

Lecture 5

The Poisson Distribution

5.1 Introduction

Example 5.1: Drownings in Malta

The book [Mou98] cites data from the St. Luke's Hospital Gazette, on the monthly number of drownings on Malta, over a period of nearly 30 years (355 consecutive months). Most months there were no drownings. Some months there was one person who drowned. One month had four people drown. The data are given as counts of the number of months in which a given number of drownings occurred, and we repeat them here as Table 5.1.

Looking at the data in Table 5.1, we might suppose that one of the following hypotheses is true:

- Some months are particularly dangerous;
- Or, on the contrary, when one person has drowned, the surrounding publicity makes others more cautious for a while, preventing drownings?
- Or, drownings are simply independent events?

How can we use the data to decide which of these hypotheses is true? We might reasonably suppose that the first hypothesis would predict that there would be more months with high numbers of drownings than the independence hypothesis; the second

Table 5.1: Monthly counts of drownings in Malta.

No. of drowning deaths per month	Frequency (No. months observed)
0	224
1	102
2	23
3	5
4	1
5+	0

hypothesis would predict fewer months with high numbers of drownings. The problem is, we don't know how many we should expect, if independence is correct.

What we need is a **model**: A sensible probability distribution, giving the probability of a month having a certain number of drownings, under the independence assumption. The standard model for this sort of situation is called the **Poisson distribution**. ■

The Poisson distribution is used in situations when 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 given 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 (Table 1.2), that we saw in Lecture 1. The birth times of the babies throughout the day are shown in Figure 5.1(a). 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 5.1(b). 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 5.1(c). The histogram of these birth times per hour is shown in Figure 5.1(d). 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.

5.2 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

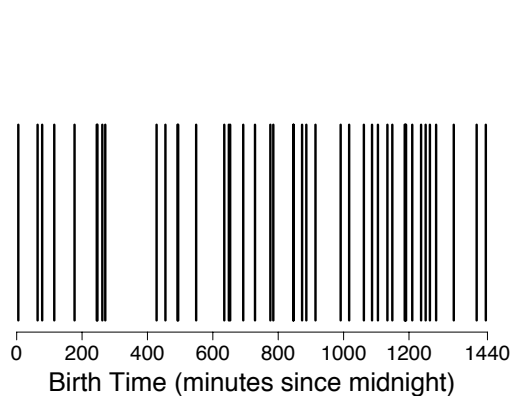
$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad x = 0, 1, 2, 3, 4, \dots$$

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

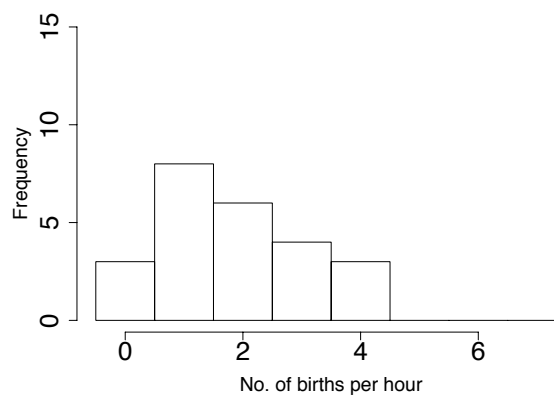
If the probabilities of X are distributed in this way, we write

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

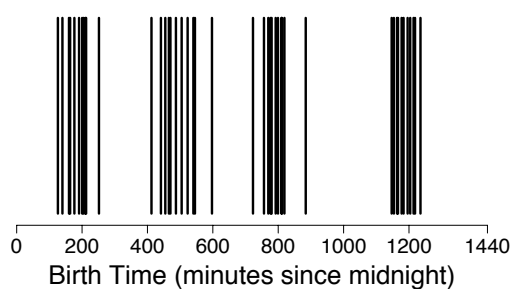
λ is the **parameter** of the distribution. We *say* X follows a Poisson distribution with parameter λ



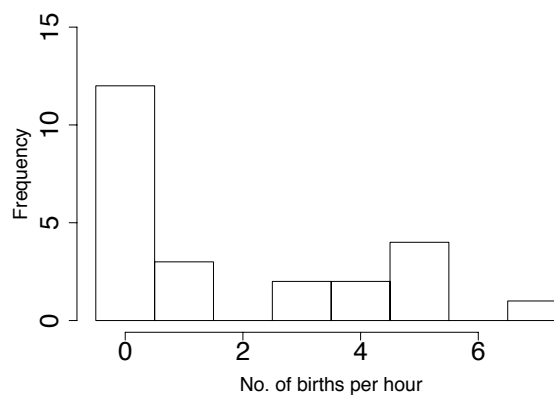
(a) Babyboom data birth times



(b) Histogram of Babyboom birth times



(c) Nonrandom birth times



(d) Histogram of nonrandom birth times

Figure 5.1: Representing the babyboom data set (upper two) and a nonrandom hypothetical collection of birth times (lower two).

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

Example 5.2: Hospital births

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

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

■

Example 5.3: Disease incidence

Suppose there is a disease, whose average incidence is 2 per million people. What is the probability that a city of 1 million people has at least twice the average incidence?

Twice the average incidence would be 4 cases. We can reasonably suppose the random variable $X = \#$ cases in 1 million people has Poisson distribution with parameter 2. Then

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \left(e^{-2} \frac{2^0}{0!} + e^{-2} \frac{2^1}{1!} + e^{-2} \frac{2^2}{2!} + e^{-2} \frac{2^3}{3!} \right) = 0.143.$$

■

5.3 The shape of the Poisson distribution

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 5.2 shows 3 different Poisson distributions. We observe that the distributions

- (i). are unimodal;
- (ii). exhibit positive skew (that decreases as λ increases);
- (iii). are centred roughly on λ ;
- (iv). have variance (spread) that increases as λ increases.

5.4 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

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

5.5 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.

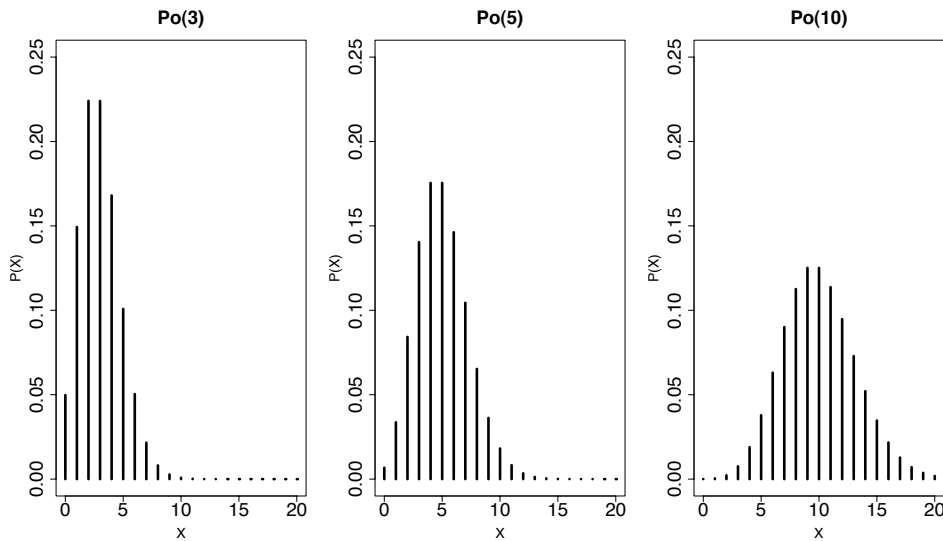


Figure 5.2: Three different Poisson distributions.

5.6 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 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 $X + Y \sim \text{Po}(5.4)$

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

Example 5.4: Disease Incidence, continued

Suppose disease A occurs with incidence 1.7 per million, and disease B occurs with incidence 2.9 per million. Statistics are compiled, in which these diseases are not distinguished, but simply are all called cases of disease “AB”. What is the probability that a city of 1 million people has at least 6 cases of AB?

If $Z = \#$ cases of AB, then $P \sim \text{Po}(4.6)$. Thus,

$$\begin{aligned} P(Z \geq 6) &= 1 - P(Z \leq 5) \\ &= 1 - e^{-4.6} \left(\frac{4.6^0}{0!} + \frac{4.6^1}{1!} + \frac{4.6^2}{2!} + \frac{4.6^3}{3!} + \frac{4.6^4}{4!} + \frac{4.6^5}{5!} \right) \\ &= 0.314. \end{aligned}$$

■

5.7 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.

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

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

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 8 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.

5.8 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 5.3. 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 $\lambda = np$

Example 5.5: Counting lefties

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 $\lambda = 100 \times 0.05 = 5$

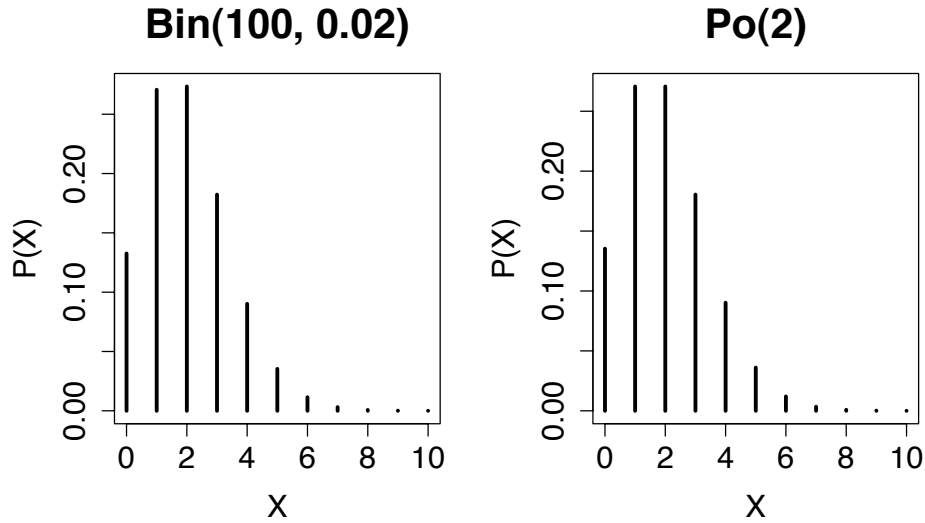


Figure 5.3: A Binomial and Poisson distribution that are very similar.

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

$$\begin{aligned}
 P(X \geq 2) &= 1 - P(X < 2) \\
 &= 1 - \left(P(X = 0) + P(X = 1) \right) \\
 &\approx 1 - \left(e^{-5} \frac{5^0}{0!} + e^{-5} \frac{5^1}{1!} \right) \\
 &\approx 1 - 0.040428 \\
 &\approx 0.9596
 \end{aligned}$$

If we use the exact Binomial distribution we get the answer 0.9629. ■

The idea of using one distribution to approximate another is widespread throughout statistics and one we will meet again. Why would we use an approximate distribution when we actually know the exact distribution?

- The exact distribution may be hard to work with.
- The exact distribution may have too much detail. There may be some features of the exact distribution that are irrelevant to the questions

we want to answer. By using the approximate distribution, we focus attention on the things we're really concerned with.

For example, consider the Babyboom data, discussed in Example 5.2. We said that "random" birth times should yield numbers of births in each hour that are Poisson distributed. Why? Consider the births between 6 am and 7 am. When we say that the births are random, we probably mean something like this: The times are independent of each other, and have equal chances of happening at any time. Any given one of the 44 births has 24 hours when it could have happened. The probability that it happens during *this* hour is $p = 1/24 = 0.0417$. The births between 6 am and 7 am should thus have about the $\text{Bin}(44, 0.0417)$ distribution. This distribution is about the same as $\text{Po}(1.83)$, since $1.83 = 44 \times 0.0417$.

Example 5.6: Drownings in Malta, continued

We now analyse the data on the monthly numbers of drowning incidents in Malta. Under the hypothesis that drownings have nothing to do with each other, and have causes that don't change in time, we would expect the probability the random number X of drownings occur in a month to have a Poisson distribution? Why is that? We might imagine that there are a large number n of people in the population, each of whom has an unknown probability p of drowning in any given month. Then the number of drownings in a month has $\text{Bin}(n, p)$ distribution. In order to use this model, we need to know what n and p are. That is, we need to know the size of the population, which we don't really care about.

On the other hand, the expected (mean) number of monthly drownings is np , and that can be estimated from the observed mean number of drownings. If we approximate the binomial distribution by $\text{Po}(\lambda)$, where $\lambda = np$, then we don't have to worry about

We estimate λ as total number of drownings/number of months. The total number of drownings is $0 \cdot 224 + 1 \cdot 102 + 2 \cdot 23 + 3 \cdot 5 + 4 \cdot 1 = 167$, so we estimate $\lambda = 167/355 = 0.47$. We show the probabilities for the different possible outcomes in the last column of Table 5.2. In the third column we show the *expected* number of months with a given number of drownings, assuming

Table 5.2: Monthly counts of drownings in Malta, with Poisson fit.

No. of drowning deaths per month	Frequency (No. months observed)	Expected frequency Poisson $\lambda = 0.47$	Probability
0	224	221.9	0.625
1	102	104.3	0.294
2	23	24.5	0.069
3	5	3.8	0.011
4	1	0.45	0.001
5+	0	0.04	0.0001

the independence assumption — and hence the Poisson model — is true. This is computed by multiplying the last column by 355. After all, if the probability of no drownings in any given month is 0.625, and we have 355 months of observations, we expect $0.625 \cdot 355$ months with 0 drownings.

We see that the observations (in the second column) are pretty close to the predictions of the Poisson model (in the third column), so the data do not give us strong evidence to reject the neutral assumption, that drownings are independent of one another, and have a constant rate in time. In Lecture 8 we will describe one way of testing this hypothesis formally. ■

Example 5.7: Swine flu vaccination

In 1976, fear of an impending swine flu pandemic led to a mass vaccination campaign in the US. The pandemic never materialised, but there were concerns that the vaccination may have led to an increase in a rare and serious neurological disease, Guillain-Barré Syndrome (GBS). It was difficult to determine whether the vaccine was really at fault, since GBS may arise spontaneously — about 1 person in 100,000 develops GBS in a given year — and the number of cases was small.

Consider the following data from the US state of Michigan: Out of 9 million residents, about 2.3 million were vaccinated. Of

those, 48 developed GBS between July 1976 and June 1977. We might have expected

$$2.3 \text{ million} \times 10^{-5} \text{ cases/person-year} = 23 \text{ cases.}$$

How likely is it that, purely by chance, this population would have experienced 48 cases in a single year? If Y is the number of cases, it would then have Poisson distribution with parameter 23, so that

$$P(Y \geq 48) = 1 - \sum_{i=0}^{47} e^{-23} \frac{23^i}{i!} = 3.5 \times 10^{-6}.$$

So, such an extreme number of cases is likely to happen less than 1 year in 100,000. Does this prove that the vaccine caused GBS?

The people who had the vaccine are people who **chose** to be vaccinated. They may differ from the rest of the population in multiple ways in addition to the elementary fact of having been vaccinated, and some of those ways may have predisposed them to GBS. What can we do? The paper [BH84] takes the following approach: If the vaccine were not the cause of the GBS cases, we would expect no connection between the timing of the vaccine and the onset of GBS. In fact, though, there seemed to be a particularly large number of cases in the six weeks following vaccination. Can we say that this was more than could reasonably be expected by chance?

The data are given in Table 5.3. Each of the 40 GBS cases was assigned a time, which is the number of weeks after vaccination when the disease was diagnosed. (Thus “week 1” is a different calendar week for each subject.) If the cases are evenly distributed, the number in a given week should be Poisson distributed with parameter $40/30 = 1.33$. Using this parameter, we compute the probabilities of 0, 1, 2, ... cases in a week, which we give in row 3 of Table 5.3. Multiplying these numbers by 40 gives the expected frequencies in row 4 of the table. It is clear that the observed and expected frequencies are very different. One way of seeing this is to consider the standard deviation. The Poisson distribution has SD $\sqrt{1.33} = 1.15$ (as discussed in section 5.4,

while the data have SD

$$s = \sqrt{\frac{1}{30-1} \left(16 \cdot (0 - 1.33)^2 + 7 \cdot (1 - 1.33)^2 + 3 \cdot (2 - 1.33)^2 + 2 \cdot (4 - 1.33)^2 + 1 \cdot (9 - 1.33)^2 + 1 \cdot (10 - 1.33)^2 \right)} = 2.48.$$

Table 5.3: Cases of GBS, by weeks after vaccination

# cases per week	0	1	2	3	4	5	6+
observed frequency	16	7	3	0	2	0	2
probability	0.264	0.352	0.234	0.104	0.034	0.009	0.003
expected frequency	10.6	14.1	9.4	4.1	1.4	0.4	0.1

■