Introduction to Data Science - 1MS041

Benny Avelin
Department of Mathematics

HT 2023

Recall from last time

- A random variable is a function from the sample space to a number (or vector).
- A distribution function is $F(x) = \mathbb{P}(X \leq x)$.
- A discrete random variable takes discrete values, i.e. $0, 1, 2, 3, \ldots$. The probability mass function is defined as $f(x) = \mathbb{P}(X = x)$.
- A random variable is called continuous if the distribution function F can be written as

$$F(x) = \int_{-\infty}^{x} f(v)dv$$

for a piecewise continuous function f. f is called the density function.

Compare and contrast

Discrete	Continuous
$F(x) = \sum_{x_i \leq x} f(x_i)$	$F(x) = \int_{-\infty}^{x} f(v)dv$
$F(b) - F(a) = \sum_{a < x_i \leq b} f(x_i)$	$F(b) - F(a) = \int_a^b f(x)dx$
$\mathbb{P}(X = x) = f(x)$	$\mathbb{P}(X = x) = 0$
$\sum_x f(x) = 1$	$\int_{-\infty}^{\infty} f(x)dx = 1.$

Joint Distribution Function

Definition (JDF)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be an $\mathbb{R}^m$ valued RV. Then the **joint distribution function (JDF)** or **joint cumulative distribution function (JCDF)**, $F_X(x) : \mathbb{R}^m \to [0, 1]$ is defined as

$$\begin{aligned} F_X(x) &= \mathbb{P}(\cap_{i=1}^{m}(X_i \leq x_i)) \ = \ \mathbb{P}(X_1 \leq x_1, \ldots, X_m \leq x_m) \\ &= \mathbb{P}\left(\{\omega : X_1(\omega) \leq x_1, \ldots, X_m \leq x_m\}\right), \end{aligned}$$

where $X = (X_1, \ldots, X_m)$ and each $X_i \in \mathbb{R}$, and $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$.

See example in notebook.

Marginal

Consider a JDF of two random variables X, Y, $F_{X,Y}$. The marginal distribution for X is defined as

$$F_X(x) := F_{X,Y}((x, \infty)) = \mathbb{P}(X \leq x, Y \leq \infty) = \mathbb{P}(X \leq x).$$

Simply put

The marginal distribution for X is what we get when ignoring the value of Y.

Recall

We say that two events A and B are **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

or equivalently

$$\mathbb{P}(A \mid B) = \mathbb{P}(A).$$

Independence

Definition (Independence of Two RVs)

Consider an $\mathbb{R}^2$-valued RV $X := (X_1, X_2)$. Then the $\mathbb{R}$-valued RVs X_1 and X_2 are said to be independent or independently distributed if and only if

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) = \mathbb{P}(X_1 \leq x_1)\mathbb{P}(X_2 \leq x_2)$$

or equivalently,

$$F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2),$$

for any pair of real numbers $(x_1, x_2) \in \mathbb{R}^2$.

Independence

Consider two random variables X_1, X_2 and let $X = (X_1, X_2)$ be a random variable.

Discrete	Continuous
Joint Distribution Function	Joint Distribution Function
Joint Probability Mass Function	Joint Probability Density Function
$F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$	$F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$
$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$	$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$

Sequences

Definition

- A sequence of random variables $X_1, \ldots, X_n$ is nothing but a random vector $X = (X_1, \ldots, X_n)$.

Sequences

Definition

- A sequence of random variables $X_1, \ldots, X_n$ is nothing but a random vector $X = (X_1, \ldots, X_n)$.
- We say that the sequence is **independent** if all the components are mutually independent, i.e.
  $F(X \leq x) = F_1(X_1 \leq x_1)F_2(X_2 \leq x_2) \cdots F_n(X_n \leq x_n)$.

Sequences

Definition

- A sequence of random variables $X_1, \ldots, X_n$ is nothing but a random vector $X = (X_1, \ldots, X_n)$.
- We say that the sequence is **independent** if all the components are mutually independent, i.e.
  $F(X \leq x) = F_1(X_1 \leq x_1)F_2(X_2 \leq x_2) \cdots F_n(X_n \leq x_n)$.
- We say that the sequence is **identically distributed** if all the marginal distributions are the same, i.e. $F_{X_i} = F_{X_j}$ for all pairs.

Sequences

Definition

- A sequence of random variables $X_1, \ldots, X_n$ is nothing but a random vector $X = (X_1, \ldots, X_n)$.
- We say that the sequence is **independent** if all the components are mutually independent, i.e.
  $F(X \leq x) = F_1(X_1 \leq x_1)F_2(X_2 \leq x_2) \cdots F_n(X_n \leq x_n)$.
- We say that the sequence is **identically distributed** if all the marginal distributions are the same, i.e. $F_{X_i} = F_{X_j}$ for all pairs.
- We say that the sequence is **independent and identically distributed** i.i.d. (or IID) if they are both independent and identically distributed.

Conditional

Definition

Let (X, Y) be a $\mathbb{R}^2$ valued random variable, and let $A \subset \mathbb{R}$ be a Borel ("think interval") set such that $\mathbb{P}(Y \in A) > 0$ then define the conditional distribution function of X given that $Y \in A$ as

$$F_{X|Y}(x \mid A) := \frac{\mathbb{P}(X \leq x, Y \in A)}{\mathbb{P}(Y \in A)}.$$

Conditional

Definition

Let (X, Y) be a $\mathbb{R}^2$ valued random variable, and let $A \subset \mathbb{R}$ be a Borel ("think interval") set such that $\mathbb{P}(Y \in A) > 0$ then define the conditional distribution function of X given that $Y \in A$ as

$$F_{X|Y}(x \mid A) := \frac{\mathbb{P}(X \leq x, Y \in A)}{\mathbb{P}(Y \in A)}.$$

If Y is a discrete random variable and $f_Y(y) > 0$ the above definition is well defined for $A = \{y\}$ and we can write

$$F_{X|Y}(x \mid y) := \frac{\mathbb{P}(X \leq x, Y = y)}{\mathbb{P}(Y = y)}.$$

Conditional

Definition (Conditional PDF or PMF)

Let (X, Y) be a $\mathbb{R}^2$ valued RV. Then the **conditional probability mass / density function** is defined as

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

if $f_Y(y) > 0$

Conditional

Definition (Conditional PDF or PMF)

Let (X, Y) be a $\mathbb{R}^2$ valued RV. Then the **conditional probability mass / density function** is defined as

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

if $f_Y(y) > 0$

Lemma

Let (X, Y) be a $\mathbb{R}^2$ valued RV. Then

$$f_{X|Y}(x \mid y) f_Y(y) = f_{X,Y}(x,y)$$

where the left hand side is interpreted as 0 if $f_Y(y) = 0$.

Functions of RVs

Definition (Expectation of a function of a RV)

The **Expectation** of a function $g(X)$ of a random variable X is defined as:

$$\mathbb{E}(g(X)) := \int g(x)dF(x) = \begin{cases} \sum_x g(x)f(x) & \text{discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{continuous} \end{cases}$$

provided $\mathbb{E}(g(X))$ exists, i.e., $\int |g(x)|dF(x) < \infty$.

Moments

The simplest form of functions of interest is powers. The common statistical quantities that we often look at is the so called central moments

$$p\text{:th central moment: } \mathbb{E}[(X - \mathbb{E}[X])^p]$$

where $p = 1, 2, 3, \ldots$. There is also the p:th standarized moment

$$p\text{:th standarized moment: } \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^p\right]$$

Moments

The simplest form of functions of interest is powers. The common statistical quantities that we often look at is the so called central moments

$$p\text{:th central moment: } \mathbb{E}[(X - \mathbb{E}[X])^p]$$

where $p = 1, 2, 3, \ldots$. There is also the p:th standarized moment

$$p\text{:th standarized moment: } \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^p\right]$$

- $p = 2$ central moment: Variance (Measures spread)
- $p = 3$ standardized moment: Skewness (Measures lopsidedness)
- $p = 4$ standardized moment: Kurtosis (Measure heavy taildness)

Sample versions

For each of the moments we can use data to try to estimate it, lets consider the sample variance of $X_1, \dots, X_n$ an i.i.d. sequence of random variables

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

L^p, not all random variables are created equal

- Not all random variables have a finite variance. Ex. the Pareto distribution
- Not all random variables have a finite expectation. Ex. The Cauchy distribution.
- If the p:th moment of a random variable X exists we will say that $X \in L^p(\mathbb{P})$. (We will mostly be working with $p = 1$ or $p = 2$).

Properties of the expectation

1. If $X \in L^1(\mathbb{P})$ is an $\mathbb{R}$ valued RV and $\alpha \in \mathbb{R}$, then

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$$

Properties of the expectation

1. If $X \in L^1(\mathbb{P})$ is an $\mathbb{R}$ valued RV and $\alpha \in \mathbb{R}$, then

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$$

2. If $X, Y \in L^1(\mathbb{P})$ are $\mathbb{R}$ valued RV, then

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

Properties of the expectation

1. If $X \in L^1(\mathbb{P})$ is an $\mathbb{R}$ valued RV and $\alpha \in \mathbb{R}$, then

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$$

2. If $X, Y \in L^1(\mathbb{P})$ are $\mathbb{R}$ valued RV, then

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

3. If $X, Y \in L^2(\mathbb{P})$ are independent $\mathbb{R}$ valued RV, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Properties of the expectation

1. If $X \in L^1(\mathbb{P})$ is an $\mathbb{R}$ valued RV and $\alpha \in \mathbb{R}$, then
$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$$

2. If $X, Y \in L^1(\mathbb{P})$ are $\mathbb{R}$ valued RV, then
$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

3. If $X, Y \in L^2(\mathbb{P})$ are independent $\mathbb{R}$ valued RV, then
$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

4. $A \subset \mathbb{R}$
$$\mathbb{E}[\mathbb{1}_A(X)] = \mathbb{P}(X \in A)$$

Conditional expectation

We can also construct conditional expectations, like

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx.$$

Conditional expectation

We can also construct conditional expectations, like

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx.$$

But it also allows for the following very useful property

Theorem (The tower property)

Let (X, Y) be a $\mathbb{R}^2$ valued RV. Then

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

We introduced a new notation, namely $\mathbb{E}[X \mid Y]$, what is this?
Denote $g(y) = \mathbb{E}[X \mid Y = y]$, then define

$$\mathbb{E}[X \mid Y] := g(Y).$$

Note

By our definition, $\mathbb{E}[X \mid Y]$ is now another random variable. I.e. we have averaged out everything w.r.t. X but Y is still not averaged over.

Transformations

Consider a random variable X with distribution function F. Also consider another function $g : \mathbb{R} \to \mathbb{R}$, what is the distribution function of $g(X)$?

Transformations

Consider a random variable X with distribution function F. Also consider another function $g : \mathbb{R} \to \mathbb{R}$, what is the distribution function of $g(X)$? To answer this question we must first observe that the inverse image $g^{[-1]}$ satisfies the following properties:

- $g^{[-1]}(\mathbb{Y}) = \mathbb{X}$

Transformations

Consider a random variable X with distribution function F. Also consider another function $g : \mathbb{R} \to \mathbb{R}$, what is the distribution function of $g(X)$? To answer this question we must first observe that the inverse image $g^{[-1]}$ satisfies the following properties:

- $g^{[-1]}(\mathbb{Y}) = \mathbb{X}$
- For any set A, $g^{[-1]}(A^c) = \left(g^{[-1]}(A)\right)^c$

Transformations

Consider a random variable X with distribution function F. Also consider another function $g : \mathbb{R} \to \mathbb{R}$, what is the distribution function of $g(X)$? To answer this question we must first observe that the inverse image $g^{[-1]}$ satisfies the following properties:

- $g^{[-1]}(\mathbb{Y}) = \mathbb{X}$
- For any set A, $g^{[-1]}(A^c) = \left(g^{[-1]}(A)\right)^c$
- For any collection of sets $\{A_1, A_2, \ldots\}$,

$$g^{[-1]}\left(A_1 \cup A_2 \cup \cdots\right) = g^{[-1]}(A_1) \cup g^{[-1]}(A_2) \cup \cdots \quad .$$

Consequentially,

$$\mathbb{P}_g(A) = P\left(g(X) \in A\right) = P\left(X \in g^{[-1]}(A)\right) \tag{1}$$

For a discrete random variable X with probability mass function f_X we can obtain the probability mass function f_Y of $Y = g(X)$ as follows:

$$
\begin{aligned}
f_Y(y) &= \mathbb{P}(Y = y) = \mathbb{P}(Y \in \{y\}) \\
&= P\left(g(X) \in \{y\}\right) = P\left(X \in g^{[-1]}(\{y\})\right) \\
&= P\left(X \in g^{[-1]}(y)\right) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ .
\end{aligned}
$$

For a discrete random variable X with probability mass function f_X we can obtain the probability mass function f_Y of $Y = g(X)$ as follows:

$$\begin{aligned} f_Y(y) &= \mathbb{P}(Y = y) = \mathbb{P}(Y \in \{y\}) \\ &= P\left(g(X) \in \{y\}\right) = P\left(X \in g^{[-1]}(\{y\})\right) \\ &= P\left(X \in g^{[-1]}(y)\right) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ . \end{aligned}$$

This gives the formula:

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ . \tag{2}$$

Transformations of continuous random variables

In the continuous context, we have to really care about a function not being $1-1$ for instance. For more information, see the lecture notes.