

MEI CONFERENCE 2006

At READING UNIVERSITY

STATISTICS

Fun with the Poisson Distribution

Tuesday 11th July 2006

Philip Stockton
(Member of the MEI Statistics Development Group)
Prince William School
Herne Road
Oundle
PETERBOROUGH
PE8 4BS
Tel: 01832 272881
Email: stockton_phil@hotmail.com

SWIMMING WITH SHARKS

Fancy a holiday in Florida?

What if you saw the headline from the National Post “Increase in shark attacks”?

Is this caused by changing currents?

Or dwindling food supplies?

Or rise in shark-feeding tourist operations?

Or just a statistical blip.

Can Statistics really explain the increase in shark attacks?

The formula for the Poisson distribution is:

$$P(X=r) = \frac{e^{-\lambda} \lambda^r}{r!}$$

for $r = 0, 1, 2, 3, \dots$

The Poisson distribution is useful in many situations and the standard conditions are that events occur

- (i) randomly
 - (ii) independently
- and
- (iii) uniformly over an interval of time

There are many practical situations that can be modelled by a Poisson distribution:

- The number of phone calls on a randomly chosen day
- Insurance claims made by motorists in a given amount of time
- Particles emitted by a radioactive source in a given amount of time
- The number of cars passing in a randomly chosen 10 minute period on a road with no traffic problems, eg no traffic lights
- The number of accidents in a factory per month
- The number of typing errors on a randomly chosen page from a large document

Sometimes these situations may not be appropriate, just because you can calculate a mean rate of something occurring does not mean the Poisson distribution is appropriate.

Consider the following situations.

Insurance claims in a town will not occur randomly after a period of flooding.

The number of cars passing in say a 1-minute would not be uniform if controlled by sets of traffic lights.

Now returning to the example on SHARKS.

If there are an average two shark attacks per summer then the chance of having six shark attacks in the next summer can be calculated by using

$$\lambda = 2 \qquad r = 6$$

Let X be the number of shark attacks in a summer.

$$\begin{aligned} P(X = 6) &= \frac{e^{-2}2^6}{6!} \\ &= 0.01203 \end{aligned}$$

which is a little more than a 1% chance. Six shark attacks are quite unlikely to happen in any one year, though it is likely to happen about once every 85 years.

If you are going to Florida let us try and be a bit more positive.

What is the probability of no attacks?

$$\begin{aligned} \lambda &= 2 \qquad r = 0 \\ P(X = 0) &= \frac{e^{-2}2^0}{0!} \\ &= e^{-2} = 0.13533 \end{aligned}$$

A summer without a shark attack will occur every seven or eight years.

Collecting data in France

We have already seen that the mean of a Poisson distribution with parameter λ is equal to λ .

The Poisson distribution is unusual in that the parameter λ is also equal to the variance.

So the Poisson distribution has equal values of the mean and variance.

This property can help us decide if a Poisson distribution is a suitable model.

Imagine that you are at a small French town aiming to collect some data.

The following data was collected at the entrance to a Tourist Information office in a French town Villeneuve.

x	f
0	18
1	12
2	9
3	4
Totals	43

X is the distribution of the number of people visiting per minute.

Up to 3 people visited per minute, during the data collection of 43 minutes.

It would seem to be reasonable that X would be a Poisson distribution, but let us check the mean and variance of the sample data.

x	f	xf	x^2f
0	18	0	0
1	12	12	12
2	9	18	36
3	4	12	36
Totals	43	42	84

$$\text{Mean} = 0.977$$

$$\begin{aligned}\text{Variance} &= \frac{84 - 43 \times 0.977^2}{42} \\ &= 1.022\end{aligned}$$

Clearly the mean is approximately equal to the variance, so a Poisson model would appear to be a good fit.

If we now want to work out individual probabilities it is normal convention to take λ as the value of the mean.

In this case take $\lambda = 0.977$

$$\begin{aligned}P(X=0) &= \frac{e^{-0.977} 0.977^0}{0!} \\ &= 0.376\end{aligned}$$

$$\begin{aligned}P(X=1) &= \frac{e^{-0.977} 0.977^1}{1!} \\ &= 0.368\end{aligned}$$

$$\begin{aligned}P(X=2) &= \frac{e^{-0.977} 0.977^2}{2!} \\ &= 0.180\end{aligned}$$

$$\begin{aligned}P(X=3) &= \frac{e^{-0.977} 0.977^3}{3!} \\ &= 0.0585\end{aligned}$$

As we had values 0 to 3 in the table you may expect these probabilities to add up to 1.

$$\text{But } P(X \leq 3) = 0.9825$$

You must remember that X follows a Poisson distribution, which is an infinite distribution. If you continue to take S3 this is a very important step.

The expected frequencies can be found by multiplying the probabilities by 43 and the last interval is changed to 3+.

x	Expected frequency
0	16.2
1	15.8
2	7.7
3 or more	3.3
Totals	43

Note the expected frequencies should not be rounded to the nearest integer.

Example:

The following data was collected at the entrance to a Church in a French town Briancon.

x	f
0	23
1	7
2	5
3	5
4	2
5	0
6	2
7	0
8	1
Total	45

X is the distribution of the number of people visiting per minute.

Up to 8 people visited per minute, during the data collection of 45 minutes.

It would seem to be reasonable that X would be a Poisson distribution, but let us check the mean and variance of the sample data.

x	f	xf	x ² f
0	23	0	0
1	7	7	7
2	5	10	20
3	5	15	45
4	2	8	32
5	0	0	0
6	2	12	72
7	0	0	0
8	1	8	64
Total	45	60	240

Mean = 1.333

$$\text{Variance} = \frac{240 - 45 \times 1.333^2}{44}$$

$$= 3.64$$

Clearly the mean is not approximately equal to the variance, so a Poisson model would not appear to be a good fit.

Why could this happen?

In this case it was noted that there were several groups entering the church together. In the tourist information example the numbers of people entering the building were small so appeared to be entering independently. This does not seem the case for visits to the church.

WORLD CUP MATHEMATICS

You could draw a bar chart,

A pie chart,

Even a scatter diagram

Calculate mean, median, mode of:

Fouls

Corners

Possession

Yellow Cards

Shots

Or even goals

But why not do some modelling of a Poisson distribution?

WORLD CUP QUOTES

The World Cup is a truly international event.

Nearly all the Brazilian supporters are wearing yellow shirts – it's a fabulous kaleidoscope of colour.

The goals made such a difference to the way the game went.

The unexpected is always likely to happen.

Is John Motson a better statistician than a commentator?

Does he know all about Poisson events?

Let us look at some world cup data:

Over a long period of time analyse the number of goals per game.

1990 – 2002

232 matches

575 goals

2.4784 goals per game

(Okay even Rooney struggles with 0.4784 goals)

Goals	Games
0	19
1	49
2	60
3	47
4	32
5	18
6 or more	7
Total	232

Mean = 2.4784

Variance = 2.4584

Good agreement, that for a Poisson distribution the mean and the variance are equal.

Going into S3 we can work out the expected number of games using $\lambda = 2.4784$

Goals	Probability	Expected Games
0	0.0839	19.46
1	0.2079	48.23
2	0.2576	59.76
3	0.2128	49.37
4	0.1319	30.60
5	0.0654	15.17
6 or more	0.0406	9.41
		<hr/> 232

A chi-square test accepts the validity of a Poisson fit.

What about this year's World Cup? Try it with $\lambda = 2.4784$

What if you analyse in a different time interval?

Try 45 minutes
 232 matches
 464 intervals of 45 minutes
 575 goals
 1.2392 goals per 45 minutes

Variance = 1.258

Goals	Actual	Expected
0	141	134
1	157	167
2	100	103
3	48	43
4	16	13
5+	2	3

Again seems a good fit. Could do a chi-squared test again.

Data set of Von Borikewicz (1898) for the chance of a Prussian Cavalryman being kicked by the kick of a horse.

Ten Army Corps, over 20 years.

Total deaths is 122.

Mean number of deaths per year per army corps = $\frac{122}{200} = 0.61$

Using this we get

Deaths	P	Expected	Actual
0	0.54335	108.67	109
1	0.33145	66.29	65
2	0.10110	20.22	22
3	0.02055	4.11	3
4	0.00315	0.63	1
5	0.00040	0.08	0
6	0.00005	0.01	0

LIGHTNING STRIKES

1950 – 1998 Average number of deaths is approximately 5

(82% male)

1989 – 1998 Average number of deaths is approximately 3

Contrast with

1852 – 1898 Average number of deaths is approximately 19

(84% male)

1852 45 deaths

1872 46 deaths

1895 43 deaths

Since 1960 only two years when over 10 deaths

1970 11 deaths

1982 14 deaths

More people: the population has trebled since 1850, but there are now less deaths, why????