4 One Dimensional Random Variables

In this chapter we will revise some of the material on discrete random variables and their distributions which you have seen in Probability I. We will also consider the statistical question of deciding whether a sample of data may reasonably be assumed to come from a particular discrete distribution.

First some revision:

Definition 4.1
If $E$ is an experiment having sample space $S$, and $X$ is a function that assigns a real number $X(e)$ to every outcome $e \in S$, then $X(e)$ is called a random variable (r.v.)

Definition 4.2
Let $X$ denote a r.v. and $x$ its particular value from the whole range of all values of $X$, say $R_X$. The probability of the event $(X \leq x)$ expressed as a function of $x$:

$$F_X(x) = P_X(X \leq x)$$

is called the Cumulative Distribution Function (cdf) of the r.v. $X$.

Properties of cumulative distribution functions

- $0 \leq F_X(x) \leq 1$, $-\infty < x < \infty$
- $\lim_{x \to -\infty} F_X(x) = 1$
- $\lim_{x \to -\infty} F_X(x) = 0$
- The function is nondecreasing.
  That is if $x_1 \leq x_2$ then $F_X(x_1) \leq F_X(x_2)$.

4.1 Discrete Random Variables

Values of a discrete r.v. are elements of a countable set \{x_1, x_2, \ldots, x_n, \ldots\}. We associate a number $p_X(x_i) = P_X(X = x_i)$ with each outcome $x_i$, $i = 1, 2, \ldots$, such that:

1. $p_X(x_i) \geq 0$ for all $i$
2. $\sum_{i=1}^{\infty} p_X(x_i) = 1$

Note that

$$F_X(x_i) = P_X(X \leq x_i) = \sum_{x \leq x_i} p_X(x) \quad (4.2)$$

$$p_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \quad (4.3)$$

The function $p_X$ is called the \textit{Probability Function} of the random variable $X$, and the collection of pairs

$$\{(x_i, p_X(x_i)), \ i = 1, 2, \ldots\} \quad (4.4)$$

is called the \textit{Probability Distribution} of $X$. The distribution is usually presented in either tabular, graphical or mathematical form.

\textbf{Example 4.1}

$$X \sim \text{Binomial}(8, 0.4)$$

That is $n = 8$, and the probability of success $p$ equals 0.4. Present the distribution in a mathematical, tabular and graphical way. Also, draw the cdf of the variable $X$.

Mathematical form:

$$\{(k, P(X = k) = ^nC_k p^k (1 - p)^{n-k}), \ k = 0, 1, 2, \ldots 8\} \quad (4.5)$$

Tabular form:

<table>
<thead>
<tr>
<th>$k$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X=k)$</td>
<td>0.0168</td>
<td>0.0896</td>
<td>0.2090</td>
<td>0.2787</td>
<td>0.2322</td>
<td>0.1239</td>
<td>0.0413</td>
<td>0.0079</td>
<td>0.0007</td>
</tr>
<tr>
<td>$P(X \leq k)$</td>
<td>0.0168</td>
<td>0.1064</td>
<td>0.3154</td>
<td>0.5941</td>
<td>0.8263</td>
<td>0.9502</td>
<td>0.9915</td>
<td>0.9993</td>
<td>1</td>
</tr>
</tbody>
</table>

Other important discrete distributions are:

- \textit{Bernoulli}(p)
- \textit{Geometric}(p)
- \textit{Hypergeometric}(n, M, N)
- \textit{Poisson}(\lambda)

For their properties see Probability I course lecture notes.
4.2 Goodness of fit tests for discrete random variables

4.2.1 A straight forward example

Suppose we wish to test the hypothesis (or assumption) that a set of data follows a binomial distribution.

For example suppose we throw three drawing pins and count the number of ups. We want to test the hypothesis that a drawing pin is equally likely to land up or down. We do this 120 times and get the following data

<table>
<thead>
<tr>
<th>Ups</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed frequency</td>
<td>10</td>
<td>35</td>
<td>54</td>
<td>21</td>
</tr>
</tbody>
</table>

Is there any evidence to suggest that the drawing pin is not equally likely to land up or down?

Suppose it was equally likely then the number of ups in a single throw, assuming independent trials, would have a binomial distribution with \( n = 3 \) and \( p = \frac{1}{2} \). So writing \( Y \) as the number of ups we would have \( P[Y = 0] = \frac{1}{8} \), \( P[Y = 1] = \frac{3}{8} \), \( P[Y = 2] = \frac{3}{8} \), \( P[Y = 3] = \frac{1}{8} \). Thus in 120 trials our expected frequencies under a binomial model would be

<table>
<thead>
<tr>
<th>Ups</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed frequency</td>
<td>15</td>
<td>45</td>
<td>45</td>
<td>15</td>
</tr>
</tbody>
</table>

Now our observed frequencies are not the same as our expected frequencies. But this might be due to random variation. We know a random variable doesn’t always take its mean value. But how surprising is the amount of variation we have here?

We make use of a test statistic \( X^2 \) defined as follows

\[
X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}
\]

where \( O_i \) are the observed frequencies, \( E_i \) are the expected frequencies and \( k \) is the number of classes, or values that \( Y \) can take.

Now it turns out that if we find the value of \( X^2 \) for lots of samples for which our hypothesis is true it has a particular distribution called a \( \chi^2 \) or chi-squared distribution. We can calculate the value of \( X^2 \) for our sample.
If this value is big, i.e. it is in the right tail of the \( \chi^2 \) distribution we might regard this as evidence that our hypothesis or assumption is false. (Note if the value of \( X^2 \) was very small we might regard this as evidence that the agreement was “too good” and that some cheating had been going on.)

In our example

\[
X^2 = \frac{(10 - 15)^2}{15} + \frac{(35 - 45)^2}{45} + \frac{(54 - 45)^2}{45} + \frac{(21 - 15)^2}{15}
\]

\[
= \frac{25}{100} \cdot 15 + \frac{81}{45} + \frac{36}{15} = \frac{75 + 100 + 81 + 108}{45} = \frac{364}{45} = 8.08
\]

Now look at Table 7, p37 in the New Cambridge statistical tables. This gives the distribution function of a \( \chi^2 \) random variable. It depends on a parameter \( \nu \) which is called the degrees of freedom. For our goodness of fit test the value of \( \nu \) is given by \( k - 1 \). So \( \nu = 3 \). For 8.0 the distribution function value is 0.9540. For 8.2 it is 0.9579. If we interpolate linearly we will get

\[
0.9540 + 0.08/0.20 \times (.9579 - .9540) = .9556
\]

Thus the area to the right of 8.08 is 1 - 0.9556 = 0.0444. This is quite a small value. It represents the probability of obtaining an \( X^2 \) value of 8.08 or more if we carry out this procedure repeatedly on samples which actually do come from a binomial distribution with \( p = 0.5 \). It is called the P value of the test. A P value of 0.0444 would be regarded by most statisticians as moderate evidence against the hypothesis.

An alternative approach to testing is to make a decision to accept or reject the hypothesis. This is done so that there is a fixed probability of rejecting the hypothesis when it is true. This probability is often chosen as 0.05. (Note: there is no good reason for picking this value rather than some other value; also we ought to choose a smaller probability as \( n \) increases. We will return to these ideas later in the course.) If we did choose 0.05 Table 8 shows us that for \( \nu = 3 \) the corresponding value of the \( \chi^2 \) distribution is
7.815. If the value of $X^2 \leq 7.815$ we accept the hypothesis if $X^2 > 7.815$ we reject the hypothesis. As $X^2 = 8.08$ we reject the hypothesis. To make it clear we have chosen 0.05 as our probability of rejecting the hypothesis when it is true, we say we reject the hypothesis at a 5% significance level. We call the value 7.815 the critical value.

### 4.2.2 Complicating factors

There are a couple of factors to complicate the goodness of fit test. Firstly if any of the expected frequencies ($E_i$) are less than 5 then we must group adjacent classes so that all expected frequencies are greater than 5. Secondly if we need to estimate any parameters from the data then the formula for the degrees of freedom is amended to read

$$\nu = k - p - 1$$

where $k$ is the number of classes and $p$ is the number of parameters estimated from the data.

We can illustrate both these ideas in the following example.

It is thought that the number of accidents per month at a junction follows a Poisson distribution. The frequency of accidents in 120 months was as follows

<table>
<thead>
<tr>
<th>Accidents</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed frequency</td>
<td>41</td>
<td>40</td>
<td>22</td>
<td>10</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

To find the Poisson probabilities we need the mean $\mu$. Since this isn’t specified in the question we will have to estimate it from the data. A reasonable estimate is the sample mean of the data. This is

$$\frac{0 \times 41 + 1 \times 40 + 2 \times 22 + \cdots + 6 \times 1}{120} = 1.2$$

Now using the Poisson formula

$$P[Y = y] = \frac{e^{-\mu} \mu^y}{y!}$$

or Table 2 in New Cambridge Statistical Tables we can complete the probabilities in the following table

5
<table>
<thead>
<tr>
<th>Accidents</th>
<th>Probability</th>
<th>$E_i$</th>
<th>$O_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.3012</td>
<td>36.14</td>
<td>41</td>
</tr>
<tr>
<td>1</td>
<td>0.3614</td>
<td>43.37</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>0.2169</td>
<td>26.03</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>0.0867</td>
<td>10.40</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>0.0261</td>
<td>3.13</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>0.0062</td>
<td>0.74</td>
<td>0</td>
</tr>
<tr>
<td>6+</td>
<td>0.0015</td>
<td>0.18</td>
<td>1</td>
</tr>
</tbody>
</table>

Now the last three expected frequencies are all less than 5. If we group them together into a class 4+ the expected frequency will be 4.05, still less than 5. So we group the last four classes into a class 3+ with expected frequency 14.45 and observed frequency 17. We find $X^2$ as before.

$$
X^2 = \frac{(36.14 - 41)^2}{36.14} + \frac{(43.37 - 40)^2}{43.37} + \frac{(26.03 - 22)^2}{26.03} + \frac{(14.45 - 17)^2}{14.45}
$$

$$
= 0.65 + 0.26 + 0.62 + 0.45
$$

$$
= 1.98
$$

Now after our grouping there are four classes so $k = 4$ and we estimated one parameter, the mean, from the data so $p = 1$. So $\nu = 4 - 1 - 1 = 2$. Looking in Table 7 the distribution function for 1.9 is 0.6133 and for 2.0 is 0.6321. So the interpolated value for 1.98 is 0.6133 + 0.08/0.10 $\times$ (0.6321 - 0.6133) = 0.6283. Thus the P value is 1 - 0.6283 = 0.3717. Such a large P value is regarded as showing no evidence against the hypothesis that the data have a Poisson distribution.

Alternatively for a significance test at the 5% level the critical value is 5.991 from table 8 and as 1.98 is smaller than this value we accept the hypothesis that the data have a Poisson distribution.

### 4.3 Continuous Random Variables

Values of a continuous r.v. are elements of an uncountable set, for example a real interval. The c.d.f. of a continuous r.v. is a continuous, nondecreasing, differentiable function. An interesting difference from a discrete r.v. is that for $\delta > 0$

$$
P_X(X = x) = \lim_{\delta \to 0} (F_X(x + \delta) - F_X(x)) = 0
$$

6
We define the *Density Function* of a continuous r.v. as:

\[ f_X(x) = \frac{d}{dx} F_X(x) \]  

(4.6)

Hence

\[ F_X(x) = \int_{-\infty}^{x} f_X(t)dt \]  

(4.7)

Similarly to the properties of the probability distribution of a discrete r.v. we have the following properties of the density function:

1. \( f_X(x) \geq 0 \) for all \( x \in \mathbb{R} \)

2. \( \int_{\mathbb{R}} f_X(x)dx = 1 \)

Probability of an event \((X \in A)\), where \(A\) is an interval, is expressed as an integral

\[ P_X(-\infty < X < a) = \int_{-\infty}^{a} f_X(x)dx = F_X(a) \]  

(4.8)

or for a bounded interval

\[ P_X(b < X < c) = \int_{b}^{c} f_X(x)dx = F_X(c) - F_X(b) \]  

(4.9)

**Example 4.2** Normal Distribution \(N(\mu, \sigma^2)\)

The density function is given by:

\[ f_X(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \]  

(4.10)

There are two parameters which tell us about the position and the shape of the density curve: the expected value \( \mu \) and the standard deviation \( \sigma \).

You have already seen in Probability I how to use Tables to find probabilities about the normal distribution.

Other important continuous distributions are

- \( Uniform(a, b) \)

- \( Exponential(\lambda) \)

For their properties see Probability I course lecture notes.

Note that all distributions you have come across depend on one or more parameters, for example \( p, \lambda, \mu, \sigma \). These values are usually unknown and their estimation is one of the important problems in statistical analysis.
4.3.1 A goodness of fit test for a continuous random variable

Consider the following example.

Traffic is passing freely along a road. The time interval between successive vehicles is measured (in seconds) and recorded below.

<table>
<thead>
<tr>
<th>Time interval</th>
<th>0-20</th>
<th>20-40</th>
<th>40-60</th>
<th>60-80</th>
<th>80-100</th>
<th>100-120</th>
<th>120+</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of cars</td>
<td>54</td>
<td>28</td>
<td>12</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Test whether an exponential distribution provides a good fit to these data.

We need to estimate the parameter $\lambda$ of the exponential distribution. Since $\lambda^{-1}$ is the mean of the distribution it seems reasonable to put $\lambda = 1/\bar{x}$. (We will discuss this further when we look at estimation). Now the data are presented as intervals so we will have to estimate the sample mean. It is common to do this by pretending that all the values in an interval are actually at the mid-point of the interval. We will do this whilst recognising that for the exponential distribution, which is skewed, it is a bit questionable.

The calculation for the sample mean is given below.

<table>
<thead>
<tr>
<th>Midpoint $x$</th>
<th>Frequency $f$</th>
<th>$fx$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>54</td>
<td>540</td>
</tr>
<tr>
<td>30</td>
<td>28</td>
<td>840</td>
</tr>
<tr>
<td>50</td>
<td>12</td>
<td>600</td>
</tr>
<tr>
<td>70</td>
<td>10</td>
<td>700</td>
</tr>
<tr>
<td>90</td>
<td>4</td>
<td>360</td>
</tr>
<tr>
<td>110</td>
<td>2</td>
<td>220</td>
</tr>
<tr>
<td></td>
<td>110</td>
<td>3260</td>
</tr>
</tbody>
</table>

thus the estimated mean is $3260/110 = 29.6$. Thus we test if the data are from an exponential distribution with parameter $\lambda = 1/29.6$.

We must calculate the probabilities of lying in the intervals given this distribution.

\[
P[X < 20] = \int_0^{20} \lambda e^{-\lambda x} dx
\]

\[
= 1 - e^{-20\lambda}
\]

\[
= 0.4912
\]
\[ P[20 < X < 40] = \int_{20}^{40} \lambda e^{-\lambda x} dx \]
\[ = e^{-20\lambda} - e^{-40\lambda} \]
\[ = 0.2499 \]

Similarly
\[ P[40 < X < 60] = e^{-40\lambda} - e^{-60\lambda} = 0.1272 \]
\[ P[60 < X < 80] = e^{-60\lambda} - e^{-80\lambda} = 0.0647 \]
\[ P[80 < X < 100] = e^{-80\lambda} - e^{-100\lambda} = 0.0329 \]
\[ P[100 < X] = e^{-100\lambda} = 0.0341 \]

Multiplying these probabilities by 110 we find the expected frequencies as given in the table below.

<table>
<thead>
<tr>
<th>Time interval</th>
<th>0-20</th>
<th>20-40</th>
<th>40-60</th>
<th>60-80</th>
<th>80-100</th>
<th>100+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed frequency</td>
<td>54</td>
<td>28</td>
<td>12</td>
<td>10</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Expected frequency</td>
<td>54.03</td>
<td>27.49</td>
<td>13.99</td>
<td>7.12</td>
<td>3.62</td>
<td>3.75</td>
</tr>
</tbody>
</table>

We must merge the final two classes so that the expected values are greater than 5. Thus for 80+ we have 6 observed and 7.37 expected.

We find
\[ X^2 = \sum \frac{(O - E)^2}{E} = 1.71. \]

Now \( \nu = 5 - 1 - 1 = 3 \) since after grouping there were 5 classes and we estimated one parameter from the data. From Table 7 the P value is thus \( 1 - 0.3653 = 0.6347 \) and there is no evidence against the hypothesis that the data follows an exponential distribution.