8: Random Vectors / Multivariate Distributions

PSTAT 120A: Summer 2022

Ethan P. Marzban

July 12, 2022

University of California, Santa Barbara

Where We've Been

- Axioms of Probability, Probability Spaces, Counting
- Conditional Probabilities, independence, etc.
- Basics of Random Variables (classification, p.m.f., c.m.f., moments)
- Discrete Distributions
- Continuous Distributions
- Transformations of Random Variables
- Double Integrals

Random Vectors

- Consider the following experiment: suppose I pick a point P at random from the interior of the unit disk, and I let X denote the $x-$coordinate and Y denote the $y-$coordinate.
- We could investigate X and Y separately, but we have this intuitive sense that these two random variables are in some way related.
- We will quantify this relationship in an upcoming lecture. For now, we will simply say: "let's try and consider X and Y *together*, as a pair (X, Y)."
- We can imagine generalizing this to not just two random variables, but a collection of n random variables! $(X_1, X_2, \cdots, X_n)$.
- Now, remember how I said a random variable X maps from Ω to $\mathbb{R}$? Well, clearly when we start to imagine pairs (or tuples) of random variables we no longer have a map from Ω to $\mathbb{R}$.
- Specifically, let's consider that "picking a point" example; Ω is simply the unit disk $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$. Additionally, this pair (X, Y) takes an element in **Ω and spits out a pair of numbers** (namely, the $x-$ and $y-$coordinates of the point, respectively). In other words,

$$(X, Y) : \Omega \to \mathbb{R}^2$$

- For this reason, we often refer to the pair (X, Y) as a **random vector** as opposed to a random variable. (Another terminology is to call them a **pair of bivariate random variables**, but this language does not generalize as nicely to more than 2

Random Vectors

- Let's start making some of this more formal.

Definition: Random Vector

Given a probabilty space $(\Omega, \mathcal{F}, \mathbb{P})$, a **random vector**

$$\vec{\boldsymbol{X}} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

is a mapping $\vec{\boldsymbol{X}} : \Omega \to \mathbb{R}^n$. We say that the **dimension** of $\vec{\boldsymbol{X}}$ is $\boldsymbol{n}$, or that $\vec{\boldsymbol{X}}$ is an $\boldsymbol{n}$**-dimensional** random vector.

- Though it is customary to write vectors in column format, often times we are lazy and simply write them as row vectors:

$$\vec{\boldsymbol{X}} = (X_1, X_2, \cdots, X_n)$$

Random Vectors

- Remember how we constructed continuous random variables? Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a [continuous] random variable $X : \Omega \to \mathbb{R}$, we argued that depending on our choice of $\mathbb{P}$ we can construct a c.d.f. $F_X(x) := \mathbb{P}(X \leq x)$, which, provided we have differentiability, gave rise to a p.d.f. that we can use to find probabilities, expectations, etc.
- We can do something similar for random vectors. We start with the notion of a:

Definition: Joint Cumulative Distribution Function

Given an $n-$dimensional random vector $\vec{\boldsymbol{X}} = (X_1, X_2, \cdots, X_n)$ we define the **joint cumulative distribution function** (or **joint c.d.f.**, for short) to be

$$F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) := \mathbb{P}(X_1 \leq x_1,\ X_2 \leq x_2,\ \cdots,\ X_n \leq x_n)$$

- Under appropriate conditions, we have the following:

Theorem

Under certain conditions (conditions over which we won't concern ourselves for the purposes of this class), we have the existence of a function $f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n)$ such that

$$F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f_{X_1, X_2, \cdots, X_n}(t_1, t_2, \cdots, t_n)\, \mathrm{d}t_1\, \mathrm{d}t_2\, \cdots\, \mathrm{d}t_n$$

Such a function is called a **joint probability density function** (a.k.a. **joint p.d.f**, or just **joint density**).

Theorem

A joint density function must satisfy the following two conditions:

(1) $f_{X_1,\cdots,X_n}(x_1,\cdots,x_n) \geq 0$ for all $(x_1,\cdots,x_n) \in \mathbb{R}^n$

(2) $\displaystyle\int \cdots \int_{\mathbb{R}^n} f_{X_1,\cdots,X_n}(x_1,\cdots,x_n)\, \mathrm{d}x_1 \ \cdots \ \mathrm{d}x_n = 1$

This also works in the other direction; that is, if we have a function $f_{X_1,\cdots,X_n}(x_1,\cdots,x_n)$ that satisfies the above two conditions then it is the joint density of some random vector $\vec{\boldsymbol{X}}$.

- The relationship between joint c.d.f's and joint p.d.f.'s is

$$f_{X_1,X_2,\cdots,X_n}(x_1,x_2,\cdots,x_n) = \frac{\partial^n}{\partial x_1 \ \partial x_2 \ \cdots \ \partial x_n} F_{X_1,X_2,\cdots,X_n}(x_1,x_2,\cdots,x_n)$$

Random Vectors

- This is perhaps a good time to introduce some simplifying notation.
- When dealing with random vectors in generality, we often will need to write $n-$dimensional integrals.
- I shall adopt the following notation, which I borrow from Physics:

$$\int_{\mathbb{R}^n} f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})\ \mathrm{d}\vec{\boldsymbol{x}}$$

  shall mean

$$\int \cdots \int_{\mathbb{R}^n} f_{X_1,\cdots,X_n}(x_1,\cdots,x_n)\ \mathrm{d}x_1\ \cdots\ \mathrm{d}x_n$$

- So, for instance, the second condition above can be written as $\int_{\mathbb{R}^n} f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})\ \mathrm{d}\vec{\boldsymbol{x}} = 1$.
- By the way: in the subscript I'm using a capital X ($\vec{\boldsymbol{X}}$) and in the argument I'm using a lowercase x ($\vec{\boldsymbol{x}}$).

Random Vectors

- Okay, I admit that dealing with random vectors in generality can get a bit pesky.
- When you start talking about "sampling" in 120B, you'll see why random vectors arise **extremely** often throughout statistics. (Loosely speaking: Statisticians like to collect a *lot* of data, which can be modeled nicely using random vectors; a random variable for each observation!)
- For the purposes of this class, we will primarily restrict our considerations to $n = 2$, which gives rise to so-called **bivariate** random variables and distributions. But let's quickly run through some generalities first:

Multivariate distributions

Multivariate Distributions

- Much like we had distributions in the case of random variables, we also have distributions in the case of random vectors. These distributions are often referred to as **multivariate** distributions.
- Unlike with univariate distributions, however, there aren't a whole lot that have specific names associated with them.
- There are two exceptions; we will discuss one of them in a bit, and time permitting we will discuss the second a little later.

- Let's return to our "picking a point" example. More generally, we could consider the following situation: from a region $\mathcal{R}$ in $\mathbb{R}^n$, pick a point at random.
- Associated with this experiment, we could utilize the following choice of probability measure:

$$\mathbb{P}(A) = \frac{\text{volume}(A)}{\text{volume}(\Omega)}$$

In the case of $n = 2$, this is equivalently written as

$$\mathbb{P}(A) = \frac{\text{area}(A)}{\text{area}(\Omega)}$$

Multivariate Distributions

- Letting $\vec{\boldsymbol{X}} = (X_1, \cdots, X_n)$ denote the coordinates of the selected points, one can find (through a similar argument we used to derive the p.d.f. of the Unif$[a, b]$ distribution) that the joint density of $\vec{\boldsymbol{X}}$ is

$$f_{\vec{\boldsymbol{X}}}(\vec{x}) = \frac{1}{\text{area}(\Omega)} \cdot \mathbb{1}_{\{\vec{x} \in \Omega\}} \tag{1}$$

- So, for instance, in our "picking a point from the unit disc" problem the joint density of (X, Y) is

$$f_{X,Y}(x, y) = \frac{1}{\pi} \cdot \mathbb{1}_{\{(x,y): x^2+y^2 \leq 1\}} = \begin{cases} \frac{1}{\pi} & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

  - You can check that this is in fact a valid joint probability density function!

- This distribution (i.e. the one with p.d.f. listed in equation (1) above) doesn't have a standard name, but I will often refer to this as a **multivariate uniform** distribution, due to its similarity to our familiar Unif$[a, b]$ distribution (note that an interval $[a, b]$ is nothing but a "region" in $\mathbb{R}^1$!)

Bivariate Random Variables/Distributions

Bivariate Random Variables

- Given a pair of random variables (X, Y), we have the notion of a bivariate density function: a function $f_{X,Y}(x, y)$ that is nonnegative over $\mathbb{R}^2$ and also integrates to unity (when integrated over $\mathbb{R}^2$.
- With such a function, we find that a great many of our familiar functions have nice bivariate analogs: for example, the LOTUS becomes

$$\mathbb{E}[g(X, Y)] = \iint_{\mathbb{R}^2} g(x, y) \cdot f_{X,Y}(x, y) \, \mathrm{d}A$$

- Additionally, just like we found probabilities in the univariate case by integrating the density, we get probabilities in the bivariate case by integrating the bivariate density:

$$\mathbb{P}((X, Y) \in \mathcal{R}) = \iint_{\mathcal{R}} f_{X,Y}(x, y) \, \mathrm{d}A$$

- Maybe now you see why we did that whole double integral review...
- One new piece of terminology: the region over which a joint density is nonzero is called the **support** of the random vector. It will almost always be a good idea to sketch the support of a random vector!

Marginals

- One more piece of terminology that is unique to random vectors is that of the **marginal** density/distribution:

Definition: Marginals

Given a random vector $\vec{\boldsymbol{X}} = (X_1, \cdots, X_n)$ with joint p.d.f. $f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})$, the **marginal density of $\boldsymbol{X_i}$** is given by integrating out all other random variables from the joint density.

In the Bivariate case, for instance,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, \mathrm{d}y$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, \mathrm{d}x$$

- Note that, since the joint density is often only nonzero over a portion of $\mathbb{R}^2$, the limits of the integrals above likely involve variables.

Joints

- Given higher-dimensional random vectors, we can get more and more quantities by integrating out various random variables.
- For instance, given a random vector (X, Y, Z) with joint p.d.f. $f_{X,Y,Z}(x, y, z)$, in addition to the marginal densities of X, Y, and Z we can also get various joint densities as well:

$$f_{X,Y}(x, y) = \iint_{\mathbb{R}^2} f_{X,Y,Z}(x, y, z)\ \mathrm{d}z$$

$$f_{X,Z}(x, z) = \iint_{\mathbb{R}^2} f_{X,Y,Z}(x, y, z)\ \mathrm{d}y$$

$$f_{Y,Z}(y, z) = \iint_{\mathbb{R}^2} f_{X,Y,Z}(x, y, z)\ \mathrm{d}x$$

Example

Suppose (X, Y) is a pair of random variables with joint density given by

$$f_{X,Y}(x,y) = \begin{cases} c \cdot e^{-(x+y)} & \text{if } x \leq y < \infty,\ 0 \leq x < \infty \\ 0 & \text{otherwise} \end{cases}$$

where $c > 0$ is an as-of-yet undetermined constant.

(a) Find the value of c that ensures $f_{X,Y}(x, y)$ is a valid joint p.d.f..

(b) Compute $\mathbb{P}(X \geq 0.5,\ Y \geq 0.5)$

(c) Compute $\mathbb{E}[XY]$

(d) Find $f_X(x)$, the marginal density of X.

Discrete?

- So far we've dealt only with continuous random vectors. What about discrete ones?
- Well, the primary difference is that instead of a joint p.d.f. we have a (perhaps more easily intuitable) joint probability *mass* function

$$p_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = \mathbb{P}(X_1 = x_1,\ \cdots,\ X_n = x_n)$$

  that obeys:

  (1) $0 \leq p_{X_1,\cdots,X_n}(x_1,\cdots,x_n) \leq 1$ for all $\vec{\boldsymbol{x}} \in \mathbb{R}^n$
  (2) $\sum_{\mathbb{R}^n} p_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = 1$

- Familiar analogies apply:

$$\mathbb{P}(\vec{\boldsymbol{X}} \in A) = \sum_{\vec{\boldsymbol{x}} \in A} p_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})$$

  and the LOTUS becomes

$$\mathbb{E}[g(\vec{\boldsymbol{X}})] = \sum_{\mathbb{R}^n} g(\vec{\boldsymbol{x}}) \cdot p_{\vec{\boldsymbol{x}}}(\vec{\boldsymbol{x}})$$

  [note that both summations above are really $n-$summations; that is, they are n sums in one]

Example

Let (X, Y) be a pair of bivariate discrete random variables with joint p.m.f.

$$p_{X,Y}(x, y) = \begin{cases} c \cdot xy & \text{if } x \in \{1, 2, 3, 4\},\ y \in \{1, 2, 3\} \\ 0 & \text{otherwise} \end{cases}$$

where $c > 0$ is an as-of-yet undetermined constant.

(a) Find the value of c

(b) Compute $\mathbb{E}[XY]$

A Useful Result

Theorem: Linearity of Expectation

Given a collection of random variables $X_1, \cdots, X_n$ and a collection of constants $a_1, \cdots, a_n$, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

Proof.

- We use the multidimensional LOTUS:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] &= \int_{\mathbb{R}^n} \left(\sum_{i=1}^{n} a_i x_i\right) f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})\ \mathrm{d}x \\
&= \int_{\mathbb{R}^n} \sum_{i=1}^{n} (a_i x_i f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}}))\ \mathrm{d}\vec{\boldsymbol{x}} \\
&= \sum_{i=1}^{n} \left[\int_{\mathbb{R}^n} a_i x_i f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}}) \mathrm{d}\vec{\boldsymbol{x}}\right] \\
&= \sum_{i=1}^{n} \mathbb{E}[a_i X_i] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]
\end{aligned}$$

A Useful Result

- If the vector notation on the previous slide is too confusing, you can think of things in terms of $n = 2$; the proof for general n follows analogously.

$$\begin{aligned}
\mathbb{E}[a_1 X_1 + a_2 X_2] &= \iint\limits_{\mathbb{R}^2} (a_1 x_1 + a_2 x_2) f_{X_1,X_2}(x_1, x_2) \, \mathrm{d}A \\
&= \iint\limits_{\mathbb{R}^2} \left[a_1 x_1 f_{X_1,X_2}(x_1, x_2) + a_2 x_2 a_1 x_1 f_{X_1,X_2}(x_1, x_2) \right] \, \mathrm{d}A \\
&= \iint\limits_{\mathbb{R}^2} a_1 x_1 f_{X_1,X_2}(x_1, x_2) \, \mathrm{d}A + \iint\limits_{\mathbb{R}^2} a_2 x_2 f_{X_1,X_2}(x_1, x_2) \, \mathrm{d}A \\
&= \mathbb{E}[a_1 X_1] + \mathbb{E}[a_2 X_2] = a_1 \mathbb{E}[X_1] + a_2 \mathbb{E}[X_2]
\end{aligned}$$