University of Warwick
Math for Economics, 2009 - 10

Lecture 9: Probability Theory

1 Measure Theory

Suppose that we have fixed a universe S. Denote by $\mathcal{P}(S)$ the set of all subsets of S (that is, $E \in \mathcal{P}(S)$ iff $E \subseteq S$. Obviously, $\varnothing \in \mathcal{P}(S)$ and $S \in \mathcal{P}(S)$. One can also say that $\{\varnothing, S\} \subseteq \mathcal{P}(S)$ and $\varnothing \subseteq \mathcal{P}(S)$, but it would be a mistake to say that $S \subseteq \mathcal{P}(S)$.

1.1 Algebras and σ-algebras:

Our problem now is to define, in a consistent manner, the size of (some of) the subsets of S. The consistency of our definition will require some "structure" on the family of subsets whose size we define: (i) we should be able to tell the size of the set with no elements in it; (ii) if we are able to measure a set, we should also be able to measure the rest of the universe; and (iii) if we are able to measure a series of sets, then we should also be able to measure their union.

Definition 1. *A family of subsets of S, $\Sigma \subseteq \mathcal{P}(S)$, is* an algebra *if:*

1. *it contains the empty set: $\varnothing \in \Sigma$;*
2. *it is closed under complement: if $A \in \Sigma$, then $S \setminus A \in \Sigma$; and*
3. *it is closed under finite union: if $\{A_n\}_{n=1}^N \subseteq \Sigma$ is a finite set, then $\cup_{n=1}^N A_n \in \Sigma$.*

The following theorem is almost immediate from the definition.

Theorem 1. *If Σ is an algebra, then: (1) it contains S: $S \in \Sigma$; and (2) it is closed under finite intersection: if $\{A_n\}_{n=1}^N \subseteq \Sigma$ is a finite set, then $\cap_{n=1}^N A_n \in \Sigma$.*

Notice that the conditions of the definition of algebra have the intuition we wanted. For some purposes, however, we need to strengthen the third property:

Definition 2. *A family of subsets of S, $\Sigma \subseteq \mathcal{P}(S)$, is* a σ-algebra *if:*

1. *it contains the empty set: $\varnothing \in \Sigma$;*
2. *it is closed under complement: if $A \in \Sigma$, then $S \setminus A \in \Sigma$; and*
3. *it is closed under countable union: if $\{A_n\}_{n=1}^\infty \subseteq \Sigma$, then $\cup_{n=1}^N A_n \in \Sigma$.*

The following theorems are left as exercises.

Theorem 2. *$\mathcal{P}(S)$ is a σ-algebra.*

Theorem 3. *If Σ is a σ-algebra, then it is an algebra. When S is finite, if Σ is an algebra, then it is a σ-algebra.*

Theorem 4. *If Σ is a σ-algebra, then: (1) it contains S: $S \in \Sigma$; and (2) it is closed under countable intersection: if $\{A_n\}_{n=1}^\infty \subseteq \Sigma$, then $\cap_{n=1}^N A_n \in \Sigma$.*

If Σ is a σ-algebra for S, then (S, Σ) is said to be *a measurable space*. The idea here is that Σ is the collection of subsets of S that we can "measure." Now, what do we understand by "measuring"? Intuitively, what we want to do is to associate each set to a number. Of course, this assignment cannot be arbitrary: (i) sizes cannot be negative, and we must consider the possibility of finding an "infinitely large" set; (ii) a set that contains nothing must have zero measure; and (iii) if we take a collection of mutually separated sets and we measure them, and we then measure their union, the sum of the first measures must equal the last measure.

Formally, denote $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, \infty\}$, which is usually called the *extended real line*, and let us take its positive orthant: $\mathbb{R}^*_+ = \mathbb{R}_+ \cup \{\infty\}$. Let Σ be an algebra for S and let $\mu : \Sigma \to \mathbb{R}^*$. Function μ is said to be *finitely additive* if for any finite collection of mutually disjoint sets in Σ, $\{A_n\}_{n=1}^N$, one has that $\mu(\cup_{n=1}^N A_n) = \sum_{n=1}^N \mu(A_n)$. It is said to be *$\sigma$-additive* if for any sequence $\{A_n\}_{n=1}^\infty$ of mutually disjoint sets in Σ, similarly, $\mu(\cup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mu(A_n)$.

Obviously, if μ is σ-additive, then it is also finitely additive. It is also immediate that if S is finite, Σ is an algebra for S and $\mu : \Sigma \to \mathbb{R}^*$is finitely additive, then (S, Σ) is a measurable space and μ is σ-additive.

It is obvious that the structure imposed when we consider arbitrary sequences is more than when we consider finite sequences only. But, is the extra complication necessary? To see that it is, consider the following experiment: a coin is tossed until if comes head. Suppose that we want to measure the probability that the experiment stops in an even toss. We need to consider countable but infinite sequences! Notice that σ-additivity corresponds to the condition (iii) that we want to impose to our measures.

Let (S, Σ) be a measurable space and let $\mu : \Sigma \to \mathbb{R}^*$. Function μ is said to be *a measure* if it is σ-additive and satisfies that $\mu(\varnothing) = 0$ and $\mu(A) \in \mathbb{R}^*_+$ for all $A \in \Sigma$. A *measure space* is (S, Σ, μ), where (S, Σ) is a measurable space and $\mu : \Sigma \to \mathbb{R}^*_+$ is a measure.

EXERCISE 1. *Prove the following results:*

1. *If $X : S \to \mathbb{R}$ is $\{\varnothing, S\}$-measurable, then $X(s) = X(s')$ for all $s, s' \in S$.*
2. *Any function $X : S \to \mathbb{R}$ is $\mathcal{P}(S)$-measurable.*
3. *$\sigma(\{S\}) = \{\varnothing, S\}$.*
4. *If Σ is a σ-algebra, then $\sigma(\Sigma) = \Sigma$.*

2 Probability

When the space S we are dealing with is the space of all possible results of an experiment, the subsets we want to measure are called events and their measures are understood as probabilities (which can be understood from either a frequentist or a likelihood perspective). Now, if (S, Σ, p) is a measure space and $\mathrm{p}(S) = 1$, we say that S is a *sample space*, that (S, Σ, p) is a *probability space* and that p is a *probability measure*. So defined, the properties we impose for p to be considered a probability measure are known as *Kolmogorov's axioms*; their arguments are left as exercises.

THEOREM 5. *Let (S, Σ, p) be a probability space, let $E, E' \in \Sigma$, and let $(E_n)_{n=1}^\infty$ be a sequence in Σ. Then,*

1. $\mathrm{p}(E) \leq 1$ *and* $\mathrm{p}(E^c) = 1 - \mathrm{p}(E)$, *where* $E^c = S \setminus E$;
2. $\mathrm{p}(E' \cap E^c) = \mathrm{p}(E') - \mathrm{p}(E' \cap E)$ *and* $\mathrm{p}(E \cup E') = \mathrm{p}(E) + \mathrm{p}(E') - \mathrm{p}(E \cap E')$;

3. $p(E) = \sum_{n=1}^{N} p(E \cap E_n)$, *if* $(E_n)_{n=1}^{\infty}$ *is a partition of* S;[1]

4. Bonferroni's simple inequality: $\mathrm{p}(E \cap E') \geq \mathrm{p}(E) + \mathrm{p}(E') - 1$;

5. Boole's inequality: $\mathrm{p}(\cup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} \mathrm{p}(E_n)$;

6. *if* $E_n \subseteq E_{n+1}$ *for each* n, *then* $\mathrm{p}(\cup_{n=1}^{\infty} E_n) = \lim_{n\to\infty} \mathrm{p}(E_n)$; *and*

7. *if* $E_n \supseteq E_{n+1}$ *for each* n, *then* $\mathrm{p}(\cap_{n=1}^{\infty} E_n) = \lim_{n\to\infty} \mathrm{p}(E_n)$.

3 Conditional Probability and Independence

Henceforth, we maintain a probability space, (S, Σ, p), fixed.

3.1 Conditional Probability

Let $E^* \in \Sigma$ be such that $\mathrm{p}(E^*) > 0$. The *probability measure given (or conditional on)* E^* is defined by $\mathrm{p}(\cdot|E^*) : \Sigma \to [0, 1]$, with $\mathrm{p}(E|E^*) = \mathrm{p}(E \cap E^*)/\mathrm{p}(E^*)$.

THEOREM 6. *Let* $E, E' \in \Sigma$ *and suppose that* $\mathrm{p}(E) \in (0, 1)$. *Then,*

$$\mathrm{p}(E') = \mathrm{p}(E)\mathrm{p}(E'|E) + (1 - \mathrm{p}(E))\mathrm{p}(E'|E^c).$$

Proof: By definition,

$$\begin{aligned} \mathrm{p}(E)\mathrm{p}(E'|E) + (1 - \mathrm{p}(E))\mathrm{p}(E'|E^c) &= \mathrm{p}(E)\frac{\mathrm{p}(E' \cap E)}{\mathrm{p}(E)} + (1 - \mathrm{p}(E))\frac{\mathrm{p}(E' \cap E^c)}{\mathrm{p}(E^c)} \\ &= \mathrm{p}(E' \cap E) + \mathrm{p}(E' \cap E^c) \\ &= \mathrm{p}(E'), \end{aligned}$$

because of σ-additivity. *Q.E.D.*

The previous theorem, in fact, admits the following generalization, whose proof is left as an exercise.

THEOREM 7. *Let* $(E_n)_{n=1}^{N}$ *be a partition of* S *such that every* $\mathrm{p}(E_n) > 0$. *Then, for any* $E' \in \Sigma$, *one has that* $\mathrm{p}(E') = \sum_{n=1}^{N} \mathrm{p}(E_n)\mathrm{p}(E'|E_n)$.

THEOREM 8. *Let* $(E_n)_{n=1}^{N}$ *be a finite sequence of sets in* Σ, *such that* $\mathrm{p}(\cap_{n=1}^{N-1} E_n) > 0$. *Then,*

$$\mathrm{p}(\cap_{n=1}^{N} E_n) = \mathrm{p}(E_1)\mathrm{p}(E_2|E_1)\mathrm{p}(E_3|E_1 \cap E_2)...\mathrm{p}(E_N| \cap_{n=1}^{N-1} E_n).$$

Proof: The proof is left as an exercise. (Hint: recall mathematical induction!) *Q.E.D.*

[1] A partition of S is a sequence $(E_n)_{n=1}^{N}$, with N finite or equal to infinity, of pairwise disjoint sets in Σ such that $\cup_{n=1}^{N} E_n = S$.

3.2 Independence

A family of events $\mathcal{E} \subseteq \Sigma$ is *pairwise independent* if $\mathrm{p}(E \cap E') = \mathrm{p}(E)\mathrm{p}(E')$ for any two distinct $E, E' \in \mathcal{E}$. It is *independent* if

$$\mathrm{p}(\cap_{n=1}^{N} E_n) = \prod_{n=1}^{N} \mathrm{p}(E_n)$$

for any finite subfamily $\{E_n\}_{n=1}^{N}$ of distinct sets in $\mathcal{E}$.

Example 1. *Notice that pairwise independence does not suffice for independence. Let* $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ *and* $\Sigma = \mathcal{P}(S)$, *and suppose that* : $\mathrm{p}(\{s\}) = 1/9$ *for each* $s \in S$. *Let* $E_1 = \{1, 2, 7\}$, $E_2 = \{3, 4, 7\}$ *and* $E_3 = \{5, 6, 7\}$, *so* $\mathrm{p}(E_1) = \mathrm{p}(E_2) = \mathrm{p}(E_3) = 1/3$. *Now, if* $i, j \in \{1, 2, 3\}$, $i \neq j$, *then* $\mathrm{p}(E_i \cap E_j) = 1/9 = \mathrm{p}(E_i)\mathrm{p}(E_j)$, *but*

$$\mathrm{p}(E_1 \cap E_2 \cap E_3) = 1/9 \neq 1/27 = \mathrm{p}(E_1)\mathrm{p}(E_2)\mathrm{p}(E_3),$$

so this family is pairwise independent, but not independent.

Notice that $\cap_{E \in \mathcal{E}} E = \varnothing$ is neither necessary nor sufficient for independence.

Theorem 9. *If* $\{E, E'\}$ *is independent, then so is* $\{E^c, E'\}$.

4 Random Variables, Moments and Independence

Fix a measurable space (S, Σ).

4.1 Random Variables

Function $X : S \to \mathbb{R}$ is *measurable with respect to* Σ *(or* Σ*-measurable)* if for every $x \in \mathbb{R}$, one has that

$$\{s \in S | X(s) \leq x\} \in \Sigma.$$

Theorem 10. *If* X *is* Σ*-measurable, then for all* $x \in \mathbb{R}$, *the following sets lie in* Σ: $\{s \in S | X(s) \geq x\}$, $\{s \in S | X(s) < x\}$, $\{s \in S | X(s) > x\}$ *and* $\{s \in S | X(s) = x\}$.

A *random variable (in* $\mathbb{R}$*)* is a Σ-measurable function $X : S \to \mathbb{R}$. Let us now endow the measurable space with a probability measure p and fix a random variable $X : S \to \mathbb{R}$. The *distribution function of* X is $F_X : \mathbb{R} \to [0, 1]$, defined by $F_X(x) = \mathrm{p}(\{s \in S | X(s) \leq x\})$. (Note that F_X is well defined because X is Σ-measurable.)

Theorem 11. *Let* F_X *be the distribution function of* X. *Then,*

1. $\lim_{x \to -\infty} F_X(x) = 0$ *and* $\lim_{x \to \infty} F_X(x) = 1$;
2. $F_X(x) \geq F_X(x')$ *whenever* $x \geq x'$;
3. F_X *is right continuous: for all* $x \in \mathbb{R}$, $\lim_{h \downarrow 0} F_X(x + h) = F_X(x)$.
4. $\mathrm{p}(\{s \in S | X(s) > x\}) = 1 - F_X(x)$;
5. $\mathrm{p}(\{s \in S | x < X(s) \leq x'\}) = F_X(x') - F_X(x)$, *whenever* $x \leq x'$;
6. $\mathrm{p}(\{s \in S | X(s) = x\}) = F_X(x) - \lim_{h \uparrow 0} F_X(x + h)$.

Notice that it is not necessarily true that $\lim_{h\uparrow 0} F_X(x+h) = F_X$, so we cannot guarantee that F is continuous. It is a good exercise to find a case in which $\lim_{h\uparrow 0} F_X(x+h) \neq F_X$. It is also important to see which step in the obvious proof of left-continuity would fail:

$$\{s \in S | \exists n \in \mathbb{N} : X(s) \leq x - 1/n\} = \{s \in S | X(s) < x\},$$

which may be a proper subset of $\{s \in S | X(s) \leq x\}$.

The distribution function of a random variable characterizes (totally defines) its associated probability measure. Random variable X is said to be *continuous* if F_X is continuous. X is said to be *absolutely continuous* if there exists an integrable function $f_X : \mathbb{R} \to \mathbb{R}_+$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

for all $x \in \mathbb{R}$; in this case, f_X is said to be a *density function* of X.

EXAMPLE 2 (Standard Uniform Distribution). *Suppose that the distribution function of X is*

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0; \\ x, & \text{if } 0 \leq x \leq 1; \\ 1, & \text{if } x > 1. \end{cases}$$

Notice that X is absolutely continuous, and one density function is

$$f_X(x) = \begin{cases} 0, & \text{if } x < 0; \\ 1, & \text{if } 0 \leq x \leq 1; \\ 0, & \text{if } x > 1. \end{cases}$$

The closure of the set on which F is increasing is known as the *support of X*.

4.2 Moments

Henceforth, we assume that X is an absolutely continuous random variable, with density f_X. We assume that for every set $D \subseteq \mathbb{R}$ such that $\int_D f_X(x)\mathrm{d}x$ exists, it is true that

$$\mathrm{p}(\{s \in S | X(s) \in D\}) = \int_D f_X(x)\mathrm{d}x.$$

Moreover, for simplicity, the notation $\mathrm{p}(X \in D)$ will replace $\mathrm{p}(\{s \in S | X(s) \in D\})$ from now on.

Let $g : \mathbb{R} \to \mathbb{R}$ and define the random variable $g \circ X : S \to \mathbb{R}$. The *expectation of $g \circ X$* is defined as

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\mathrm{d}x,$$

if the integral exists.

THEOREM 12. *Let $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ and $a, b, c \in \mathbb{R}$. Then,*

1. $\mathrm{E}(ag_1(X) + bg_2(X) + c) = a\mathrm{E}(g_1(X)) + b\mathrm{E}(g_2(X)) + c$;
2. *if $g_1(x) \geq 0$ for all $x \in \mathbb{R}$, then* $\mathrm{E}(g_1(X)) \geq 0$;

3. if $g_1(x) \geq g_2(x)$ for all $x \in \mathbb{R}$, then $\mathrm{E}(g_1(X)) \geq \mathrm{E}(g_2(X))$;

4. if $a \leq g_1(x) \leq b$ for all $x \in \mathbb{R}$, then $a \leq \mathrm{E}(g_1(x)) \leq b$.

THEOREM 13 (Chebychev's Inequality). *Let $g : \mathbb{R} \to \mathbb{R}_+$ be such that $\mathrm{E}(g(X)) \in \mathbb{R}$. Then, for all $r > 0$,*

$$\mathrm{p}(g(X) \geq r) \leq \frac{\mathrm{E}(g(X))}{r}.$$

Proof: By definition,

$$\begin{aligned}
\mathrm{E}(g(X)) &= \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x \\
&\geq \int_{\{x \in \mathbb{R} | g(x) \geq r\}} g(x) f_X(x) \mathrm{d}x \\
&\geq \int_{\{x \in \mathbb{R} | g(x) \geq r\}} r f_X(x) \mathrm{d}x \\
&= r \int_{\{x \in \mathbb{R} | g(x) \geq r\}} f_X(x) \mathrm{d}x \\
&= r \mathrm{p}(g(X) \geq r),
\end{aligned}$$

where the first inequality follows since, for all $x \in \mathbb{R}$, $g(x) \geq 0$. *Q.E.D.*

For a finite integer k, the *k-th (non-central) moment of X* is $\mathrm{E}(X^k)$, whenever it exists. If $\mathrm{E}(X)$ exists in $\mathbb{R}$, the *k-th central moment of X* is $\mathrm{E}((X - \mathrm{E}(X))^k)$, whenever this integral exists. The first moment of X is its *expectation* or *mean*, and its second central moment is its *variance* and is denoted $\mathrm{V}(X)$.

EXERCISE 2. *Prove the following: Suppose that $\mathrm{E}(X)$ and $\mathrm{V}(X) > 0$ exist; then, for all $t > 0$ one has that*

$$\mathrm{p}(|X - \mathrm{E}(X)| \geq t\sqrt{\mathrm{V}(X)}) \leq \frac{1}{t^2}.$$

For instance, if $E(X)$ and $V(X) > 0$ exist, then.

$$p(|X - E(X)| < 2\sqrt{V(X)}) \geq 0.75.$$

Notice the surprising implication of the previous exercise: the probability that the realization of a random variable be at least two standard deviations from its mean is at least 0.75, regardless of its distribution!

4.3 Independence of Random Variables

Let $(X_n)_{n=1}^N$ be a finite sequence of random variables. The *joint distribution of* $(X_n)_{n=1}^N$ is $F_{(X_n)_{n=1}^N} : \mathbb{R}^N \to [0,1]$, defined by

$$F_{(X_n)_{n=1}^N}(x) = \mathrm{p}(\{s \in S | (X_n(s))_{n=1}^N \leq x\}).$$

Sequence $(X_n)_{n=1}^N$ is said to be *absolutely continuous* if there exists a function $f_{(X_n)_{n=1}^N} : \mathbb{R}^N \to \mathbb{R}_+$ such that

$$F_{(X_n)_{n=1}^N}(x) = \int_{v \leq x} f_{(X_n)_{n=1}^N}(u) \mathrm{d}u$$

for all $x \in \mathbb{R}^N$; in this case, function $f_{(X_n)_{n=1}^N}$ is said to be a *joint density function* for $(X_n)_{n=1}^N$. If the sequence is absolutely continuous, and if each random variable X_n has function f_{X_n} as a density function, then $(X_n)_{n=1}^N$ is said to be *independent* if $f_{(X_n)_{n=1}^N} = \prod_{n=1}^N f_{X_n}$. A sequence $(X_n)_{n=1}^\infty$ of random variables is said to be *independent* if every finite sequence $(X_{n_k})_{k=1}^K$ constructed with elements of $(X_n)_{n=1}^\infty$ is independent.

As before, if $(X_n)_{n=1}^N$ is a sequence of absolutely continuous random variables and $g : \mathbb{R}^N \to \mathbb{R}$, the *expectation of* $g((X_n)_{n=1}^N)$ is defined as

$$\mathrm{E}(g((X_n)_{n=1}^N)) = \int_{-\infty}^{\infty} g(x) f_{(X_n)_{n=1}^N}(x)\mathrm{d}x,$$

whenever the integral exists.

Theorem 14. *Let* $(X_n)_{n=1}^N$ *be a finite sequence of random variables. If* $(X_n)_{n=1}^N$ *is independent, then* $\mathrm{E}(\prod_{n=1}^N X_n) = \prod_{n=1}^N \mathrm{E}(X_n)$.

Exercise 3. *Prove the following corollary: if* (X_1, X_2) *is independent, then*

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}((X_1 - \mathrm{E}(X_1))(X_2 - \mathrm{E}(X_2))) = 0.$$

Perhaps because of the previous example, the idea of "no correlation" is oftentimes confused with independence. One must be careful about this: is two random variables are independent, then their correlation is zero; but the other causality is not true: if X is normal, then X and X^2 are uncorrelated, but they certainly are not independent.

5 Conditional Density and Conditional Expectation

For the purposes of this subsection, let (X_1, X_2) be a pair of random variables, assume that it is absolutely continuous and denote by $f_{(X_1,X_2)}$ its density function. The *marginal density of* X_1 is the function defined by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{(X_1,X_2)}(x_1, x_2)\mathrm{d}x_2,$$

for each $x_1 \in \mathbb{R}$. If $f_{X_1}(x_1) > 0$, the *conditional density of* X_2 *given that* $X_1 = x_1$ is given by the function

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{(X_1,X_2)}(x_1, x_2)}{f_{X_1}(x_1)},$$

for each $x_2 \in \mathbb{R}$.

These definitions are useful when one needs to "decompose" the bivariate problem: if one needs to know the probability that $X_1 \in [a, b]$, by definition one needs to compute

$$\int_a^b \int_{-\infty}^{\infty} f_{(X_1,X_2)}(x_1, x_2)\mathrm{d}x_2\mathrm{d}x_1 = \int_a^b f_{X_1}(x_1)\mathrm{d}x_1;$$

and, if one knows that the realization of X_1 is x_1, and one needs to compute the probability that $X_2 \in [a, b]$ given that knowledge, one simply needs to know $\int_a^b f_{X_2|X_1}(x_2|x_1)\mathrm{d}x_2$, as the conditional density given that $X_1 = x_1$ re-normalizes the "prior" density $f_{(X_1,X_2)}$ to take into account the knowledge on X_1.

Theorem 15. *If (X_1, X_2) is independent, then, for any x_1 such that $f_{X_1}(x_1) > 0$, one has that $f_{X_2|X_1}(x_2|x_1) = f_{X_2}(x_2)$.*

Let $g : \mathbb{R} \to \mathbb{R}$. The *conditional expectation of $g(X_2)$ given that $X_1 = x_1$* is

$$\mathrm{E}(g(X_2)|X_1 = x_1) = \int_{-\infty}^{\infty} g(x_2) f_{X_2|X_1}(x_2|x_1)\mathrm{d}x_2.$$

Similarly, the *conditional expectation of $g(X_2)$ given X_1*, which we denote $\mathrm{E}(g(X_2)|X_1)$, is the random variable that takes value $\mathrm{E}(g(X_2)|X_1 = x_1)$ in any state in which $X_1 = x_1$; in particular, the probability that $\mathrm{E}(g(X_2)|X_1 = x_1) \in [a, b]$ equals

$$\int_{\{x_1 : \mathrm{E}(g(X_2)|X_1 = x_1) \in [a,b]\}} f_{X_1}(x_1)\mathrm{d}x_1.$$

Theorem 16 (The Law of Iterated Expectations). *Let (X_1, X_2) be a pair of random variables. $\mathrm{E}(X_2) = \mathrm{E}(\mathrm{E}(X_2|X_1))$.*

Proof: By direct computation,

$$\begin{aligned}
\mathrm{E}(X_2) &= \int\int x_2 f_{(X_1,X_2)}(x_1, x_2)\mathrm{d}x_2\mathrm{d}x_1 \\
&= \int_{\{x_1|f_{X_1}(x_1)\}} (\int x_2 f_{X_2|X_1}(x_2|x_1)\mathrm{d}x_2) f_{X_1}(x_1)\mathrm{d}x_1 \\
&= \int \mathrm{E}(X_2|X_1 = x_1) f_{X_1}(x_1)\mathrm{d}x_1 \\
&= \mathrm{E}(\mathrm{E}(X_2|X_1)).
\end{aligned}$$

Q.E.D.

6 The Fundamental Theorems of Probability

6.1 Convergence of Random Variables

There are several concepts of convergence for random variables. We consider two of them: a sequence $(X_n)_{n=1}^{\infty}$ of random variables
(a) *converges in probability to the random variable X* if for all $\varepsilon > 0$ one has that

$$\lim_{n\to\infty} \mathrm{p}(|X_n - X| < \varepsilon) = 1;$$

(b) it *converges in distribution to the random variable X* if at all $x \in \mathbb{R}$ at which F_X is continuous, one has that

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x),$$

where each F_{X_n} is the distribution function of X_n and F_X is the one of X.

For notational conciseness, we denote the two types of convergence by $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{d} X$, respectively. It is important to understand the relationship between these three concepts: if $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{d} X$, but the opposite causality is not true. When the limit variable is constant, convergence in distribution implies convergence in probability, however.

6.2 The Law of Large Numbers

A sequence $(X_n)_{n=1}^{\infty}$ of random variables is said to be *i.i.d.* if it is independent and for every n and n', $\mathrm{p}(X_n \in D) = \mathrm{p}(X_{n'} \in D)$ for all $D \subseteq \mathbb{R}$.

Theorem 17 (The Law of Large Numbers). *Let $(X_n)_{n=1}^{\infty}$ be an i.i.d. sequence of random variables, and suppose that, for all n, $\mathrm{E}(X_n) = \mu \in \mathbb{R}$ and $\mathrm{V}(X_n) = \sigma^2 \in \mathbb{R}_{++}$. The sequence of random variables given by $\bar{X}_n = \sum_{k=1}^{n} X_k/n$ converges in probability to μ.*

Proof: Given that $(X_n)_{n=1}^{\infty}$ is i.i.d. $\mathrm{E}(\bar{X}_n) = \mu$ and $\mathrm{V}(\bar{X}_n) = \sigma^2/n$.[2] Now, by Exercise 2, for $\varepsilon > 0$ one has that

$$\mathrm{p}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2},$$

so $\lim_{n\to\infty} \mathrm{p}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \lim_{n\to\infty} \sigma^2/(n\varepsilon^2) = 0$. *Q.E.D.*

Exercise 4. *Prove the following result:* let X be an absolutely continuous random variable, and fix $\Omega \subseteq \mathbb{R}$ such that $\mathrm{p}(X \in \Omega) \in (0, 1)$. Consider the experiment where $n \in \mathbb{N}$ realizations of X are taken independently, and let G_n be the relative frequency with which a realization in Ω is obtained in the experiment. Then, $G_n \xrightarrow{p} \mathrm{p}(X \in \Omega)$.

6.3 The Central Limit Theorem

Theorem 18 (The Central Limit Theorem). *Let $(X_n)_{n=1}^{\infty}$ be an i.i.d. sequence of random variables, and suppose that, for every n, $\mathrm{E}(X_n) = \mu \in \mathbb{R}$, $\mathrm{V}(X_n) = \sigma^2 \in \mathbb{R}_{++}$ and $M_{X_n} = M_X$ is defined in an open neighborhood of 0. Define the sequence $(\bar{X}_n)_{n=1}^{\infty}$ as in Theorem 17. Then,*

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x.$$

Exercise 5. *How can both the law of large numbers and the central limit theorem be true? That is, if the law says that $\bar{X}_n$ converges in probability to a constant(μ), and convergence in probability implies convergence in distribution, then how can $\bar{X}_n$ also converge in distribution to the standard normal?*

Take a good look at what the Central Limit Theorem implies. In particular, it does not imply that every "large" sample of realizations of a random variable "tends" to be distributed normally!

[2] Showing this is left as an exercise.