Lecture 17 (Chapter 12): The Poisson distribution

Corinne Riddell (Instructor: Alan Hubbard and Tomer Altman)

October 9, 2024

Learning objectives for today

- Learn what is a Poisson random variable and the types of random phenomena they represent
- Calculate probabilities by hand using the probability distribution function for Poisson random variables
- Calculate probabilities using the R functions `ppois` and `dpois` and know which one is used to compute exact vs. cumulative probabilities

Readings

- Chapter 12 of Baldi and Moore
- Online resource: Poisson Distribution in R

The Poisson distribution

- Last class we covered the binomial distribution
- Ch.12 covers another distribution for discrete variables: the Poisson distribution
- The main distinction between the Binomial and the Poisson distributions is that Poisson random variables have no upper bound, whereas the upper bound of a binomial random variable X is n, the size of the sample
- The most common usage of the Poisson distribution is to model rare events

The Poisson distribution

A Poisson distribution describes the count of occurrences of a defined event in fixed, finite intervals of time or space where:

1. Occurrences are all independent. That is, knowing that one event has occurred does not change the probability that another event may occur.
2. The probability of an occurrence is the same over all possible intervals of the same size

This count is denoted by the random variable X, which, as a count, can take the values 0, 1, 2, and so on, with no upper bound.

Example 1 of the Poisson distribution

The Poisson distribution can be used to model rare, but infectious diseases. For example, the number of deaths X attributed to typhoid fever over 100 years follows a Poisson distribution if:

a) The probability of a new death from typhoid fever in any one day is very small
b) The number of cases reported in any two distinct periods of time are independent random variables

Each year, the number of deaths from typhoid fever in the US could be recorded. This sequence of deaths over time might look like this: 0, 1, 0, 0, 1, 1, 0, 2, and so on.

Citation: https://ani.stat.fsu.edu/~debdeep/p4_s14.pdf

Example 2 of the Poisson distribution

The Poisson distribution can also be used to model rare events occurring on a surface area. For example, the distribution of number of bacterial colonies growing on an agar plate. The number of bacterial colonies over the entire agar plate follows a Poisson distribution if:

a) The probability of finding any bacterial colonies in a small area is very small
b) The events of finding bacterial colonies in any two areas are independent

The agar plates surface can be divided into several small areas. For each area, you could count the number of bacterial colonies and record this information in a variable in R and it might look like this: 0, 1, 0, 0, 1, 1, 0, 2, and so on.

Citation: https://ani.stat.fsu.edu/~debdeep/p4_s14.pdf

Poisson probabilities

X follows the Poisson distribution with a parameter of μ, which is the mean number of occurrences per interval. The possible values of X are 0, 1, 2, and so on. If k is any one of these values, then the probability that X is equal to k can be written as:

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}$$

The above formula is the probability distribution function for a Poisson distribution.

For example,

$$P(X = 2) = \frac{e^{-\mu}\mu^2}{2!}$$

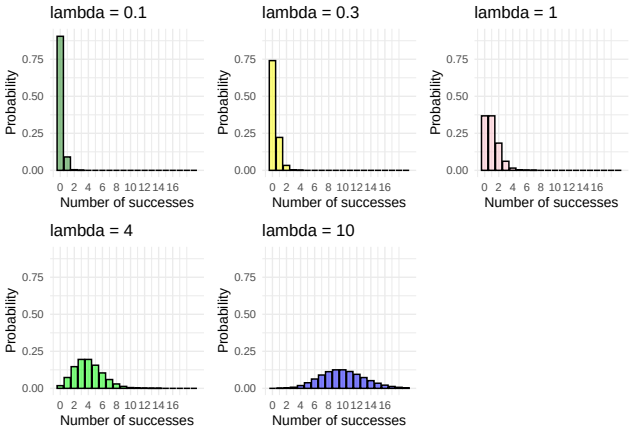
will calculate the probability of observing two events when X follows a Poisson distribution with an average of μ, or $X \sim Pois(\mu)$

- How would you calculate $P(X < 3)$?
- How would you calculate $P(X >= 2)$?

Mean and standard deviation (SD) of a Poisson random variable

- The mean of a Poisson random variable is equal to μ
- The variance is also equal to μ, and thus the SD is equal to $\sqrt{\mu}$
- When the mean is large, so is the SD, and this makes for a flat and wide probability distribution
- Poisson distributions are most commonly used to describe rare, random events (or random events examined over small time intervals).
- The parameter of the Poisson distribution is often called `lambda` (λ), instead of μ
- In R, the function to calculate $P(X = x)$ is `dpois(x=?, lambda=?)`

Probability distribution of a Poisson random variable



Example: Mumps

In Iowa, the average monthly number of reported cases of mumps per year is 0.1. If we assume that cases of mumps are random and independent, the number X of monthly mumps cases in Iowa has approximately a Poisson distribution with $\mu = 0.1$. The probability that in a given month there is no more than 1 mumps case is:

$P(X \leq 1) = P(X = 0) + P(X = 1)$

$= \frac{e^{-0.1}0.1^0}{0!} + \frac{e^{-0.1}0.1^1}{1!}$ (note that $0! = 1$, by definition, and $x^0 = 1$, for any value of x)

$= 0.9048 + 0.0905 = 0.9953$

Thus, we expect to see exactly 0 cases in 90.5% of the months, and exactly 1 case in 9.05% of the months. Overall, in 99.5% of the months, we expect to see less than 2 cases.

Example: Mumps calculated using R using `ppois()` and `dpois()`

- `ppois(q = y, lambda = 0.1)` is the probability of y events **or less**. Thus `ppois` returns a cumulative probability of the lower tail by default.
- `dpois(x = y, lambda = 0.1)` is the probability of **exactly** y events.

So to calculate the probability of one event or less we could use `ppois()` like this:

```
ppois(q = 1, lambda = 0.1) # notice that lambda is the parameter
```

```
## [1] 0.9953212
```

Or `dpois()` twice, like this:

```
dpois(x = 0, lambda = 0.1) + dpois(x = 1, lambda = 0.1)
```

```
## [1] 0.9953212
```

Example: Mumps, continued

Suppose you saw 4 cases of Mumps in a given month. What are the chances of seeing 4 or more cases in any given month?

```
1 - ppois(q = 3, lambda = 0.1) #careful, we used q = 3 here, why 3 and not 4?
```

```
## [1] 3.846834e-06
```

Could you have performed this calculation using `dpois()`?

According to our calculation, seeing 4 or more cases in any given month is extremely unlikely. This suggests a substantial departure from the model, suggesting a contagious outbreak where the assumption of independence of events is no longer met. At this point, the data would no longer be Poisson distributed because in future months you might expect more and more cases of Mumps (exponential growth), which is not modelled by a Poisson random variable.

Example: Polydactyly

In the US, 1 in every 500 babies is born with an extra finger or toe. These events are random and independent. Suppose that the local hospital delivers an average of 268 babies per month. This means that for each month we expect to see 0.536 babies born with an extra finger or toe at that hospital (how do you calculate 0.536 here?). Let X be the count of babies born with an extra finger or toe in a month at that hospital.

a) What values can X take?
b) What distribution might X follow?
c) Give the mean and standard deviation of X

Example: Polydactyly, continued

To get a sense of what the data might look like, use R to simulate data across five years (60 months) for this hospital.

```
rpois(n = 12*5, lambda = 0.536)
```

```
##  [1] 4 0 0 0 1 0 0 1 0 0 1 2 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 1 0 0 2 0 0 0
## [39] 1 2 0 0 0 1 0 0 1 1 0 1 0 0 0 2 1 0 0 1 1 0
```

More random number generation

Examining a stream of Poisson-distributed random numbers helps us get a sense of what these data look like. Can you think of a variable that might be Poisson-distributed according to one of these distributions?

```
rpois(100, lambda = 0.1)
```

```
##   [1] 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0
```

```
rpois(100, lambda = 0.5)
```

```
##   [1] 0 1 0 0 0 1 0 0 2 1 1 0 0 2 0 2 0 0 0 2 1 0 0 1 1 1 0 1 1 1 0 0 0 2 2 0 1
##  [38] 0 2 0 0 0 1 0 1 0 1 0 0 2 1 0 1 1 0 2 0 1 1 1 0 1 2 0 1 0 0 0 1 0 0 0 0 0
##  [75] 0 0 0 0 1 1 0 1 0 0 0 3 0 1 1 0 0 0 0 0 0 1 0 0 0 0
```

```
rpois(100, lambda = 1)
```

```
##   [1] 2 0 0 1 0 0 0 2 0 1 1 0 2 3 1 0 3 1 0 0 1 1 1 2 0 0 0 0 0 0 1 4 0 1 0 1 0
##  [38] 2 3 1 1 1 1 1 2 1 2 2 1 1 1 1 1 0 1 2 1 1 0 0 0 1 1 1 0 2 0 0 2 0 1 3 1 0
##  [75] 0 0 0 0 0 0 0 0 0 1 0 2 2 0 1 0 1 3 1 0 3 1 2 3 1 2 1
```