Introduction to Data Science - 1MS041

Benny Avelin

Department of Mathematics

HT 2023

Recall from last time

- Experiment: is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**.
- The set of all outcomes is called the **sample space**, and is denoted by Ω.
- Trial: doing the experiment once and getting an outcome.
- The subsets of Ω are called **events** events.
- Given an outcome $\omega \in \Omega$ we say that the event $E \subset \Omega$ **occured** if $\omega \in E$.

Random variables

- When working with experiments with outcomes, they can be anything. Like strings or just representations of whatever.

Random variables

- When working with experiments with outcomes, they can be anything. Like strings or just representations of whatever.
- When working with data, we would like to say things like "What is the average (expected) outcome?"

Random variables

- When working with experiments with outcomes, they can be anything. Like strings or just representations of whatever.
- When working with data, we would like to say things like "What is the average (expected) outcome?"
- To do this, we should really be working with numbers.

Random variables

- When working with experiments with outcomes, they can be anything. Like strings or just representations of whatever.
- When working with data, we would like to say things like "What is the average (expected) outcome?"
- To do this, we should really be working with numbers.
- Recall: When we simulated the coin toss, we assigned 1 to Heads and 0 to Tails, this allowed us to take the average!

Random variables

Definition (Random Variable)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability triple. Then, a **Random Variable (RV)**, say X, is a function from the sample space Ω to the set of real numbers $\mathbb{R}$

$$X : \Omega \to \mathbb{R}$$

Random variables

Definition (Random Variable)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability triple. Then, a **Random Variable (RV)**, say X, is a function from the sample space Ω to the set of real numbers $\mathbb{R}$

$$X : \Omega \to \mathbb{R}$$

such that for every $x \in \mathbb{R}$, the inverse image of the half-open real interval $(-\infty, x]$ is an element of the collection of events $\mathcal{F}$, i.e.:

Random variables

Definition (Random Variable)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability triple. Then, a **Random Variable (RV)**, say X, is a function from the sample space Ω to the set of real numbers $\mathbb{R}$

$$X : \Omega \to \mathbb{R}$$

such that for every $x \in \mathbb{R}$, the inverse image of the half-open real interval $(-\infty, x]$ is an element of the collection of events $\mathcal{F}$, i.e.:

$$\text{for every } x \in \mathbb{R}, \qquad X^{[-1]}(\ (-\infty, x]\ ) := \{\omega : X(\omega) \le x\} \in \mathcal{F}\ .$$

Random variables

Definition (Random Variable)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability triple. Then, a **Random Variable (RV)**, say X, is a function from the sample space Ω to the set of real numbers $\mathbb{R}$

$$X : \Omega \rightarrow \mathbb{R}$$

such that for every $x \in \mathbb{R}$, the inverse image of the half-open real interval $(-\infty, x]$ is an element of the collection of events $\mathcal{F}$, i.e.:

$$\text{for every } x \in \mathbb{R}, \qquad X^{[-1]}(\ (-\infty, x]\ ) := \{\omega : X(\omega) \leq x\} \in \mathcal{F}\ .$$

We assign probability to the RV X as follows:

$$\mathbb{P}(X \leq x) = \mathbb{P}(\ X^{[-1]}(\ (-\infty, x]\ )\ ) := \mathbb{P}(\ \{\omega : X(\omega) \leq x\}\ )\ . \tag{1}$$

Lets unpack!

A random variable is a function from the sample space to a value!

Lets unpack!

A random variable is a function from the sample space to a value!
Consider the coin-toss: We had $\Omega = \{\mathtt{H}, \mathtt{T}\}$, and define

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = \mathtt{H} \\ 0 & \text{if } \omega = \mathtt{T} \end{cases}$$

Then as we observe $\mathtt{H}, \mathtt{T}, \mathtt{H}, \mathtt{H}, \mathtt{T}$, we observe for X, $1, 0, 1, 1, 0$.

More unpacking!

For every $x \in \mathbb{R}$, the inverse image of the half-open real interval $(-\infty, x]$ is an element of the collection of events $\mathcal{F}$.

More unpacking!

For every $x \in \mathbb{R}$, the inverse image of the half-open real interval $(-\infty, x]$ is an element of the collection of events $\mathcal{F}$. To understand the above, let us first unpack the inverse image

$$X^{[-1]}(\ (-\infty, x]\ ) := \{\omega : X(\omega) \leq x\} = \text{"}X \text{ is less than or equal to } x\text{"}$$

The inverse image is the event "X is less than or equal to x".

Example

Consider again the coin toss, where $X = 1$ for Heads and 0 for Tails

$$X^{[-1]}(\ (-\infty, 0]\ ) = \{\mathtt{T}\}$$
$$X^{[-1]}(\ (-\infty, 1]\ ) = \{\mathtt{H}, \mathtt{T}\}$$
$$X^{[-1]}(\ (-\infty, 2]\ ) = \{\mathtt{H}, \mathtt{T}\}$$

Last step!

For every $x \in \mathbb{R}$, the inverse image of the half-open real interval $(-\infty, x]$ is an element of the collection of events $\mathcal{F}$.
We now know what the inverse image is, the last requirement is that this is in our $\mathcal{F}$, i.e. our sigma-algebra.

Conclusion

A function that assigns a value to the outcome of the trial is a random variable if we can observe it!

Examples

- If our experiment is that we are checking if a light bulb is defective or not, we had $\Omega = \{\text{Defective}, \text{Non Defective}\}$. We could create a random variable X such that $X(\text{Defective}) = 1$ and $X(\text{Non Defective}) = 0$.

Examples

- If our experiment is that we are checking if a light bulb is defective or not, we had $\Omega = \{\text{Defective}, \text{Non Defective}\}$. We could create a random variable X such that $X(\text{Defective}) = 1$ and $X(\text{Non Defective}) = 0$.
- We could also assign a cost to the lightbulb, if Defective we lose money and if non defective, we can sell it for money. $X(\text{Defective}) = -1$ and $X(\text{Non Defective}) = 2$.

Examples

- If our experiment is that we are checking if a light bulb is defective or not, we had $\Omega = \{\text{Defective}, \text{Non Defective}\}$. We could create a random variable X such that $X(\text{Defective}) = 1$ and $X(\text{Non Defective}) = 0$.
- We could also assign a cost to the lightbulb, if Defective we lose money and if non defective, we can sell it for money. $X(\text{Defective}) = -1$ and $X(\text{Non Defective}) = 2$.
- If the experiment is to select a random person in this classroom. Then the sample space is $\Omega = \{p_1, p_2, \ldots, p_n\}$. We could measure each persons length, then call that $X(p_i)$. Then X is a random variable.

Simplification

- Quite quickly one would get sick of having to all the time figure out what the sample space is and how the events look like.

Simplification

- Quite quickly one would get sick of having to all the time figure out what the sample space is and how the events look like.
- So we usually dont specify Ω exactly, but we just make the assumption that it can be defined if we wanted to. Actually it quickly becomes very complicated.

Simplification

- Quite quickly one would get sick of having to all the time figure out what the sample space is and how the events look like.
- So we usually dont specify Ω exactly, but we just make the assumption that it can be defined if we wanted to. Actually it quickly becomes very complicated.
- Instead we focus our attention on the random variables themselves.

Lets work a bit with random variables: discrete

Definition

We say that a real valued random variable X is discrete if it takes discrete values. For instance $(0, 1, 2, 3, \ldots)$.

Now consider this

Definition

Let X be a $\mathbb{R}$-valued discrete RV. We define the **probability mass function** (PMF) f of X to be the function $f : \mathbb{R} \to [0, 1]$ defined as follows:

$$f(x) := \mathbb{P}(X = x) = \mathbb{P}(\ \{\omega : X(\omega) = x\}\ ) = \begin{cases} \theta_i & \text{if } x = x_i \in \mathbb{X}. \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Distribution function

Definition (Distribution Function)

The **Distribution Function (DF)** or **Cumulative Distribution Function (CDF)** of any RV X, over a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, denoted by F is:

$$F(x) := \mathbb{P}(X \leq x) = \mathbb{P}(\ \{\omega : X(\omega) \leq x\}\ ), \qquad \text{for any} \quad x \in \mathbb{R}\ . \tag{3}$$

Thus, $F(x)$ or simply F is a non-decreasing, right continuous, $[0, 1]$-valued function over $\mathbb{R}$. When a RV X has DF F we write $X \sim F$.

Expectations

In the coin toss experiment H is 1 and T is 0 we said was that in average you would expect to see roughly half 1 and half 0. The average value would tend to 0.5 in a fair coin.

Expectations

In the coin toss experiment H is 1 and T is 0 we said was that in average you would expect to see roughly half 1 and half 0. The average value would tend to 0.5 in a fair coin. Once we have the PMF we can actually compute the theoretical average, called an **expectation or mean**:

$$\mathbb{E}[X] = \sum_x x f(x)$$

Example

In the $X \sim$Bernoulli(p) case we get

$$\mathbb{E}[X] = p + 0(1-p) = p.$$

For a fair coin, $p = 0.5$ we get $\mathbb{E}[X] = 0.5$.

Learning from data

The simplest form of learning is to estimate the mean from data. When working with data, we have a sequence of outcomes $\omega_1, \ldots, \omega_n$ and the values we observe of the random variable X is $X(\omega_1), \ldots, X(\omega_n)$.

Important

It is an important distinction between a random variable X, which is a function (or you can think procedure) and the observation of X which is a value.
It is like the difference between a computer program and the result of running it once. Or simply the difference between an Experiment and a Trial.

Learning from data

The most natural way to estimate the expectation from data is to take the empirical mean. Denote $X(\omega_i) = x_i$, then we can consider

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

this is the so called observed empirical mean.

Lets get mathematical!

Definition

An **n-product experiment** is obtained by repeatedly performing n trials of some experiment.

Lets get mathematical!

Definition

An **n-product experiment** is obtained by repeatedly performing n trials of some experiment.

If we have a random variable X on the original experiment we can list all the n values as $Z = (X_1, X_2, \ldots, X_n)$ where we consider Z as being a single random variable with n values. This is called a multivariate random variable.

Lets get mathematical!

- This allows us to have a way of representing a repeated experiment mathematically, before we do it. In the same way that a single real valued random variable represented a single experiment.

Lets get mathematical!

- This allows us to have a way of representing a repeated experiment mathematically, before we do it. In the same way that a single real valued random variable represented a single experiment.
- The empirical mean is defined as

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Lets get mathematical!

- This allows us to have a way of representing a repeated experiment mathematically, before we do it. In the same way that a single real valued random variable represented a single experiment.
- The empirical mean is defined as

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- WARNING: This is again a random variable, the observed empirical mean is an observation of the empirical mean!!! That is, a single observation of $\overline{X}_n$ is

$$\frac{1}{n}\sum_{i=1}^{n} x_i.$$

Lets get mathematical!

- This allows us to have a way of representing a repeated experiment mathematically, before we do it. In the same way that a single real valued random variable represented a single experiment.
- The empirical mean is defined as

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- WARNING: This is again a random variable, the observed empirical mean is an observation of the empirical mean!!! That is, a single observation of $\overline{X}_n$ is

$$\frac{1}{n} \sum_{i=1}^{n} x_i .$$

- Lets simulate

Continuous random variables

Definition (Continuous random variable)

Let X be a $\mathbb{R}$-valued random variable with distribution function F. We say that X is a **continuous** RV if there exists a piecewise-continuous function $f : \mathbb{R} \to [0, \infty]$, called the **probability density function (PDF)** of X, such that

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(v)\, dv. \tag{4}$$

Compare and contrast

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