Paper Name: Basic Statistics

Paper Code: MS 204

Unit II

Discrete Probability Distributions

Basics of Probability Distributions

As a reminder, a variable or what will be called the random variable from now on, is represented by the letter *x* and it represents a quantitative (numerical) variable that is measured or observed in an experiment.

Also remember there are different types of quantitative variables, called discrete or continuous. What is the difference between discrete and continuous data? **Discrete** data can only take on particular values in a range. **Continuous** data can take on any value in a range. Discrete data usually arises from counting while continuous data usually arises from measuring.

Examples of each:
How tall is a plant given a new fertilizer? Continuous. This is something you measure.
How many fleas are on prairie dogs in a colony? Discrete. This is something you count.

If you have a variable, and can find a probability associated with that variable, it is called a **random variable**. In many cases the random variable is what you are measuring, but when it comes to discrete random variables, it is usually what you are counting. So for the example of how tall is a plant given a new fertilizer, the random variable is the height of the plant given a new fertilizer. For the example of how many fleas are on prairie dogs in a colony, the random variable is the number of fleas on a prairie dog in a colony.

Now suppose you put all the values of the random variable together with the probability that that random variable would occur. You could then have a distribution like before, but now it is called a probability distribution since it involves probabilities. A **probability distribution** is an assignment of probabilities to the values of the random variable. The abbreviation of pdf is used for a probability distribution function.

For probability distributions, $0 \le P(x) \le 1$ and $\sum P(x) = 1$

Example 1: Probability Distribution

The 2010 U.S. Census found the chance of a household being a certain size. The data is in table #5.1.1 ("Households by age," 2013).

Table 1: Household Size from U.S. Census of 2010

Size of household	1	2	3	4	5	6	7 or more
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

Solution:
In this case, the random variable is x = number of people in a household. This is a discrete random variable, since you are counting the number of people in a household.

This is a probability distribution since you have the *x* value and the probabilities that go with it, all of the probabilities are between zero and one, and the sum of all of the probabilities is one.

You can give a probability distribution in table form (as in table #5.1.1) or as a graph. The graph looks like a histogram. A probability distribution is basically a relative frequency distribution based on a very large sample.

Example 2: Graphing a Probability Distribution

The 2010 U.S. Census found the chance of a household being a certain size. The data is in the table ("Households by age," 2013). Draw a histogram of the probability distribution.

Table 2: Household Size from U.S. Census of 2010

Size of household	1	2	3	4	5	6	7 or more
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

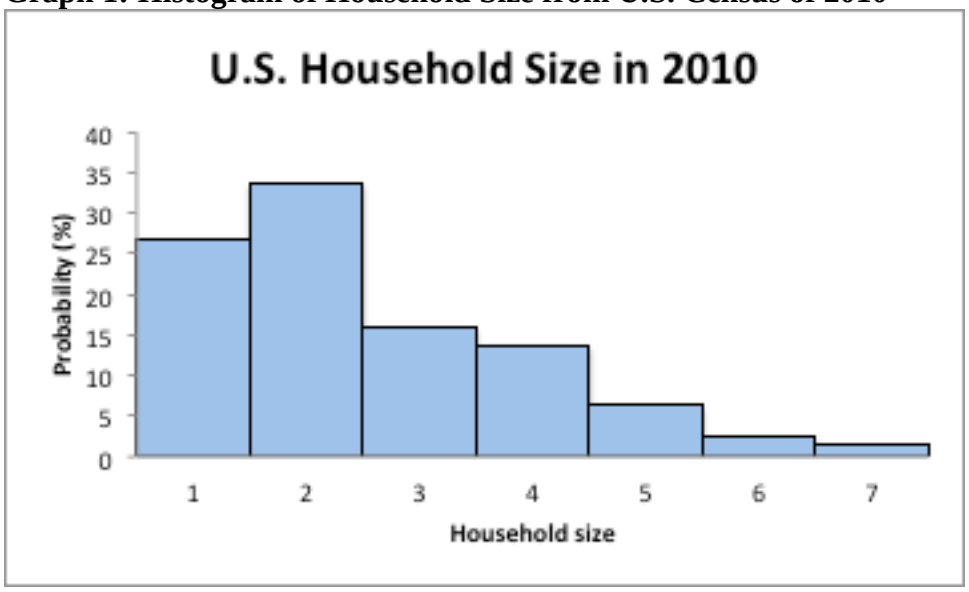
Solution:

State random variable:

x = number of people in a household

You draw a histogram, where the *x* values are on the horizontal axis and are the *x* values of the classes (for the 7 or more category, just call it 7). The probabilities are on the vertical axis.

Graph 1: Histogram of Household Size from U.S. Census of 2010



Notice this graph is skewed right.

Just as with any data set, you can calculate the mean and standard deviation. In problems involving a probability distribution function (pdf), you consider the probability distribution the population even though the pdf in most cases come from repeating an experiment many times. This is because you are using the data from repeated experiments to estimate the true probability. Since a pdf is basically a population, the mean and standard deviation that are calculated are actually the population parameters and not the sample statistics. The notation used is the same as the notation for population mean and population standard deviation that was used in chapter 3. Note: the mean can be thought of as the **expected value**. It is the value you expect to get if the trials were repeated infinite number of times. The mean or expected value does not need to be a whole number, even if the possible values of *x* are whole numbers.

For a discrete probability distribution function,

The mean or expected value is $\mu = \sum xP(x)$

The variance is $\sigma^2 = \sum (x-\mu)^2 P(x)$

The standard deviation is $\sigma = \sqrt{\sum (x-\mu)^2 P(x)}$

where x = the value of the random variable and $P(x)$ = the probability corresponding to a particular x value.

Binomial Probability Distribution

Section 5.1 introduced the concept of a probability distribution. The focus of the section was on discrete probability distributions (pdf). To find the pdf for a situation, you usually needed to actually conduct the experiment and collect data. Then you can calculate the experimental probabilities. Normally you cannot calculate the theoretical probabilities instead. However, there are certain types of experiment that allow you to calculate the theoretical probability. One of those types is called a **Binomial Experiment**.

Properties of a **binomial experiment** (or Bernoulli trial):
1) Fixed number of trials, n, which means that the experiment is repeated a specific number of times.
2) The n trials are independent, which means that what happens on one trial does not influence the outcomes of other trials.
3) There are only two outcomes, which are called a success and a failure.
4) The probability of a success doesn't change from trial to trial, where p = probability of success and q = probability of failure, $q = 1 - p$.

If you know you have a binomial experiment, then you can calculate binomial probabilities. This is important because binomial probabilities come up often in real life. Examples of binomial experiments are:

Toss a fair coin ten times, and find the probability of getting two heads.

Question twenty people in class, and look for the probability of more than half being women?

Shoot five arrows at a target, and find the probability of hitting it five times?

To develop the process for calculating the probabilities in a binomial experiment, consider example.1.

Example 1: Deriving the Binomial Probability Formula

Suppose you are given a 3 question multiple-choice test. Each question has 4 responses and only one is correct. Suppose you want to find the probability that you can just guess at the answers and get 2 questions right. (Teachers do this all the time when they make up a multiple-choice test to see if students can still pass without studying. In most cases the students can't.) To help with the idea that you are going to guess, suppose the test is in Martian.

a.) What is the random variable?

Solution:

x = number of correct answers

b.) Is this a binomial experiment?

Solution:

1.) There are 3 questions, and each question is a trial, so there are a fixed number of trials. In this case, $n = 3$.
2.) Getting the first question right has no affect on getting the second or third question right, thus the trials are independent.
3.) Either you get the question right or you get it wrong, so there are only two outcomes. In this case, the success is getting the question right.
4.) The probability of getting a question right is one out of four. This is the same for every trial since each question has 4 responses. In this case, $p = \frac{1}{4}$ and $q = 1 - \frac{1}{4} = \frac{3}{4}$

This is a binomial experiment, since all of the properties are met.

c.) What is the probability of getting 2 questions right?

Example 2: Calculating Binomial Probabilities

When looking at a person's eye color, it turns out that 1% of people in the world has green eyes ("What percentage of," 2013). Consider a group of 20 people.
a.) State the random variable.

Solution:

x = number of people with green eyes

b.) Argue that this is a binomial experiment

Solution:

1.) There are 20 people, and each person is a trial, so there are a fixed number of trials. In this case, $n = 20$.
2.) If you assume that each person in the group is chosen at random the eye color of one person doesn't affect the eye color of the next person, thus the trials are independent.
3.) Either a person has green eyes or they do not have green eyes, so there are only two outcomes. In this case, the success is a person has green eyes.
4.) The probability of a person having green eyes is 0.01. This is the same for every trial since each person has the same chance of having green eyes. $p = 0.01$ and $q = 1 - 0.01 = 0.99$

Find the probability that
c.) None have green eyes.

Solution:

$$P(x = 0) = {}_{20}C_0 (0.01)^0 (0.99)^{20-0} \approx 0.818$$

d.) Nine have green eyes.

Solution:

$$P(x=9) = {}_{20}C_{9}(0.01)^{9}(0.99)^{20-9} \approx 1.50 \times 10^{-13} \approx 0.000$$

e.) At most three have green eyes.

Solution:

At most three means that three is the highest value you will have. Find the probability of *x* is less than or equal to three.

$$P(x \le 3) = P(x=0) + P(x=1) + P(x=2) + P(x=3)$$
$$= {}_{20}C_0 (0.01)^0 (0.99)^{20} + {}_{20}C_1 (0.01)^1 (0.99)^{19}$$
$$+ {}_{20}C_2 (0.01)^2 (0.99)^{18} + {}_{20}C_3 (0.01)^3 (0.99)^{17}$$
$$\approx 0.818 + 0.165 + 0.016 + 0.001 > 0.999$$

The reason the answer is written as being greater than 0.999 is because the answer is actually 0.9999573791, and when that is rounded to three decimal places you get 1. But 1 means that the event will happen, when in reality there is a slight chance that it won't happen. It is best to write the answer as greater than 0.999 to represent that the number is very close to 1, but isn't 1.

f.) At most two have green eyes.

Solution:

$$P(x \le 2) = P(x=0) + P(x=1) + P(x=2)$$
$$= {}_{20}C (0.01)^0 (0.99)^{20} + C (0.01)^1 (0.99)^{19} + C (0.01)^2 (0.99)^{18}$$

$\approx 0.818 + 0.165 + 0.016 \approx 0.999$

g.) At least four have green eyes.

Solution:
At least four means four or more. Find the probability of x being greater than or equal to four. That would mean adding up all the probabilities from four to twenty. This would take a long time, so it is better to use the idea of complement. The complement of being greater than or equal to four is being less than four. That would mean being less than or equal to three. Part (e) has the answer for the probability of being less than or equal to three. Just subtract that number from 1.
$P(x \geq 4) = 1 - P(x \leq 3) = 1 - 0.999 = 0.001$.
Actually the answer is less than 0.001, but it is fine to write it this way.

The Poisson distribution

The Poisson distribution

THE AVONFORD STAR

Full moon madness hits Avonford bypass.

Since opening two years ago, Avonford bypass has seen more than its fair share of accidents but last night was way beyond that ever experienced before. There were no less than 4 separate accidents during the hours of darkness. And it was full moon!!
Was it, we wonder, full moon madness? Or was it just one of those statistical quirks that happen from time to time?
Our Astrology expert, Jessie Manning told us that this was only to be expected when the moon dominates Saturn.
However, the local vicar, the Rev Paul Cheney took a different view. "We must be careful of jumping to the wrong conclusions" he said when we telephoned him this morning. "This is a load of dangerous rubbish that will lead more vulnerable people to believe dangerous things. I am not a Statistician so I cannot tell you what the chances are of there being 4 accidents in one evening, but I reckon that it is a statistical possibility".

How would you decide whether four accidents in a night are reasonably likely?
The first thing is to look at past data, and so learn about the distribution of accidents.
Since the bypass was opened nearly two years ago, the figures (not including the evening described in the article) are as follows:
(A day is taken to run from one midday to the next.)

Number of accidents per day, x	0	1	2	3	> 3
Frequency, f	395	235	73	17	0

These figures look as though the data could be drawn from a Poisson distribution. This distribution gives the probability of the different possible number of occurrences of an event in a given time

interval under certain conditions. If you are thinking of using a Poisson distribution, here is a check list to see if it is suitable.

- The events occur independently
- The events occur at random
- The probability of an event occurring in a given time interval does not vary with time

In this case, the given time interval is one day, or 24 hours. An event is an accident.

The total number of accidents has been $0 \times 395 + 1 \times 235 + 2 \times 73 + 3 \times 17 = 432$
The number of days has been $395 + 235 + 73 + 17 = 720$
So the mean number of accidents per day has been $\frac{432}{720} = 0.6$

The Poisson distribution is an example of a probability model. It is usually defined by the mean number of occurrences in a time interval and this is denoted by λ.

The probability that there are r occurrences in a given interval is given by $\frac{\lambda^r}{r!}e^{-\lambda}$.

The value of e is 2.71281 828 459…… There is a button for it on your calculator.

So, the probability of

0 occurrences is $e^{-\lambda}$

1 occurrence is $\lambda e^{-\lambda}$

2 occurrences is $\frac{\lambda^2}{2!}e^{-\lambda}$

3 occurrences is $\frac{\lambda^3}{3!}e^{-\lambda}$

and so on.

In this example, $\lambda = 0.6$ and so the probabilities and expected frequencies in 720 days are as follows

Number of accidents per day	0	1	2	3	4	5	> 5
Probability (4 d.p.)	0.5488	0.3293	0.0988	0.0197	0.0030	0.0004	0
Expected frequency (1 d.p.)	395.1	237.1	71.1	14.2	2.1	0.3	0

? Explain where the various figures in this table have come from.

? Compare the expected frequencies with those observed. Is the Poisson distribution a good model?

The table shows that with this model you would expect 2.4 days in 720 (i.e. just over 1 a year) where there would be 4 or more accidents. It would seem as though The Rev. Paul Cheney was right; the seemingly high number of accidents last night could be just what the statistical model would lead you to expect. There is no need to jump to the conclusion that there was another factor, such as full moon, that influenced the data.

ACTIVITY 1

Use your calculator to find the probability of 0, 1, 2, 3 occurrences of an event which has a Poisson distribution with mean $\lambda = 2.5$.

Use of tables

Another way to find probabilities in a Poisson distribution is to use tables of *Cumulative Poisson probabilities,* like those given in the MEI Students' Handbook.
In these tables you are not given P($X = r$) but P($X \leq r$). This means that it gives the sum of all probabilities from 0 up to r.

In the example of the accidents on Avonford Bypass the mean, $\bar{x}$, was 0.6 and probabilities of $X = 0$ and 1 were calculated to be 0.5488 and 0.3293.

To find these values in the tables, look at the column for $\lambda = 0.6$. The first entry in this column is 0.5488, representing the probability that there are no accidents.

x \ λ	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865
4	.....	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977
5	.....	.....	.....	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997
6	.....	.....	.....	.....	.....		1.0000	1.0000	1.0000
7	.....	.....	.....	.....	.....	.....	.....	.....	.....

Fig 1.1

The second entry is 0.8781. This is the probability that there will be 0 or 1 accidents.
To find the probability that there is one accident, subtract these two values giving 0.8781 – 0.5488 = 0.3293.
In the same way, the probability that there are 2 accidents is found by taking the second entry from the third.
Continuing the process gives the following.

Number of accidents	Probability	Number of accidents	Probability
0	0.5488	0	0.5488
0 or 1	0.8781	1	0.8781 – 0.5488 = 0.3293
0, 1 or 2	0.9769	2	0.9769 – 0.8781 = 0.0988
0, 1, 2 or 3	0.9966	3	0.9966 – 0.9769 = 0.0197
0, 1, 2, 3 or 4	0.9996	4	0.9996 – 0.9966 = 0.0030

? How can you use the tables to find the probability of exactly 5 accidents in any night?
Check that you get the same answer entering $\frac{(0.6)^5}{5!}e^{-0.6}$ into your calculator.

! You can see that the probability of having 4 accidents in one night is 0.0030. The probability of having 4 *or more* accidents in one night is 1 – probability of having 3 or fewer accidents, which is 0.9966. So the probability of having 4 or more accidents is 1 – 0.9966 =0.0034.
In other words 34 in every 10 000 days or roughly 2.5 days in 720. This confirms that it is not necessary to look for other explanations for the 4 accidents in the same night.

! You will see that the tables in the Students' Handbook cover values of λ from 0.01 to 8.90. You will clearly have a problem if you are trying to calculate probabilities with a value of λ that is not given in the tables. In such cases you will need to use the formula.

ACTIVITY 2

You were asked to find the probabilities of 0, 1, 2, 3 occurrences λ =2.5 in the previous activity. Now use the cumulative tables to find these probabilities.

Example 8.1

The mean number of typing errors in a document is 1.5 per page.
Find the probability that on a page chosen at random there are

(i) no mistakes,

(ii) more than 2 mistakes.

SOLUTION

If you assume that spelling mistakes occur independently and at random then the Poisson distribution is a reasonable model to use.

(i) For λ = 1.5 the tables give P(0 mistakes) = 0.2231.

x \ λ	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80	1.90
0	0.3679	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496
1	0.7358	0.6990	0.6626	0.6268	0.5918	0.5578	0.5249	0.4932	0.4628	0.4337
2	0.9197	0.9004	0.8795	0.8571	0.8335	0.8088	0.7834	0.7572	0.7306	0.7037
3	0.9810	0.9743	0.9662	0.9569	0.9463	0.9344	0.9212	0.9068	0.8913	0.8747
4	0.9963	0.9946	0.9923	0.9893	0.9857	0.9814	0.9763	0.9704	0.9636	0.9559
5	0.9994	0.9990	0.9985	0.9978	0.9968	0.9955	0.9940	0.9920	0.9896	0.9868
6	0.9999	0.9999	0.9997	0.9996	0.9994	0.9991	0.9987	0.9981	0.9974	0.9966
7	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9996	0.9994	0.9992
8	.....	.....	.....	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998
9	.....	.....	.....	.....	.....	.....	.....	1.0000	1.0000	1.0000

Fig 1.2

(ii) P(more than 2 mistakes) = 1 – P(up to 2 mistakes)
= 1 – 0.8088
= 0.1912.

? How would you answer this question using the Poisson Formula? Check that you get the same answers.

Historical note

Simeon Poisson was born in France in 1781. He worked as a mathematician in Paris for most of his life after giving up the study of medicine. His contribution to mathematics embraced electricity, magnetism and planetary orbits and ideas in integration as well as in statistics. He wrote over 300 papers and articles.
The modelling distribution that takes his name was originally derived as an approximation to the binomial distribution.

Exercise 8A

1 The number of cars passing a point on a country lane has a mean 1.8 per minute. Using the Poisson distribution, find the probability that in any one minute there are

(i) no cars, **(ii)** 1 car, **(iii)** 2 cars, **(iv)** 3 cars, **(v)** more than 3 cars.

2 A fire station experiences an average call-out rate of 2.2 every period of three hours. Using the Poisson distribution, find the probability that in any period of 3 hours there will be

(i) no callouts, **(ii)** 1 callout, **(iii)** 2 callouts, **(iv)** 3 callouts, **(v)** 4 callouts, **(v)** more than 4 callouts.

3 The number of radioactive particles emitted in a minute from a meteorite is recorded on a Geiger counter. The mean number is found to be 3.5 per minute. Using the Poisson distribution, find the probability that in any one minute there are

(i) no particles,

(ii) 2 particles,

(iii) at least 5 particles.

4 Bacteria are distributed independently of each other in a solution and it is known that the number of bacteria per millilitre follows a Poisson distribution with mean 2.9. Find the probability that a sample of 1 ml of solution contains

(i) 0, **(ii)** 1, **(iii)** 2, **(iv)** 3, **(v)** more than 3 bacteria.

5 The demand for cars from a car hire firm may be modelled by a Poisson distribution with mean 4 per day.

(i) Find the probability that in a randomly chosen day the demand is for

(***A***) 0, (***B***) 1, (***C***) 2, (***D***) 3 cars.

(ii) The firm has 5 cars available for hire. Find the probability that demand exceeds the number of cars available.

6 A book of 500 pages has 500 misprints. Using the Poisson distribution, estimate to three decimal places the probabilities that a given page contains

(i) exactly 3 misprints,

(ii) more than 3 misprints.

7 190 raisins are put into a mixture which is well stirred and made into 100 small buns. Which is the most likely number of raisins found in a bun?

8 Small hard particles are found in the molten glass from which glass bottles are made. On average 20 particles are found in 100 kg of molten glass. If a bottle made of this glass contains one or more such particles it has to be discarded.
Bottles of mass 1 kg are made using this glass.

(i) Criticise the following argument:
Since the material for 100 bottles contains 20 particles, approximately 20% will have to be discarded.

(ii) Making suitable assumptions, which should be stated, develop a correct argument using a Poisson model and find the percentage of faulty 1 kg bottles to 3 significant figures.

9 A hire company has two lawnmowers which it hires out by the day. The number of demands per day may be modelled by a Poisson distribution with mean 1.5. In a period of 100 working days, how many times do you expect

(i) neither lawnmower to be used,

(ii) some requests for a lawnmower to have to be refused?

Conditions for modelling data with a Poisson distribution

You met the idea of a probability model in Z1. The binomial distribution is one example. The Poisson distribution is another model. A model in this context means a theoretical distribution that fits your data reasonably well.

You have already seen that the Poisson distribution provides a good model for the data for the Avonford Star article on accidents on the bypass.

Here are the data again.

Number of accidents per day, x	0	1	2	3	> 3
Frequency, f	395	235	73	17	0

For these data, $n = 720$, $\sum xf = 432$ and $\sum x^2 f = 680$

So $\bar{x} = \dfrac{432}{720} = 0.6$

$S_{xx} = \sum x^2 f - n\bar{x}^2 = 680 - 720 \times 0.6^2 = 420.8$

So the variance = $\dfrac{S_{xx}}{n-1} = 0.585$

You will notice that the mean, 0.6, and the variance, 0.585, are very close in value. This is a characteristic of the Poisson distribution and provides a check on whether it is likely to provide a good model for a particular data set.

In the theoretical Poisson distribution, the mean and the variance are equal. However, it is usual to call λ the **parameter** of a Poisson distribution, rather than either the mean or the variance. The common notation for describing a Poisson distribution is Poisson(λ); so Poisson(2.4) means the Poisson distribution with parameter 2.4.

You should check that the conditions on page 1 apply - that the events occur at random, independently and with fixed probability.

Example 2

A mail order company receives a steady supply of orders by telephone. The manager wants to investigate the pattern of calls received so he records the number of calls received per day over a period of 40 days as follows.

Number of calls per day	0	1	2	3	4	5	> 5
Frequency of calls	8	13	10	6	2	1	0

(i) Calculate the mean and variance of the data. Comment on your answers.

(ii) State whether the conditions for using the Poisson distribution as a model apply.

(iii) Use the Poisson distribution to predict the frequencies of 0, 1, 2, 3… calls per hour.

(iv) Comment on the fit.

SOLUTION

(i) Summary statistics for these data are:

$n = 40, \sum xf = 64, \sum x^2 f = 164$

So mean, $\bar{x} = \dfrac{\sum xf}{n} = \dfrac{64}{40} = 1.6$

$S_{xx} = 164 - 40 \times 1.6^2 = 61.6$

So variance, $s^2 = \dfrac{61.6}{39} = 1.5795$

The mean is close to the variance, so it may well be appropriate to use the Poisson distribution as a model.

(ii) It is reasonable to assume that

- the calls occur independently
- the calls occur at random
- the probability of a call being made on any day of the week does not vary with time, given that there is a steady supply of orders.
-

(iii) From the cumulative tables with $\lambda = 1.6$ gives the following.

Calls	Probability	Calls	Probability	Expected frequency (probability × 40)
0	0.2019	0	0.2019	8.1
0 or 1	0.5249	1	0.5249 – 0.2019 = 0.3230	12.9
0, 1 or 2	0.7834	2	0.7834 – 0.5249 = 0.2585	10.3
0, 1, 2 or 3	0.9212	3	0.9212 – 0.7834 = 0.1378	5.5
0, 1, 2, 3 or 4	0.9763	4	0.9763 – 0.9212 = 0.0551	2.2
0, 1, 2, 3, 4 or 5	0.9940	5	0.9940 – 0.9763 = 0.0177	0.7
0, 1, 2, 3, 4, 5 or 6	0.9987	6	0.9987 – 0.9940 = 0.0047	0.2

(iv) This is the table showing the comparisons.

.

Number of calls per day	0	1	2	3	4	5	> 5
Actual frequency of calls	8	13	10	6	2	1	0
Theoretical frequency of calls (1 d.p.)	8.1	12.9	10.3	5.5	2.2	0.7	0.2

The fit is very good, as might be expected with the mean and variance so close together.

Example 3

Avonford Town Football Club recorded the number of goals scored in each one of their 30 matches in one season as follows.

Goals, x	0	1	2	3	4	> 4
Frequency, f	12	12	4	1	1	0

(i) Calculate the mean and variance for this set of data.

(ii) State whether the conditions for using the Poisson distribution apply.

(iii) Calculate the expected frequencies for a Poisson distribution having the same mean number of goals per match.

(iv) Comment on the fit.

SOLUTION

(i) For this set of data

$n = 30, \quad \sum xf = 27, \quad \sum x^2 f = 53$

So mean, $\bar{x} = \dfrac{\sum xf}{n} = \dfrac{27}{30} = 0.9,$

$S_{xx} = \sum x^2 f - \bar{n}x^2 = 53 - 30 \times 0.9^2 = 28.7$

So variance, $s^2 = \dfrac{28.7}{29} = 0.9897$

(ii) It is reasonable to assume that

- The goals are scored independently
- The goals are scored at random.
- the probability of scoring a goal is constant from one match to the next.

In addition, the value of the mean is close to the value of the variance. Hence, the Poisson distribution can be expected to provide a reasonably good model.

(iii) From the cumulative probability tables for $\lambda = 0.9$.

Goals	Probability	Goals	Probability	Expected frequency (probability × 30)
0	0.4066	0	0.4066	12.2
0 or 1	0.7725	1	0.7725 – 0.4066 = 0.3659	11.0
0, 1 or 2	0.9371	2	0.9371 – 0.7725 = 0.1646	4.9
0, 1, 2 or 3	0.9865	3	0.9865 – 0.9371 = 0.0494	1.5
0, 1, 2, 3 or 4	0.9977	4	0.9977 – 0.9865 = 0.0112	0.3

(iv) As expected from the closeness of the mean and the variance values, the fit is very good.

This is the table showing the comparisons.

Goals, x	0	1	2	3	4	> 4
Actual frequency, f	12	12	4	1	1	0
Theoretical frequency	12.3	11.0	4.9	1.5	0.3	0

? In example 8.2 above, it was claimed that the goals scored in a match were independent of each other. To what extent do you think this is true?

Exercise 8B

1 The number of bacteria in 50 100cc samples of water are given in the following table.

Number of bacteria per sample	0	1	2	3	4 or more
Number of samples	23	16	9	2	0

(i) Find the mean number and the variance of bacteria in a 100cc sample.

(ii) State whether the conditions for using the Poisson distribution as a model apply.

(iii) Using the Poisson distribution with the mean found in part (i), estimate the probability that another 100cc sample will contain

(A) no bacteria,

(B) more than 4 bacteria.

2 Avonford Town Council agree to install a pedestrian crossing near to the library on Prince Street if it can be shown that the probability that there are more than 4 accidents per month exceeds 0.1.
The accidents recorded in the last 10 months are as follows:

3 2 2 1 0 2 5 4 3 1

(i) Calculate the mean and variance for this set of data.

(ii) Is the Poisson distribution a reasonable model in this case?

(iii) Using the Poisson distribution with the mean found in part (i), find the probability that, in any month taken at random, there are more than 4 accidents.
Hence say whether Avonford Town Council should install the pedestrian crossing.

3 The numbers of customers entering a shop in forty consecutive periods of one minute are given below

3	0	0	1	0	2	1	0	1	1
0	3	4	1	2	0	2	0	3	1
1	0	1	2	0	2	1	0	1	2
3	1	0	0	2	1	0	3	1	2

(i) Draw up a frequency table and illustrate it by means of a vertical line graph.

(ii) Calculate values of the mean and variance of the number of customers

entering the shop in a one minute period.

(iii) Fit a Poisson distribution to the data and comment on the degree of agreement between the calculated and observed values.

4 A machine in a factory produces components continuously. Each day a sample of 20 components are selected and tested. Over a period of 30 days the number of defective components in the sample is recorded as follows.

Number of defectives per sample	0	1	2	3	4	> 4
Number of samples	8	9	8	3	2	0

The quality Control Inspector says that he will stop the production if any sample contains 5 or more defective components.

(i) Find the mean and variance of the number of defectives per sample.

(ii) State whether the data can be modelled by the Poisson distribution.

(iii) Using the Poisson distribution with the mean found in part (i), find the probability that on any one day the quality control inspector will stop the production.

5 In a college, the number of accidents to students requiring hospitalisation in one year of 30 weeks is recorded as follows.

Number of accidents requiring hospitalisation each week	0	1	2	3 or more
Frequency	25	4	1	0

The Principal uses these data to assesses the risk of such accidents.

(i) Is the Poisson distribution a suitable model for this assessment?
State the assumptions that need to be made about the data provided for this to be so.

(ii) Assuming that the Poisson distribution is a suitable model calculate the probability that

(A) In any one week there will be 3 accidents requiring hospitalisation,

(B) In a term of 8 weeks there will be no accidents.