Lecture 2: Discrete Random Variables

Michaelmas 2002
Introduction

In this world nothing can be said to be certain, except death and taxes.

- Benjamin Franklin

We use probabilities all the time

- Gambling (not recommended)
- Weather forecasting (not very successfully)
- Insurance (risk assessment)
- Stock market

For the purpose of handling probability the world can be partitioned into two distinct areas

1. **Discrete**: there are a finite number of events that can occur (for example rolling a die).

2. **Continuous**: there is a continuum of events (for example exactly measuring the height of an individual).

This lecture considers discrete events. The problem of continuous events is considered in the next lecture.
Discrete Probabilities

Discrete events are the simplest to interpret. For example What’s the probability of:

• it raining tomorrow?
• Red Rum winning the Grand National?
• a 6 being thrown on a die?

Probability can be thought of as the chance of a particular event occurring. We limit the range of our probability measure to lie in the range 0 to 1, where

• Lower numbers indicate that the event is less likely to occur, 0 indicates it will never occur.

• Higher numbers indicate that the event is more likely to occur, 1 indicates that the event will definitely occur.

We like to think that we have a good grasp of both estimating and using “probability”. For simple cases “will it rain tomorrow?” we can do reasonably well. However as situations get more complicated things are not always so clear.

The aim of probability theory is to give us a mathematically sound way of inferring information using probabilities.
In order to make inference using probability we need to define some notation:

Let some event have have $M$ possible outcomes. We are interested in the probability of each of these outcomes occurring. Let the set of possible outcomes be

$$\mathcal{X} = \{v_1, v_2, \ldots, v_M\}$$

We will want to compute the probability of a particular event occurring

$$p_i = \Pr(x = v_i)$$

where $x$ is a discrete random variable.

For a single die there are 6 possible events ($M = 6$)

$$\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$$

For a fair die we know that (for example)

$$p_1 = \frac{1}{6}, \quad p_2 = \frac{1}{6}$$

and so on.
Probability Mass Function

It is more convenient to express the set of probabilities:

\[ \{p_1, p_2, \ldots, p_M\} \]

as a *probability mass function*, \( P(x) \).

Attributes of a probability mass function (PMF):

\[ P(x) \geq 0 \]

and

\[ \sum_{x \in \mathcal{X}} P(x) = 1 \]

In words:

1. The first constraint means that probabilities must always be positive (what would a negative probability mean?).

2. The second constraint states that one of the set of possible must occur.

From these constraints it is simple to obtain

\[ 0 \leq P(x) \leq 1 \]

We will normally use PMFs in the lectures.
**Expected Values**

If we have a PMF there are some useful *statistics* that may be obtained from it.

We will often use *expected values* (think of it as an average)

\[ E \{ x \} = \sum_{x \in X} xP(x) = \sum_{i=1}^{M} v_i p_i \]

This is also known as the *mean* (this will be used alot in this course)

\[ \mu = E \{ x \} = \sum_{x \in X} xP(x) \]

The mean value from a single roll of a die is

\[ \mu = \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5 \]

We can also take expected values over *functions* of the random variable

\[ E \{ f(x) \} = \sum_{x \in X} f(x)P(x) \]

An important attribute when \( f(x) \) is *linear* is that

\[ E \{ \alpha_1 f_1(x) + \alpha_2 f_2(x) \} = E \{ \alpha_1 f_1(x) \} + E \{ \alpha_2 f_2(x) \} \]
Moments of a PMF

We will sometimes need to calculate the moments of a PMF. The $n^{th}$ moment is defined as

$$\mathcal{E} \{ x^n \} = \sum_{x \in \mathcal{X}} x^n P(x)$$

We will often need the second moment and the variance. The variance is defined as

$$\text{Var} \{ x \} = \sigma^2$$
$$= \mathcal{E} \{(x - \mu)^2\}$$
$$= \mathcal{E} \{ x^2 \} - (\mathcal{E} \{ x \})^2$$

i.e. it is simply the difference between the second moment and the first moment squared.

An attribute of the variance is

$$\sigma^2 \geq 0$$

Again taking the example of the die, the second moment is

$$\mathcal{E} \{ x^2 \} = \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6} = 15.1667$$

Therefore the variance is

$$\sigma^2 = 15.1667 - 3.5 \times 3.5 = 2.9167$$
Pairs of Random Variables

For many problems we will need to handle situations where there is more than a single random variable. Consider the case of 2 discrete random variables, $x$ and $y$. Here $y$ may take any of the values of the set $\mathcal{Y}$.

Now instead of having PMFs of a single variable, we want the \textit{joint PMF}, $P(x, y)$. This may be viewed as the probability of $x$ taking a particular values \textbf{and} $y$ taking a particular value.

This joint PMF must still satisfies the rules

$$P(x, y) \geq 0$$

and

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$$

Take a simple example of the weather. We are interested in whether it rains or not on 2 particular days. So $x$ is the random variable associated with it raining on day 1, $y$ on day 2. The joint PMF may be described by the table

<table>
<thead>
<tr>
<th>$P(x, y)$</th>
<th>rain</th>
<th>sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>rain</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>sun</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>total</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>
### Marginal Distributions

Given the joint distribution we are not always interested in the joint event, we may just be interested in the probability of a single event.

From the joint PMF we can derive *marginal* distributions

\[
P_x(x) = \sum_{y \in Y} P(x, y)
\]

and

\[
P_y(y) = \sum_{x \in X} P(x, y)
\]

For ease of notation we will write \(P_x(x)\) as \(P(x)\) where the context makes it clear.

Take the rain and sun example

<table>
<thead>
<tr>
<th>(P(x, y))</th>
<th>rain</th>
<th>sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>rain</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>sun</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>total</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

So

\[
\Pr(x = \text{rain}) = 0.6
\]
\[
\Pr(x = \text{sun}) = 0.4
\]

and similarly for the marginal distribution for \(y\).
Independence

An important concept in probability is *independence*. Two variables are statistically independent if

\[ P(x, y) = P(x)P(y) \]

This is very important since it is only necessary to know the individual PMFs to obtain the joint PMF.

Take the “sun and rain” example. Is whether it rains or no on the second day independent of the first day. Take the example of raining on both days and assume independence

\[
\Pr(x = \text{rain}, y = \text{rain}) = \Pr(x = \text{rain}) \Pr(y = \text{rain}) \\
= 0.6 \times 0.7 = 0.42 \\
\neq 0.4
\]

So from the joint PMF we know that the two random variables are *not* independent of one another.
Expected Values

We must also be able to handle expected values of two variables

\[ \mathcal{E} \{ f(x, y) \} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x, y) P(x, y) \]

This follows directly from the single variable case.

What are the moments of a joint distribution?

Using vector notation where

\[ x = \begin{bmatrix} x \\ y \end{bmatrix} \]

We get the first moment as

\[ \mu = \begin{bmatrix} \mathcal{E} \{x\} \\ \mathcal{E} \{y\} \end{bmatrix} = \mathcal{E} \{x\} = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} x P(x) \]

and the covariance matrix as

\[ \Sigma = \begin{bmatrix} \mathcal{E} \{(x - \mu_x)^2\} & \mathcal{E} \{(x - \mu_x)(y - \mu_y)\} \\ \mathcal{E} \{(x - \mu_x)(y - \mu_y)\} & \mathcal{E} \{(y - \mu_y)^2\} \end{bmatrix} 
\]

\[ = \mathcal{E} \{(x - \mu)(y - \mu)\}' \]

\[ = \mathcal{E} \{xx'\} - \mu \mu' \]

Covariance matrices are always symmetric.
Correlation

If we have two random variables $x$ and $y$ the covariance matrix may be written as

$$
\Sigma = \begin{bmatrix}
\sigma_{xx} & \sigma_{xy} \\
\sigma_{xy} & \sigma_{yy}
\end{bmatrix} = \begin{bmatrix}
\sigma_x^2 & \sigma_{xy} \\
\sigma_{xy} & \sigma_y^2
\end{bmatrix}
$$

where

$$
\sigma_{xy} = \mathbb{E} \{ (x - \mu_x)(y - \mu_y) \}
$$

and $\sigma_x^2 = \sigma_{xx}$.

The correlation coefficient, $\rho$, is defined as

$$
\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}
$$

This takes the values

$$
-1 \leq \rho \leq 1
$$

In general when $\rho = 0$ the two random variables are said to be uncorrelated. Note that independent random variables are always uncorrelated (you should be able to simply show this).
Conditional Probability

We are also interested in the probability of an event occurring given that some event has already happened. This is called the conditional probability.

\[
P(y|x) = \frac{P(x, y)}{P(x)}
\]

This is simple to illustrate with an example. From the “sun and rain” example: “What is the probability that it will rain on the second day given it rained on the first day?”

From the above equation and using the joint and marginal distributions

\[
\Pr(y = \text{rain}|x = \text{rain}) = \frac{\Pr(x = \text{rain}, y = \text{rain})}{\Pr(x = \text{rain})} = \frac{0.4}{0.6} = 0.6667
\]

It is worth noting that when the \( x \) and \( y \) are independent

\[
P(y|x) = \frac{P(x, y)}{P(x)} = P(y)
\]
Bayes’ Rule

One important rule that will be used often in the course is **Bayes’ Rule**. This is a very useful way of manipulating probabilities.

\[
P(x|y) = \frac{P(x, y)}{P(y)} = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(x)}{\sum_{x \in \mathcal{X}} P(y|x)P(x)} = \frac{P(y|x)P(x)}{\sum_{x \in \mathcal{X}} P(y|x)P(x)}
\]

We can express the conditional probability \( P(x|y) \) in terms of

- \( P(y|x) \) the conditional probability of \( y \) given \( x \)
- \( P(x) \) the probability of \( x \)

We will make extensive use of Bayes’ rule when look at statistical pattern classification.
Sum of Random Variables

We have so far considered a single random variable and joint distributions. What happens if we sum two random variables?

![Bar charts](a) Single Die (b) Sum of 2 Dice

- **Mean**: the mean of the sum of 2 RVs is
  \[ \mathcal{E} \{ x + y \} = \mathcal{E} \{ x \} + \mathcal{E} \{ y \} = \mu_x + \mu_y \]

- **Variance**: The variance of the sum of 2 independent RVs is
  \[
  \mathcal{E} \{(x + y - \mu_x - \mu_y)^2\} = \mathcal{E} \{(x - \mu_x)^2\} + \mathcal{E} \{(y - \mu_y)^2\} \\
  + 2 \mathcal{E} \{(x - \mu_x)(y - \mu_y)\} \\
  = \sigma_x^2 + \sigma_y^2
  \]

What happens to the variance if the 2 RVs are not independent?
Entropy

It would be useful to have a measure of how “random” a distribution is. Entropy, $H$, is defined as

$$H = - \sum_{x \in \mathcal{X}} P(x) \log_2(P(x))$$

$$= \mathcal{E} \left\{ \log_2 \left( \frac{1}{P(x)} \right) \right\}$$

Note:

1. $\log_2()$ is log base 2, not base 10 or natural log.

2. By definition $0 \log_2(0) = 0$.

For discrete distributions the entropy is usually measured in bits. One bit corresponds to the uncertainty that can be resolved with a simple yes/no answer.

For any set of $M$ possible symbols

- the PMF which has the maximum entropy is the uniform distribution
  $$P(x) = \frac{1}{M}$$

- the PMF which has the minimum entropy is the distribution where only a single probability is non zero (and so must be one).