experiment. It will also be helpful to let $\mathcal{A}$ denote the set of all events (i.e., $\mathcal{A}$ is the set containing all subsets of Ω).

- The sample space and the events are the fundamental building blocks for defining probabilities.

Random Experiments: Examples

- To make these definitions more salient, let's consider two quick examples.

- Example 1: Suppose we survey 10 randomly selected people on their employment status and count how many are unemployed - that is, our outcome is the count of people who are unemployed.

  1. What is the sample space Ω?

  2. What is the event A such that more than 30% of those surveyed are unemployed?

Random Experiments: Examples

- To make these definitions more salient, let's consider two quick examples.

- Example 1: Suppose we survey 10 randomly selected people on their employment status and count how many are unemployed.

  1. What is the sample space Ω?
     $\Omega = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
  2. What is the event A such that more than 30% of those surveyed are unemployed?
     $A = \{4, 5, 6, 7, 8, 9, 10\}$

Random Experiments: Examples

- To make these definitions more salient, let's consider two quick examples.

- Example 1: Suppose we survey 10 randomly selected people on their employment status and count how many are unemployed.

  1. What is the sample space Ω?
     $\Omega = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
  2. What is the event A such that more than 30% of those surveyed are unemployed?
     $A = \{4, 5, 6, 7, 8, 9, 10\}$

- Example 2: Suppose I survey a random person on their income (Raj was unavailable).

  1. What is the sample space Ω?

  2. How would I write down the event A such that a person earns between \$30k and \$40k?

Random Experiments: Examples

- To make these definitions more salient, let's consider two quick examples.

- Example 1: Suppose we survey 10 randomly selected people on their employment status and count how many are unemployed.

  1. What is the sample space Ω?
     $\Omega = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
  2. What is the event A such that more than 30% of those surveyed are unemployed?
     $A = \{4, 5, 6, 7, 8, 9, 10\}$

- Example 2: Suppose I survey a random person on their income (Raj was unavailable).

  1. What is the sample space Ω?
     $\Omega = \mathbb{R}$
  2. How would I write down the event A such that a person earns between \$30k and \$40k?
     $A = [\$30000, \$40000]$

Probability

- Now that we have developed some definitions to characterize random experiments, our goal is to sensibly define the probability of an event.

- Before we begin, let's philosophize - when we refer to the probability of an event, what do you think we are trying to communicate?

- Why is this not trivial? If the sample space Ω is finite (example #1), probability is easy to define. But when the sample space is not finite - for instance, if a person's income can take on any real number (see example #2) - then we need to think more carefully.

- It turns out that measure theory is the appropriate mathematical framework to define probabilities in a unified way.

Probability: Defining σ-algebras

- Let Ω be a set and $\mathcal{A} \subseteq 2^{\Omega}$ be a family of its subsets. We say that $\mathcal{A}$ is a σ-algebra if (and only if) it satisfies:
  1. $\Omega \in \mathcal{A}$
  2. Closure under complements: If $A \in \mathcal{A}$, then $A^c = \Omega \setminus A \in A$.
  3. Closure under countable unions: If $A_n \in \mathcal{A}$ for $n = 1, 2, ...$, then $\cup_n A_n \in \mathcal{A}$
- Note that we will adopt a little bit of set theory jargon. Let X^c denote the complement of a set X, and let $\cup$ and $\cap$ denote the union and intersection operators, respectively.
- If Ω is a set and $\mathcal{A} \subseteq 2^{\Omega}$ is a σ-algebra, we say that $(\Omega, \mathcal{A})$ is a measurable space.
- If $(\Omega, \mathcal{A})$ is a measurable space, we say $A \in \mathcal{A}$ is measurable with respect to $\mathcal{A}$.

Probability: Properties of σ-algebras

- σ-algebras have lots of useful properties that follow almost immediately from the definition. Here are two immediately useful ones:

  1. $\emptyset \in \mathcal{A}$
  2. Closure under countable intersections: If $A_n \in \mathcal{A}$ for $n = 1, 2, ...$, then $\cap_n A_n \in \mathcal{A}$.

- You will prove these properties with your classmates in our first breakout session shortly.

Probability: Measures

- Let $(\Omega, \mathcal{A})$ be a measurable space. A measure is a function, $\mu : \mathcal{A} \to \mathbb{R}$, that satisfies:
  1. $\mu(\emptyset) = 0$
  2. $\mu(A) \geq 0$ for all $A \in \mathcal{A}$
  3. If $A_n \in \mathcal{A}$ for $n = 1, 2, ...$ with $A_i \cap A_j = \emptyset$ for $i \neq j$, then $\mu(\cup_n A_n) = \sum_n \mu(A_n)$
- If $\mu(\Omega) = 1$, we say that μ is a probability measure, denoted $P : \mathcal{A} \to [0, 1]$.

Probability: Probability Spaces

- Congrats! We have now developed all of the building blocks we need to characterize the probability of any random experiment.
- A random experiment is characterized by a probability space $(\Omega, \mathcal{A}, P)$, where:
  - Ω: The sample space, or set of outcomes
  - $\mathcal{A}$: The set of events, assumed to admit a σ-algebra representation
  - P: A probability measure defined on the σ-algebra

Probability and Measure Theory

- Why did we need to go through such abstract machinery to think about the probability of events? Familiar laws of probability 'pop out' from the structure we imposed on a probability space (i.e. σ-algebras, measures) almost immediately.

- In addition, your intuitive understanding of a probability (i.e. a long-run average or subjective beliefs about a process) can be shown to be interpreted through this setup.

- We will not discuss measure theory much more in math camp or in the econometrics sequence (at other programs, they will spend more time on this). Having some understanding that probability is built on the fundamentals of measure theory is a powerful idea you should remember.

Basic Laws of Probability

- Consider a probability space $(\Omega, \mathcal{A}, P)$. The following laws always hold:

  1. For any $A \in \mathcal{A}$, we have $P(A^c) = 1 - P(A)$
  2. $P(\Omega) = 1$
  3. If $A_1, A_2 \in \mathcal{A}$ with $A_1 \subseteq A_2$, then $P(A_1) \leq P(A_2)$
  4. For all $A \in \mathcal{A}$, $0 \leq P(A) \leq P(\Omega)$
  5. If $A_1, A_2 \in \mathcal{A}$, then $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

- These laws follow directly from the definitions of probability measures and σ-algebras. You will prove them with your classmates in our first breakout session shortly.

Conditional Probability

- Given a random experiment and the information that event B has occurred, what is the probability that the outcome also belongs to event A?

- Let $A, B \in \mathcal{A}$ with $P(B) > 0$. The conditional probability of A given B is:

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

- Note that all of our usual basic laws of probability (and algebra) apply to $P(A|B)$.

- We will see later on that conditional probabilities are really important, but they also motivate three really important rules that come up when manipulating conditional probabilities.

Conditional Probability: Multiplication Rule

- Consider n events $A_1, ..., A_n$. The multiplication rule relates the probability of all events A_i occurring jointly to conditional probabilities.

- In general, we can express the multiplication rule as:

$$P(\cap_{i=1}^{n} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \cdots P(A_n| \cap_{i=1}^{n-1} A_i)$$

- When $n = 2$, we have $P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$.
  - Observe this is just rearranging the definition of a conditional probability.

- When $n = 3$, we have $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$.

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 (aka 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:

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

- The easiest proof relies on the multiplication rule. You will prove Bayes' rule with your classmates shortly.

Independence

- How can we characterize the extent to which two events are related?

- We say that two events A and B are independent if $P(A|B) = P(A)$, or equivalently, $P(B|A) = P(B)$, or $P(A \cap B) = P(A)P(B)$.

- Let $E_1, ..., E_n$ be events. $E_1, ..., E_n$ are said to be jointly independent if for any $i_1, ..., i_k$:

$$P(E_{i_1}|E_{i_2} \cap \cdots \cap E_{i_k}) = E_{i_1}$$

- Given an event C, we say that two events A and B are conditionally independent if:

$$P(A \cap B|C) = P(A|C)P(B|C)$$

Break-Out Session #1

Break-Out Session #1

- We will now split out into breakout groups to work through proving some of the laws of probability using the mathematical machinery we built up.

- Please see the attached Metrics Math Camp Worksheet #1 for the problems.

- Take 15 minutes to work through these problems with your small group. Raise your hand on Zoom if you get stuck or have a question.

- Important note: Proving these results is less important than taking this as an opportunity to work with some new folks in your class.

Lunch Break

- This concludes our brief primer on the basics of probability. So far, we have:
  1. Developed a measure-theoretic model of probability
  2. Derived rules for manipulating probabilities from this model
- After lunch, we'll discuss how to formalize random variables, expectations, and conditional expectations.

Random Variables

Random Variables

- You're going to hear the term 'random variable' quite a bit in the first year.
- What is a random variable, though? You probably have a pretty good idea.
- Consider the random experiment we thought about earlier: a random variable might be thought as representing the outcomes of a random experiment.
- ... But, since you're now a first-year grad student, it will pay to be more precise in defining what exactly we mean. We will need a couple more pieces of mathematical machinery from measure theory.

Random Variables: Building Blocks

- The first building block we need is a particular σ-algebra, the Borel σ-algebra, often denoted $\mathcal{B}$.

- Let $\Omega = \mathbb{R}$, $\mathcal{A} =$ collection of all open intervals in $\mathbb{R}$. The "smallest" σ-algebra containing all open sets is the Borel σ-algebra.

- More rigorously, $\mathcal{B}$ is the collection of all Borel sets, which is any set in $\mathbb{R}$ formed by countable union, countable intersection, relative complement.

Random Variables: Building Blocks

- The second building block we need is the idea of a measurable function between two measure spaces.

- Let $(\Omega, \mathcal{A}, \mu)$ and $(\Omega', \mathcal{A}', \mu')$ two measure spaces. Let $f : \Omega \to \Omega'$ be a function. We say that f is measurable if (and only if) $f^{-1}(A') \in \mathcal{A}$ for all $A' \in \mathcal{A}'$.

- What does this even mean? Loosely, a measurable function can be thought of as a function between the sets underneath two measure spaces that preserves the structure of the measure spaces: the preimage of any measurable set is measurable.

Random Variables: Definition

- Let $(\Omega, \mathcal{A}, P)$ denote a probability space and let $X : \Omega \to \mathbb{R}$ denote a real-valued function.
- We say that X is a random variable if (and only if) X is P-measurable. That is, $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$ where $\mathcal{B}$ is the Borel σ-algebra.
- In this sense, a random variable is simply a mapping between events and probabilities.

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 Function: Properties

- The cumulative distribution function F_X has many handy properties. Among them:
  1. For $x_1 \leq x_2$, $F_X(x_2) - F_X(x_1) = P(x_1 < X < x_2)$
  2. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty}(x) = 1$
  3. F_X is non-decreasing.
  4. F_X is right-continuous.

- Cool fact: for any function F that satisfies these properties, we can construct a random variable whose cumulative distribution function is F. The details of this construction are left as an exercise for us in the next break-out section.

Cumulative Distribution Functions: Quantiles

- The quantiles of a random variable X are a way to characterize the range of a cdf.

- The quantile function can be defined very generally:

$$Q(u) = \inf\{x : F_X(x) \geq u\}$$

  If F_X is invertible, then the quantile function can be written in terms of the inverse CDF:

$$Q(u) = F_X^{-1}(u)$$

- Interpretation: $Q(u)$ tells you the value of the random variable such that a fraction u of observations have a value less than u. For instance, if $u = 0.5$, the quantile function returns the median.

Discrete Random Variables

- Let X be a random variable. We say that X is a discrete random variable if (and only if) F_X is constant except at a countable number of points (i.e. F_X is a step function).

$$p_i = P(X = x_i) = F_X(x_i) - \lim_{x \to x_i^-} F_X(x)$$

Use this to define the probability mass function of X:

$$f_X(x) = \begin{cases} p_i & \text{if } x = x_i,\ i = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

- We can write:

$$P(x_1 < X \leq x_2) = \sum_{x_1 < x \leq x_2} f_X(x)$$

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 $S_X = \{x : f_X(x) > 0\}$

Continuous Random Variables: Notes

- There are two useful facts to remember about continuous random variables (actually, probably more).

- First, note that for $x_2 \geq x_1$, we have:

$$\begin{aligned} P(x_1 < X \leq x_2) =& F_X(x_2) - F_X(x_1) \\ =& \int_{x_1}^{x_2} f_X(t)dt \end{aligned}$$

- Second, note that $P(X = x) = 0$; that is, the probability that a continuous random variable takes on any particular value x is 0. At a deep level, this fact is why we needed measure theory to describe random variables!

Joint Distributions

- Let X and Y be two (scalar) random variables. A random vector (X, Y) is a measurable mapping from Ω to $\mathbb{R}^2$.

- The joint cumulative distribution function of (X, Y) is:

$$\begin{aligned} F_{X,Y}(x, y) =& P(X \leq x, Y \leq y) \\ =& P(\{\omega : X(\omega) \leq x\} \cap \{\omega : Y(\omega) \leq y\}) \end{aligned}$$

  We say that (X, Y) is a discrete random vector if:

$$F_{X,Y}(x, y) = \sum_{u \leq x} \sum_{v \leq y} f_{X,Y}(u, v)$$

  where $f_{X,Y}(x, y) = P(X = x, Y = y)$ is the joint probability mass function of (X, Y).

Joint Distributions

- Let X and Y be two (scalar) random variables. We say that (X, Y) is a continuous random vector if:

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) dv du$$

  where $f_{X,Y}(x, y)$ is the joint probability density function of (X, Y).

- As before, at points where $F_{X,Y}$ is continuous, we have:

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$$

Joint Distributions to Marginal Distributions

- If we know the joint cdf of (X, Y), we can recover the marginal *cdfs* of X and Y:

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(X \leq x, Y \leq \infty) \\ &= \lim_{y \to \infty} F_{X,Y}(x, y) \end{aligned}$$

- We can also recover the marginal pdfs from the joint pdf:

$$f_X(x) = \sum_y f_{X,Y}(x, y) \qquad \text{(discrete)}$$

$$f_X(x) = \int_{S_y} f_{X,Y}(x, y) dy \qquad \text{(continuous)}$$

Conditioning with Discrete Variables

- Consider x with $f_X(x) > 0$. The conditional pmf of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

  where $f_{Y|X}(y|x)$ satisfies $f_{Y|X}(y|x) \geq 0$ and $\sum_y f_{Y|X}(y|x) = 1$.

- The conditional cdf of Y given $X = x$ is defined:

$$F_{Y|X}(y|x) = P(Y \leq y|X = x) = \sum_{v \leq y} f_{Y|X}(v|x)$$

Conditioning with Continuous Variables

- Consider x with $f_X(x) > 0$. The conditional pdf of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

- The conditional cdf of Y given $X = x$ is defined:

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(v|x)dv$$

Independence of Random Variables

- Let X and Y be random variables. We say that X and Y are independent if (and only if):

$$F_{Y|X}(y|x) = F_Y(y)$$

  or, equivalently:

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

- This can also be defined in terms of densities - replace F with f above.

Transformations of Random Variables

- Let X be a random variable with a cdf F_X. Define a random variable $Y = h(x)$, where h is a one-to-one function whose inverse h^{-1} exists. What is the distribution of Y?

- First, let's tackle the discrete case. Suppose that X is discrete with values $x_1, ..., x_n$. Y is also discrete with the values $y_i = h(x_i)$ for $i = 1, ..., n$, and the pmf of Y is given by:

$$P(Y = y_i) = P(X = h^{-1}(x_i))$$
$$f_Y(y) = f_X(h^{-1}(y_i))$$

Transformations of Random Variables

- Next, let's consider the case where X is a continuous random variable. It helps to consider two cases:
  1. First, suppose h is increasing. Then we have:

$$F_Y(y) = P(Y \leq y) = P(X \leq h^{-1}(y)) = F_X(h^{-1}(y))$$

     and so $f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}$.
  2. Second, suppose h is decreasing. Then we have:

$$f_Y(y) = -f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}$$

- Combining these cases, we have that in general:

$$f_Y(y) = f_X(h^{-1}(y))\left|\frac{dh^{-1}(y)}{dy}\right|$$

Expectations of Random Variables

Expectations of Discrete Random Variables

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

$$E[X] = \sum_x x f_X(x)$$

  if $\sum_x |x| f_X(x) < \infty$. Otherwise, the expectation does not exist.

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

$$E[g(X)] = \sum_x g(x) f_X(x)$$

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.

Conditional Expectations as Optimal Forecasts

- Conditional expectation functions are incredibly useful objects, and they have several useful interpretations in econometrics.

- One interpretation is that conditional expectations are the solution to an optimal forecasting problem. Suppose you want to forecast the value of a random variable Y. More precisely, suppose you want to pick some $h \in \mathbb{R}$ that minimizes the expected mean squared error:

$$E[(Y-h)^2] = \int (y-h)^2 f_Y(y) dy$$

The first-order condition for this problem is:

$$\int y f_Y(y) dy = \int h f_Y(y) dy \implies h^* = E[Y]$$

Conditional Expectations as Orthogonal Projections

- Another perspective is that we can interpret the conditional expectation of Y given X as the orthgonal projection of Y onto the space of functions of the random variable X, i.e., L^2 space.

- This is the focus of the first two-ish weeks of Econ 2120.

- It provides a unifying perspective on much of regression analysis, and this is really the central focus of the first half of your econometrics sequence. It is an important idea that you'll spend a couple weeks thinking about.

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.

- Intuitively: the outer expectation integrates out the conditional information in the inner expectation, so we're left with an unconditional expectation.

Moments

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 covariance of X and Y tells us about the extent to which X and Y move together (covary), and is defined as:

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

- 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.

Break-Out Session #2

Break-Out Session #2

- We will now split out into breakout groups to work through a couple additional problems.
- Please see Metrics Math Camp Worksheet #2 in our GitHub repo: github.com/mdroste/metrics-mathcamp-2021
- Take some time to work through these problems with your small group. We'll re-convene to discuss any issues you might have had afterwards.

Day 1 Wrap-Up

- We covered a great deal of content today - you all earned a break!
- We will re-convene tomorrow to cover asymptotic theory and the least squares estimator through the lens of the projection theorem.
- I will stick around for questions!