

Lecture 5 : The Poisson Distribution

Jonathan Marchini

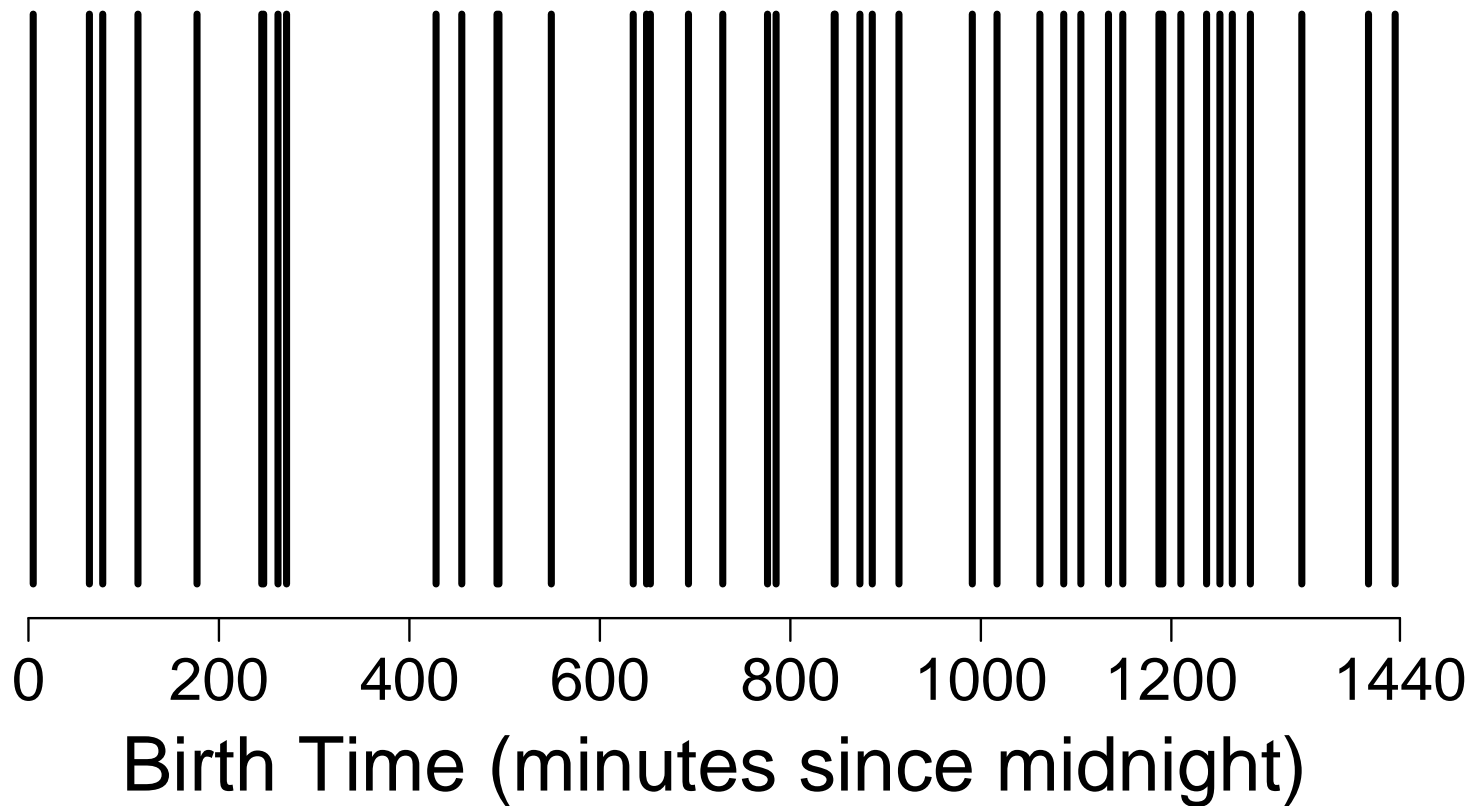
Random events in time and space

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

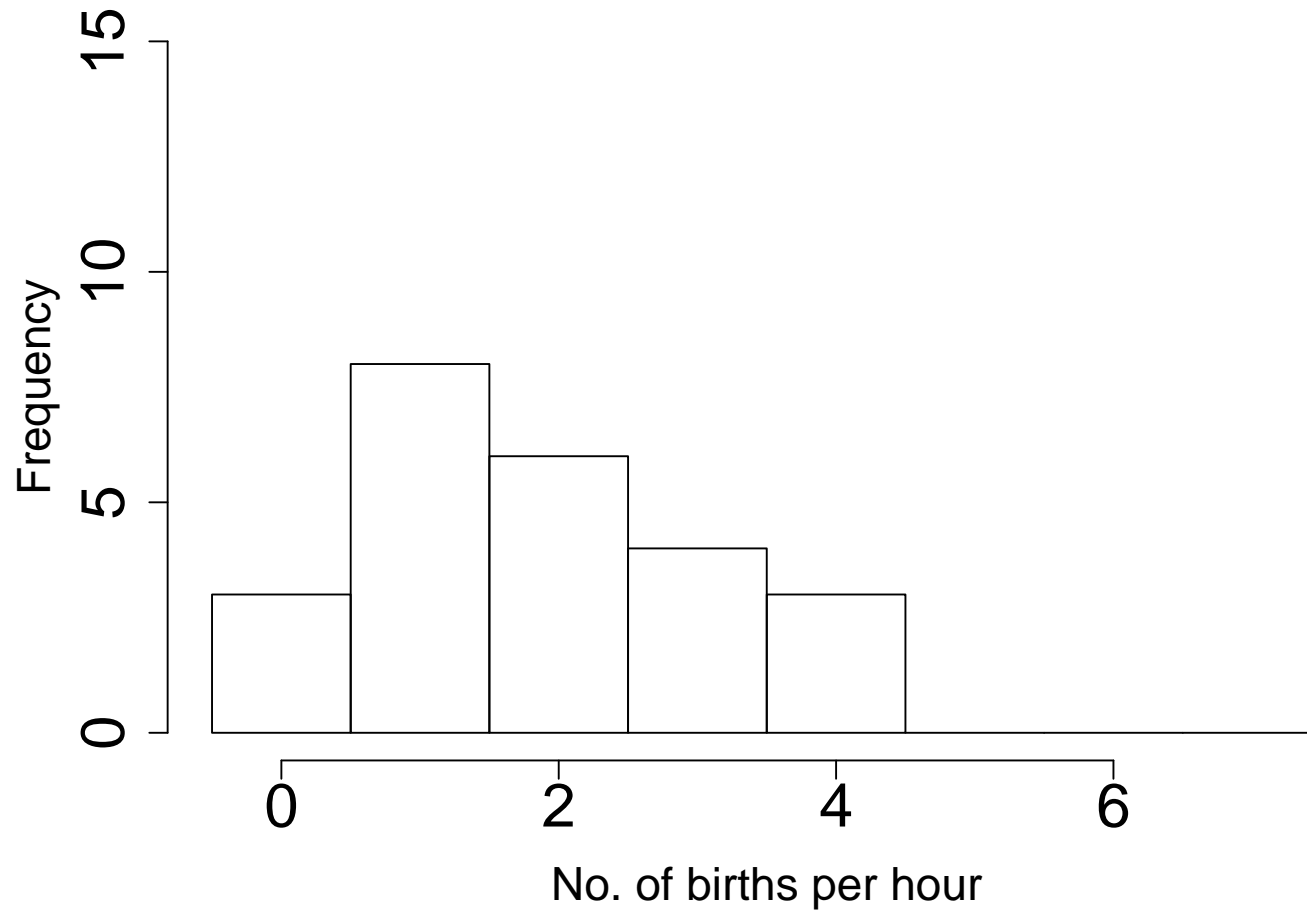
- The number of cases of a disease in different towns
- The number of mutations in set sized regions of a chromosome

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

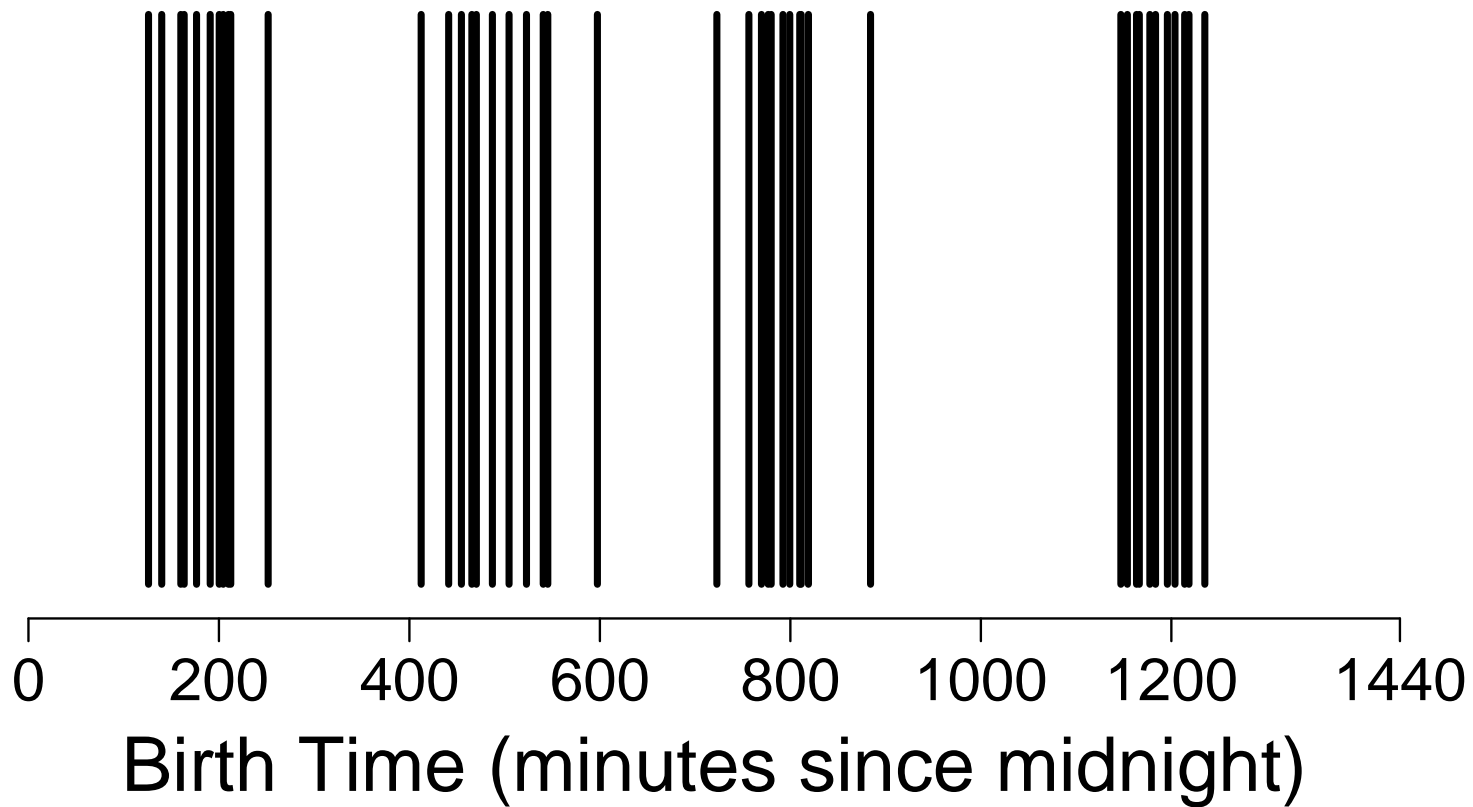
Consider the birth times from the Babyboom dataset we saw in Lecture 2.



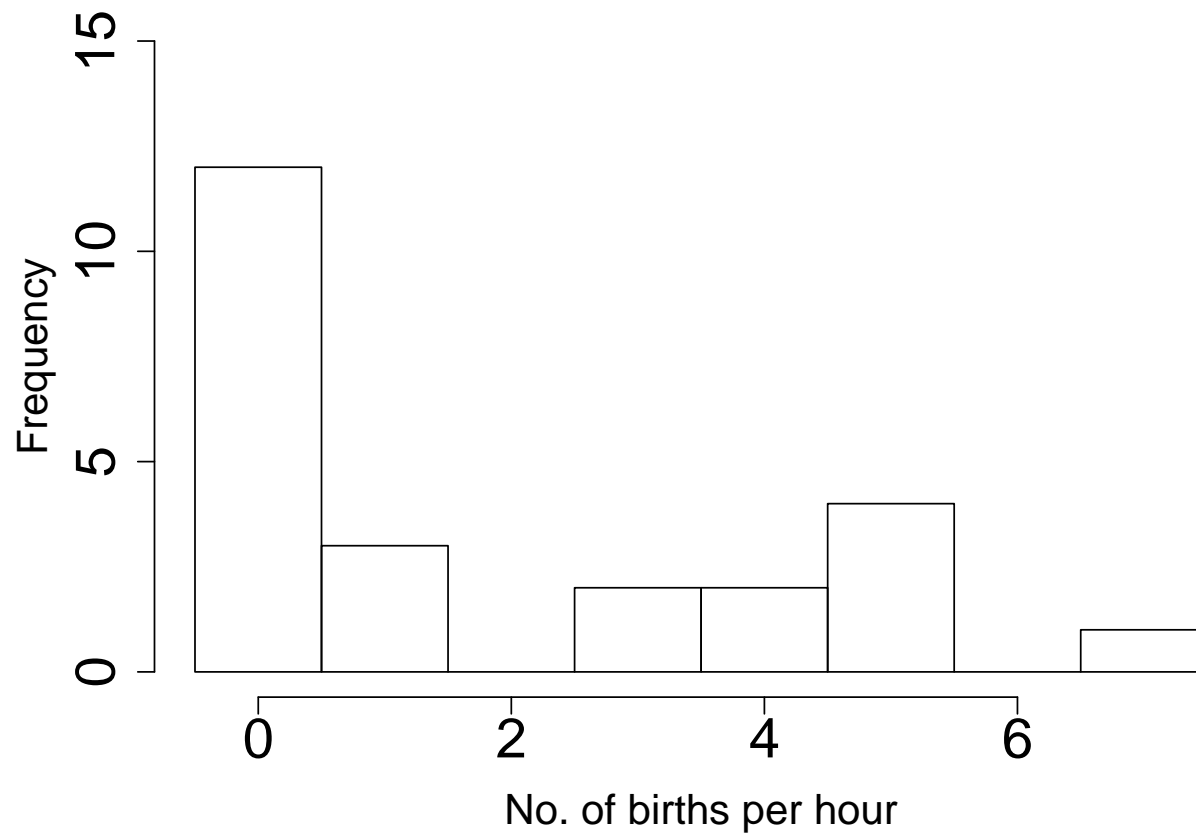
We can plot a histogram of births per hour.



Consider the following sequence of birth times that are obviously not random.



We observe a very different pattern in the histogram of these birth times per hour.



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.

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).

X = The number of events in a given interval,

λ = mean number of events per interval

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

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

Example 1

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

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

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

Example 2

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

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

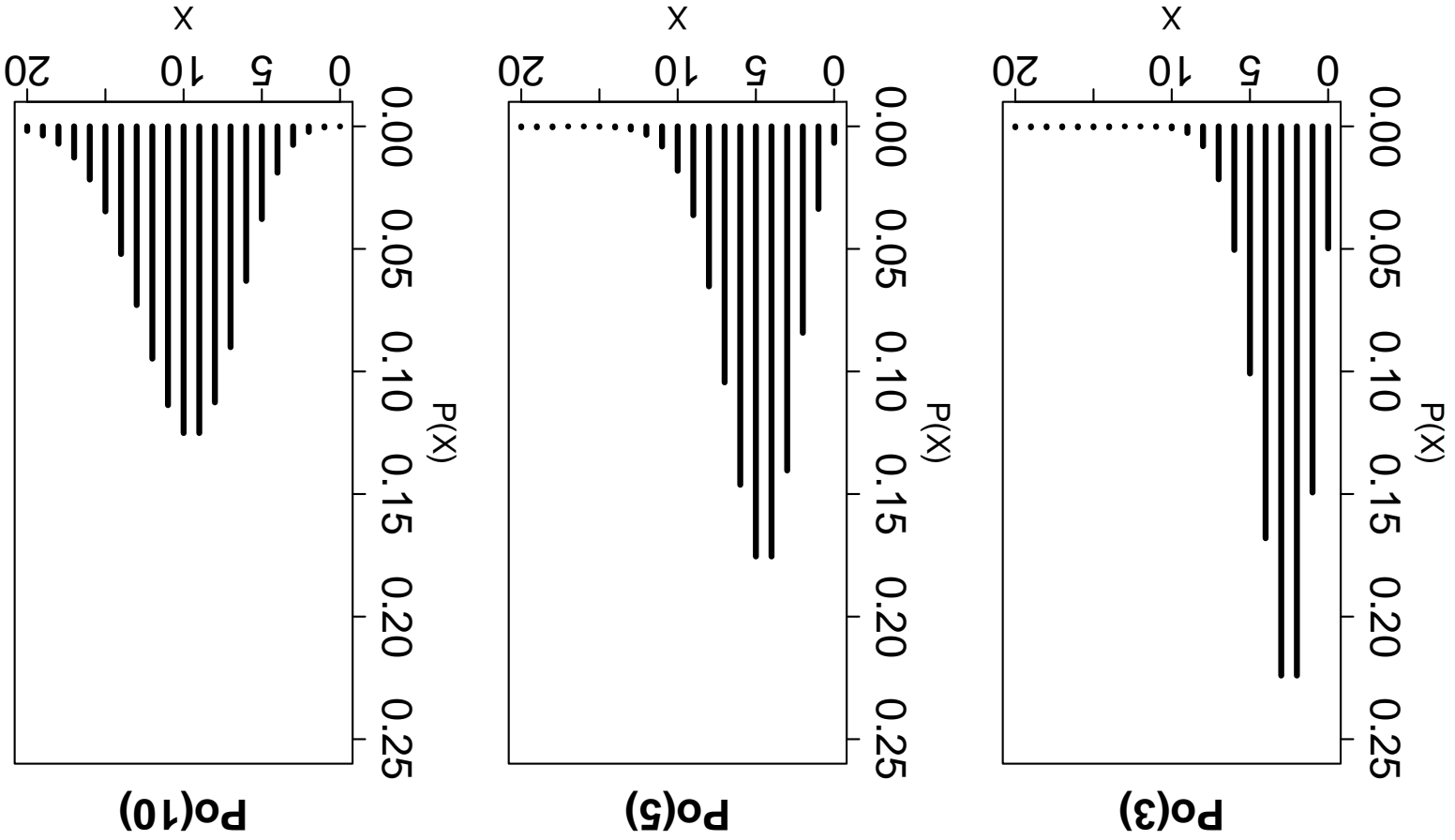
Example 3

Jinkinson and Slater (1981) observed the nature of queues at a London Underground station. They counted the number of women present in queues of length ten. The data for 100 such queues are presented below:

No. women	0	1	2	3	4	5	6	7	8	9	10
Frequency	1	3	4	23	25	19	18	5	1	1	0

Which is model is appropriate for this data
(a) Binomial, or (b) Poisson?

Shape of the Poisson



Poisson 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

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

$$\mu = \lambda$$

$$\sigma = \sqrt{\lambda}$$

$$\sigma^2 = \lambda$$

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?

1.8 births per 1 hour interval

⇒

3.6 births per 2 hour interval

Let $Y = \text{No. of births in a 2 hour period}$

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

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

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.

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.

$X =$ No. of births in a given hour at hospital A

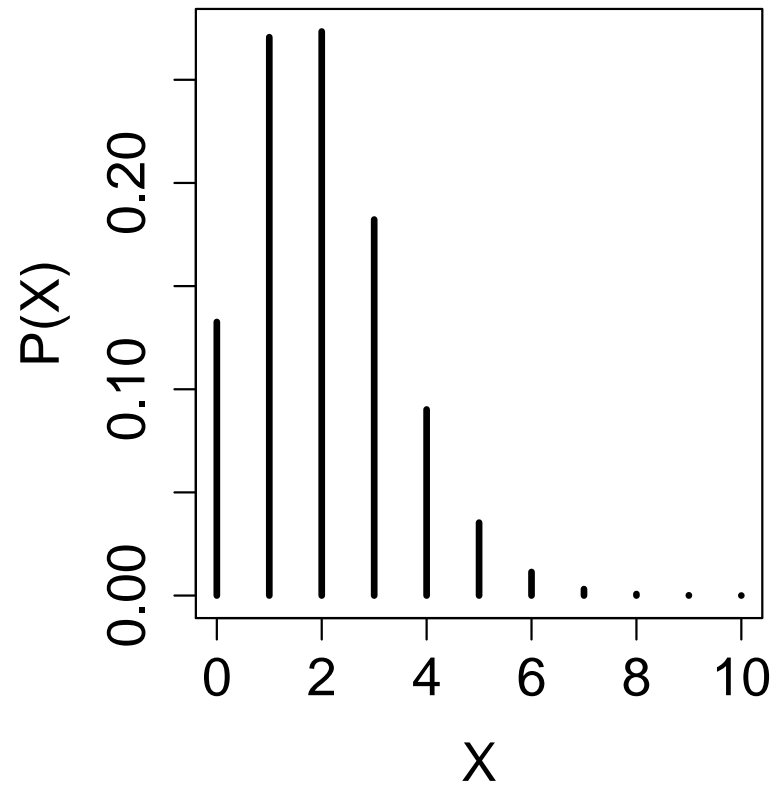
$Y =$ No. of births in a given hour at hospital B

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

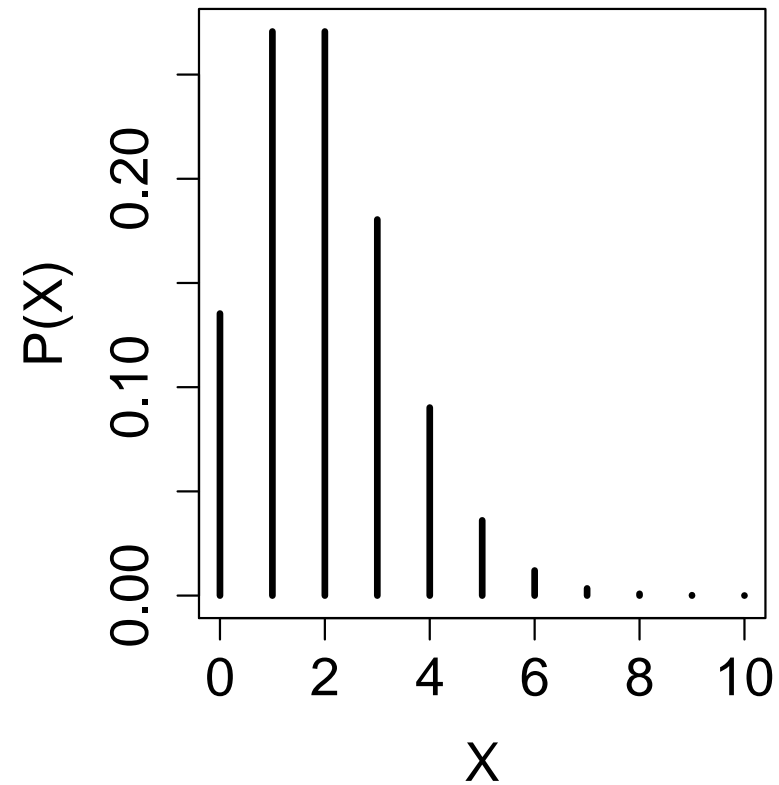
$$\Rightarrow P(X + Y = 7) = \mathbf{e}^{-5.4} \left(\frac{5.4^7}{7!} \right) = 0.11999$$

Using the Poisson to approximate the Binomial

Bin(100, 0.02)



Po(2)



In general,

If n is large (say > 50) and p (say < 0.1) is small 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 out 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 $P(X \geq 2)$?

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

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

Fitting a Poisson distribution

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

x	0	1	2	3	4	5	≥ 6
$P(X = x)$	0.159	0.293	0.269	0.164	0.075	0.028	0.011

Note We group all of the values ≥ 6 together because they have very small probabilities of occurring.

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 $24 \times P(X = x)$	3.84	7.04	6.45	3.94	1.81	0.66	0.27

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

Fitting discrete distributions

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.

x	0	1	2	3	4	5	≥ 6
Expected	3.84	7.04	6.45	3.94	1.81	0.66	0.27
Observed	3	8	6	4	3	0	0

The agreement is quite good.

For the second sequence of birth times there is much less agreement.

x	0	1	2	3	4	5	≥ 6
Expected	3.84	7.04	6.45	3.94	1.81	0.66	0.27
Observed	12	3	0	2	2	4	1

In Lecture 7 we will see how to formally test this discrepancy.