Chapter 3
Integration Theory

In the previous chapter we learned about random variables and their distributions. This distribution completely characterizes a random variable. But in general distributions are very complex functions. The human brain cannot comprehend such things easily. So the human brain wants to talk about one typical value. For example, one can give a distribution for the random variable representing player salaries in the NBA. Here the variability (probability space) is represented by the specific player chosen. However, probably one is not interested in such a distribution. One simply wants to know what is the typical salary in the NBA. The person probably contemplates a career in sports and wants to find out if as an athlete should go for basketball or baseball, therefore he is much better serve by comparing only two numbers. Calculating such a number is hard (which number?). In this chapter we create a theory to calculate any numbers that the person wishes. Paradoxically, to calculate a simple number we need to understand a very complex theory.

3.1 Integral of measurable functions

Recall that the random variables are nothing more than measurable functions. Let $(\Omega, \mathscr{F}, P)$ be a probability space. We wish to define for any measurable function f an integral of f with respect to the measure P.

Notation. We shall use the following notations for this integral:

$$\int_\Omega f(\omega)\mathbf{P}(d\omega) = \int f d\mathbf{P}$$

$$\text{for } A \in \mathscr{F} \text{ we have } \int_A f(\omega)\mathbf{P}(d\omega) = \int_A f d\mathbf{P} = \int f \mathbf{1_A} d\mathbf{P}$$

Recall the Dirac Delta we have defined previously? With its help summation is another kind of integral. Let $\{a_n\}$ be a sequence of real numbers. Let $\Omega = \mathbb{R}, \mathscr{F} = \mathscr{B}(\mathbb{R})$ and the measure on this set is $\delta(A) = \sum_{i=1}^{\infty} \delta_i(A)$.

Then the function $i \mapsto a_i$ is integrable if and only if $\sum a_i < \infty$ and in this case we have:

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \int_{-\infty}^{\infty} a_x d\delta_n(x) = \int_{-\infty}^{\infty} a_x \sum_{n=1}^{\infty} d\delta_n(x) = \int_{-\infty}^{\infty} a_x d\delta(x)$$

What is the point of this? The simple argument above shows that any "discrete" random variable (in the undergraduate text definition) may be treated as a "continuous" random variable. Not that there was any doubt after all the big fuss we made about it in the previous chapter.

Integral of Simple (Elementary) Functions

If $A \in \mathscr{F}$ we know that we can define a measurable function by its indicator $\mathbf{1}_A$. We define the integral of this measurable function $\int \mathbf{1}_A d\mathbf{P} = \mathbf{P}(A)$. We note that this variable has the same distribution as that of the Bernoulli random variable. The variable takes values 0 and 1 and we can easily calculate the probability that the variable is 1 as:

$$\mathbf{P} \circ \mathbf{1}_A^{-1}(\{1\}) = \mathbf{P}\{\omega : \mathbf{1}_A(\omega) = 1\} = \mathbf{P}(A).$$

Therefore the variable is distributed as a Bernoulli random variable with parameter $p = \mathbf{P}(A)$.

Definition 3.1 (Simple function). f is called a *simple* (elementary) function if and only if f can be written as a finite linear combination of indicators or, more specifically there exist sets $A_1, A_2, \dots, A_n$ all in $\mathscr{F}$ and constants $a_1, a_2, \dots, a_n$ in $\mathbb{R}$ such that:

$$f(\omega) = \sum_{k=1}^{n} a_k \mathbf{1}_{A_k}(\omega)$$

If the constants a_k are all positive, then f is a positive simple function.

Note that the sets A_i do not have to be disjoint but an easy exercise (Problem 3.1) shows that f could be written in terms of disjoint sets.

For any simple function f we define its integral:

$$\int f d\mathbf{P} = \sum_{k=1}^{n} a_k \mathbf{P}(A_k) < \infty$$

We adopt the conventions $0 * \infty = 0$ and $\infty * 0 = 0$ in the above summation.

We need to check that the above definition is proper. For there exist many representations of a simple function and we need to make sure that any such representation produces the same integral value. Furthermore, the linearity and monotonicity properties of the integral may be proven. We skip these results since they are simple to prove and do not bring any additional insight.

Integral of positive measurable functions

For every f positive measurable function $f:\Omega \longrightarrow [0,\infty)$ we define:

$$\int f d\mathbf{P} = \sup\left\{\int h d\mathbf{P} : h \text{ is a simple function, } h \leq f\right\}$$

For a given positive measurable function can we find a sequence of simple functions that converge to it? The answer is yes and is provided by the next simple exercise:

Exercise 3.1. Let $f:\Omega \to [0,\infty]$ be a positive, measurable function. For all $n \geq 1$, we define:

$$f_n(\omega) := \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \mathbf{1}_{\left\{\frac{k}{2^n} \leq f(\omega) < \frac{k+1}{2^n}\right\}}(\omega) + n\mathbf{1}_{\{f(\omega)\geq n\}} \tag{3.1}$$

1. Show that f_n is a simple function on $(\Omega,\mathscr{F})$, for all $n \geq 1$.
2. Show that the sets present in the indicators in equation (3.1) form a partition of Ω, for all $n \geq 1$.
3. Show that the sequence of simple functions is increasing $g_n \leq g_{n+1} \leq f$, for all $n \geq 1$.
4. Show that $g_n \uparrow f$ as $n \to \infty$. Note that this is not an a.s. statement, it is true for all $\omega \in \Omega$.

The solution to this exercise is not complicated and in fact it is an assigned problem (Problem 3.3).

The following lemma is a very easy to understand and useful tool.

Lemma 3.1. *If f is a positive measurable function and $\int f d\mathbf{P} = 0$ then $\mathbf{P}\{f > 0\} = 0$ (or $f = 0$ a.s.).*

Proof. We have $\{f > 0\} = \bigcup_{n\geq 0}\{f > \frac{1}{n}\}$. Since the events are increasing by the monotone convergence property of measure we must have $\mathbf{P}\{f > 0\} = \lim_{n\to\infty}\mathbf{P}\{f > \frac{1}{n}\}$. If we assume by absurd that $\mathbf{P}\{f > 0\} > 0$ then there must exist an n such that $\mathbf{P}\{f > \frac{1}{n}\} > 0$. However, in this case by the definition of the integral of positive measurable functions:

$$\int f d\mathbf{P} \geq \int \frac{1}{n}\mathbf{1}_{\{f>\frac{1}{n}\}} d\mathbf{P} > 0,$$

contradiction. □

The next theorem is one of the most useful in probability theory. In our immediate context it tells us that the integral for positive measurable functions is well defined.

Theorem 3.1 (Monotone Convergence Theorem). *If f is a sequence of measurable positive functions such that $f_n \uparrow f$ then:*

$$\int_\Omega f_n(\omega)\mathbf{P}(d\omega) \uparrow \int_\Omega f(\omega)\mathbf{P}(d\omega)$$

Note: This is all there is to integration theory. The proof of the monotone convergence theorem is not difficult, you may want to look at it.

Proof. **Ion: Write the proof**

Integral of measurable functions

Let f be any measurable function. Then we write $f = f^+ - f^-$ where:

$$f^+(s) = \max\{f(s), 0\}$$
$$f^-(s) = \max\{-f(s), 0\}$$

Then f^+ and f^- are positive measurable functions and $|f| = f^+ + f^-$. Since they are positive measurable their integrals are well defined by the previous part.

Definition 3.2. We define $L^1(\Omega, \mathscr{F}, P)$ as being the space of all functions f such that:

$$\int |f| d\mathbf{P} = \int f^+ d\mathbf{P} + \int f^- d\mathbf{P} < \infty$$

For any f in this space which we will shorten to $L^1(\Omega)$ or even simpler to L^1 we define:

$$\int f d\mathbf{P} = \int f^+ d\mathbf{P} - \int f^- d\mathbf{P}$$

Note: With the above it is trivial to show that $|\int f d\mathbf{P}| \leq \int |f| d\mathbf{P}$

Linearity:

If $f, g \in L^1(\Omega)$ with $a, b \in \mathbb{R}$, then:

$$af + bg \in L^1(\Omega)$$
$$\int (af + bg) d\mathbf{P} = a \int f d\mathbf{P} + b \int g d\mathbf{P}$$

Lemma 3.2 (Fatou's Lemma for measurable functions). *If one of the following is true:*

a) $\{f_n\}_n$ is a sequence of positive measurable functions or
b) $\{f_n\} \subset L^1(\Omega)$

then:

$$\int \liminf_n f_n d\mathbf{P} \leq \liminf_n \int f_n d\mathbf{P}$$

Proof. Note that $\liminf_n f_n = \lim_{m\to\infty} \inf_{n\geq m} f_n$, where $\lim_{m\to\infty} \inf_{n\geq m} f_n$ is an increasing sequence.

Let $g_m = \inf_{n\geq m} f_n$, and $n \geq m$:

$$f_n \geq \inf_{n\geq m} f_m = g_m \Rightarrow \int f_n d\mathbf{P} \geq \int g d\mathbf{P} \Rightarrow \int g_m d\mathbf{P} \leq \inf_{n\geq m} \int f_n d\mathbf{P}$$

Now g_m increases so we may use the Monotone Convergence Theorem and we get:

$$\int \lim_{m\to\infty} g_m d\mathbf{P} = \lim_{m\to\infty} \int g_m d\mathbf{P} \leq \lim_{m\to\infty} \inf_{n\geq m} \int f_n d\mathbf{P} = \liminf_{n} \int f_n d\mathbf{P}$$

Theorem 3.2 (Dominated Convergence Theorem). *If f_n, f are measurable, $f_n(\omega) \to f(\omega)$ for all $\omega \in \Omega$ and the sequence f_n is dominated by $g \in L^1(\Omega)$:*

$$|f_n(\omega)| \leq g(\omega), \qquad \forall \omega \in \Omega, \forall n \in \mathbb{N}$$

then:

$$f_n \to f \text{ in } L^1(\Omega) \qquad \left(i.e. \int |f_n - f| d\mathbf{P} \to 0 \right)$$

Thus $\int f_n d\mathbf{P} \to \int f d\mathbf{P}$ and $f \in L^1(\Omega)$.

The Standard Argument:

This argument is the most important argument in the probability theory. Suppose that we want to prove that some property holds for all functions h in some space such as $L^1(\Omega)$ or the space of measurable functions.

1. Show that the result is true for all indicator functions.
2. Use linearity to show the result holds true for all f simple functions.
3. Use the Monotone Convergence Theorem to obtain the result for measurable positive functions.
4. Finally from the previous step and writing $f = f^+ - f^-$ we show that the result is true for all measurable functions.

3.2 Expectations

Since a random variable is just a measurable function we just need to particularize the results of the previous section. An integral with respect to a probability measure is called an expectation. Let $(\Omega, \mathscr{F}, P)$ be a probability space.

Definition 3.3. For X a r.v. in $L^1(\Omega)$ define:

$$\mathbf{E}(X) = \int_\Omega X d\mathbf{P} = \int_\Omega X(\omega) d\mathbf{P}(\omega) = \int_\Omega X(\omega) \mathbf{P}(d\omega)$$

This expectation has the same properties of the integral defined before and some extra ones since the space has finite measure.

Convergence Theorems:

(i) *Monotone Convergence Theorem:* If $X_n \geq 0$, $X_n \in L^1$ and $X_n \uparrow X$ then $\mathbf{E}(X_n) \uparrow \mathbf{E}(X) \leq \infty$.
(ii) *Fatou:* $\mathbf{E}(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} \mathbf{E}(X_n)$
(iii) *Dominated Convergence Theorem:* If $|X_n(\omega)| \leq Y(\omega)$ on Ω with $Y \in L^1(\Omega)$ and $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$ then $\mathbf{E}(|X_n - X|) \to 0$.

Now let us present specific properties of the expectation. This is to be expected since the space has finite measure therefore we can obtain more specific properties.

Markov Inequality:

Let Z be a r.v. and let $g : \mathbb{R} \longrightarrow [0,\infty]$ be an *increasing* measurable function. Then:

$$\mathbf{E}\left[g(Z)\right] \geq \mathbf{E}\left[g(Z)\mathbf{1}_{\{Z\geq c\}}\right] \geq g(c)\mathbf{P}(Z \geq c)$$

Thus

$$\mathbf{P}(Z \geq c) \leq \frac{\mathbf{E}[g(Z)]}{g(c)}$$

for all g increasing functions and $c > 0$.

Example 3.1 (Special cases of the Markov inequality). If we take $g(x) = x$ an increasing function and X a positive random variable then we obtain:

$$\mathbf{P}(Z \geq c) \leq \frac{\mathbf{E}(Z)}{c}.$$

To get rid of the necessity that $X \geq 0$ take $Z = |X|$. Then we obtain the classical form of the Markov inequality:

$$\mathbf{P}(|X| \geq c) \leq \frac{\mathbf{E}(|X|)}{c}.$$

If we take $g(x) = x^2$, $Z = |X - \mathbf{E}(X)|$ and we use the variance definition (which we will see in a minute), we obtain the Chebyshev inequality:

$$\mathbf{P}(|X - \mathbf{E}(X)| \geq c) \leq \frac{Var(X)}{c^2}.$$

If we denote $\mathbf{E}(X) = \mu$ and $Var(X) = \sigma$ and we take $c = k\sigma$ in the previous inequality we will obtain the classical Chebyshev inequality presented in undergraduate courses:

$$\mathbf{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

If $g(x) = e^{\theta x}$, with $\theta > 0$ then

$$\mathbf{P}(Z \geq c) \leq e^{-\theta c}\mathbf{E}(e^{\theta z}),$$

This inequality states that the tail of the distribution decays exponentially in c if Z has finite exponential moments. With simple manipulations one can obtain Chernoff's inequality using it.

Jensen's Inequality for convex functions:

This is just a reminder.

Definition 3.4. A function $g : I \longrightarrow \mathbb{R}$ is called a convex function on I (where I is any open interval in $\mathbb{R}$, if its graph lies below any of its chords. Mathematically: for any $x, y \in I$ and for any $\alpha \in (0,1)$ we have

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

Some examples of convex functions on the whole $\mathbb{R}$: $|x|$, x^2 and $e^{\theta x}$, with $\theta > 0$.

Lemma 3.3 (Jensen's Inequality). *Let f be a convex function and let X be a r.v. in $L^1(\Omega)$. Assume that $\mathbf{E}(f(X)) \leq \infty$ then:*

$$f(\mathbf{E}(X)) \leq \mathbf{E}(f(X))$$

Proof. Skipped. The classical approach indicators → simple functions → positive measurable → measurable is a standard way to prove Jensen.

L^p spaces.

We generalize the L^1 notion presented earlier in the following way. For $1 \leq p \leq \infty$ we define the space:

$$L^p(\Omega, \mathscr{F}, P) = L^p(\Omega) = \left\{ X : \Omega \longrightarrow \mathbb{R} : \mathbf{E}\left[|X|^p\right] = \int |X|^p d\mathbf{P} < \infty \right\},$$

On this space we define a norm called the p-norm as:

$$||X||_p = \mathbf{E}\left[|X|^p\right]^{1/p}$$

Lemma 3.4 (Properties of L^p spaces).

(i) L^p is a vector space. (i.e., if $X, Y \in L^p$ and $a, b \in \mathbb{R}$ then $aX + bY \in L^p$).
(ii) L^p is complete (every Cauchy sequence in L^p is convergent)

Lemma 3.5 (Cauchy-Bunyakovsky-Schwarz inequality). *If $X, Y \in L^2(\Omega)$ then $X, Y \in L^1(\Omega)$ and*

$$|\mathbf{E}[XY]| \leq \mathbf{E}[|XY|] \leq ||X||_2 ||Y||_2$$

A historical remark. This inequality, one of the most famous and useful un any area of analysis (not only probability) is usually credited to Cauchy for sums and Schwartz for integrals and is usually known as the Cauchy-Schwartz inequality. However,the Russian mathematician Victor Yakovlevich Bunyakovsky (1804-1889) discovered and first published the inequality for integrals in 1859 (when Schwartz was 16). Unfortunately, he was born in eastern Europe... However, all who are born in eastern Europe (including myself) learn the inequality by its proper name.

Proof. The first inequality is clear by Jensen inequality. We need to show

$$\mathbf{E}[|XY|] \leq (\mathbf{E}[X^2])^{1/2}(\mathbf{E}[Y^2])^{1/2}$$

Let $W = |X|$ and $Z = |Y|$ then $W, Z \geq 0$.
Truncation:
Let $W_n = W \bigwedge n$ and $Z_n = Z \bigwedge n$ that is

$$W_n(\omega) = \begin{cases} W(\omega), & \text{if } W(\omega) < n \\ n, & \text{if } W(\omega) \geq n \end{cases}$$

Clearly, defined in this way W_n, Z_n are bounded. Let $a, b \in \mathbb{R}$ two constants. Then:

$$0 \leq \mathbf{E}[(aW_n + bZ_n)^2] = a^2\mathbf{E}(W_n^2) + 2ab\mathbf{E}(W_nZ_n) + b^2\mathbf{E}(Z_n^2)$$

If we let $a/b = c$ we get:

$$c^2\mathbf{E}(W_n^2) + 2c\mathbf{E}(W_nZ_n) + \mathbf{E}(Z_n^2) \geq 0 \qquad \forall c \in \mathbb{R}$$

This means that the quadratic function in c has to be positive. But this is only possible if the determinant of the equation is negative and the leading coefficient $\mathbf{E}(W_n^2)$ is strictly positive, the later condition is obviously true. Thus we must have:

$$\begin{aligned} &4(\mathbf{E}(W_nZ_n))^2 - 4\mathbf{E}(W_n^2)\mathbf{E}(Z_n^2) \leq 0 \\ &\Rightarrow (\mathbf{E}(W_nZ_n))^2 \leq \mathbf{E}(W_n^2)\mathbf{E}(Z_n^2) \leq \mathbf{E}(W^2)\mathbf{E}(Z^2) \qquad \forall n \end{aligned}$$

If we let $n \uparrow \infty$ and use the monotone convergence theorem we get:

$$(\mathbf{E}(WZ))^2 \leq \mathbf{E}(W^2)\mathbf{E}(Z^2).$$

□

A more general inequality is:

Lemma 3.6 (Hölder inequality). *If $1/p + 1/q = 1$, $X \in L^p(\Omega)$ and $Y \in L^q(\Omega)$ then $XY \in L^1(\Omega)$ and:*

$$\mathbf{E}|XY| \leq \|X\|_p\|Y\|_q = (\mathbf{E}|X|^p)^{\frac{1}{p}}(\mathbf{E}|Y|^q)^{\frac{1}{q}}$$

Proof. The proof is simple and uses the following inequality (Young inequality): if a and b are positive real numbers and p, q are as in the theorem then:

$$ab \le \frac{a^p}{p} + \frac{b^q}{q},$$

with equality if and only if $a^p = b^q$.

Taking this inequality as given (not hard to prove) define:

$$f = \frac{|X|}{\|X\|_p}, \quad g = \frac{|Y|}{\|Y\|_p}.$$

Note that the Hölder inequality is equivalent with $\mathbf{E}[fg] \le 1$ ($\|X\|_p$ and $\|Y\|_q$ are just numbers that can be taken in and out of integral by the linearity property). To prove this apply the Young inequality to $f \ge 0$ and $g \ge 0$ and then integrate to obtain:

$$\mathbf{E}[fg] \le \frac{1}{p}\mathbf{E}[f^p] + \frac{1}{q}\mathbf{E}[g^q] = \frac{1}{p} + \frac{1}{q} = 1$$

$\mathbf{E}[f^p] = 1$ and similarly for g may be easily checked. Finally, the extreme cases ($p = 1$, $q = \infty$, etc.) may be treated separately. □

Lemma 3.7 (Minkowski Inequality). *If* $X, Y \in L^p$ *then* $X + Y \in L^p$ *and:*

$$\|X+Y\|_p \le \|X\|_p + \|Y\|_p$$

Proof. We clearly have:

$$|X+Y|^p \le 2^{p-1}(|X|^p + |Y|^p).$$

For example use the definition of convexity for the function x^p with $x = |X|$ and $y = |Y|$ and $\alpha = 1/2$. Now integrating implies that $X + Y \in L^p$. Now we can write:

$$\begin{aligned}
\|X+Y\|_p^p &= \mathbf{E}[|X+Y|^p] \le \mathbf{E}\left[(|X|+|Y|)|X+Y|^{p-1}\right] \\
&= \mathbf{E}\left[|X||X+Y|^{p-1}\right] + \mathbf{E}\left[|Y||X+Y|^{p-1}\right] \\
&\overset{\text{Hölder}}{\le} \left(\mathbf{E}\left[|X|^p\right]\right)^{1/p}\left(\mathbf{E}\left[|X+Y|^{(p-1)q}\right]\right)^{1/q} + \left(\mathbf{E}\left[|Y|^p\right]\right)^{1/p}\left(\mathbf{E}\left[|X+Y|^{(p-1)q}\right]\right)^{1/q} \\
&\overset{\left(q=\frac{p}{p-1}\right)}{=} \left(\|X\|_p + \|Y\|_p\right)\left(\mathbf{E}\left[|X+Y|^p\right]\right)^{1-\frac{1}{p}} \\
&= \left(\|X\|_p + \|Y\|_p\right)\frac{\mathbf{E}\left[|X+Y|^p\right]}{\|X+Y\|_p}
\end{aligned}$$

Now, identifying the left and right hand after simplifications we obtain the result. □

Example 3.2 (due to Erdós). Suppose there are 17 fence posts around the perimeter of a field and exactly 5 of them are rotten. Show that irrespective of which of these

5 are rotten, there should exist a row of 7 consecutive posts of which at least 3 are rotten.

Proof (Solution). First we label the posts $1,2\cdots 17$. Now define :

$$I_k = \begin{cases} 1 & \text{if post } k \text{ is rotten} \\ 0 & \text{otherwise} \end{cases}$$

For any fixed k, let R_k denote the number of rotten posts among $k+1,\cdots,k+7$ (starting with the next one). Note that when any of $k+1,\cdots,k+7$ are larger than 17 we start again from 1 (i.e., modulo 17 +1).

Now pick a post at random this obviously can be done in 17 ways with equal probability. Then after we pick this post we calculate the number of rotten boards. We have:

$$\begin{aligned}
\mathbf{E}(R_k) &= \sum_{k=1}^{17}(I_{k+1}+\cdots+I_{k+7})\frac{1}{17} \\
&= \frac{1}{17}\sum_{k=1}^{17}\sum_{j=1}^{7} I_{k+j} == \frac{1}{17}\sum_{j=1}^{7}\sum_{k=1}^{17} I_{j+k} \\
&= \frac{1}{17}\sum_{j=1}^{7} 5 \qquad \text{(the sum is 5 since we count all the rotten posts in the fence)} \\
&= \frac{35}{17}
\end{aligned}$$

Now, $35/17 > 2$ which implies $\mathbf{E}(R_k) > 2$. Therefore, $\mathbf{P}(R_k > 2) > 0$ (otherwise the expectation is necessarily bounded by 2) and since R_k is integer valued $\mathbf{P}(R_k \geq 3) > 0$. So there exists some k such that $R_k \geq 3$.

Of course now that we see the proof we can play around with numbers and see that there exists a row of 4 consecutive posts in which at least two are rotten, or that there must exist a row of 11 consecutive posts in which at least 4 are rotten and so on (row of 14 containing all 5 rotten ones).

3.3 Variance and the correlation coefficient

Definition 3.5. The variance or the Dispersion of a random variable $X \in L^2(\Omega)$ is:

$$V(X) = \mathbf{E}[(X-\mu)^2] = \mathbf{E}(X^2) - \mu^2$$

Where $\mu = \mathbf{E}(X)$.

Definition 3.6. Given two random variables X, Y we call the covariance between X and Y the quantity:

$$Cov(X,Y) = \mathbf{E}[(X-\mu_X)(Y-\mu_Y)]$$

Where $\mu_X = \mathbf{E}(X)$ and $\mu_Y = \mathbf{E}(Y)$.

Definition 3.7. Given random variables X, Y we call the correlation coefficient:

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\mathbf{E}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\mathbf{E}[(X-\mu_X)^2]\mathbf{E}[(Y-\mu_Y)^2]}}$$

From the Cauchy-Schwartz inequality applied to $X - \mu_X$ and $Y - \mu_Y$ we get $|\rho| < 1$ or $\rho \in [-1,1]$.

The variable X and Y are called **uncorrelated** if the covariance (or equivalently the correlation) between them is zero.

Proposition 3.1 (Properties of expectation). *The following are true:*

(i) If X and Y are integrable r.v.'s then for any constants α and β the r.v. $\alpha X + \beta Y$ is integrable and $\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}X + \beta \mathbf{E}Y$.
(ii) $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X,Y)$
(iii) If X,Y are independent then $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$ and $Cov(X,Y) = 0$.
(iv) If $X(\omega) = c$ with probability 1 and $c \in \mathbb{R}$ a constant then $\mathbf{E}X = c$.
(v) If $X \geq Y$ a.s. then $\mathbf{E}X \geq \mathbf{E}Y$. Furthermore, if $X \geq Y$ a.s. and $\mathbf{E}X = \mathbf{E}Y$ then $X = Y$ a.s.

Proof. Exercise. Please note that the reverse of the part (iii) above is not true, if the two variables are uncorrelated this does not mean that they are independent. In fact in Problem 3.5 you are required to provide a counterexample.

3.4 Functions of random variables. The Transport Formula.

In Section 2.4 on page 49 we showed how to calculate distributions and in particular p.d.f.'s for continuous random variables. We have also promised a more general result. Well, here it is. This general result allows to construct random variables and in particular distributions in any space. This is the result that allows us to claim that studying random variables on $([0,1], \mathscr{B}([0,1]), \lambda)$ is enough. We had to postpone presenting the result until this point since we had to learn first how to integrate.

Theorem 3.3 (General Transport Formula). *Let $(\Omega, \mathbb{R}, P)$ be a probability space. Let f be a measurable function such that:*

$$(\Omega, \mathscr{F}) \xrightarrow{f} (S, \mathscr{G}) \xrightarrow{\varphi} (\mathbb{R}, \mathscr{B}(\mathbb{R})),$$

where $(S, \mathscr{G})$ is a measurable space. Assuming that at least one of the integrals exists we then have:

$$\int_\Omega \varphi \circ f d\mathbf{P} = \int_S \varphi d\mathbf{P} \circ f^{-1},$$

for all φ measurable functions.

Proof. We will use the standard argument technique discussed above.

1. Let φ be the indicator function. $\varphi = \mathbf{1}_A$ for $A \in \mathscr{G}$:

   $$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

   Then we get:

   $$\begin{aligned} \int_\Omega \mathbf{1_A} \circ f d\mathbf{P} &= \int_\Omega \mathbf{1}_A(f(\omega)) d\mathbf{P}(\omega) = \int_\Omega \mathbf{1}_{f^{-1}(A)}(\omega) d\mathbf{P}(\omega) \\ &= \mathbf{P}(f^{-1}(A)) = \mathbf{P} \circ f^{-1}(A) = \int_S \mathbf{1}_A d(\mathbf{P} \circ f^{-1}) \end{aligned}$$

   recalling the definition of the integral of an indicator.
2. Let φ be a simple function $\varphi = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ where a_i's are constant and $A_i \in \mathscr{G}$.

   $$\begin{aligned} \int_\Omega \varphi \circ f d\mathbf{P} &= \int_\Omega \left(\sum_{i=1}^n a_i \mathbf{1}_{A_i} \right) \circ f d\mathbf{P} \\ &= \int_\Omega \sum_{i=1}^n a_i (\mathbf{1}_{A_i} \circ f) d\mathbf{P} = \sum_{i=1}^n a_i \int_\Omega \mathbf{1}_{A_i} \circ f d\mathbf{P} \\ &\overset{(\text{part 1})}{=} \sum_{i=1}^n a_i \int_S \mathbf{1}_{A_i} d\mathbf{P} \circ f^{-1} = \int_S \sum_{i=1}^n a_i \mathbf{1}_{A_i} d\mathbf{P} \circ f^{-1} = \int_S \varphi d\mathbf{P} \circ f^{-1} \end{aligned}$$

3. Let φ be a positive measurable function and let φ_n be a sequence of simple functions such that $\varphi_n \nearrow \varphi$ then:

   $$\begin{aligned} \int_\Omega \varphi \circ f d\mathbf{P} &= \int_\Omega (\lim_{n\to\infty} \varphi_n) \circ f d\mathbf{P} \\ &= \int_\Omega \lim_{n\to\infty} (\varphi_n \circ f) d\mathbf{P} \overset{\text{monotone convergence}}{=} \lim_{n\to\infty} \int \varphi_n \circ f d\mathbf{P} \\ &\overset{(\text{part 2})}{=} \lim_{n\to\infty} \int \varphi_n d\mathbf{P} \circ f^{-1} \overset{\text{monotone convergence}}{=} \int \lim_{n\to\infty} \varphi_n d\mathbf{P} \circ f^{-1} \\ &= \int_S \varphi d(\mathbf{P} \circ f^{-1}) \end{aligned}$$

4. Let φ be a measurable function then $\varphi^+ = \max(\varphi, 0)$, $\varphi^- = \max(-\varphi, 0)$. Which then gives us $\varphi = \varphi^+ - \varphi^-$. Since at least one integral is assumed to exist we get that $\int \varphi^+$ and $\int \varphi^-$ exist. Also note that:

   $$\begin{aligned} \varphi^+ \circ f(\omega) &= \varphi^+(f^{-1}(\omega)) = \max(\varphi(f(\omega)), 0) \\ \max(\varphi \circ f(\omega), 0) &= (\varphi \circ f)^+(\omega) \end{aligned}$$

   Then:

$$\int \varphi^+ d\mathbf{P}\circ f^{-1} = \int \varphi^+ \circ f d\mathbf{P} = \int (\varphi\circ f)^+ d\mathbf{P}$$
$$\int \varphi^- d\mathbf{P}\circ f^{-1} = \int \varphi^- \circ f d\mathbf{P} = \int (\varphi\circ f)^- d\mathbf{P}$$

These equalities follow from part 3 of the proof. After subtracting both:

$$\int \varphi d\mathbf{P}\circ f^{-1} = \int \varphi\circ f d\mathbf{P}$$

Exercise 3.2. If X and Y are independent random variables defined on $(\Omega,\mathbb{R},P)$ with $X,Y\in L^1(\Omega)$ then $XY\in L^1(\Omega)$:

$$\int_\Omega XY d\mathbf{P} = \int_\Omega X d\mathbf{P} \int_\Omega Y d\mathbf{P} \qquad (\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y))$$

Proof (Solution). This is an exercise that you have seen before, here is presented to exercise the standard approach.

Example 3.3. Let us solve the previous exercise using the transport formula. Let us take $f:\Omega\to\mathbb{R}^2$, $f(\omega)=(X(\omega),Y(\omega))$; and $\varphi:\mathbb{R}^2\to\mathbb{R}$, $\varphi(x,y)=xy$. Then we have from the transport formula:

$$\int_\Omega X(\omega)Y(\omega)dP(\omega) \overset{(T)}{=} \int_{\mathbb{R}^2} xydP\circ(X,Y)^{-1}$$

The integral on the left is $\mathbf{E}(XY)$, while the integral on the right can be calculated as:

$$\begin{aligned}\int_{\mathbb{R}^2} xyd(P\circ X^{-1},P\circ Y^{-1}) &= \int_{\mathbb{R}} xdP\circ X^{-1}\int_{\mathbb{R}} ydP\circ Y^{-1}\\ &\overset{(T)}{=} \int_\Omega X(\omega)dP(\omega)\int_\Omega Y(\omega)dP(\omega) = \mathbf{E}(X)\mathbf{E}(Y)\end{aligned}$$

Example 3.4. Finally we conclude with an application of the transport formula which will produce one of the most useful formulas. Let X be a r.v. defined on the probability space $(\Omega,\mathscr{F},\mathbf{P})$ with distribution function $F(x)$. Show that:

$$\mathbf{E}(X) = \int_{\mathbb{R}} xdF(x),$$

where the integral is understood in Riemann-Stieltjes sense.

Proving the formula is immediate. Take $f:\Omega\to\mathbb{R}$, $f(\omega)=X(\omega)$ and $\varphi:\mathbb{R}\to\mathbb{R}$, $\varphi(x)=x$. Then from the transport formula:

$$\mathbf{E}(X) = \int_\Omega X(\omega)d\mathbf{P}(\omega) = \int_\Omega x\circ X(\omega)d\mathbf{P}(\omega) \overset{(\mathrm{T})}{=} \int_{\mathbb{R}} xd\mathbf{P}\circ X^{-1}(x) = \int_{\mathbb{R}} xdF(x)$$

Clearly if the distribution function $F(x)$ is derivable with $\frac{dF}{dx}(x)=f(x)$ or $dF(x)=f(x)dx$ we obtain the lower level classes formula for calculating expectation of a “continuous” random variable:

$$\mathbf{E}(X) = \int_{\mathbb{R}} x f(x) dx$$

3.5 Applications. Exercises in probability reasoning.

The next two theorems are presented to observe the proofs. They are both early exercises in probability. We will present later much stronger versions of these theorems (and we will also see that these convergence types have very precise definitions), but for now we lack the tools to give general proofs to these stronger versions.

Theorem 3.4 (Law of Large Numbers). *Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $\{X_n\}_n$ be a sequence of i.i.d random variables with $\mathbf{E}(X_i) = \int_\Omega X_i d\mathbf{P} = \mu$. Assume that the fourth moment of these variables is finite and $\mathbf{E}(X_i^4) = K_4$ for all i. Then:*

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{a.s} \mu$$

Proof. Recall what it means for a statement to hold almost surely (a.s.). In our specific context if we denote $S_n = X_1 + \cdots + X_n$ then we need to show that $\mathbf{P}(S_n/n \to \mu) = 1$.

First step. Let us show that we can reduce to the case of $\mathbf{E}(X_i) = \mu = 0$. Take $Y_i = X_i - \mu$. If we prove that $\frac{Y_1+\cdots+Y_n}{n} \to 0$ then substituting back we shall obtain $\frac{S_n - n\mu}{n} \to 0$, or $\frac{S_n}{n} \to \mu$. Which gives our result. Thus we assume that $\mathbf{E}(X_i) = \mu = 0$.

Second step. We want to show that $\frac{S_n}{n} \xrightarrow{\text{a.s}} 0$. We have:

$$\mathbf{E}\left(S_n^4\right) = \mathbf{E}\left((X_1 + \cdots + X_n)^4\right) = \mathbf{E}\left(\sum_{i,j,k,l} X_i X_j X_k X_l\right)$$

If any factor in the sum above appears with power one, from independence we will have $\mathbf{E}(X_i X_j X_k X_l) = \mathbf{E}(X_i)\mathbf{E}(X_j X_k X_l) = 0$. Thus, the only terms remaining in the sum above are those with power larger than one.

$$\begin{aligned}\mathbf{E}\left(\sum_{i,j,k,l} X_i X_j X_k X_l\right) &= \mathbf{E}\left(\sum_i X_i^4 + \sum_{i<j} \binom{4}{2} X_i^2 X_j^2\right)\\ &= \sum_i \mathbf{E}(X_i)^4 + 6 \sum_{i<j} \mathbf{E}(X_i^2 X_j^2)\end{aligned}$$

Using the Cauchy-Schwartz inequality we get:

$$\mathbf{E}(X_i^2 X_j^2) \le \mathbf{E}(X_i^4)^{1/2} \mathbf{E}(X_j^4)^{1/2} = K_4 < \infty$$

Then:

$$\begin{aligned}\mathbf{E}(S_n^4) &= \sum_{i=1}^{n}\mathbf{E}(X_i)^4 + 6\sum_{i<j}\mathbf{E}(X_i^2X_j^2) \le nK_4 + 6\binom{n}{2}\cdot K_4 \\ &= (n+3n(n-1))K_4 = (3n^2-2n)K_4 \le 3n^2K_4\end{aligned}$$

Therefore:

$$\mathbf{E}\left(\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4\right) = \sum_{n=1}^{\infty}\frac{\mathbf{E}(S_n^4)}{n^4} \le \sum_{n=1}^{\infty}\frac{3n^2K}{n^4} = \sum_{n=1}^{\infty}\frac{3K}{n^2} < \infty$$

Since the expectation of the random variable is finite then we must have the random variable finite with the exception of a set of measure 0 (otherwise the expectation will be infinite). This implies:

$$\sum_n \left(\frac{S_n}{n}\right)^4 < \infty \quad \text{a.s.}$$

But a sum can only be convergent if the term under the sum converges to zero. Therefore:

$$\lim_{n\to\infty}\left(\frac{S_n}{n}\right)^4 = 0 \quad \text{a.s.}$$

and consequently:

$$\frac{S_n}{n} \xrightarrow{\text{a.s}} 0$$

□

Example 3.5. I cannot resist giving a simple application of this theorem. Let A be an event that appears with probability $\mathbf{P}(A) = p \in (0,1]$. For example, roll a fair six sided die and let A be the event roll a 1 or a 6 ($\mathbf{P}(\{1,6\}) = 1/3$). Let γ_n denote the number of times A appears in *n independent* repetitions of the experiment. Then :

$$\lim_{n\to\infty}\frac{\gamma_n}{n} = p$$

This is an important example for statistics. Suppose for instance that we do not know that the die is fair but we have our suspicions. How do we test? All we have to do is roll the die many times ($n \to \infty$) and look at the average number of times 1 or 6 appears. If this number stabilizes around a different value than $1/3$ then the die is tricked. The next theorem will also tell how many times to roll the dies to be confident in our assessment.

To prove the result we simply apply the previous theorem. Define X_i as:

$$X_i = \begin{cases} 1 \text{ if event } A \text{ appears in repetition } i \\ 0 \text{ otherwise}\end{cases}$$

Then $\mathbf{P}(X_i = 1) = p$ and $\mathbf{P}(X_i = 0) = 1 - p$ so that $\mathbf{E}(X_i) = 1 \cdot p + 0 \cdot (1 - p) = p$. Clearly, the fourth moment is finite as well and applying the theorem: $\gamma_n = \sum_{i=1}^{n} X_i$. will converge to the stated value.

A Basic Central Limit Theorem: *The DeMoivre-Laplace Theorem:*

In order to prove the theorem we need:

Lemma 3.8 (Stirling's Formula). *For large n it can be shown that:*

$$n! \sim \sqrt{2\pi n} \cdot n^n e^{-n}$$

The proof of this theorem is only of marginal interest to us.

Theorem 3.5 (DeMoivre-Laplace). *Let $\xi_1 \cdots \xi_n$ be n independent r.v.'s each taking value 1 with probability p and 0 with probability $1 - p$ (Binomial(p) random variables). Let*

$$S_n = \sum_{i=1}^{n} \xi_i$$

and

$$S_n^* = \frac{S_n - \mathbf{E}(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - np}{\sqrt{np(1-p)}}$$

then for any $x_1, x_2 \in \mathbb{R}$, $x_1 < x_2$:

$$\begin{aligned} \lim_{n\to\infty} \mathbf{P}(x_1 \le S_n^* \le x_2) &= \Phi(x_2) - \Phi(x_1) \\ &= \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \end{aligned}$$

Note that Φ is the distribution function of a $N(0,1)$ random variable. This is exactly the statement of the regular Central Limit Theorem applied to Bernoulli random variables.

Proof. Notice that $S_n \sim$ Binomial(n, p) and $S_n^* = (S_n - np)/\sqrt{np(1-p)}$ is distributed equidistantly in the total interval $[\frac{-np}{\sqrt{np(1-p)}}, \frac{n-np}{\sqrt{np(1-p)}}]$. The length between two such consecutive values is $\Delta x = 1/\sqrt{np(1-p)}$.

For k large and $n - k$ large:

$$
\begin{aligned}
\mathbf{P}(S_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^k \\
&= \frac{\sqrt{2\pi n}\cdot n^n e^{-n}}{\sqrt{2\pi k}\cdot k^k e^{-k}\sqrt{2\pi(n-k)}\cdot(n-k)^{n-k} e^{-(n-k)}} p^k (1-p)^{n-k} \qquad (3.2)\\
&= \underbrace{\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}}_{\text{Term I}}\ \underbrace{\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}}_{\text{Term II}}
\end{aligned}
$$

(3.2) follows from Stirling's Formula. Remember that for $S_n = k$ the x value of $S_n^* = (S_n - np)/\sqrt{np(1-p)}$ is:

$$
\begin{aligned}
x = \frac{k-np}{\sqrt{np(1-p)}} &\Rightarrow k = np + x\sqrt{np(1-p)} \\
&\Rightarrow \frac{k}{np} = 1 + x\sqrt{\frac{1-p}{np}}
\end{aligned}
$$

Likewise we may express:

$$
\begin{aligned}
n-k = n - np - x\sqrt{np(1-p)} &\Rightarrow n-k = n(1-p) - x\sqrt{np(1-p)} \\
&\Rightarrow \frac{n-k}{n(1-p)} = 1 - x\sqrt{\frac{p}{n(1-p)}}
\end{aligned}
$$

Using these two expressions in the Term II of equation (3.2):

$$
\begin{aligned}
\log\left(\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}\right) &= -k\log\frac{k}{np} - (n-k)\log\frac{n-k}{n(1-p)} \\
= -k\log\left(1 + x\sqrt{\frac{1-p}{np}}\right) &- (n-k)\log\left(1 - x\sqrt{\frac{p}{n(1-p)}}\right)
\end{aligned}
$$

If we approximate $\log(1+\alpha) \simeq \alpha - \frac{\alpha^2}{2}$ we continue:

$$
\simeq -k\left(x\sqrt{\frac{1-p}{np}} - \frac{x^2}{2}\frac{1-p}{np}\right) - (n-k)\left(-x\sqrt{\frac{p}{n(1-p)}} - \frac{x^2}{2}\frac{p}{n(1-p)}\right) \qquad (3.3)
$$

Finally, we substitute k and $n-k$ and after calculations (skipped) we obtain:

$$
\lim_{n\to\infty} \log\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} = -\frac{x^2}{2}
$$

Also note that:

$$\sqrt{\frac{n}{k(n-k)}} \simeq \sqrt{\frac{n}{np \cdot n(1-p)}} = \frac{1}{\sqrt{np(1-p)}}$$

Putting both terms together we obtain:

$$\lim_{n\to\infty} \mathbf{P}(S_n^* = x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Delta x$$

where $\Delta x = \frac{1}{\sqrt{np(1-p)}}$

Thus:

$$\begin{aligned}\lim_{n\to\infty} \mathbf{P}(x_1 \leq S_n^* \leq x_2) &= \lim_{n\to\infty} \sum_{x_1 \leq x \leq x_2} \mathbf{P}(S_n^* = x) = \lim_{n\to\infty} \sum \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Delta x \\ &= \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-x^2/2} dx\end{aligned}$$

□

Problems

3.1. It is well-known that 23 “random” people have a probability of about 1/2 of having at least 1 shared birthday. There are 365 x 24 x 60 = 525,600 minutes in a year. (We’ll ignore leap days.) Suppose each person is labeled by the minute in which the person was born, so that there are 525,600 possible labels. Assume that a “random” person is equally likely to have any of the 525,600 labels, and that different “random” people have independent labels.

a) About how many random people are needed to have a probability greater than 1/2 of at least one shared birth-minute? (A numerical value is required.)
b) About how many random people are needed to have a probability greater than 1/2 of at least one birth-minute shared by three or more people? (Again, a numerical value is required. You can use heuristic reasoning, but explain your thinking.)

3.2. Show that any simple function f can be written as $\sum_i b_i \mathbf{1}_{B_i}$ with B_i disjoint sets (i.e. $B_i \cap B_j = \emptyset$, if $i \neq j$).

3.3. Prove the 4 assertions in Exercise 3.1 on page 61.

3.4. Give an example of two variables X and Y which are uncorrelated but not independent.

3.5. Prove the properties (i)-(v) of the expectation in Proposition 3.1 on page 69.