Lecture 5: The Poisson distribution

11th of November 2015

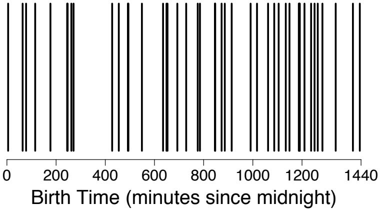
Many experimental situation occur in which we observe the counts of events within a set unit of time, area, volume, length etc. For example,

- The number of cases of a disease in different towns;
- The number of mutations in 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.

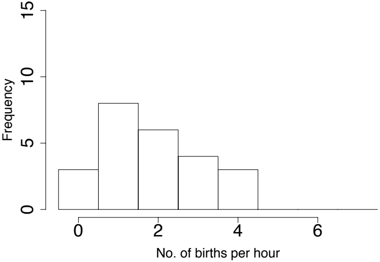
In such situations we are often interested in whether the events occur randomly in time or space, or not.

Example: baby boom dataset

Consider the Babyboom dataset, that we saw in the first lecture,

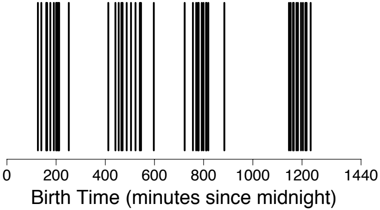


and the histogram of these birth times per hour.

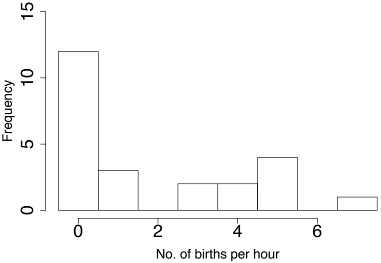


How does this compare to the histogram of counts for a process that isn't random?

Suppose the 44 birth times were distributed in time as shown here.



Remark: there are more hours with zero births and more hours with large numbers of births than the real birth times histogram.



Does birth occur randomly in time?

- Simply looking at the histogram isn't sufficient 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.

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

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

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

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

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

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.

Example: 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

$$\left.\begin{array}{l}\text{(i) Events occur randomly} \\ \text{(ii) Mean rate } \lambda = 1.8\end{array}\right. \Rightarrow X \sim \mathrm{Po}(1.8)$$

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

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

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

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

Example: 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

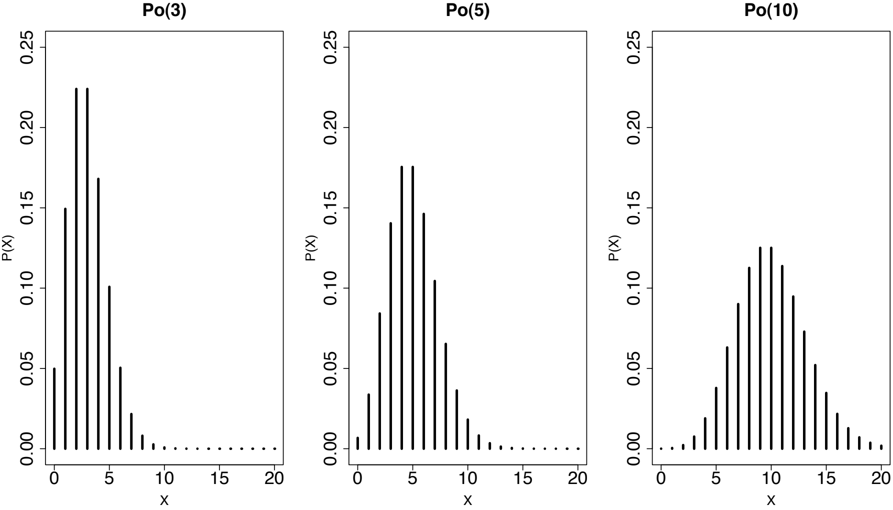
$$X = \text{number of cases in 1 million people}$$

has Poisson distribution with parameter 2.

Then

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

The shape of the Poisson distribution



The shape of the Poisson distribution

We observe that the Poisson distributions

1. are unimodal;
2. exhibit positive skew (that decreases as λ increases);
3. are centred roughly on λ;
4. have variance (spread) that increases as λ increases.

Mean and Variance of the Poisson distribution

The expected mean μ and the expected standard deviation, σ of a Poisson are as follows:

If X $\sim$ Po(λ) then

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

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 $Y \sim$ Po(3.6)

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

Changing the size of the interval

This example illustrates the following rule

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

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
- 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

$$\boxed{\begin{array}{ll} \text{If} & X \sim \text{Po}(\lambda_1) \text{ on 1 unit interval,} \\ \text{and} & Y \sim \text{Po}(\lambda_2) \text{ on 1 unit interval,} \\ \text{then} & X + Y \sim \text{Po}(\lambda_1 + \lambda_2) \text{ on 1 unit interval.} \end{array}}$$

So if we let

$$X = \text{No. of births in a given hour at hospital A}$$

and

$$Y = \text{No. of births in a given hour at hospital B}$$

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

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

Example: Disease Incidence

Suppose

- disease A occurs with incidence 1.7 per million,
- 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=Number of cases of AB, then $P \sim Po(4.6)$. Thus,

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

Fitting a Poisson distribution

Consider the two sequences of birth times we saw at the beginning. 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 value.

In practice we group values with low probability into one category.

x	0	1	2	3	4	5	≥ 6
$P(X = x)$	0.159	0.293	0.268	0.164	0.075	0.027	0.011

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 freq.	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.

Fitting a Poisson distribution

This consists of 3 steps

1. Estimating the parameters of the distribution from the data
2. Calculating the probability distribution
3. 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.

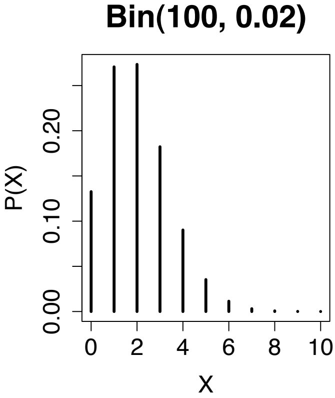
x	0	1	2	3	4	5	≥ 6
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	3	8	6	4	3	0	0

When we compare the expected frequencies to those observed from the non-random clustered sequence in slide 4 we see that there is much less agreement.

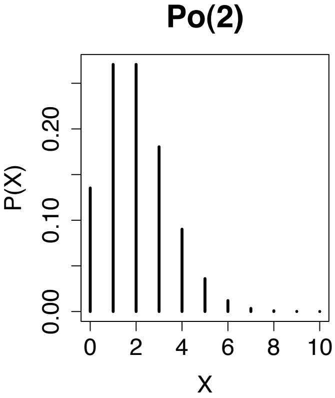
x	0	1	2	3	4	5	≥ 6
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	12	3	0	2	2	4	1

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.



Bin(100, 0.02)



Po(2)

Using the Poisson to approximate the Binomial

In general,

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

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

Examples: Drownings in Malta

The data are given as counts of the number of months in which a given number of drownings occurred.

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

Do these drowning events occur randomly in time?

Assume these events are independent and occur randomly in time.

Notations:

- We imagine 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 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.

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 $Po(\lambda)$, where $\lambda = np$, then we don't have to worry about the size of the population.

We compute the probabilities for the different possible outcomes assuming the independence assumption — and hence the Poisson model– with mean

$$\lambda = \frac{\text{total number of drownings}}{\text{number of months}} = 0.47$$

No. of drowning 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 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.