Harvard Economics Math Camp 2018: Econometrics, Probability Review

Ashesh Rambachan[1]

August 2018

DISCLAIMERS:

1. There is *absolutely no* expectation for you to read these notes prior to math camp. Maximize utility as you see fit.
2. This is intended to provide a brief refresher on some concepts and preview some material that will be covered in the first year econometrics sequence. If some of the material is unfamiliar, *do not worry*.
3. These notes contain more content than we will have time to cover during math camp. This is intentional. Hopefully these notes can be a useful reference material for you throughout the year.

[1] These notes are pull heavily from materials in many econometrics and statistics textbooks (see references below) and draw heavily upon notes from other econometrics and statistics courses (Max Kasy's notes from Ec 2110 at Harvard, Gary Chamberlain's notes from Ec 2120, Mark Watson's notes from ECON 517 at Princeton, Ramon van Handel's notes from ORFE 309 at Princeton and the University of Minnesota's math camp notes from Fall 2015). I do not provide references in the text. And so, I take ZERO credit and all errors are my own.

Contents

Principles of Probability

A RANDOM EXPERIMENT is an experiment whose outcome cannot be predicted beforehand. How do we model a random experiment? There are three key elements: The sample space, the events and the probability measure. Each of these are described in turn.[2]

[2] There are going to be a lot of definitions that seem overly complex to describe something that seems fairly simple... Bear with me and welcome to a Ph.D. in economics.

Definition 0.1. *The **sample space** Ω is the set of all possible outcomes of a random experiment. We denote an outcome as $\omega \in \Omega$.*

Definition 0.2. *An **event** A is a subset of the sample space, $A \subseteq \Omega$. Let $\mathcal{A}$ denote the family of all events.*

Example 0.1. *Suppose we survey 10 randomly selected people on their employment status and count how many are unemployed.*

$$\Omega = \{0, 1, 2, \ldots, 10\}$$

A is the event that more than 30% of those surveyed are unemployed.

$$A = \{4, 5, 6, \ldots, 10\}$$

Example 0.2. *Suppose we ask a random person what is their income.*

$$\Omega = \mathbb{R}_+$$

A is the event that the person earns between \$30,000 and \$40,000.

$$A = [30,000, 40,000]$$

We place additional restrictions on $\mathcal{A}$. These impose sufficient structure that will allow us to consistently define the probabilities of events.

Definition 0.3. *Let Ω be a set and $\mathcal{A} \subseteq 2^\Omega$ be a family of its subsets. $\mathcal{F}$ is a **σ-algebra** if and only if it satisfies the following*

1. *$S \in \mathcal{F}$.*
2. *$\mathcal{F}$ is closed under complements: $A \in \mathcal{F}$ implies that $A^C = S - A \in \mathcal{F}$.*
3. *$\mathcal{F}$ is closed under countable union: If $A_n \in \mathcal{F}$ for $n = 1, 2, \ldots$, then $\cup_n A_n \in \mathcal{F}$.*

Remark 0.1. *Properties 1 and 2 of a σ-algebra implies that $\emptyset \in \mathcal{F}$. Properties 2 and 3 imply that $\mathcal{F}$ is closed under countable intersection by DeMorgan's Law. That is, if $A_n \in \mathcal{F}$ for $n = 1, 2, \ldots$, then $\cap_n A_n \in \mathcal{F}$.*

We assume that $\mathcal{A}$, the family of all events on Ω, is a σ-algebra. We say that $(\Omega, \mathcal{A})$ is *measurable space* and that $A \in \mathcal{A}$ is *measurable* with respect to the σ-algebra $\mathcal{A}$.

Definition 0.4. *Let $(\Omega, \mathcal{A})$ be a measurable space. A **measure** is a function, $\mu : \mathcal{A} \to \mathbb{R}$ such that*

1. $\mu(\emptyset) = 0$.
2. $\mu(A) \geq 0$ *for all* $A \in \mathcal{A}$.
3. *If* $A_n \in \mathcal{A}$ *for* $n = 1, 2, \ldots$ *with* $A_i \cap A_j = \emptyset$ *for* $i \neq j$, *then*

$$P(U_n A_n) = \sum_n P(A_n)$$

*If $\mu(\Omega) < \infty$, we call μ a **finite measure**, If $\mu(\Omega) = 1$, we call μ a **probability measure**. We denote a probability measure as $P : \mathcal{A} \to [0, 1]$.*

Definition 0.5. *A triple $(\Omega, \mathcal{A}, \mu)$ where Ω is a set, $\mathcal{A}$ is a σ-algebra and μ is a measure on $\mathcal{A}$ is a **measure space**. If μ is a probability measure, it is **probability space**.*

We've now defined all components needed to model a random experiments. A random experiment is characterized by its probability space, $(\Omega, \mathcal{A}, P)$. With the definition of a probability space that we laid out above, we can prove all of the usual probability facts.

Proposition 0.1. *Consider a probability space $(\Omega, \mathcal{A}, P)$. The following hold:*

1. *For all* $A \in \mathcal{A}$, $P(A^C) = 1 - P(A)$.
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* $A \in \mathcal{A}$, $0 \leq P(A) \leq P(1)$.
5. *If* $A_1, A_2 \in \mathcal{A}$, *then*

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

Exercise 0.1. *Prove these properties from the definition of a probability space.*

Conditional Probability

Conditional probability gives us a way to model the outcome of a random experiment conditional on some partial information. For instance, given a random experiment and the information that event B has occurred, what is the probability that the outcome also belongs to event A? To do so, we define a new probability measure on Ω.

Definition 0.6. *Let $A, B \in \mathbb{A}$ with $P(B) > 0$. The **conditional probability of** A **given** B is*

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

$P(A|B)$ is a probability measure.

Remark 0.2. *We can think about $P(A|B)$ as part of a new probability space with $\Omega = B$ and $P(S) = P(S|B)$ for $S \subseteq B$.*

Remark 0.3. *Because the conditional probability is a probability measure, all of the usual properties of probability measures in Proposition 0.1 apply.*

The definition of conditional probability implies the following useful formula. We have that

$$P(A \cap B) = P(A|B)P(B) \tag{1}$$

We next list several important results about conditional probabilities.

Theorem 0.1. *The multiplication rule*

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

Proof. This follows via repeated application of the definition of conditional probability. □

Theorem 0.2. *Law of total probability*

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

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

Proof. We have that

$$\begin{aligned} C &= C \cap \Omega \\ &= C \cap (\cup_{i=1}^{K} C_i) \\ &= (C \cap C_1) \cup \dots \cup (C \cap C_K) \end{aligned}$$

It follows

$$P(C) = \sum_{i=1}^{k} P(C \cap C_i)$$

and the result follows from the definition of conditional probability.

□

Theorem 0.3. *Bayes' Rule*

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

Exercise 0.2. *Prove Bayes' Rule from the results presented.*

Example 0.3. *Suppose you survey 2 randomly selected individuals. What is the probability that both are female given that at least one is female? Assume that all outcomes are equally likely.*

Solution. The sample space is $\Omega = \{MM, MF, FM, FF\}$. The conditioning event is $B = \{MF, FM, FF\}$ and $A = \{FF\}$. Therefore,

$$P(A|B) = \frac{P(\{FF\})}{P(\{MF, FM, FF\})} = 1/3$$

■

As mentioned, we use conditioning to describe the partial information that an event B gives about another event A. What if B provides no information about A?

Definition 0.7. *Two events A, B are **independent** if*

$$P(A|B) = P(A)$$

*Equivalently, they are **independent** if*

$$P(B|A) = P(B)$$

or

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

Remark 0.4. *If events A, B are independent, then so are A^C, B, A, B^C and A^C, B^C.*

We can extend the definition of independence to collections of events.

Definition 0.8. *Let $E_1, \ldots, E_n$ be events. $E_1, \ldots, E_n$ are **jointly independent** if for any $i_1, \ldots, i_k$*

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

Moreover, since conditional probabilities are probability measures, we can define independence with respect to a conditional probability.

Definition 0.9. *Given an event C, events A, B are **conditionally independent** if*

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

*Equivalently, A, B are **independent conditional on** C if*

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

Random Variables

Borel σ-algebra

Earlier in these notes, we defined a σ-algebra. This was a collection of sets that satisfied some additional restrictions that helped us consistently define the probability of each set. A particularly important σ-algebra is called the **Borel σ-algebra**. This is a σ-algebra over the real line.

Definition 0.10. *Let $\Omega = \mathbb{R}$. Let $\mathcal{A}$ be the collection of all open intervals. The smallest σ-algebra containing all open sets is the **Borel σ-algebra**. It is typically denoted as $\mathcal{B}$.*

Note that the Borel σ-algebra contains all closed intervals as well and could have been equivalently defined as the smallest σ-algebra that contains all closed sets. Moreover, we can extend the Borel σ-algebra to higher dimensions - it is the smallest σ-algebra that contains the open balls. The Borel σ-algebra will be useful later on in this section.

Measurable functions

A measurable function is a function that maps from one measure space to another. Measurable functions are useful because for a given set of values in the function's range, we can measure the subset of the function's domain upon which these values occur.

Definition 0.11. *Let $(\Omega, \mathcal{A}, \mu)$ and $(\Omega', \mathcal{A}', \mu')$ be two measure spaces. Let $f : \Omega \to \Omega'$ be a function. f is **measurable** if and only if $f^{-1}(A') \in \mathcal{A}$ for all $A' \in \mathcal{A}'$.*

That is, $\mu'(f^{-1}(A'))$ is well-defined for a measurable function f. A particularly useful case occurs when

$$(\Omega', \mathcal{A}', \mu') = (\mathcal{R}, \mathcal{B}, \lambda)$$

where λ is the lebesgue measure. That is, f is a real-valued function. We say that f is **μ-measurable** if and only if

$$f^{-1}((-\infty, c)) = \{\omega \in \Omega : f(\omega) < c\} \in \mathcal{A} \quad \forall c \in \mathbb{R}$$

We could also state this definition in terms of $>, \leq$ or $\geq$. With these definitions, we are now ready to define a random variable

Random variables

Consider a probability space $(\Omega, \mathcal{A}, P)$. A **random variable** is simply a measurable function from the sample space Ω to the real-line.

Definition 0.12. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and $X : \Omega \to \mathbb{R}$ is a function. 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.*

Definition 0.13. *Let X be a random variable. The **cumulative distribution function** (cdf) $F : \mathbb{R} \to [0, 1]$ of X is defined as*

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

For simplicity, we often write

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

Note that $(\mathbb{R}, \mathbb{B}, F_X)$ form a probability space. The cumulative distribution function F_X has the following properties:

1. *For $x_1 \leq x_2$,*

$$F_X(x_2) - F_X(x_1) = P(x_1 < X < x_2).$$

2. $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$.
3. *$F_X(x)$ is non-decreasing.*
4. *$F_X(x)$ is right-continuous:* $\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$.

Remark 0.5. *Quantiles The **quantiles** of a random variable X are given by the inverse of its cumulative distribution function. Generally, the **quantile function** is*

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

If F_X is invertible, then

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

Remark 0.6. *For any function F that satisfies the properties of a cdf listed above, we can construct a random variable whose cdf is F. Let U be uniformly distributed on $[0, 1]$. That is, $F_U(u) = u$ for all $u \in [0, 1]$. Define $Y = Q(U)$, where Q is the quantile function associated with F. In the case, where F is invertible, we have*

$$F_Y(y) = P(F^{-1}(U) \leq y) = P(U \leq F(y)) = F(y)$$

Discrete random variables

If F_X is constant except at a countable number of points (i.e. F_X is a step function), then we say that X is a **discrete random variable**. The size of the jump at x_i

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

is the probability that X takes on the value x_i. That is,

$$P(X = x_i) = p_i$$

The **probability mass function** (pmf) of X is defined as

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

It follows that we can write

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

Continuous random variables

If F_X can be written as

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

where $f_X(x)$ satisfies

$$f_X(x) \geq 0 \quad \forall x \in \mathbb{R}$$
$$\int_{-\infty}^{\infty} f_X(t)dt = 1,$$

we say that X is a **continuous random variable**. By the fundamental theorem of calculus, at the points where f_X is continuous,

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

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

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

the **support** of X.

Note that for $x_2 \geq x_1$,

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

and that

$$P(X = x) = 0$$

for a continuous random variable.

Remark 0.7. *Do not interpret the pdf of a continuous random variable as expressing a probability ($f_X(x) \neq P(X = x)$). The proper interpretation is that $f_X(x)$ expresses the probability that X falls in some small interval $(x, x + \Delta x)$. That is,*

$$P(X \in (x, x + \Delta x)) \approx f(x)\Delta x$$

Joint distributions

Let X, Y be two scalar random variables. A **random vector** (X, Y) is a mapping from Ω to $\mathbb{R}^2$.[3] The **joint cumulative distribution function** of X, Y is

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

[3] We can extend the formal definitions for random variables to random vectors. (X, Y) is a random vector if and only if (X, Y) is P-measurable. That is, $(X, Y)^{-1}(B) \in A$ for all $B \in \mathbb{B}$, where $\mathbb{B}$ is now the Borel σ-algebra on $\mathbb{R}^2$.

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). 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 in the univariate case,

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

at the points of continuity of $F_{X,Y}$. From the joint cdf of (X, Y), we can recover the **marginal cdfs**. We have that

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

We can also recover the **marginal pdfs** from the joint pdf using

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

, and

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

Consider the discrete case. Let x be such that $f_X(x) > 0$. Then, the **conditional pmf of Y given $X = x$** is

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

This satisfies the following two properties:

$$\begin{aligned} f_{Y|X}(y|x) &\geq 0 \\ \sum_y f_{Y|X}(y|x) &= 1. \end{aligned}$$

That is, $f_{Y|X}(y|x)$ is a well-defined pmf for a discrete random variable. The **conditional cdf** of Y given $X = x$ is then

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

.

Next, consider the continuous case. It is analogous. For any x such that $f_X(x) > 0$, the **conditional pdf** of Y given $X = x$ is

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

Provided that $f_X(x) > 0$, this is a well-defined pdf for a continuous random variable. The **conditional cdf** is

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

Remark 0.8. *The conditional pmf for two discrete random variables can be interpreted as a probability. That is, for the discrete random vector* (X, Y)*,*

$$f_{Y|X}(y|x) = P(Y = y|X = x).$$

However, this is not true for continuous random variables because if X is continuous, $P(X = x) = 0$*. Instead, think about it as*

$$\begin{aligned} f_{Y|X}(x,y) &= P(X \in dx|Y = y) \\ &= \lim_{\Delta y \to 0} P(X \in dx|y \leq Y \leq y + dy) \end{aligned}$$

Finally, we extend the definition of independence to random variables. The random variables X, Y are **independent** if $F_{Y|X}(y|x) = F_Y(y)$ or equivalently, if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$. We can also define this in terms of the densities. X, Y are independent if $f_{Y|X}(y|x) = f_Y(y)$ or equivalently, if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

All of these definitions and results extend to n-dimensional random variables in a straightforward manner.

Transformations of Random Variables

Let X be a random variable with cdf F_X. Define the 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, suppose that X is discrete and takes on values $x_1, \ldots, x_n$. Y is also discrete and takes on the values

$$y_i = h(x_i), \quad \text{for} \quad i = 1, \ldots, n.$$

We have that the pmf of Y is given by

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

$$f_Y(y) = f_X(h^{-1}(y_i))$$

Next, suppose that X is continuous. Consider the case where h is increasing. We have that

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(X \leq h^{-1}(y)) = F_X(h^{-1}(y)). \end{aligned}$$

It follows directly that

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= f_X(h^{-1}(y)\frac{dh^{-1}(y)}{dy} \end{aligned}$$

In the case were h is decreasing, we can analogously show that

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

Combining these two cases, we have that, in general,

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

Example 0.4. *Suppose $X \sim U[0,1]$ and $Y = X^2$. Over the support of X, this is a one-to-one transformation. We have that*

$$X = \sqrt{Y}, \quad dX/dY = \frac{1}{2}y^{-1/2}.$$

Applying the formula above, we have that $S_Y = [0,1]$ and

$$f_Y(y) = \frac{1}{2}y^{-1/2}.$$

This can extended to the multivariate case. Let X be a random vector and as before, define $Y = h(X)$. You can show that

$$f_Y(y) = f_X(h^{-1}(x))|J|$$

where $|J|$ is the absolute value of the determinant of the Jacobian matrix of the inverse transformation. That is, $|J|$ is the absolute value of the determinant of the matrix whose i,j-th entry is $\partial x_i/\partial y_j$.

Expectations

Suppose X is 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, its expectation is said to not exist. Suppose X is a continuous random variable. Its expectation is defined as

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

if $\int_{S_X} |x| f_X(x) dx < \infty$. Otherwise, its expectation is said to not exist.[4] We can also define the expectation of functions of random variables. Let $g : \mathbb{R} \to \mathbb{R}$. Then, if X is discrete,

[4] Formally, the expectation is defined using the Lebesgue-Stieltjes integral.

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

and if X is continuous, then

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

The following are useful properties of the expectation operator. Suppose $a, b \in \mathbb{R}$ and $g_1(\cdot), g_2(\cdot)$ are real-valued functions.

1. $E[a] = a$.
2. $E[a g_1(X)] = a E[g_1(X)]$.
3. $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$.

These properties together imply that the expectation is a *linear operator*.

We can use the expectation operator to express probabilities. An **indicator function** $1(A)$ is a function that is equal to one if condition A is true and zero otherwise. For example, if X is a random variable, then

$$1(X \leq x) = \begin{cases} 1 & \text{if} \quad X \leq x \\ 0 & \text{otherwise} \end{cases}$$

Note that (for the continuous case)

$$\begin{aligned} E[1(X \leq x)] &= \int_{-\infty}^{\infty} 1(X \leq x) f_X(x) dx \\ &= \int_{-\infty}^{x} f_X(x) dx \\ &= F_X(x) = P(X \leq x). \end{aligned}$$

More generally, if $A_X \subseteq \mathbb{R}$, we have that

$$E[1(X \in A_X)] = P(X \in A_X)$$

This is a very useful result.

Suppose X, Y are random variables with joint density $f_{X,Y}(x, y)$. Let $g(x, y) : \mathbb{R}^2 \to \mathbb{R}$. We have that

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

Note that by linearity of the expectation, for $a, b \in \mathbb{R}$,

$$E[aX + bY] = aE[X] + bE[Y].$$

Finally, if X, Y are independent, then for any functions $h_1(\cdot), h_2(\cdot)$,

$$E[h_1(X)h_2(Y)] = E[h_1(X)]E[h_2(Y)].$$

All of these results generalize directly to higher dimensions.

Conditional expectations

Given a pair of random variables (X, Y) with a joint density $f_{X,Y}(x, y)$, we can define the **conditional expectation** of Y given $X = x$ as

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

Note that this is a function of x. It is sometimes denote $\mu_Y(x)$ and called the **regression function**. In particular, this means we can view this a random function $E[Y|X]$. The following theorem is extremely useful. [5]

[5] This might be the most important thing we cover in math camp!

Theorem 0.4. *The law of iterated expectations* [6]

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

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.

[6] This is also called the Tower Property.

Proof. We have that

$$\begin{aligned} E_X E_{Y|X}[Y] &= \int \Big(\int y f_{Y|X}(y)dy \Big) f_X(x)dx \\ &= \int \int y f_{Y|X}(y) f_X(x)dydx \\ &= \int y \Big(\int f_{X,Y}(x, y)dx \Big) dy \\ &= \int y f_Y(y)dy = E[Y] \end{aligned}$$

□

What are some ways to interpret the conditional expectation? We provided a formal definition but we also want to provide some intuition. First, the conditional expectation is the solution to an *optimal forecasting* problem. Suppose you wish to forecast the value of a random variable Y. That is, we wish to pick $h \in \mathbb{R}$ that minimizes the expected mean-square error

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

The first-order condition is

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

That is, the optimal prediction of Y is $E[Y]$. [7] Now, suppose that we observe another random variable X and see that $X = x$. We wish to forecast Y as a function of x. That is, we wish to minimize

$$E[(Y - h(X))^2].$$

[7] Optimal with respect to expected mean-square error. If we changed the objective function to expected mean-absolute error, $E[|Y - h|]$, the solution is the median of Y, $h^* = median(Y)$.

Note that we can always write any function of X as

$$h(x) = \mu_Y(x) + g(x)$$

by defining $g(x) = h(x) - \mu_Y(x)$. So choosing h to minimize expected mean-square error is equivalent to choosing g. We can then write

$$(Y - h(X))^2 = (Y - \mu_Y(X))^2 - 2g(X)(Y - \mu_Y(x)) + g(X)^2.$$

I claim that [8]

$$E_{Y|X}[g(X)(Y - \mu_Y(x))] = 0$$

[8] Can you show these steps?

and so,

$$E[(Y - h(X))^2] = E[(Y - \mu_Y(X))^2 + g(X)^2]$$

. It then follows immediately that expected mean-squared error is minimized with $g(x) = 0$ and so,

$$h^*(x) = \mu_Y(x).$$

That is, the conditional expectation of Y given X is the optimal predictor of Y given X.[9]

[9] And with that, you have learned a good chunk of machine learning. I am not joking.

Second, we can interpret the conditional expectation of Y given X as the orthogonal projection of Y onto the space of functions of the random variable X i.e. L^2 space. Since this interpretation of the conditional expectation is the focus of the first several lectures of Econ 2120, I will not cover it here.

Moments and moment generating functions (MGFs)

Consider a random variable X. The ***k*-th moment of** X is defined as $E[X^k]$. The first moment of X is its **mean**, $E[X]$. The ***k*-th centered moment of** X is $E[(X - E[X])^k]$. The second centered moment of X is its **variance**, $V(X) = E[(X - E[X])^2]$. The **standard deviation** of X is $\sqrt{V(X)}$.

Remark 0.9. *Suppose X has mean μ_X and variance σ_X^2. Let $a, b \in \mathbb{R}$ and define $Y = a + bX$. Then,*

$$\mu_Y = a + b\mu_X, \quad \sigma_Y^2 = b^2\sigma_X^2.$$

Definition 0.14. *The **moment generating function** (MGF) of a random variable X is defined as*

$$\mu_X(t) = E[e^{tX}] = \int e^{tx} f_X(x)dx.$$

The MGF of X is useful because it allows us to easily compute all of the moments of a random variable. Note that

$$\mu'_X(t) = \int xe^{tx} f_X(x)dx, \quad \mu'_X(0) = \int x f_X(x)dx = E[X],$$
$$\mu''_X(t) = \int x^2 e^{tx} f_X(x)dx, \quad \mu''_X(0) = \int x^2 f_X(x)dx = E[X^2].$$

In general, we can show that

$$\mu_X^{(j)}(0) = E[X^j] \quad \text{for} \quad j = 1, 2, \ldots$$

Moreover, the MGF of a random variable completely characterizes the distribution of a random variable. If X, Y are two random variables with the same MGF, then they have the same distribution.

Remark 0.10. *The MGF may not always exist for a random variable. For example, e^{tX} may blow up for large realizations of X. However, the **characteristic function** of X is guaranteed to exist. It is defined as*

$$E[e^{itx}], \quad i = \sqrt{-1}.$$

The characteristic function is guaranteed to exist and it completely characterizes the distributed of X.

Moments for random vectors

Suppose X, Y are two random variables with joint density $f_{X,Y}(x, y)$. The **covariance** between X, Y 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 is a linear operator. That is,

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

Moreover, suppose $Z = aX + bY$ for $a, b \in \mathbb{R}$. Then,

$$V(Z) = a^2 V(X) + b^2 V(Y) + 2abCov(X, Y).$$

Now suppose that X is an n-dimensional random vector with $X = (X_1, \ldots, X_n)$. Its **mean vector** is

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

and its **covariance matrix** is

$$V(X) = \Sigma$$

where Σ is an $n \times n$ matrix whose ij-th entry is $\Sigma_{ij} = Cov(X_i, X_j)$. Σ is a positive semi-definite matrix. Why? Let $\alpha \in \mathbb{R}^n$ and define $Y = \alpha^T X$. Then,

$$V(Y) = \alpha^T \Sigma \alpha \geq 0.$$

This must hold for all $\alpha \in \mathbb{R}^n$.

Useful Probability Distributions

Bernoulli distribution

X is a discrete random variable that can only take on two values: $0, 1$. We write $f_X(1) = p$, $f_X(0) = 1 - p$ and so,

$$f_X(x) = p^x (1-p)^{1-x}.$$

Note that

$$\begin{aligned} E[X^k] &= p, \quad k \geq 1 \\ V(X) &= p(1-p), \\ \mu_X(t) &= (1-p) + pe^t. \end{aligned}$$

We say that X has a **Bernoulli distribution**.

Binomial distribution

Suppose that X_i for $i = 1, \ldots, n$ are i.i.d Bernoulli random variables with $P(X_i = 1) = p$. Define $X = \sum_{i=1}^n X_i$. We say that X follows a **binomial distribution** with parameters n and p. X takes on values $1, 2, \ldots, n$. Its pmf is

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

and

$$E[X] = np, \quad V(X) = np(1-p).$$

Poisson distribution

Suppose that X is a discrete random variables and takes on values $1, 2, 3, \ldots$. Its pmf is

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \lambda > 0$$

We say that X is a **Poisson** random variable with parameter $\lambda > 0$. Note that

$$E[X] = \lambda, \quad V(X) = \lambda.$$

Poisson random variables are typically used to model the number of discrete "successes" that occur over a time period.

Remark 0.11. *Note that if X_n is binomially distributed with parameters $n, p = \lambda/n$, then $X_n \xrightarrow{d} X$, where X is a Poisson random variable.*[10]

[10] The notation $X_n \xrightarrow{d} X$ means that the sequence of random variables X_n "converges in distribution" to the random variable X. We have not formally defined this yet but intuitively, it means that as n gets large, X_n behaves as if it were distributed like X.

Uniform distribution

Suppose that X is a continuous random variable with $f_X(x) = \frac{1}{b-a}$ for $x \in [a, b]$ and 0 otherwise. We say that X is **uniformly distributed on [a, b]** and write $X \sim U[a, b]$.

Univariate normal distribution

Suppose Z is continuously distributed with support over $\mathbb{R}$. We say that X follows a **standard normal distribution** if

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

We write $Z \sim N(0,1)$ where $E[Z] = 0, V(Z) = 1$. We say that $X \sim N(\mu, \sigma^2)$ if $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ Note that $E[X] = \mu, V(X) = \sigma^2$ and $X = \mu + \sigma Z$, where $Z \sim N(0,1)$.

The MGF of a standard normal random variable is incredibly useful. [11] It is worth memorizing. If $Z \sim N(0,1)$, then

$$M_Z(t) = e^{\frac{1}{2}t^2}.$$

[11] For example, you'll run into it all the time in macro/finance.

Why? Here's the calculation: [12]

[12] It's straightforward provided you remember how to complete the square.

$$\begin{aligned}
M_Z(t) &= E[e^{tZ}] \\
&= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tz - \frac{1}{2}z^2} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} - \frac{1}{2} e^{-\frac{1}{2}(z^2 - 2tz)} dz \\
&= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2 - 2tz + t^2)} dz \\
&= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz \\
&= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(w)^2} dz, \quad w = z - t \\
&= e^{\frac{1}{2}t^2}
\end{aligned}$$

We can use this to derive the MGF for $X \sim N(\mu, \sigma^2)$. We have that

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= E[e^{t(\mu+\sigma Z)}] \\ &= e^{t\mu} E[e^{t\sigma Z}] \\ &= e^{t\mu} M_Z(t\sigma) \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \end{aligned}$$

Therefore, we have that

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

Chi-squared distribution

Let $Z_i \sim N(0,1)$ i.i.d. for $i = 1, \ldots, n$. Let

$$X = \sum_{i=1}^{n} Z_i^2.$$

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

$$E[X] = n, \quad V(X) = 2n$$

.

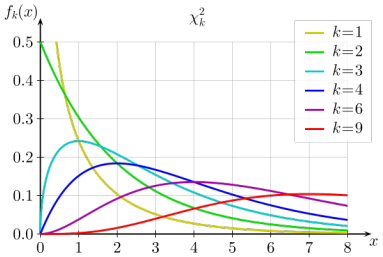

Figure 1: Density of χ^2 as degree of freedom varies. (Source: Wikipedia)

F-distribution

Let $Y_1 \sim \chi_k^2$, $Y_2 \sim \chi_l^2$ with $Y_1 \perp Y_2$. Define

$$Q = \frac{Y_1/k}{Y_2/l}.$$

We say that Q follows an **F-distribution** with k, l degrees of freedom. We write $Q \sim F_{k,l}$.

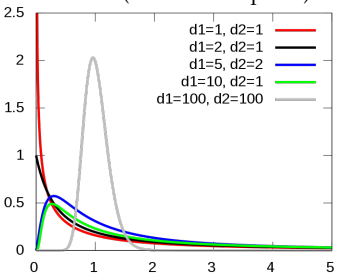

Figure 2: Density of $F_{k,l}$ as degrees of freedom vary. (Source: Wikipedia)

Student t-distribution

Let $Z \sim N(0,1)$ and $Y \sim \chi_n^2$ with $Z \perp Y$. Define

$$T = Z/\sqrt{Y/n}$$

We say that T is **student t-distributed** with n **degrees of freedom** and write $T \sim t_n$. We have that

$$E[T] = 0$$

$$V(T) = \begin{cases} \frac{n}{n-2}, & \text{if} \quad n > 2, \\ \infty, & n = 1, 2 \end{cases}$$

Remark 0.12. *As $n \to \infty$, $T_n \xrightarrow{d} Z \sim N(0,1)$. This result is the foundation of asymptotic inference in econometrics.*

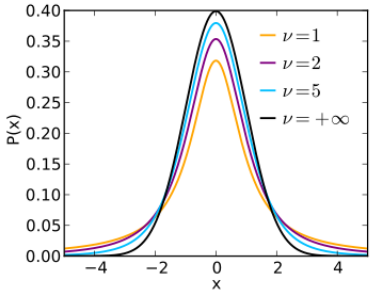

Figure 3: Density of t_n as degree of freedom varies. (Source: Wikipedia)

Exponential distribution

Suppose that X is a continuous random variable with support over $\mathbb{R}_+$. X is **exponentially distributioned** with parameter $\lambda > 0$ if

$$f_X(x) = \lambda e^{-\lambda x}.$$

We write $X \sim exp(\lambda)$ and have that

$$E[X] = \frac{1}{\lambda}$$

$$V(X) = \frac{1}{\lambda^2}$$

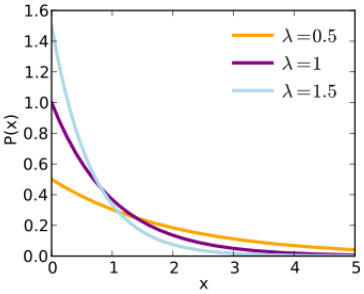

Figure 4: Distribution of $exp(\lambda)$ as λ. (Source: Wikipedia)

Multivariate Normal Distribution

Consider the random vector $Z = (Z_1, \ldots, Z_m)'$, where each $Z_i \sim N(0,1)$ i.i.d. The joint density of Z is given by

$$\begin{aligned} f_Z(z) &= \Pi_{i=1}^m \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} \\ &= (1/2\pi)^{n/2} \exp(-\frac{1}{2}\sum_{i=1}^m z_i^2) \\ &= (2\pi)^{n/2} \exp(-\frac{1}{2}z'z) \end{aligned}$$

Moreover, note that $E[Z] = 0$ and $V(Z) = I_m$. Finally, the MGF of Z is given by

$$\begin{aligned} M_Z(t) &= E[e^{t'Z}] \\ &= E[\Pi_{i=1}^m e^{t_i z_i}] \\ &= \Pi_{i=1}^m E[e^{t_i z + i}] = e^{\frac{1}{2}t't} \end{aligned}$$

This is a useful reference point as we develop some results about the multivariate normal distribution.

Definition 0.15. *A m-dimensional random vector X follows a* ***m-dimensional multivariate normal distribution*** *if and only if*

$$a^T X$$

is normally distributed for all $a \in \mathbb{R}^m$. *We write* $X \sim N_m(\mu, \Sigma)$, *where* $E[X] = \mu$ *is the m-dimensional mean vector and* $V(X) = \Sigma$ *is the* $m \times m$ *dimensional covariance matrix.*[13]

Remark 0.13. *If X follows a multivariate normal distribution, then each element* X_i *follows a univariate normal distribution with mean* μ_i *and variance* Σ_{ii}.

[13] Typically, the dimensional is suppressed in the notation. That is, if X is m-dimensional and follows a multivariate normal distribution, we will typically write $X \sim N(\mu, \Sigma)$. The dimensions of μ, Σ are implied by the context.

Remark 0.14. *It is important to have a good handle on the properties of the univariate and multivariate normal distributions. When we use asymptotics to approximate the finite-sample distribution of estimators and test-statistics in econometrics, everything "becomes" normally distributed by the central theorem.* [14]

[14] Not literally everything but you get the point.

The next two results will allow us to derive the distribution of a multivariate normal. We first derive its MGF.

Proposition 0.2. *Suppose $X \sim N(\mu, \Sigma)$. Then,*

$$M_X(t) = e^{t'\mu + \frac{1}{2}t'\Sigma t}.$$

Proof. Note that $t'X \sim N(t'\mu, t'\Sigma t)$. Therefore,

$$\begin{aligned} M_X(t) &= E[e^{t'X}] \\ &= E[e^Y], \quad Y \sim N(t'\mu, t'\Sigma t) \\ &= M_Y(1) \end{aligned}$$

and the result follows. □

Recall that for a univariate normal distribution, if $X \sim N(\mu, \sigma^2)$, then $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$. The same property holds for the multivariate normal distribution.

Proposition 0.3. *Suppose $X \sim N_m(\mu, \Sigma)$. Define*

$$Y = AX + b,$$

where $A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n$. Then,

$$Y \sim N_n(A\mu + b, A\Sigma A').$$

Proof. For $t \in \mathbb{R}^n$,

$$\begin{aligned} M_Y(t) &= E[e^{t'Y}] \\ &= E[e^{t'(AX+b)}] \\ &= e^{t'b}E[e^{(A't)'X} \\ &= e^{t'b}e^{(A't)'\mu + \frac{1}{2}(A't)'\Sigma(A't)'} \\ &= e^{t'(A\mu+b)+\frac{1}{2}t'(A\Sigma A')t} \end{aligned}$$

□

We'll now use the two previous results to derive the density of a multivariate normal distribution.

Proposition 0.4. *Suppose $X \sim N(\mu, \Sigma)$ and Σ has full column rank. Then, the density of X is given by*

$$f_X(x) = (2\pi)^{-m/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)).$$

Proof. Let Z be a m-dimensional random vector of i.i.d. standard normal random variables. At the beginning of this section, we derived that $M_Z(t) = e^{\frac{1}{2}t't}$. Therefore, $Z \sim N_m(0, I_m)$. We also derived that the density of Z is

$$f_Z(z) = (2\pi)^{-m/2} e^{-\frac{1}{2}z'z}.$$

Let $X = \mu + \Sigma^{1/2}Z$. We can show that $X \sim N_m(\mu, \Sigma)$. From the multivariate transformation of random variables formula from an earlier section,

$$f_X(x) = |\Sigma|^{-1/2} f_Z(\Sigma^{-1/2}(x - \mu))$$

and the result follows. □

The rest of this section lists additional useful properties of the multivariate normal distribution that will appear from time to time. It's useful to be familiar with them.

Proposition 0.5. *If $X_1 \sim N_m(\mu_1, \Sigma_1), X_2 \sim N_n(\mu_2, \Sigma_2)$ and $X_1 \perp X_2$, then*

$$X = (X_1', X_2')' \sim N_{m+n}(\mu, \Sigma)$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}$$

Proposition 0.6. *Let $X \sim N_m(\mu, \Sigma)$. Let X_1 be a p-dimensional sub-vector of X with $p < m$. Write*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

and

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then, $X_1 \sim N_p(\mu_1, \Sigma_{11})$.

Proposition 0.7. *Let $X \sim N_m(\mu, \Sigma)$. Partition X into two sub-vectors. That is, write*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

and

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then, $X_1 \perp X_2$ if and only if $\Sigma_{12} = \Sigma_{21} = 0$.

Proposition 0.8. *Let $X \sim N_m(\mu, \Sigma)$. If*

$$Y = AX + b, \quad V = CX + d,$$

where $A, C \in \mathbb{R}^{n \times m}$ and $b, d \in \mathbb{R}^n$, then

$$Cov(Y, V) = A\Sigma C'.$$

Moreover, $Y \perp V$ if and only if

$$A\Sigma C' = 0.$$

Exercise 0.3. *Prove these properties of the multivariate normal distribution.*

Proposition 0.9. *Let $X \sim N_m(\mu, \Sigma)$ with $X = (X_1', X_2')'$, $\mu = (\mu_1', \mu_2')'$ and*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Provided that Σ_{22} has full rank, the conditional distribution of X_1 given $X_2 = x_2$ is

$$X_1 | X_2 = x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Remark 0.15. *What's the intuition of this? We have that*

$$E[X_1 | X_2 = x_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2).$$

This formula will look more familiar if everything is one-dimensional. It becomes

$$E[X_1 | X_2 = x_2] = E[X_1] + \frac{Cov(X_1, X_2)}{V(X_2)}(x_2 - E[X_2])$$

Is this starting to look more familiar? Not yet? Ok, let's relabel $Y = X_1, X = X_2$ and re-arrange. Then,

$$E[Y | X = x] = (E[Y] - \frac{Cov(Y, X)}{V(X)}E[X]) + \frac{Cov(Y, X)}{V(X)}x.$$

This is simply the linear regression formula![15] *For a multivariate normal random distribution, conditional expectations are exactly linear. As a result, linear regression exactly returns the conditional expectation function.*

[15] Set $\beta_0 = E[Y] - \frac{Cov(Y,X)}{V(X)}E[X]$ and $\beta_1 = \frac{Cov(Y,X)}{V(X)}$. Then, $E[Y|X = x] = \beta_0 + \beta_1 x$.

This final result provides the conditional distribution of a multivariate normal distribution. This appears at random points throughout the first year and so, it is useful to keep in your back pocket.

Quadratic forms of normal random vectors

Recall that a **quadratic form** is a quantity of the form $y'Ay$, where A is a symmetric matrix. Suppose that $Z_i \sim N(0, 1)$ i.i.d. for $i = 1, \ldots, n$. We already know that $\sum_{i=1}^n Z_i^2 = Z'Z \sim \chi_n^2$.

Proposition 0.10. *If $X \sim N_m(\mu, \Sigma)$ and Σ has full rank, then*

$$(X - \mu)'\Sigma^{-1}(X - \mu) \sim \chi_m^2.$$

Proof. Let $Z = \Sigma^{-1/2}(X - \mu) \sim N_m(0, I_m)$. Then, $Z'Z \sim \chi_m^2$. □

Jensen, Markov and Chebyshev, Oh My!

The following are some useful inequalities that pop up in a variety of contexts in econometrics and other areas of economics. These are especially useful in asymptotics.

Theorem 0.5. *Jensen's inequality*
Let $h(\cdot)$ be a convex function and X be a random variable. Then,

$$E[h(X)] \geq h(E[X]).$$

Proof. Recall that if $h\cdot$ is a convex function, then $\forall x_0$ in its domain, there exists a line through $(x_0, h(x_0))$ such that $h(x)$ never falls below the line. That is, there exists some constant a such that

$$h(x) \geq h(x_0) + a(x - x_0) \quad \forall x$$

Set $x_0 = E[x]$. It follows that

$$h(X) \geq h(E[X]) + a(x - E[X])$$

holds for all x. Taking expectations, we have that

$$E[h(X)] \geq h(E[X]).$$

□

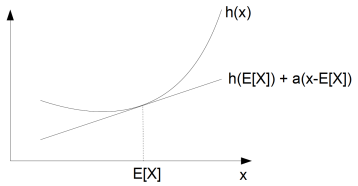

Figure 5: Picture proof of Jensen's inequality.

Remark 0.16. *If $h(\cdot)$ is concave, the opposite inequality holds. That is,*

$$E[h(X)] \leq h(E[X])$$

The next inequality (Markov's inequality) provides a bound tail behavior of a random variable as a function of its expectation.

Theorem 0.6. *Markov's inequality*
Suppose X is a random variable with $X \geq 0$ with $E[X] < \infty$. [16] *Then, for all $M > 0$,*

$$P(X \geq M) \leq \frac{E[X]}{M}.$$

[16] $X \geq 0$ for a random variable means that $P(\{\omega : X(\omega) < 0\}) = 0$.

Proof. The proof is straightforward. Because $X \geq 0$,

$$X \geq M1(X \geq M).$$

Taking expectations of both sides, we have that

$$E[X] \geq ME[1(X \geq M) = MP(X \geq M)$$

annd re-arrange to arrive at the result. □

Figure: Proof of Markov's inequality

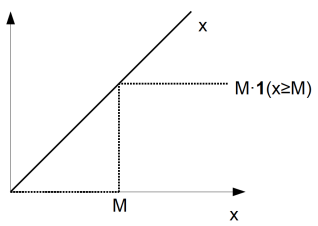

Figure 6: Picture proof of Markov's inequality.

Example 0.5. *Suppose that household income is non-negative. By Markov's inequality, no more than* $1/5$ *of households can have an income that is greater than five times the average household income.*

The final inequality (Chebyshev's inequality) is a corollary of Markov's inequality. It provides an upper bound on the probability that a random variable falls a certain distance from its expectation.

Theorem 0.7. *Chebyshev's inequality*

Suppose that X *is a random variable such that* $\sigma^2 = Var[X] < \infty$*. Then, for all* $M > o$*,*

$$P(|X - E[X]| > M) \leq \frac{\sigma^2}{M^2}.$$

Proof. Let $Y = (X - E[X])^2$. Apply Markov's inequality to Y and the cutoff M^2 to get

$$P(Y \geq M^2) \leq \frac{E[Y]}{M^2}.$$

Rewrite to get that

$$P(|X - E[X]| \geq M) \leq \frac{\sigma^2}{M^2}$$

□

Example 0.6. *Chebyshev's inequality is used in a proof of the weak law of large numbers (WLLN).* [17] *For now, WLLN states: as the sample size gets very large, the sample average of a random variable "converges" to the expectation of the random variable. One proof begins by showing that the variance of the sample average converges to zero and then uses Chebyshev's inequality to prove the result.*[18]

[17] The weak law of large numbers will be introduced in detail later.

[18] If that made no sense, don't worry about it. We'll go through this together.

References

Billingsley, P. (2012). *Probability and Measure*.
Blitzstein, J. and J. Hwang. (2014). *Introduction to Probability*.
Casella, G. and R. Berger. (2001). *Statistical Inference*.
Hogg, R., McKean, J. and A. Craig. (2012). *Introduction to Mathematical Statistics*.
Kolmogorov, A. and Fomin, S. (2012). *Introductory Real Analysis*.
Stokey, N., Lucas, R. and E. Prescott. (1989). *Recursive Methods in Economic Dynamics*.