A SUMMARY OF BASIC TOPICS IN PROBABILITY THEORY

EDWARD CHERNYSH

This document is intended to be a quick and dirty summary of the topics covered in the second half of Math 356 at McGill university. This summary will be done using the lecture notes from the fall semester of 2016. Unless otherwise stated, $(\Omega, \mathfrak{M}, \mathsf{P})$ will denote a probability space. Given a topological space $(X, \mathfrak{T})$, we shall denote by $\mathcal{B}_X$ the Borel σ-algebra on X generated by $\mathfrak{T}$. In the case of the real line, we have for every $n \in \mathbb{N}$ the relation

$$\mathcal{B}_{\mathbb{R}^n} = \bigotimes_{j=1}^{n} \mathcal{B}_{\mathbb{R}} \neq \prod_{j=1}^{n} \mathcal{B}_{\mathbb{R}}.$$

A random variable is a function $X : \Omega \longrightarrow \mathbb{R}$ with the property that $X^{-1}(B) \in \mathfrak{M}$ for every $B \in \mathcal{B}_{\mathbb{R}}$, i.e. a real valued measurable function on $(X, \mathfrak{M})$. With these notions out of the way, we are free to "jump" into the things we wish to review. Henceforth, all functions are assumed to be real valued.

Proposition 1. *Let X be a random variable on $(\Omega, \mathfrak{M}, \mathsf{P})$ and let g be Borel measurable on $\mathbb{R}$. Then $g \circ X$ is again a random variable on $(\Omega, \mathfrak{M}, \mathsf{P})$.*

Proof. The composition $g \circ X$ is measurable since for every $B \in \mathcal{B}_{\mathbb{R}}$ it is clear that

$$(g \circ X)^{-1}(B) = X^{-1}(g^{-1}(B))$$

which lives in $\mathfrak{M}$ since X is measurable and $g^{-1}(B)$ is Borel. □

1. Single Variable Topics

Let us fix a probability space $(\Omega, \mathfrak{M}, \mathsf{P})$ and discuss random variables defined on the space. There will ultimately be two important cases we must distinguish: the cases where X is *discrete* or *continuous*; random variables that are neither continuous nor discrete will not be of much importance to us.

Date: May 1, 2018.

1.1. **Discrete Random Variables.** A random variable X on $(\Omega, \mathfrak{M})$ is called discrete provided $X(\Omega)$ is countable[1]. Given a random variable X, we may speak of the distribution of X. This is a monotone function $f_X : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$F_X(x) := \mathsf{P}(X \leq x) = \mathsf{P}\left(\{x \in X : X(x) \leq x\}\right).$$

In like, the probability mass function of X, denoted f_X, is the mapping given by

$$f_X(x) := \mathsf{P}(X = x), \quad \forall x \in \mathbb{R}.$$

Definition 1. Let X be a discrete random variable on $(\Omega, \mathfrak{M}, \mathsf{P})$ and let g be Borel measurable. Denote by $S = \{x_n\}$ the image of X. The expectation of $g \circ X$ is defined as

$$\mathsf{E}[g \circ X] := \sum_n g(x_n)\mathsf{P}(X = x_n),$$

when the above exists.

To be more precise, when we speak of the expectation of X, we are referring to the quantity $\mathsf{E}[X]$, which exists if and only if $\mathsf{E}[|X|] < \infty$. Moreover, $\mathsf{E}[g \circ X]$ exists if and only if $\mathsf{E}[|g \circ X|] < \infty$.

Proposition 2. *Let X be a discrete random variable and suppose $\mathsf{E}[X]$ exists. If $a, b \in \mathbb{R}$ then $\mathsf{E}[ax + b]$ exists and is equal to $a\mathsf{E}[X] + b$.*

Proof. Obviously, letting S (as above) be the image of X we find that

$$\mathsf{E}[|aX + b|] = \sum_n |ax_n + b|\, \mathsf{P}(X = x_n) \leq |a|\, \mathsf{E}[|X|] + |b| \sum_n \mathsf{P}(X = x_n)$$

which is finite because $\sum_n \mathsf{P}(X = x_n) = \mathsf{P}(\Omega) = 1$. This argument also proves the stated equality. ☐

Furthermore, we find that there holds the following.

Proposition 3. *Let X and Y be discrete random variables on a probability space $(\Omega, \mathfrak{M}, \mathsf{P})$. When all terms are well defined,*

$$\mathsf{E}[X + Y] = \mathsf{E}[X] + \mathsf{E}[Y].$$

---

[1] This could be finite or countably infinite.

Proof. Enumerate the image of X by $\{x_n\}$ and that of Y with $\{y_m\}$. It is clear from the definition and the countable additivity of $\mathsf{P}(\cdot)$ that

$$\begin{aligned}
\mathsf{E}[X+Y] &= \sum_n \sum_m (x_n + y_m)\mathsf{P}(X = x_n, Y = y_n) \\
&= \sum_n \sum_m x_n \mathsf{P}(X = x_n, Y = y_n) + \sum_m \sum_n y_m \mathsf{P}(X = x_n, Y = y_n) \\
&= \sum_n x_n \mathsf{P}(X = x_n) + \sum_m y_m \mathsf{P}(Y = y_m) \\
&= \mathsf{E}[X] + \mathsf{E}[Y].
\end{aligned}$$

This concludes the proof. □

Definition 2. Let X be a random variable such that $\mathsf{E}[X^2]$ exists. Then the variance of X is defined as

$$\mathsf{Var}(X) := \mathsf{E}\left[(X - \mathsf{E}[X])^2\right] = \mathsf{E}[X^2] - \mathsf{E}[X]^2.$$

Notice also that $\mathsf{Var}(X) \geq 0$.

Definition 3. If X is a discrete random variable and there exists $s_0 > 0$ such that $\mathsf{E}[e^{s_0 X}]$ exists, then we can define the moment generating function of X as follows:

$$M_X(t) := \mathsf{E}[e^{tX}]$$

which exists for all $t \in (-s_0, s_0)$.

We accept the following theorem without proof.

Theorem 4. *If X is a discrete random variable and $M_X(t)$ exists in a neighbourhood of zero, then $\mathsf{E}[X^k]$ exists for all $k \in \mathbb{N}$. Moreover, for t in this neighbourhood,*

$$M^{(k)}(0) = \frac{\mathrm{d^k}}{\mathrm{d}t^k} M(s)\bigg|_{t=0}$$

and

$$\mathsf{E}[X^k] = M^{(k)}(0)$$

for all $k \in \mathbb{N}$.

1.2. **Continuous Random Variables.** A natural counterpart to discrete random variables are those that are continuous. Let X be a random variable on $(\Omega, \mathfrak{M}, \mathsf{P})$. One can always define a distribution function for X by setting

$$F_X(x) = \mathsf{P}(X \leq x), \quad \forall x \in \mathbb{R}.$$

With this notion, we are free to define what we mean by a *continuous random variable.*

Definition 4. A random variable X on a probability space $(X, \mathfrak{M}, \mathsf{P})$ is called *continuous* provided F_X is absolutely continuous.

Therefore, we have the following "simpler" characterization of continuous random variables.

Corollary 5. *A random variable on* $(\Omega, \mathfrak{M}, \mathsf{P})$ *is continuous if and only if there exists* $f_X \in L^1(\mathbb{R})$ *such that*

$$\mathsf{P}(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(\sigma)\,\mathrm{d}m(\sigma)$$

for all $x \in \mathbb{R}$*. We call this* f_X *the probability mass function of* X*.*

In this way, we may proceed to give a definition of expectation for which the results of the previous subsection continue to hold. Let X be a continuous random variable on $(\Omega, \mathfrak{M}, \mathsf{P})$ and let f_X be its probability mass function. If $g : \mathbb{R} \longrightarrow \mathbb{R}$ is Borel measurable, $g \circ X$ is once again a random variable and we define its expectation to be

$$\mathsf{E}[g \circ X] := \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}m.$$

whenever $\mathsf{E}[|g \circ X|] < \infty$. Once again, the expectation of X is simply

$$\mathsf{E}[X] := \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}m$$

whenever it exists. One defines variance in a similar way and the stuff covered above follows once again.

Let us now conclude with the following observation, which applies to discrete and continuous random variables.

> If X and Y are random variables of the same type and $a, b \in \mathbb{R}$, then
>
> $$\mathsf{E}[aX + bY] = a\mathsf{E}[X] + b\mathsf{E}[Y]$$
>
> whenever all is well defined.

2. Higher Dimensions and Independence

Let us fix n random variables $X_1, X_2, \dots, X_n$ on some probability space $(\Omega, \mathfrak{M}, \mathsf{P})$ and consider the tuple $(X_1, \dots, X_n)$ which is a function

$$\Omega \times \cdots \times \Omega \longrightarrow \mathbb{R}^n.$$

We call this tuple a random variable in n-dimensions provided $(X_1, \dots, X_n)^{-1}(B)$ is an element of $\bigotimes_{j=1}^{n} \mathfrak{M}$ for all $B \in \mathcal{B}_{\mathbb{R}^n} = \bigotimes_{j=1}^{n} \mathcal{B}_{\mathbb{R}}$. For the sake of simplicity, we will focus on a pair of random variables (X, Y) throughout this section. The following holds

Proposition 6. *A pair* (X, Y) *is a 2 dimensional random variable if and only if* X *and* Y *are both random variables on* $(\Omega, \mathfrak{M}, \mathsf{P})$.

We may now introduce the so-called *joint distribution function* of a pair (X, Y). This distribution function of the pair (X, Y) is given by

$$F_{(X,Y)}(x, y) := \mathsf{P}\left(X \leq x, Y \leq y\right).$$

2.1. **More on Distributions and Mass Functions.**

There are certain subtleties that arise when considering pairs of random variables. Let $(\Omega, \mathfrak{M}, \mathsf{P})$ be a probability space and suppose X, Y are random variables on this space. We will only consider cases where either both X and Y are discrete, or both are continuous.

2.1.1. *The Discrete Case.* We say that the pair (X, Y) is a discrete two dimensional random variable provided (X, Y) takes countably many values in $\mathbb{R}^2$. This then implies that both X and Y are discrete random variables $\Omega \longrightarrow \mathbb{R}$. Conversely, if X and Y are discrete random variables $\Omega \longrightarrow \mathbb{R}$ one can form a discrete bivariate random variable (X, Y) in the natural way.

Suppose now that we are given a discrete random variable (X, Y) from $\Omega \times \Omega$ into $\mathbb{R}^2$. Then X and Y are both discrete one-dimensional variables who take values in countable sets $\{x_n\}$ and $\{y_k\}$, respectively. This gives us a convenient form for the probability mass function associated to either X or Y. Indeed, if f_X denotes the *marginal* probability mass function of X we have

$$f_X(x) = \mathsf{P}(X = x) = \sum_k \mathsf{P}(X = x, Y = y_k) = \sum_k f_{(X,Y)}(x, y_k)$$

where $f_{(X,Y)}$ is the joint probability mass function of the pair (X, Y). In the discrete case, this is given by $\mathsf{P}(X = x, Y = y)$ at (x, y).

2.1.2. *The Continuous Case.* Here we choose the same setting as above, although we assume that (X, Y) is a continuous random variable, i.e. that there exists a non-negative function $f_{(X,Y)} \in L^1(\mathbb{R}^2)$ such that for all $(x, y) \in \mathbb{R}^2$

$$F_{(X,Y)}(x, y) = \mathsf{P}(X \leq x, Y \leq y) = \iint\limits_{\substack{(\sigma,\tau) \\ \sigma \leq x, \tau \leq y}} f_{(X,Y)}(\sigma, \tau)\, \mathrm{d}m(x, y).$$

By Fubini's theorem, this is to say that

$$F_{(X,Y)}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{(X,Y)}(\sigma, \tau)\, \mathrm{d}\sigma\, \mathrm{d}\tau.$$

2.2. **Independence of Random Variables.** We generalize the notion of independence to the case where multiple random variables are given. To this end, we let $(\Omega, \mathfrak{M}, \mathsf{P})$ be a probability space and fix random variables $X_1, \ldots, X_n$ on Ω.

Definition 5. These random variables are called *independent* if and only if for every finite sequence $\{B_1, \ldots, B_n\} \subseteq \mathcal{B}_{\mathbb{R}}$ one has

$$\mathsf{P}\left(\bigcap_{j=1}^{n} X_j^{-1}(B_j)\right) = \prod_{j=1}^{n} \mathsf{P}\left(X_j^{-1}(B_j)\right).$$

This may be extended as follows.

Definition 6. Assume $(\Omega, \mathfrak{M}, \mathsf{P})$ is a probability space and $\mathcal{F}$ is a family of random variables on Ω. The elements of $\mathcal{F}$ are said to be independent if every finite subfamily of $\mathcal{F}$ is independent.

The following basic fact regarding the transformation of random variables holds. That is, independence is preserved by Borel mappings.

Proposition 7. *Let X and Y be independent random variables on $(\Omega, \mathfrak{M}, \mathsf{P})$ and suppose $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ are Borel measurable. Then $f(X)$ and $g(Y)$ are again independent.*

Proof. Let us fix $B_1, B_2 \in \mathcal{B}_{\mathbb{R}}$. Then,

$$\begin{aligned} \mathsf{P}\left((f \circ X)^{-1}(B_1) \cap (g \circ Y)^{-1}(B_2)\right) &= \mathsf{P}\left(X^{-1}(f^{-1}(B_1)) \cap Y^{-1}(g^{-1}(B_2))\right) \\ &= \mathsf{P}\left(X^{-1}(f^{-1}(B_1))\right) \cdot \mathsf{P}\left(Y^{-1}(g^{-1}(B_2))\right); \end{aligned}$$

which concludes the proof. □

We shall state, without proof, the following identity regarding independent random variables.

Proposition 8. *Let $X_1, \ldots, X_n$ be independent random variables.*

$$\mathsf{Var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \mathsf{Var}(X_j).$$

One more definition is now in order.

Definition 7. Let $\{X_j : j \in \mathbb{N}\}$ be a family of independent random variables on $(\Omega, \mathfrak{M}, \mathsf{P})$. We say that these are identically distributed if

$$F_{X_1} \equiv F_{X_2} \equiv \cdots \equiv F_n \equiv \cdots$$

2.3. **Sums of I.I.D. Random Variables.** Here we list facts regarding the sum of finitely many independent, identically distributed, random variables.

(A) We denote by $\mathcal{N}(\mu,\sigma^2)$ the set of all normal random variables with mean μ and variance σ^2. That is, if $X \sim \mathcal{N}(\mu,\sigma^2)$ then X has probability mass function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

This is to say that for every $x \in \mathbb{R}$ one has

$$\mathsf{P}(X \le x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)\,\mathrm{d}m(t).$$

Now, suppose we are given n independent random variables $X_j \sim \mathcal{N}(\mu_j,\sigma_j^2)$. Then

$$\textstyle\sum_{j=1}^{n} X_j \sim \mathcal{N}\left(\sum_{j=1}^{n}\mu_j, \sum_{j=1}^{n}\sigma_j^2\right).$$

This can be shown by studying the MGF of the sum and invoking the uniqueness.

(B) A Poisson random variable in $\mathcal{P}(\lambda)$ is a measurable function X on Ω having the probability mass function

$$\mathsf{P}(X=k) = f_X(k) := e^{-\lambda}\frac{\lambda^k}{k!}, \quad \forall k \in \mathbb{N}_0.$$

By countably additivity of a measure, it follows that the probability distribution of X is given by

$$\mathsf{P}(X \le k) = e^{-\lambda}\sum_{j=0}^{k}\frac{\lambda^j}{j!}.$$

In this case, we shall write $X \sim \mathcal{P}(\lambda)$ if $\lambda > 0$ is given. Suppose now that we are given a finite family $\{X_1,\dots,X_n\}$ with $X_j \in \mathcal{P}(\lambda_j)$ for some $\lambda_j > 0$. Then, algebraic manipulations (which we omit) shows that

$$\textstyle\sum_{j=1}^{n} X_j \sim \mathcal{P}\left(\sum_{j=1}^{n}\lambda_j\right).$$

(C) If $\alpha,\beta > 0$ a $\Gamma(\alpha,\beta)$ random variable is some measurable X on Ω having the probability density function

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$$

for $x > 0$.

(D) An exponential random variable X, written $X \sim \mathcal{E}(\lambda)$ for $\lambda > 0$, is a measurable function on Ω having probability density function

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Now, suppose we are given n random variables $X_j \sim \mathcal{E}(\lambda)$. Then,

$$\textstyle\sum_{j=1}^n X_j \sim \Gamma\left(n, \lambda\right).$$

(E) We now turn towards the χ^2_ν distributions/random variables. A random variable is called a χ^2_ν-variable (with ν degrees of freedom) if

$$X \sim \Gamma\left(\frac{\nu}{2}, 2\right).$$

That is, if X has the probability density function

$$f_X(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \begin{cases} x^{\nu-1}e^{-x/2}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

3. Modes of Convergence

In this section we mention different ways in which random variables can converge. Much of this is identical in a measure-theoretic setting, but for our sake we will assume we are given a finite measure space. Henceforth, we fix a measure space $(X, \mathfrak{M}, \mu)$ with $\mu(X) < \infty$.

Definition 8. Let $\{f_n\}_{n=1}^\infty$ be a sequence of measurable functions on X. We say that f_n converges to a measurable function f in measure if for every $\varepsilon > 0$

$$\lim_{n\to\infty} \mu\left(\{|f_n - f| \geq \varepsilon\}\right) = 0.$$

It will then be customary to write

$$f_n \xrightarrow{\mu} f, \quad \text{as } n \to \infty.$$

Let us now give some results regarding this notion of convergence.

Theorem 9. *Let $\{f_n\}$ and $\{g_n\}$ be sequences of measurable functions converging to measurable f and g in measure, respectively. Then*

$$f_n \pm g_n \xrightarrow{\mu} f + g \quad \text{and} \quad f_n g_n \xrightarrow{\mu} fg.$$

Furthermore, if g is continuous,

$$g(f_n) \xrightarrow{\mu} g(f).$$

Proof. For the first property, it suffices to consider '+' only. Let $\varepsilon > 0$ be given, and note that by the triangle inequality

$$\{|f_n + g_n - (f+g)| \geq \varepsilon\} \subseteq \{|f_n - f| + |g_n - g)| \geq \varepsilon\}.$$

On the other hand,

$$\{|f_n - f| + |g_n - g)| \geq \varepsilon\} \subseteq \left\{|f_n - f| \geq \frac{\varepsilon}{2}\right\} \cup \left\{|g_n - g| \geq \frac{\varepsilon}{2}\right\}.$$

Since $f_n \xrightarrow{\mu} f$ and $g_n \xrightarrow{\mu} g$, it follows that

$$\lim_{n\to\infty} \{|f_n + g_n - (f+g)| \geq \varepsilon\} \leq \lim_{n\to\infty} \mu\left(\left\{|f_n - f| \geq \frac{\varepsilon}{2}\right\}\right) + \lim_{n\to\infty} \mu\left(\left\{|g_n - g| \geq \frac{\varepsilon}{2}\right\}\right) = 0.$$

The remaining two will not be proven here. ☐

3.1. **Convergence in Distribution.** Now let us fix a probability space $(\Omega, \mathfrak{M}, \mathsf{P})$ and a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ on this space. We say that X_n converges to a random variable X in distribution provided

$$\lim_{n\to\infty} \mathsf{P}(X_n \leq x) = \mathsf{P}(X \leq x)$$

for all continuity points x of $\mathsf{P}(X \leq x)$. In this case, we shall write

$$X_n \xrightarrow{\mathsf{dist}} X, \quad \text{as } n \to \infty.$$

We shall state, without proof, the subsequent result

Theorem 10. *Suppose that a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges to X in distribution and another sequence $\{Y_n\}_{n=1}^{\infty}$ converges to a constant c in distribution.*

(1) $X_nY_n \xrightarrow{\mathsf{dist}} xC$,

(2) $Y_n \xrightarrow{\mathsf{dist}} c$ if and only if $Y_n \xrightarrow{\mathsf{P}} c$.

3.2. **The Central Limits Theorem.** Suppose we are given a sequence of independent, identically distributed, random variables which we call $\{X_i : i \in \mathbb{N}\}$. Assume also that for every index i

$$\mathsf{E}[X_i] = \mu \quad \text{and} \quad \mathsf{Var}(X_i) = \sigma^2,$$

where $\mu, \sigma^2 < \infty$. Define for $n \in \mathbb{N}$ the quantity $S_n := \sum_{i=1}^{n} X_i$. The Central limit theorem states that

$$\frac{S_n - nm}{\sqrt{n}\sigma} \xrightarrow{\mathsf{dist}} Z \sim \mathcal{N}(0,1).$$