

Lecture 5 : The Poisson Distribution

Jonathan Marchini

November 10, 2008

1 Introduction

Many experimental situations occur in which we observe the counts of events within a set unit of time, area, volume, length etc. For example,

- The number of cases of a disease in different towns
- The number of mutations in set sized regions of a chromosome
- The number of dolphin pod sightings along a flight path through a region
- The number of particles emitted by a radioactive source in a given time
- The number of births per hour during a given day

In such situations we are often interested in whether the events occur randomly in time or space. Consider the Babyboom dataset we saw in Lecture 2. The birth times of the babies throughout the day are shown in Figure 1. If we divide up the day into 24 hour intervals and count the number of births in each hour we can plot the counts as a histogram in Figure 2. How does this compare to the histogram of counts for a process that isn't random? Suppose the 44 birth times were distributed in time as shown in Figure 3. The histogram of these birth times per hour is shown in Figure 4. We see that the non-random clustering of events in time causes there to be more hours with zero births and more hours with large numbers of births than the real birth times histogram.

This example illustrates that the distribution of counts is useful in uncovering whether the events might occur randomly or non-randomly in time (or space). Simply looking at the histogram isn't sufficient if we want to ask the question whether the events occur randomly or not. To answer this question we need a probability model for the distribution of counts of random events that dictates the type of distributions we should expect to see.

Figure 1: The birth times of the babies in the Babyboom dataset

Figure 2: Histogram of birth times per hour of the babies in the Babyboom dataset

Figure 3: The birth times of the babies in the Babyboom dataset

Figure 4: Histogram of birth times per hour of the non-random data.

2 The Poisson Distribution

The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).

If we let X = The number of events in a given interval,

Then, if the mean number of events per interval is λ

The probability of observing x events in a given interval is given by

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad x = 0, 1, 2, 3, 4, \dots$$

Note e is a mathematical constant. $e \approx 2.718282$. There should be a button on your calculator e^x that calculates powers of e .

If the probabilities of X are distributed in this way, we write

$$X \sim \text{Po}(\lambda)$$

λ is the **parameter** of the distribution. We say X follows a Poisson distribution with parameter λ

Note A Poisson random variable can take on any positive integer value. In contrast, the Binomial distribution always has a finite upper limit.

2.1 Examples

Births in a hospital occur randomly at an average rate of 1.8 births per hour.

What is the probability of observing 4 births in a given hour at the hospital?

Let X = No. of births in a given hour

- (i) Events occur randomly $\Rightarrow X \sim \text{Po}(1.8)$
- (ii) Mean rate $\lambda = 1.8$

We can now use the formula to calculate the probability of observing exactly 4 births in a given hour

$$P(X = 4) = e^{-1.8} \frac{1.8^4}{4!} = 0.0723$$

What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

We want $P(X \geq 2) = P(X = 2) + P(X = 3) + \dots$

i.e. an infinite number of probabilities to calculate

but

$$\begin{aligned}P(X \geq 2) &= P(X = 2) + P(X = 3) + \dots \\&= 1 - P(X < 2) \\&= 1 - (P(X = 0) + P(X = 1)) \\&= 1 - \left(e^{-1.8} \frac{1.8^0}{0!} + e^{-1.8} \frac{1.8^1}{1!} \right) \\&= 1 - (0.16529 + 0.29753) \\&= 0.537\end{aligned}$$

3 The shape of the Poisson distribution

`par(mfrow = c(1, 3)) plot(0:20, dpois(0:20, 3), type = "h", ylim = c(0, 0.25), xlab = "X", main = "Po(3)", ylab = "P(X)", lwd = 3, cex.lab = 1.5, cex.axis = 2, cex.main = 2) plot(0:20, dpois(0:20, 5), type = "h", ylim = c(0, 0.25), xlab = "X", main = "Po(5)", ylab = "P(X)", lwd = 3, cex.lab = 1.5, cex.axis = 2, cex.main = 2) plot(0:20, dpois(0:20, 10), type = "h", ylim = c(0, 0.25), xlab = "X", main = "Po(10)", ylab = "P(X)", lwd = 3, cex.lab = 1.5, cex.axis = 2, cex.main = 2)` Using the formula we can calculate the probabilities for a specific Poisson distribution and plot the probabilities to observe the shape of the distribution. For example, Figure 5 shows 3 different Poisson distributions. We observe that the distributions are

- (i) unimodal
- (ii) exhibit positive skew (that decreases as λ increases)
- (iii) centered roughly on λ
- (iii) the variance (spread) increases as λ increases

4 Mean and Variance of the Poisson distribution

In general, there is a formula for the mean of a Poisson distribution. There is also a formula for the standard deviation, σ , and variance, σ^2 .

If $X \sim \text{Po}(\lambda)$ then

$$\begin{aligned}\mu &= \lambda \\ \sigma &= \sqrt{\lambda} \\ \sigma^2 &= \lambda\end{aligned}$$

5 Changing the size of the interval

Suppose we know that births in a hospital occur randomly at an average rate of 1.8 births per hour.

What is the probability that we observe 5 births in a given 2 hour interval?

Well, if births occur randomly at a rate of 1.8 births per 1 hour interval
Then births occur randomly at a rate of 3.6 births per 2 hour interval

Let Y = No. of births in a 2 hour period

Then $Y \sim \text{Po}(3.6)$

$$P(Y = 5) = e^{-3.6} \frac{3.6^5}{5!} = 0.13768$$

This example illustrates the following rule

If $X \sim \text{Po}(\lambda)$ on 1 unit interval,
then $Y \sim \text{Po}(k\lambda)$ on k unit intervals.

6 Sum of two Poisson variables

Now suppose we know that in hospital A births occur randomly at an average rate of 2.3 births per hour and in hospital B births occur randomly at an average rate of 3.1 births per hour.

What is the probability that we observe 7 births in total from the two hospitals

Figure 5: Three different Poisson distributions.

in a given 1 hour period?

To answer this question we can use the following rule

If $X \sim \text{Po}(\lambda_1)$ on 1 unit interval,
and $Y \sim \text{Po}(\lambda_2)$ on 1 unit interval,
then $X + Y \sim \text{Po}(\lambda_1 + \lambda_2)$ on 1 unit interval.

So if we let $X =$ No. of births in a given hour at hospital A
and $Y =$ No. of births in a given hour at hospital B

Then $X \sim \text{Po}(2.3)$, $Y \sim \text{Po}(3.1)$ and $X + Y \sim \text{Po}(5.4)$

$$\Rightarrow P(X + Y = 7) = e^{-5.4} \frac{5.4^7}{7!} = 0.11999$$

7 Using the Poisson to approximate the Binomial

The Binomial and Poisson distributions are both discrete probability distributions. In some circumstances the distributions are very similar. For example, consider the $\text{Bin}(100, 0.02)$ and $\text{Po}(2)$ distributions shown in Figure 6. Visually these distributions are identical.

In general,

If n is large (say > 50) and p is small (say < 0.1) then a $\text{Bin}(n, p)$ can be approximated with a $\text{Po}(\lambda)$ where $\lambda = np$

The idea of using one distribution to approximate another is widespread throughout statistics and one we will meet again. In many situations it is extremely difficult to use the exact distribution and so approximations are very useful. **Example** Given that 5% of a population are left-handed, use the Poisson distribution to estimate the probability that a random sample of 100 people contains 2 or more left-handed people.

$X =$ No. of left handed people in a sample of 100

$$X \sim \text{Bin}(100, 0.05)$$

Poisson approximation $\Rightarrow X \sim \text{Po}(\lambda)$ with $\lambda = 100 \times 0.05 = 5$

We want $P(X \geq 2)$?

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - \left(P(X = 0) + P(X = 1) \right) \\ &\approx 1 - \left(e^{-5} \frac{5^0}{0!} + e^{-5} \frac{5^1}{1!} \right) \\ &\approx 1 - 0.040428 \\ &\approx 0.959572 \end{aligned}$$

Figure 6: A Binomial and Poisson distribution that are very similar.

If we use the exact Binomial distribution we get the answer 0.96292.

8 Fitting a Poisson distribution

Consider the two sequences of birth times we saw in Section 1. Both of these examples consisted of a total of 44 births in 24 hour intervals.

Therefore the mean birth rate for both sequences is $\frac{44}{24} = 1.8333$

What would be the *expected* counts if birth times were really random i.e. what is the expected histogram for a Poisson random variable with mean rate $\lambda = 1.8333$.

Using the Poisson formula we can calculate the probabilities of obtaining each possible value¹

x	0	1	2	3	4	5	≥ 6
$P(X = x)$	0.15989	0.29312	0.26869	0.16419	0.07525	0.02759	0.01127

Then if we observe 24 hour intervals we can calculate the expected frequencies as $24 \times P(X = x)$ for each value of x .

x	0	1	2	3	4	5	≥ 6
Expected frequency $24 \times P(X = x)$	3.837	7.035	6.448	3.941	1.806	0.662	0.271

We say we have fitted a Poisson distribution to the data.

¹in practice we group values with low probability into one category.

This consisted of 3 steps

- (i) Estimating the parameters of the distribution from the data
- (ii) Calculating the probability distribution
- (iii) Multiplying the probability distribution by the number of observations

Once we have fitted a distribution to the data we can compare the expected frequencies to those we actually observed from the real Babyboom dataset. We see that the agreement is quite good.

x	0	1	2	3	4	5	≥ 6
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	3	8	6	4	3	0	0

When we compare the expected frequencies to those observed from the non-random clustered sequence in Section 1 we see that there is much less agreement.

x	0	1	2	3	4	5	≥ 6
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	12	3	0	2	2	4	1

In Lecture 7 we will see how we can formally test for a difference between the expected and observed counts. For now it is enough just to know how to fit a distribution.