School of Mathematics and Statistics

Probability Crash Course: Discrete distributions

Paul Hewson

Overview: This webfile is designed as a revision aid to some introductory concepts in probability. It is intended to supplement a formal encouter with a text book or a set of lectures. These notes are meant to be slightly interactive, mysterious green dots, squares and boxes appear which you can click on to answer questions and check solutions.

Last Revision Date: 7th September 2009

paul.hewson@plymouth.ac.uk
Version 0.1

1. Pre-amble

Definition 1 *If Ω is a* sample space *with a* probability measure *and X is a* real valued function *defined over the elements of S, then X is called a random variable.*

- *(random variable X upper case)*
- *(realisation x lower case)*

Consider tossing a fair coin twice. We have $\Omega = (HH, HT, TH, TT)$. There are four events ω in the sample space. If we define $X(\omega)$ as the number of heads, then we have $X(HH) = 2$, $X(HT) = 1$, $X(TH) = 1$ and $X(TT) = 0$. It is the function $X(\omega)$ that is called a random variable.

We often abbreviate our notation and just refer to the random variable as X, but do note that $X(\omega)$ is a function, not the value of that function at any particular ω. This point can often be ignored, but it is a formality that will be observed by many textbooks.

1.1. Density and distribution

Having carried out some experiment, we know the particular outcome $\omega \in \Omega$, and can assign a value to $X(\omega)$. We wish to assign some measure of probability to this eventually, and can do this either using a density function:

- $f(x)$=probability that X is equal to x

or more usefully through a distribution function:

- $F(x)$=probability that X does not exceed x

Do note the convention $f(x)$ and $F(x)$.

So far our discussion has been in terms of discrete variables. It seems appropriate to refresh ideas about discrete and continuous variables before going further.

2. Variables

Variables can be classified in all kind of ways:

- Quantitative (where the values taken are numerical:
  - Discrete (only particular values are possible, e.g. the integers)
  - Continuous (where any value in a range is possible)
- Qualitative (where the object has some non-numerical quality)
  - Ordinal (where these qualities can be ordered)
  - Categorical (where the qualities are simply of different types, e.g., red bicycle, blue bicycle)

Quiz Variable types

Identify the type of variable in the following examples:

1. Power to a computer

(a) Qualitative (categorical) (b) Qualitative (ordinal)
(c) Quantitative (discrete) (d) Quantitative (continuous)

2. The make of a computer

(a) Qualitative (categorical) (b) Qualitative (ordinal)
(c) Quantitative (discrete) (d) Quantitative (continuous)

3. Number of icons on a desktop computer

(a) Qualitative (categorical) (b) Qualitative (ordinal)
(c) Quantitative (discrete) (d) Quantitative (continuous)

4. Wind speed

(a) Qualitative (categorical) (b) Qualitative (ordinal)
(c) Quantitative (discrete) (d) Quantitative (continuous)

5. Quality of a Hotel (e.g. Michelin 3 star)

(a) Qualitative (categorical) (b) Qualitative (ordinal)
(c) Quantitative (discrete) (d) Quantitative (continuous)

6. Number of bedrooms in a hotel

(a) Qualitative (categorical) (b) Qualitative (ordinal)
(c) Quantitative (discrete) (d) Quantitative (continuous)

7. Time spent travelling to lectures today

(a) Qualitative (categorical) (b) Qualitative (ordinal)
(c) Quantitative (discrete) (d) Quantitative (continuous)

In terms of defining probability measure, we only worry about the two types of *quantitative* variable listed; discrete and continuous. We can only work with qualitative variables for example if we measure them or count them.

- A discrete random variable is a variable that takes a countable (this can be countably finite or countably infinite) set of real numbers with associated probabilities
- A continuous random variable is a variable which takes a continuum of values in the real line according to a rule determined by a density function

Let's concentrate on discrete random variables.

3. Discrete random variables

A frequentist interpretation would suggest that a random variable (usually denoted by a capital letter) is a numerical variable "defined" by the outcome of a random experiment. Even the experiment hasn't happened, we evaluate relative to some sense of a long run of experimental outcomes. We are interested in knowing something about random variables (the outcome) $X = x$ and the probability $p[X = x]$associated with that event.

Definition 2 *A* discrete *random variable is a real valued function defined on the sample space* Ω*. For such discrete random variables* X*, then* $F_X(x)$ *will have a finitely or infinitely countable range,* $\Omega_x = x_1, x_2, \ldots$

We illustrate the difference between countably finite and countably infinite with two examples:

- Suppose that the number of working days in a year is 250. Absences from work are noted in employees records. An experiment consists of randomly drawing a record to see the number of days absent during a year. The random variable X can be defined as the number of days absent, hence $\Omega_X = 0, 1, 2, \ldots, 250$.

- A Geiger counter is connected to a gas tube in order to record the background radiation count for a selected time interval $[0, t)$ (where t is fixed). The random variable X denotes the number of counts in this time period, and the sample space is $\Omega_X = 0, 1, 2, \ldots, \infty$.

X records the event of interest, in either case we could define the event $X = 3$. We wish to associate a probability with this event, that is $p[X = 3]$.

3.1. The density function

- The function given by $f(x) = P(X = x)$ for each x within the range of X is called the *probability density* of X
- A function can serve as the probability density of a discrete random variable x if and only if its values $f(x)$ satisfy
  - (i) $f(x) \geq 0$ for each value within its domain
  - (ii) $\sum_x f(x) = 1$ where the summation extends over all values in its domain

3.2. The distribution function

- If X is a discrete r.v., the function given by

$$F(X) = P(X \leq x) = \sum_{t \leq x} f(t) \text{ for } -\infty < x < \infty$$

where $f(t)$ is the value of the probability distribution of X at t.

- $F(x)$ is called the *distribution function* or *cumulative distribution* of X

- The values $F(x)$ of the discrete function of a discrete r.v. X satisfy the conditions:

  - $F(-\infty) = 0$ and $F(\infty) = 1$
  - If $a < b$ then $F(a) \leq F(b)$ for any real number a and b.

- If the range of a random variable X consists of the values $x_1 < x_2 < x_3 < \cdots < x_n$ then:
  - $f(x_1) = F(x_1)$
  - $f(x_i) = F(x_i) - F(x_{i-1})$ for $i = 2, \ldots, n$

Quiz Density and Distribution

1. A discrete probability density function gives us:

(a) $f(x) = Prob(X = x)$ (b) $f(x) = Prob(X \geq x)$
(c) $f(x) = Prob(X \leq x)$

2. A discrete distribution function gives us:

(a) $F(x) = Prob(X = x)$ (b) $F(x) = Prob(X \geq x)$
(c) $F(x) = Prob(X \leq x)$

3.3. An example

Consider a coin tossing experiment, where $X(H) = 1$ and $X(T) = 0$. The distribution function is given by:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - \pi & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

We will find out later that this is the distribution function called Bernoulli. Bernoulli variables can also be defined by Indicator variables:

$$I_A(\omega) = \begin{cases} 1 & \text{if} \quad \omega \in A \\ 0 & \text{if} \quad \omega \in A^C \end{cases}$$

3.4. Expectation

We can define the *expectation* of a discrete random variable as follows:

Definition 3

$$E[X] = \sum_{x \in \Omega_x} xp[X = x]$$

It's easiest to explain this with an example:

Begin Quiz The manager of a bakery believes that the demand for the number of chocolate cakes he could sell on a single day is a random variable with the probability function $f(x) = \frac{1}{6}$ for $x = 0, 1, 2, 3, 4, 5$. There is a profit of €1.00 for each cake sold, and a loss of €0.40 on each cake unsold. Assume that a cake can only be sold on the day it is made. We are therefore interested in Y, the profit (or loss) on each transaction. What the expected profit $(E[Y])$ under each of the following scenarios:

1. A day on which he bakes 5 cakes (2d.p.)

2. A day on which he bakes 4 cakes (2 d.p.)

3. A day on which he bakes 3 cakes (2 d.p.)

4. Which of the three strategies maximises his expected return?

(a) 5 a day (b) 4 a day (c) 3 a day

End Quiz Correct Points: Score:

3.5. Summary: Expectation of discrete random variables

To conclude, our definition of expectation is given as follows:

Definition 4 *The expected value of a discrete random variable (and the expected value of a* function *of a random variable) as follows:*

- $E(x) = \sum_x xf(x)$
- $E[g(x)] = \sum_x g(x)f(x)$

This is useful in a number of ways, so useful we will look at it all again next week. But for the moment we shall worry about only one particular function $g(x) = x^2$, as this lets us work with the variance.

3.6. Variance

We will re-examine this point later, but the variance of a random variable X is defined as:

Definition 5

$$V[X] = E[(X - E[X])^2]$$

A little bit of algebra also tells us that:

$$\begin{aligned} V[X] &= E[(X - E[X])^2] \\ &= E\left[(X^2) - 2X(E[X]) + (E[X])^2\right] \\ &= E[X^2] - 2(E[X])^2 + (E[X])^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

which can be a useful thing to know. This can be illustrated with an example:

Begin Quiz Consider an experiment in which a coin is tossed twice. Let X be a random variable denoting the number of heads.

1. The density function of X is:

(a)

x	0	1	2
p[X=x]	0.25	0.50	0.25

(b)

x	0	1	2
p[X=x]	0.33	0.33	0.33

2. Find $E[X]$

3. Find $V[X]$

End Quiz Correct Points: Score:

3.7. Some properties of Expectation

Some properties of expectations are given without comment.

$$\begin{aligned} E[cX] &= cE[X] \\ E[X+c] &= E[X]+c \\ E[X+Y] &= E[X]+E[Y] \end{aligned}$$

It can be seen for example that the expectation is a nice linear function (look at the effect of multiplying by or adding a constant). The same is not true of the variance:

$$\begin{aligned} Var(cX) &= c^2 Var(X) \\ Var(X+c) &= Var(X) \end{aligned}$$

Also, note that $Var(X+Y) = Var(X)+Var(Y)$ iff the two variables are *independent*.

4. Empirical distribution function

The next set of slides will consider various well studied mathematical functions that provide a model for the way random variables behave. But it is worth noting the existence of an "empirical distribution function". For example, the Human Resources department of a business with 197 employees examined sickness records. It summarised the length of time each employee was absent as follows:

0 days absence	100 staff
1 days absence	80 staff
2 days absence	10 staff
3 days absence	2 staff
4 days absence	0 staff
5 days absence	5 staff
6 or more days absence	0 staff

We have an "empirical distribution" that for example suggests $p[\text{Staff taking 2 days absence in 1 year}] = \frac{10}{197}$.

We may or may not wish to fit some kind of mathematical model to this, but there are number of applications where we could work with this empirical distribution directly as it is here. We will not consider the role of the empirical density function any further, and note it here only so that we have met the terminology.

Solutions to Quizzes

Solution to Quiz:

x	p[X=x]	Profit
0	$\frac{1}{6}$	- €2.00
1	$\frac{1}{6}$	- €0.60
2	$\frac{1}{6}$	€0.80
3	$\frac{1}{6}$	€2.20
4	$\frac{1}{6}$	€3.60
5	$\frac{1}{6}$	€5.00

We need to find $E[X] = \sum_{x\in\Omega_x} xp[X = x]$ which is given by $\frac{1}{6} \times -2.00 + \frac{1}{6} \times -0.60 + \frac{1}{6} \times 0.80 + \frac{1}{6} \times 2.20 + \frac{1}{6} \times 3.60 + \frac{1}{6} \times 5.$

This suggests that $E[X] =$ €1.5

Click on that green button to return to the quiz →

Solution to Quiz:

x	p[X=x]	Profit
0	$\frac{1}{6}$	- €1.6
1	$\frac{1}{6}$	- €0.20
2	$\frac{1}{6}$	€1.20
3	$\frac{1}{6}$	€2.60
4	$\frac{1}{6}$	€4.00
5	$\frac{1}{6}$	€4.00

We need to find $E[X] = \sum_{x \in \Omega_x} xp[X = x]$ which is given by $\frac{1}{6} \times -1.60 + \frac{1}{6} \times -0.20 + \frac{1}{6} \times 1.20 + \frac{1}{6} \times 2.60 + \frac{1}{6} \times 4.00 + \frac{1}{6} \times 4.00$. This suggests that $E[X] =$ €1.67

Click on that green button to return to the quiz →

Solution to Quiz:

x	p[X=x]	Y (Profit)
0	$\frac{1}{6}$	- €1.20
1	$\frac{1}{6}$	€0.20
2	$\frac{1}{6}$	€1.60
3	$\frac{1}{6}$	€3.00
4	$\frac{1}{6}$	€3.00
5	$\frac{1}{6}$	€3.00

We need to find $E[X] = \sum_{x \in \Omega_x} xp[X = x]$ which is given by:
$\frac{1}{6} \times -1.20 + \frac{1}{6} \times 0.20 + \frac{1}{6} \times 1.60 + \frac{1}{6} \times 3.00 + \frac{1}{6} \times 3.00 + \frac{1}{6} \times 3.00.$
This suggests that $E[X] =$ €1.60

Click on that green button to return to the quiz →

Solution to Quiz: It is clear from this very simple and artificial exercises that the baker sees the best expected value for a strategy of baking 4 cakes a day. However, in a business context you might also like to consider the cost of disappointed customers who might not come back (these are minimised in strategy 1), or the risk of a negative cash-flow (which is minimised in strategy 3)

Click on that green button to return to the quiz →

Solution to Quiz: Remember the need to evaluate all possible outcomes. So we have $\Omega = (TT), (TH), (HT), (HH)$.
The value of X associated with each value in the sample space is $0, 1, 1, 2$ respectively, hence
$p[X = x]$ for $x = 0, 1, 2$ is given by $0.25, 0.50, 0.25$

Click on that green button to return to the quiz →

Solution to Quiz: We use $E[X] = \sum_{x\in\Omega_x} xp[X = x]$ which requires $\frac{1}{4} \times 0 + \frac{1}{2} \times 1 + \frac{1}{4} \times 2$ which gives $E[X] = 1$

Click on that green button to return to the quiz $\rightarrow$

Solution to Quiz: We already know $E[X] = 1$, so need to find $E[X^2]$ which can be found as $\frac{1}{4} \times 0^2 + \frac{1}{2} \times 1^2 + \frac{1}{4} \times 2^2$ which gives $E[X^2] = 1.5$. We therefore require $E[X^2] - E[X]^2 = 1.5 - 1 = 0.5$

Click on that green button to return to the quiz $\rightarrow$