Statistics for Computing MA4413

Lecture 8

The Poisson Distribution

Kevin Burke

kevin.burke@ul.ie

Poisson Approximation

We saw in the previous lecture that the binomial probability function is $\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$.

However, **when n is large and p is small**, this formula can present computational difficulties.

In this case we can use the **Poisson approximation**. Letting $\lambda = n\,p$:

$$\boxed{\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{\lambda^x}{x\,!} e^{-\lambda}}.$$

This approximation works well when $n > 20$ and $p < 0.1$.

Note: λ is the Greek letter “lambda”.

Example: Rare Disease

Let's assume that a rare disease affects 0.1% of all individuals. What is the probability that 10 individuals in a group of 5000 have this disease?

Let X = "the number of individuals who have the disease". Clearly this has a binomial distribution: $X \sim \text{Binomial}(n = 5000, p = 0.001)$

$$\Rightarrow \Pr(X = 10) = \binom{5000}{10} 0.001^{10}\, 0.999^{4990} = 0.018.$$

Since $n > 20$ and $p < 0.1$, we can use the *Poisson approximation* with $\lambda = n\,p = 5000(0.001) = 5$

$$\Rightarrow \Pr(X = 10) \approx \frac{5^{10}}{10\,!} e^{-5} = 0.018.$$

Example: Rolling Two Dice

Consider the experiment of rolling two dice and adding the numbers showing. If repeated 30 times, what is the probability that on 2 occasions the sum is equal to 3?

You should be able to calculate the probability of getting a sum of 3 in one trial: $p = \frac{2}{36} = \frac{1}{18}$.

Repeating 30 times leads to $X \sim \text{Binomial}(n = 30, p = \frac{1}{18})$

$$\Rightarrow \Pr(X = 2) = \binom{30}{2} \left(\frac{1}{18}\right)^2 \left(\frac{17}{18}\right)^{28} = 0.2709.$$

Using the *Poisson approximation* with $\lambda = n\,p = 30 \times \frac{1}{18} = \frac{30}{18} = \frac{5}{3}$

$$\Rightarrow \Pr(X = 2) \approx \frac{(\frac{5}{3})^2}{2\,!} e^{-\frac{5}{3}} = 0.2623.$$

Question 1

Using both the binomial probability function, $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, and the Poisson approximation, $p(x) \approx \frac{\lambda^x}{x!} e^{-\lambda}$ where $\lambda = np$, evaluate each of the following:

a) $\Pr(X = 2)$ when $X \sim \text{Binomial}(n = 20, p = 0.1)$.

b) $\Pr(X = 5)$ when $X \sim \text{Binomial}(n = 100, p = 0.02)$.

c) $\Pr(X = 3)$ when $X \sim \text{Binomial}(n = 1000, p = 0.005)$.

d) $\Pr(X = 1)$ when $X \sim \text{Binomial}(n = 10000, p = 0.0001)$.

Poisson Distribution

The Poisson approximation is useful. However, the **Poisson distribution** is an important probability distribution in its own right.

It is the probability distribution for *the number of events occurring within an interval* of time / area / volume etc.

For example, the number of:

- System crashes per year.
- Text messages received per hour.
- Tasks processed by a CPU per hour.
- Flaws in a sheet of metal per m^2.
- Typographical errors per page.

Note: the events *must occur independently within the interval*, i.e., the occurrence of one event has no effect on another.

Why the Poisson Distribution Arises

Assume that a system crashes on average λ times per year.

Think about the *precise moment in time* of one such crash.

- This is the result of a Bernoulli trial with $\{1 = \text{crash}, 0 = \text{work}\}$ which has generated the outcome $1 = \text{crash}$.

Now think of the *whole year*.

- Throughout the year we observe the results of a *sequence of identical Bernoulli trials* where $p = \Pr(\text{crash})$.
- Let n be the total number of Bernoulli trials during this period, i.e., there is a trial for *every single* precise moment in time.

Why the Poisson Distribution Arises

Assuming that these Bernoulli trials are *independent*, we know that the number of crashes, X, has a binomial distribution (see Lecture7):

$$X \sim \text{Binomial}(n, p).$$

What can we say about the value of n? Think - how many precise moments in time are there in the year (or any period of time)?

What about p? Think - if we pick some moment in time, what is the likelihood that the system crashes at *exactly* that moment in time?

Why the Poisson Distribution Arises

Since time is a *continuous* quantity, there are an *infinite* number of possible times that the system can crash, i.e., $n = \infty$.

For any moment in time, the probability that the system crashes at *exactly* that moment is very low, i.e., $p \approx 0$.

Since ***n* is large and *p* is small** we know from earlier that

$$\text{Binomial}(n, p) \to \text{Poisson}(\lambda).$$

Here the *average number of events per interval* is $\lambda = n\,p$.

The same arguments hold for intervals of area / volume etc. Thus, the Poisson distribution arises naturally in a variety of situations.

Poisson Distribution

The **Poisson distribution** is used for calculating the probability of x events within an interval of time / area / volume etc.:

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

$$\boxed{\Pr(X = x) = \frac{\lambda^x}{x\,!}e^{-\lambda}}$$

where $\boxed{x \in \{0, 1, 2, \ldots, \infty\}}$

$$\boxed{E(X) = \lambda}$$

$$\boxed{Var(X) = \lambda}$$

Example: System Crashes

Let's assume that a system crashes on average three times per year $\Rightarrow \lambda = 3$ and $\Pr(X = x) = \frac{\lambda^x}{x\,!} e^{-\lambda} = \frac{3^x}{x\,!} e^{-3}$.

What is the probability that:

. . . there are no crashes in a year?

$$\Pr(X = 0) = \frac{3^0}{0\,!} e^{-3} = \frac{1}{1} e^{-3} = 0.0498.$$

. . . at least one crash in a year?

$$\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - 0.0498 = 0.9502.$$

Example: System Crashes

...less than 2 crashes in a year? (since X is discrete "< 2" means "≤ 1")

$$\begin{aligned}
\Pr(X < 2) = \Pr(X \leq 1) &= p(0) + p(1) \\
&= \frac{3^0}{0\,!} e^{-3} + \frac{3^1}{1\,!} e^{-3} \\
&= 0.0498 + 0.1494 = 0.1992.
\end{aligned}$$

...between 4 and 6 crashes in a year?

$$\begin{aligned}
\Pr(4 \leq X \leq 6) &= p(4) + p(5) + p(6) \\
&= \frac{3^4}{4\,!} e^{-3} + \frac{3^5}{5\,!} e^{-3} + \frac{3^6}{6\,!} e^{-3} \\
&= 0.1680 + 0.1008 + 0.0504 = 0.3192.
\end{aligned}$$

Poisson vs Binomial

- Binomial(n, p)
  - You have the total number of Bernoulli trials, n, and the probability of an event in each trial, p.
  - $X \in \{0, 1, 2, \ldots, n\}$, i.e., the maximum number of events is n since there are n trials.
  - Usage: probability of 1 event in 4 trials, less than 2 events in 6 trials, no events in 3 trials etc.
- Poisson(λ)
  - You have the average number of events, λ, within a fixed interval of time / area / volume etc.
  - $X \in \{0, 1, 2, \ldots, \infty\}$, i.e., there is no upper limit for X since there are an infinite number of Bernoulli trials throughout the interval.
  - Usage: probability of 1 event per interval, less than 2 events per interval, no events per interval etc.

Different Time-frame

Note the following:

- λ is the average number of events per 1 interval.
- $\lambda \times 2$ is the average number of events per 2 intervals.
- $\lambda \times 3$ is the average number of events per 3 intervals.
- $\lambda \times 0.25$ is the average number of events per 0.25 intervals.
- ...etc.

In general:

- $\boxed{\lambda t \text{ is the average number of events per } t \text{ intervals}}$.

$\Rightarrow$ **the number of events per t intervals has a Poisson(λt) distribution.**

Example: System Crashes

We said that there are $\lambda = 3$ crashes per year.

What is the probability of no crashes in 2 years? $\Rightarrow \lambda t = 3(2) = 6$ crashes on average per 2 years.

$$\Pr(X = 0) = \frac{6^0}{0\,!} e^{-6} = \frac{1}{1} e^{-6} = 0.0025.$$

What is the probability of more than 2 crashes in 6 months (i.e., 0.5 years)? $\Rightarrow \lambda t = 3(0.5) = 1.5$ crashes on average per 0.5 years.

$$\begin{aligned}\Pr(X > 2) = 1 - \Pr(X \leq 2) &= 1 - [p(0) + p(1) + p(2)] \\ &= 1 - \left(\frac{1.5^0}{0\,!} e^{-1.5} + \frac{1.5^1}{1\,!} e^{-1.5} + \frac{1.5^2}{2\,!} e^{-1.5} \right) \\ &= 1 - (0.2231 + 0.3347 + 0.2510) \\ &= 1 - 0.8088 = 0.1912.\end{aligned}$$

Question 2

You receive emails at an average rate of 2 per hour. What is the probability of receiving:

a) Six emails in one hour.

b) Less than three emails in an hour.

c) No emails in two hours.

d) More than four emails in two hours.

e) At least one email in half an hour.

f) What is the value of the mean number of emails received in one hour? What is the corresponding standard deviation?

Poisson Tables

The **Poisson tables** are used in the same way as the binomial tables.

In particular, **“greater than or equal to”** probabilities are tabulated:

$$\boxed{\Pr(X \geq r)}$$

where r is the value in question.

We select the appropriate Poisson distribution by finding the λ value in the column headings (note: the tables use the symbol m rather than λ).

Example: System Crashes

With $X \sim$ Poisson($\lambda = 3$ / year), what is the probability that:

. . . there are no crashes in a year?

$$\Pr(X = 0) = \Pr(X \geq 0) - \Pr(X \geq 1) = 1.0000 - 0.9502 = 0.0498.$$

. . . at least one crash in a year?

$$\Pr(X \geq 1) = 0.9502.$$

. . . less than 2 crashes in a year?

$$\Pr(X < 2) = 1 - \Pr(X \geq 2) = 1 - 0.8009 = 0.1991.$$

. . . between 4 and 6 crashes in a year?

$$\Pr(4 \leq X \leq 6) = \Pr(X \geq 4) - \Pr(X \geq 7) = 0.3528 - 0.0335 = 0.3193.$$

Example: System Crashes

What is the probability of no crashes in 2 years? $\Rightarrow \lambda t = 3(2) = 6 = m$ in the tables.

$$\Pr(X = 0) = \Pr(X \geq 0) - \Pr(X \geq 1) = 1.0000 - 0.9975 = 0.0025.$$

What is the probability of more than 2 crashes in 6 months (i.e., 0.5 years)? $\Rightarrow \lambda t = 3(0.5) = 1.5 = m$ in the tables.

$$\Pr(X > 2) = \Pr(X \geq 3) = 0.1912.$$

Question 3

You receive emails at an average rate of 2 per hour. What is the probability of receiving:

a) Six emails in one hour.
b) Less than three emails in an hour.
c) No emails in two hours.
d) More than four emails in two hours.
e) At least one email in half an hour.

Note: you calculated these in Question 1 using the *formula* for the probability function.

R Code

As with the binomial distribution, we can calculate probabilities for the Poisson distribution also:

Examples:

`dpois(0,lambda=3)`
gives `0.04978707`,

`dpois(4:6,lambda=3)`
gives `0.16803136 0.10081881 0.05040941`

and

`sum(dpois(4:6,lambda=3))`
gives `0.3192596`

Compare the above with slides 11 and 12

R Code

Greater than probabilities, i.e., $\Pr(X > x)$, can also be calculated.

It is important to note that this differs from the Poisson tables which (as we saw) provide *greater than or equal to* probabilities.

Examples:

`ppois(0,lambda=3,lower=F)`
gives `0.9502129` which is $\Pr(X > 0) = \Pr(X \geq 1)$.

`ppois(3,lambda=3,lower=F)`
gives `0.3527681` which is $\Pr(X > 3) = \Pr(X \geq 4)$.

`ppois(6,lambda=3,lower=F)`
gives `0.03350854` which is $\Pr(X > 6) = \Pr(X \geq 7)$.

Compare this with slide 18.

R Code

We can *generate* Poisson random variables as follows:

Example:

```
rpois(100,lambda=3)
```

generates 100 Poisson($\lambda = 3$) variables.