Probability for Statistics

Kexing Ying

May 15, 2020

Contents

1 Introduction **1**
1.1 Probability Measures 1
1.2 The Lebesgue Measure 3

2 Random variables **4**
2.1 Cumulative Distribution Function 5
2.2 Types of Random Variables 6
2.3 Transformations of Random Variables 7

3 Multivariate Random Variables **8**
3.1 Covariance and Correlation 8
3.2 Transformation of Multivariate Random Variables 9

1 Introduction

1.1 Probability Measures

Last year we saw briefly constructions and definitions relevant to working with probabilities such as σ-algebras, random variables and more. We will revisit them here with a more general (and more technical) approach.

Definition 1.1 (σ-algebra)**.** Let X be a set. A σ-algebra on X, $\mathcal{A}$ is a collection of subsets of X such that

- $\varnothing \in \mathcal{A}$
- for all $A \in \mathcal{A}$, $A^C \in \mathcal{A}$
- for all $(A_n)_{n=1}^{\infty} \subseteq \mathcal{A}$, $\bigcup_n A_n \in \mathcal{A}$.

Proposition 0.1. *Let X be a set and I a non-empty collection of σ-algebras on X. Then $\bigcap I$ is also a σ-algebra on X.*

This proposition is easy to check and thus, it makes sense to consider the σ-algebra generated by some set.

Definition 1.2 (Generator of σ-algebra)**.** Let X be a set and $S \subseteq \mathcal{P}(X)$ a collection of subsets of X. Then the σ-algebra generated by S is

$$\sigma(S) := \bigcap \{\mathcal{A} \supseteq S \mid \mathcal{A} \text{ is a } \sigma\text{-algebra on } X\}$$

By the fact that the power set of X is a σ-algebra containing S, we see that $\{\mathcal{A} \supseteq S \mid \mathcal{A} \text{ is a } \sigma\text{-algebra on } X\}$ is non-empty and so for all $S \subseteq \mathcal{P}(X)$, $\sigma(S)$ a (and the smallest) σ-algebra on X.

With this, we can construct a commonly seen σ-algebra, the Borel σ-algebra. Given some topological space X, the Borel σ-algebra on X is the σ-algebra generated by $\mathcal{T}_X$, i.e. $\mathcal{B}(X) = \sigma(\mathcal{T}_X)$. We will most commonly work with the Borel σ-algebra on the real numbers $\mathcal{B}(\mathbb{R})$.

We call the ordered pair $(X, \mathcal{A})$ where $\mathcal{A}$ is a σ-algebra n X a *measurable space.*

Definition 1.3 (Measure)**.** Given a measurable space $(X, \mathcal{A})$, a measure on this measurable space $\mu : \mathcal{A} \to [0, \infty]$ is a function such that

- $\mu(\varnothing) = 0$
- for all disjoint sequence $(A_n)_{n=1}^{\infty} \subseteq \mathcal{A}$, $\mu\left(\bigsqcup_n A_n\right) = \sum_n \mu(A_n)$

With measures defined, we can add an additional restriction to create a *probability space.*

Definition 1.4 (Probability Measure)**.** Let μ be a measure on the measurable space $(X, \mathcal{A})$, then μ is a probability measure if and only if $\mu(X) = 1$. We then call the order triplet $(X, \mathcal{A}, \mu)$ a probability space.

To distinguish probability space from normal measure spaces, we will often write $(\Omega, \mathcal{F}, \mathbb{P})$ to denote a probability space. We will call Ω the *sample space*, $\mathcal{F}$ the *events* and for all $A \in \mathcal{F}$, $\mathbb{P}(A)$ the *probability* of the event A.

1.1.1 Some Properties of the Probability Measure

Theorem 1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with $(A_i)_{i=1}^{\infty}$ an increasing sequence in $\mathcal{F}$, then*

$$\mathbb{P}\left(\bigcup_i A_i\right) = \lim_{i \to \infty} \mathbb{P}(A_i).$$

Proof. Follows from additivity of the probability measure by writing $\bigcup_i A_i$ as the disjoint union $A_1 \sqcup \bigsqcup_i (A_{i+1} \setminus A_i)$. □

A corollary of the above is immediately deduced by considering the complement of a decreasing function.

Corollary 1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with $(A_i)_{i=1}^{\infty}$ a decreasing sequence in $\mathcal{F}$, then*

$$\mathbb{P}\left(\bigcap_i A_i\right) = \lim_{i \to \infty} \mathbb{P}(A_i).$$

In fact the two above propositions apply to general measures with identical proofs.

Theorem 2. *Suppose $(\Omega, \mathcal{F})$ is a measurable space with the finitely additive function $\mathbb{P} : \mathcal{F} \to [0,1]$ such that theorem 1 holds, then $\mathbb{P}$ is a probability measure.*

Proof. Let $(A_i)_{i=1}^{\infty}$ be a sequence of disjoint sequence in $\mathcal{F}$, then, let us define $B_n = \bigcup_{i=1}^{n} A_i$. As σ-algebras are closed under unions, $B_n \in \mathcal{F}$ for all n. Now, by assumption, as (B_n) is increasing, $\mathbb{P}\left(\bigsqcup_i A_i\right) = \mathbb{P}\left(\bigcup_n B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ where the second to last equality if true by finite additivity. $\square$

1.2 The Lebesgue Measure

As the point of measures in general is to assign sets (in the relevant σ-algebra) to some number, it might be useful to take a look at the most famous measure of them all – the Lebesgue measure.

In the easiest terms, the Lebesgue measure is a measure, that maps the interval $[a,b] \subseteq \mathbb{R}$ to the real number $b - a$. In probability, we can think of this as $\mathbb{P}([a,b])$, or the probability of $X \in [a,b]$ where X is a random variable with uniform distribution, (we will talk more about what this means in the next section).

In this course, we will assume the Lebesgue measure exists (and in fact, is unique which we shall prove from first principle in next term's measure theory course).

It turns out that a lot of sets are Lebesgue measurable, in fact, the set of sets that are Lebesgue measurable is greater than the Borel σ-algebra. However, unfortunately, not all sets a Lebesgue measurable. We will give an example of a non-Lebesgue measurable set here called the Vitali set.

Definition 1.5 (The Vitali Set)**.** Let $\Omega := [0, 2\pi)$, then we can some probability measure $\mathbb{P}$ such that $\mathbb{P}(s) = \frac{\beta - \alpha}{2\pi}$ corresponding to the Lebesgue measure. Now, let $\sim$ be the equivalence relation such that $x \sim y$ if and only if $x - y$ is a rational multiple of 2π. As $\sim$, is an equivalence relation, it partitions Ω, so there is a set of equivalence classes $\Omega / \sim$. Now, by using the axiom of choice, the Vitali set is defined to be the set A choosing one element from each equivalence classes in $\Omega / \sim$.

Theorem 3. *The Vitali Set is not measurable with respect to the measure in theorem 1.5.*

Proof. We suppose for contradiction that the Vitali set is measurable. As $\mathbb{Q}$ is countable, let $x_1, x_2, \cdots$ be the enumeration of all rational multiples of 2π in $[0, 2\pi)$. Now, define $A_i := A + x_i = \{a + x_i \mid a \in A\}$. We see that A_i, A_j are disjoint for all $i \neq j$ since if there exists some $a \in A + x_i \cap A + x_j$, so there exists $\alpha, \beta \in A$, $\alpha + x_i = a = \beta + x_j$, so $\alpha \sim \beta$ implying $\alpha = \beta$ by the construction of A and hence, $x_i = x_j$. Now, as $\Omega = \bigsqcup_{i=1}^{\infty} A_i$, we have $1 = \mathbb{P}(\Omega) = \mathbb{P}(\bigsqcup_{i=1}^{\infty} A_i) = \sum \mathbb{P}(A_i)$. However, as the Lebesgue measure is transitional invariant, for all i, j, $\mathbb{P}(A_i) = \mathbb{P}(A_j)$, so $1 = \sum \mathbb{P}(A_i) = \lim_{i\to\infty} i\mathbb{P}(A_1)$ which results in a contradiction by applying excluded middle on $\mathbb{P}(A_1) = 0$. $\square$

2 Random variables

Now that we have the basic notion of a probability space, we would like to play around with it using *random variables.* In the most general sense, random variables are simply functions from the probability space to another measurable space, most commonly the real numbers equipped with $\mathcal{B}(\mathbb{R})$.

Definition 2.1 (Measurable Functions)**.** Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be two measurable spaces and $f : X \to Y$ a mapping between the two. We call f measurable if and only if for all $A \in \mathcal{B}$, $f^{-1}(A) \in \mathcal{A}$.

Definition 2.2 (Random Variables)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E, \mathcal{A})$ be a measurable space. Then an E-valued random variable is a measurable function $X : \Omega \to E$.

In general, we will only be working with real valued random variables, so the image measurable space is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Often, when we have a random variable $X : \Omega \to \mathbb{R}$, we might ask questions such as "what is the probability that $X \in A$" for some $A \subseteq \mathrm{Im}X$. We now see that this question is asking for exactly $\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$ (this makes sense as X is measurable).

Theorem 4. *If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $X : \Omega \to \mathbb{R}$ is a function. Then X is a $\mathbb{R}$-valued random variable if and only if for all $x \in \mathbb{R}$,*

$$\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}.$$

Proof. The forward direction is trivial so let us consider the reverse. Suppose for all $x \in \mathbb{R}$, $\{\omega \in \Omega \mid X(\omega) \leq x\} = X^{-1}((-\infty, x]) \in \mathcal{F}$. Then, for all $a, b \in \mathbb{R}$, $a < b$, $X^{-1}((-\infty, a]), X^{-1}((-\infty, b]) \in \mathcal{F}$, so $X^{-1}((-\infty, a])^c = X^{-1}((a, \infty)) \in \mathcal{F}$, and thus, $X^{-1}((a, \infty)) \cap X^{-1}((-\infty, b]) = X^{-1}((a, b]) \in \mathcal{F}$. $\square$

Let us now consider some properties we can but on these random variables.

Definition 2.3 (Identically Distributed Random Variables)**.** Let X, Y be two real valued random variables. We say X and Y are identically distributed if for all $S \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(X \in S) = \mathbb{P}(Y \in S).$$

We note that two random variables are identically distributed does not imply they are equal, that is they are not necessarily the same function. An easy example of this if to let X, Y be the number of heads and tails of n coin flips. We see that X, Y are identically distributed by symmetry but definitely not equal.

Another property that is useful for random variables is the notion of independence.

Definition 2.4 (Independence of Events)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(A_n) \subseteq \mathcal{F}$ a sequence of events. Then (A_n) is said to be independent if and only if for all *finite* index set I,

$$\mathbb{P}\left(\bigcap_{n \in I} A_n\right) = \prod_{n \in I} \mathbb{P}(A_n).$$

Definition 2.5 (Independence of σ-algebras)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{A}_n)$ be a sequence of sub-σ-algebras of $\mathcal{F}$. Then $(\mathcal{A}_n)$ is said to be independent if and only if for all $(A_n) \subseteq \mathcal{F}$ a sequence of events such that $A_i \in \mathcal{A}_i$, (A_n) is independent.

Equipped with these two notions of independence, it makes sense to create a notion of some σ-algebra induced by arbitrary measurable functions and with that the notion of independence of random variables is also induced.

Definition 2.6 (σ-algebra Generated by Functions)**.** Let E be a set and $\{f_i : E \to \mathbb{R} \mid i \in I\}$ be an indexed family of real-valued functions. Then the σ-algebra on E generated by these functions is

$$\sigma(\{f_i \mid i \in I\}) := \sigma(\{f_i^{-1}(A) \mid A \in \mathcal{B}(\mathbb{R}), i \in I\}).$$

Note that with this definition, we created the smallest σ-algebra on E such that all f_i are measurable and for a single function f, $\sigma(\{f \mid i \in I\}) = \{f^{-1}(A) \mid A \in \mathcal{B}(\mathbb{R})\}$.

Definition 2.7 (Independence of Random Variables)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and (X_n) be a sequence of real-valued random variables. Then (X_n) is said to be independent if and only if the family of σ-algebras $\sigma(X_n)$ is independent.

We will check that this definition of independence of random variables behave as intended, that is $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$.

Theorem 5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X, Y be real-valued random variables. Then X, Y are independent if and only if for all $A, B \in \mathcal{B}(\mathbb{R})$,*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

Proof. Recall that $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}((X \in A) \cap (Y \in B)) = \mathbb{P}(X^{-1}(A) \cap Y^{-1}(B))$. Now, if $\sigma(X)$ and $\sigma(Y)$ are independent, as $X^{-1}(A) \in \sigma(X)$ and $Y^{-1}(B) \in \sigma(Y)$, by definition, we have $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$.

Similarly, if the equality in question is true for all $A, B \in \mathcal{B}(\mathbb{R})$, then the σ-algebras are independent by definition, and thus, so are the random variables. $\square$

2.1 Cumulative Distribution Function

We would now like to take a look at the cumulative distribution function of a random variable X.

Definition 2.8 (Cumulative Distribution Function)**.** Given a random variable X, the cumulative distribution function, of simply the CDF of X is

$$F_x(x) = \mathbb{P}(X \leq x).$$

This function is well defined as by our previous assertion, X is measurable on $\mathcal{B}(\mathbb{R})$ if and only if $\{\omega \in \Omega \mid X(\omega) \leq x\}$ is measurable for all x.

The CDF of a random variable is important as it characterised the random variable. Formally it can be stated as,

Theorem 6. *Let X, Y be real valued random variables. Then X, Y are identically distributed if and only if $F_X = F_Y$.*

Proof. The forward direction is trivial while the backwards direction follows from the fact that every open real set can be constructed using sets of the form $\{x \leq a | x \in \mathbb{R}\}$ from some $a \in \mathbb{R}$. $\square$

The CDF of a random variable have some nice properties which we have used throughout our first year probability course.

Proposition 6.1. *Given a random variable X, its CDF, F_X, is non-decreasing.*

Proof. This follows from the fact for all $x, y \in \mathbb{R}$, if $x < y$, then we can write $\{\omega \in \Omega \mid X(\omega) \leq y\} = \{\omega \in \Omega \mid X(\omega) \leq x\} \sqcup \{\omega \in \Omega \mid x < X(\omega) \leq y\}$, and so

$$F_X(y) = F_X(x) + \mathbb{P}(x < X \leq y) \geq F_X(x).$$

$\square$

Proposition 6.2. *Given a random variable X, $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.*

Proof. We recall that the axiom that $\mathbb{P}(\Omega) = 1$, so it suffices to prove that $X^{-1}(\lim_{x \to \infty}(-\infty, x])) = \Omega$. But this is trivial as every element of Ω is mapped to a real number so we are done. (This first claim is true by similar argument.)[1] $\square$

Proposition 6.3. *Let X be a random variable, then $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$.*

Proof. Similar proof to the previous proposition. $\square$

2.2 Types of Random Variables

The most simple random variable we have is the point mass random variable.

Definition 2.9. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, then the point mass random variable X_a at a is the function $X_a : \Omega \to \mathbb{R} : \omega \mapsto a$.

We can easily see that the point mass random variable has the CDF $\delta_a(x) =$ if $(x < a)$ then $0,$ else 1. While the point mass random variable in itself is not very interesting, it is the building blocks for discrete random variables.

Definition 2.10 (Discrete Random Variable)**.** A random variable X is a discrete random variable if and only if there exist sequences $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$ such that $\sum b_i = 1$ and $F_X(x) = \sum b_i \delta_{a_i}(x)$.

Definition 2.11 (Continuous Random Variable)**.** A random variable X is a continuous random variable if and only if F_X is continuous on $\mathbb{R}$.

[1] Note that this proof is not technically true since we can't say $\lim_{x \to \infty}(-\infty, x] = \mathbb{R}$. But this can be fixed by considering any sequence (x_n) that it increasing to ∞.

Definition 2.12 (Absolutely Continuous Random Variable)**.** A random variable X is absolutely continuous if and only if there exists some $f_X : \mathbb{R} \to \mathbb{R}$ such that $F_X(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$.

We note that continuous random variables need not be absolutely continuous (see the Cantor distribution), however, for most purposes, we can assume interchangeability.

Proposition 6.4. *Let X be any random variable and let $x_n \uparrow x \in \mathbb{R}$, then $\mathbb{P}(X < x) = \lim_{x_n \uparrow x} \mathbb{P}(X \le x_n)$.*

Proof. We define $A_n := \{\omega \in \Omega \mid X(\omega) \le x_n\}$, then $A_n \uparrow A := \{\omega \in \Omega \mid X(\omega) < x\}$. So, by taking the probability of the limit of A_n, we have $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$. □

Proposition 6.5. *Let X be a continuous random variable, then $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.*

Proof. This follows as the probability measure is continuous. □

While these are the some nicely behaving random variables, often times, random variables appears to be neither discrete or continuous. An example of this is to consider the random variable X representing the units of beer an individual within the population had consumed today.

2.3 Transformations of Random Variables

Often times, we might want to work with transformed random variables. This can be done in many ways, but the most obvious way is to work with the transformed random variable straight away. While this can work in simple cases, we might find it is normally easier to work with the transformed CDF instead. But before we can discuss the consequences of transforming random variables, we should first consider when is a transformed random variable still a random variable.

Recall, by definition, a random variable is a measurable function from the measurable set Ω to some other measurable set, most often the reals. So, for a transformed random variable to also be a random variable, we require it to be measurable as well.

Theorem 7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a real random variable. Then, for all $g : \mathbb{R} \to \mathbb{R}$ where g is measurable with respect to $\mathcal{B}(\mathbb{R})$, $g(X) := g \circ X$ is a random variable.*

Proof. Follows directly from definitions. □

It is in general very hard to construct a non-$\mathcal{B}(\mathbb{R})$-measurable function (but one example of this is the indicator function of the Vitali set), we can regard most transformations of random variables to also be a random variable[2].

Working with transformed random variables is very simple. Say X is a real random variable and g is $\mathcal{B}(\mathbb{R})$-measurable. Then to get the CDF of $g(X)$ (recall that the CDF characterises the random variable) we simply consider $F_{g(X)}(x) = \mathbb{P}(g(X) \in (-\infty, x]) = \mathbb{P}(X \in g^{-1}(-\infty, x])$ which we can obtain using the CDF of X.

[2]For one, continuity implies $\mathcal{B}(\mathbb{R})$-measurable.

3 Multivariate Random Variables

Recall the definition regarding multivariate distributions from year one and we shall in this section consider some of their properties.

Theorem 8. *Let $X_1, \cdots, X_n$ be independent random variables and $f_1, \cdots, f_n$ are Borel measurable real-valued functions, then $f_1(X_1), \cdots, f_n(X_n)$ are also independent.*

Proof. Suppose $B_1, \cdots, B_n \in \mathcal{B}(\mathbb{R})$, then we need to show that $\mathbb{P}(\bigcap_{i=1}^n \{f_i(X_i) \in B_i\}) = \prod_{i=1}^n \mathbb{P}(f_i(X_i) \in B_i)$. By considering by definition $\{f_i(X_i) \in B_i\} = f_i(X_i)^{-1}(B_i) = X_i^{-1}(f_i^{-1}(B_i))$, we have $\mathbb{P}(\bigcap_{i=1}^n \{f_i(X_i) \in B_i\}) = \mathbb{P}(\bigcap_{i=1}^n X_i^{-1}(f_i^{-1}(B_i))) = \prod_{i=1}^n \mathbb{P}(X_i^{-1}(f_i^{-1}(B_i))) = \prod_{i=1}^n \mathbb{P}(f_i(X_i) \in B_i)$ where the third equality is due to the independence of X_i. $\square$

3.1 Covariance and Correlation

While this is a nice theorem on independence of transformed random variables, it is also useful to develop some tools to help us determine whether or not two random variables are independent. Recall the definition of covariance.

Definition 3.1 (Covariance)**.** For random variables X, Y, with finite expectations μ_X, μ_Y respectively, the covariance between X and Y is defined to be

$$\mathrm{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)).$$

By expanding, we see that the covariance between X and Y is equivalently

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

Furthermore, we see that the covariance is zero for independent random variables, however, the reverse is not necessarily true. Indeed, not much can be interpreted from this value as $\mathrm{Cov}(X, Y)$ has the same dimension as XY, thus, it does not make sense the refer to the covariance as big or small, this is instead the role of the correlation.

Definition 3.2 (Correlation)**.** The correlation of the random variables X, Y is

$$\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

We recall from last year that the correlation is always between -1 and 1, so it does make sense to consider the size of the correlation.

From analysis, we recall the definition of a inner product space – that is a vector space equipped with an inner product. By checking the axioms, we find that the covariance of random variables forms an inner product over the space of random variables.

Remark. *The previous statement is not necessarily true since* $\mathrm{Cov}(X, X) = 0$ *for* $X = c$ *for some* $c \in \mathbb{R}$*; that is the covariance does not satisfy positive definiteness. To fix this, we quotient on the set of random variables with the equivalence relation* $X \sim Y$ *if and only if there exists* $c \in \mathbb{R}$, $\mathbb{P}(X = Y + c) = 1$.

3.2 Transformation of Multivariate Random Variables

Let $D \subseteq \mathbb{R}^2$ and $T : D \to \mathbb{R}^2$ be a function with range $R \subseteq \mathbb{R}^2$. Suppose the partial derivatives of T exist and are continuous. We define the Jacobian of T is

$$J(u,v) = \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}.$$

Then, if $(U,V) = T(X,Y)$is a function of the pair of random variables (X,Y) with joint probability density function f_{XY}, the joint pdf of (U,V) is

$$f_{UV}(u,v) = f_{XY}(x(u,v), y(u,v)) \left| J(u,v) \right|.$$