Harvard Economics
Econometrics Math Camp 2020

1 Break-Out Session #1

The following questions concern deriving basic laws of probability from the setup we discussed during the first hour or so of lecture. It may be helpful to refer to the econometrics math camp slides (or the much beefier econometrics math camp notes) for details on the properties of σ-algebras, measures, etcetera.

These exercises have two purposes. First, I want to demonstrate how you can take a measure-theoretic presentation of probability theory - with a very small list of mathematical axioms - and derive all of the usual laws of probability you're used to. Second, I want you to meet some more folks in your cohort and work through some problems with them. In my view, this second objective is way more important than the first.

1. Let $(\Omega, \mathcal{A}, P)$ be a probability space. Please prove the following laws of probability:

   (a) For any $A \in \mathcal{A}$, $P(A^c) = 1 - P(A)$
   Consider $A \in \mathcal{A}$. Note that $A^c = \Omega \setminus A \in \mathcal{A}$ because σ-algebras are closed under complements. Next, because P is a probability measure, we have $P(\Omega) = 1$. Then since A and A^c are disjoint by construction of A^c, $P(A \cup A^c) = P(A^c) + P(A) = P(\Omega) = 1$. Rearrange to yield $P(A^c) = 1 - P(A)$, as desired.

   (b) $P(\Omega) = 1$
   This fact is immediate from the definition of a probability measure.

   (c) For all $A \in \mathcal{A}$, $0 \leq P(A) \leq P(\Omega)$
   Let $A \in \mathcal{A}$. Recall that any measure takes on non-negative values, so $P(A) \geq 0$. Next, note that $A \cup A^c = \Omega$, with $A \cap A^c = \emptyset$ by construction of the complement $A^c = \Omega \setminus A$. The third property of measures we discussed was that the measure of any countable union of sets is equal to the sum of the measures of each set. So $P(A \cup A^c) = P(A) + P(A^c)$. Then since $\Omega = A \cup A^c$, we can write $P(A) + P(A^c) = P(\Omega)$. Finally, since measures take on non-negative values, $P(A^c) > 0$, which immediately implies $P(A) \leq P(\Omega)$. Putting everything together, we have shown $0 \leq P(A) \leq P(\Omega)$, as desired.

   (d) If $A_1, A_2 \in \mathcal{A}$ with $A_1 \subseteq A_2$, then $P(A_1) \leq P(A_2)$

There are a few ways to show this. I think this is the simplest way. Let $A_1, A_2 \in \mathcal{A}$, with $A_1 \subseteq A_2$. Observe that we can write $A_2 = A_1 \cup B$, where $B = (A_1^c \cap A_2)$. Note that $B \in \mathcal{A}$, since σ-algebras are closed under countable intersections (this was an "additonal property" of σ-algebras we discussed that comes for free from the axiom about closure of countable unions plus De Morgan's laws for set theory). Next, note that A_1 and B are disjoint, i.e. $A_1 \cap B = \emptyset$. Then we can apply the fact that for a measure, the probability of the union of disjoint events is equal to the sum of the events probabilities to obtain $P(A_1) + P(B) = P(A_2)$. Finally, note that the probability of an event is non-negative, and in particular $P(B) \geq 0$. Hence, we have $P(A_1) \leq P(A_2)$, as desired.

2. Let $(\Omega, \mathcal{A}, P)$ be a probability space. Consider K disjoint events C_k that partition the sample space Ω, and let $A \in \mathcal{A}$ be an event. Please prove the law of total probability, which states that $P(A) = \sum_{i=1}^{K} P(A|C_i)P(C_i)$.

   I am sorry if this proof was tricky for you - the easiest way to prove this (I think) uses the fact that the set union and intersection operators distribute, which is something I had not yet emphasized in the slides. I'll bring it up in our second lecture. It's not super important for the first year but does turn out to be a good thing to know.

   Let $\{C_i\}_{i=1}^{K}$ denote the set of K disjoint events that partition the sample space, so that $C_i \cap C_j =$ for $i \neq j$ and $\uplus_{i=1}^{K} C_i = \Omega$. Let $A \subset \Omega$ denote an event. I can obtain the following chain of equalities:

$$\begin{aligned} A &= A \cap \Omega \\ &= A \cap (\bigcup_i C_i) \\ &= \bigcup_i (A \cap C_i) \end{aligned}$$

   where the first line is almost trivially true since $A \subseteq \Omega$, the second uses the fact that the union of all C_i is Ω, and the third line uses the fact that the set intersection operator distributes. Now I can apply the probability measure to both sides and obtain another chain of equalities:

$$\begin{aligned} P(A) &= P\big(\bigcup_{i=1}^{K} (A \cap C_i)\big) \\ &= \sum_{i=1}^{K} P(A \cap C_i) \\ &= \sum_{i=1}^{K} P(A|C_i)P(C_i) \end{aligned}$$

   where the first line follows from the previous set of equalities, the second line uses the property of measures where the measure of the union of disjoint sets is the sum of their measures (this is useful because $A \cup C_i$ and $A \cup C_j$ are disjoint when $i \neq j$, which in turn is true because the sets C_i are disjoint), and the last line plugs in the definition of conditional probability / the multiplication rule K times.

3. Let $(\Omega, \mathcal{A}, P)$ be a probability space, and consider two events $A, B \in \mathcal{A}$. Recall that Bayes' rule can be stated as $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$. Please derive this rule.
We can start with the equation $P(A \cap B) = P(B \cap A)$, which is true because the intersection and union operators are commutative (i.e. the order of B and A does not matter). Next, apply the definition of conditional probability to both sides to obtain $P(A|B) * P(B) = P(B|A) * P(A)$. Dividing both sides by $P(A)$ yields the desired result.

2 Break-Out Session #2

The following questions concern random variables, distributions, and expectations.

1. I earlier said that any function F which satisfies the four 'handy properties' of a cumulative distribution function can be used to construct a random variable, call it Y, whose cumulative distribution function is F. Please detail this construction.
First off, I'm sorry this was a little unclear - the construction was not intended to be super rigorous. I also should have given a hint, and allowed you to assume that F is invertible (just to make things a little simpler). Loosely, the idea is simply to draw random numbers from a uniform distribution on the unit interval and plug them into F^{-1} (assumed to exist).

2. We talked a little bit about how transformations of random variables affect their distributions. We can often say something concrete if the transformation is 'nice'. Suppose that X is drawn from a uniform distribution on the unit interval and that $Y = X^2$. What is the cdf and pdf of Y?
In order to think about the distribution of Y, it is helpful to be explicit about the distribution of X. In words, it's easy to remember that the uniform distribution over $[a, b]$ is such that all values in the interval $[a, b]$ are equally likely to occur. More explicitly, the pdf of a uniform distribution over $[a, b]$ is $f(x) = \frac{1}{b-a}$ if $x \in [a, b]$ and $f(x) = 0$ otherwise. Applying this to X drawn from a uniform distribution over $[0, 1]$, this means $f_X(x) = 1$ if $x \in [0, 1]$ and $f_X(x) = 0$ otherwise. Thus, the cdf is $F_X(x) = 0$ if $x < 0$, $F_X(x) = \int_0^x 1dt = x$ if $x \in [0, 1]$, $F_X(x) = 1$ if $x > 1$.
Next, let's think about the transformation h. It is easy to see that the transformation is $h(x) = x^2$. From here, it's clear the inverse h^{-1} exists and is $h^{-1}(x) = x^2$.
Now we can think about mapping what we know about the distribution of X and the transformation h to determine the distribution of Y. Recall from the slides that in general, we have $f_Y(y) = f_X(h^{-1}(y))|dh^{-1}(y)/dy|$. Let's take this piece by piece. First, we have $f_X(h^{-1}(y)) = 1$ for $y \in [0, 1]$ and 0 otherwise. Next, we have $|dh^{-1}(y)/dy| = \frac{1}{2\sqrt{y}}$. Putting these together, we see the pdf of Y is $f_Y y = \frac{1}{2\sqrt{y}}$ if $y \in [0, 1]$ and 0 otherwise. Then we can obtain the cdf of Y by integrating, so $F_Y(y) = \int_0^y \frac{1}{2\sqrt{y}} dy = \sqrt{y}$ for $y \in [0, 1]$, $F_y(y) = 0$ for $y < 0$, and $F_y(y) = 1$ for $y > 1$.

3. The following questions concern characterizing the distribution of a variable X that is uniformly distributed on the interval $[-1, 1]$.

(a) What is the pdf of X? What about the cdf?

This is basically the same thing as the first part of the previous problem. The pdf of a uniform distribution over $[a,b]$ is $f(x) = \frac{1}{b-a}$ if $x \in [a,b]$ and $f(x) = 0$ otherwise. So when X is drawn from a uniform distribution over the interval $[-1,1]$, we say that $f_X(x) = 1/2$ if $x \in [-1,1]$ and $f_X(x) = 0$ otherwise. We obtain the *cdf* by integrating, i.e. $F_X(x) = \int_{-1}^{x} 1/2dx = \frac{1}{2}(x+1)$ for $x \in [0,1]$, $F_x(x) = 0$ if $x < 0$, and $F_x(x) = 1$ if $x > 1$.

(b) Derive an expression for the q^{th} quantile of X.

Recall from our slides that we can derive the quantile function of X very easily when the cdf is invertible. Recall the interesting part of the cdf (for $x \in [0,1]$ was $F_X(x) = \frac{1}{2}(x+1)$. Let q denote the qth quantile of X, and set $F_X(x) = q \implies x = 2q-1$. A different equivalent way to see this is that you just want to invert the cdf. If I want to find the median, for instance, I'd just want to plug 0.5 into F_X^{-1}.

(c) What is the mean of X?

Since all values are equally likely to occur, the distribution is symmetric, and it is obvious that the mean of X occurs at the mid-point of the support, i.e. $E[X] = 0.5$. It may be worthwhile to over-do this and see how one could obtain this result from the definition of $E[X]$:

$$\begin{aligned}
E[X] &= \int_0^1 xf(x)dx \\
&= \int_0^1 xdx \\
&= 0.5(1)^2 - 0.5(0)^2 \\
&= 0.5
\end{aligned}$$

where the first line is the definition of $E[X]$, the second line uses the fact that $f(x) = 1$ over the interval $[0,1]$, the third line uses the fundamental theorem of calculus to get rid of the integral, and the last line simplifies. The nice thing about this approach, explicitly using the definition of the mean, is that it generalizes to more complicated distributions where the mean is not a priori obvious (which is the case for any interesting problem).

4. Recall the CEF decomposition theorem, which states that a random variable Y can be expressed as $Y = E[Y|X] + \epsilon$, with ϵ uncorrelated with any function of another random variable X. Prove this. Hint: start by showing $E[\epsilon|X] = 0$, and then use the law of iterated expectations with this result to prove that ϵ is uncorrelated with any function of X. This is a little tricky!

If you are reading this, congratulations! This is a really useful proof to see. It looks a lot like the homework problems you'll be facing very shortly, and features and crucial use of the law of iterated expectations.

First, we start by defining $\epsilon = Y - E[Y|X]$, and let h denote any function of X. Then all I need to show is that (1) $E[\epsilon|X] = 0$ and (2) $E[\epsilon \cdot h(X)] = 0$. Let's start with by showing that $E[\epsilon|X] = 0$, i.e. ϵ is mean independent of X (think for a second about why that term makes sense). We can write the following chain of equalities:

$$\begin{aligned} E[\epsilon|X] &= E[Y - E[Y|X]|X] \\ &= E[Y|X] - E[Y|X] \\ &= 0 \end{aligned}$$

where the first line substitutes in the definition of ϵ, the second line distributes the conditional expectation across $-$ (expectations are linear operators), and the third line simplifies. Second, I need to show that ϵ is uncorrelated with any function of X, which amounts to showing that $E[h(X) \cdot \epsilon] = 0$ for any function h of X. We have:

$$\begin{aligned} E[h(X) \cdot \epsilon] &= E[h(X) \cdot E[\epsilon|X]] \\ &= E[h(X) \cdot 0] \\ &= 0 \end{aligned}$$

where the first line is a crucial application of the law of iterated expectations, the second line applies our result from earlier that $E[\epsilon|X] = 0$, and the third line simplifies. This completes the proof.

5. Let X and Y denote random variables. Recall the law of iterated expectations, $E_Y[Y] = E_X E_{Y|X}[Y]$. Prove it. It may be useful to start from the right-hand side of the equation, write it as a double integral, and rearrange terms until you find that it is equal to $E[Y]$. This is also a useful proof, because it gets into the weeds of how expectations are defined. Expectations are nice because we can often manipulate them as operators to prove stuff without explicitly invoking their definition as integrals (see the previous problem), but sometimes we still need to play around with the integrals to prove what we need to. This problem is one good example of why.

   Before proceeding, I want to note that I've been somewhat pedantic about notation in the slides/this problem. You will often see the law of iterated expectations expressed as $E[E[Y|X]] = E[Y]$, where the subscripts on the expectations are suppressed. You can write expectations in this more parsimonious way on homework, and it's almost always understood what the expectation is being 'taken over'. In short, these are two almost-equivalent choices of notation, where the more-common approach of not subscripting your expectations operators is a bit less rigorous (which is fine).

   The method of proof here is to start with $E_X E_{Y|X][Y]}$ and show that it is equal to $E_Y[Y]$ by manipulating the integrals that define an expectation. We have:

$$
\begin{aligned}
E_X E_{Y|X}[Y] &= \int E_{Y|X}[Y] f_x(x) dx \\
&= \int \left[\int y f_y(y|X=x) dy \right] f_x(x) dx \\
&= \int \left[\int y f_y(y|X=x) f_x(x) dx \right] dy \\
&= \int y \left[\int f_y(y|X=x) f_x(x) dx \right] dy \\
&= \int y \left[\int f_{x,y}(x,y) dx \right] dy \\
&= \int y \left[f_y(y) \right] dy \\
&= E_Y[Y]
\end{aligned}
$$

where the first line substitutes in the definition of the conditional expectation $E_{Y|X}[Y]$, the second line substitutes in the definition of the expectation E_X, the third line flips the order of integration (which we can do by a very under-appreciated result you might remember from multivariable calculus called Fubini's theorem), the fourth line pulls y out of the inner integral (which we can do because the inner integral is with respect to x, so y is essentially a constant), the fifth line substitutes in the definition of a joint density, the sixth line uses the fact that we can recover a marginal density f_y by "integrating out" the joint density with respect to x, and the sixth line uses the definition of the expectation $E_Y[Y]$.