Poisson and Exponential Distributions

XDASI Fall 2021

10/21/2021

Contents

Background

- **W&S Chapter 8.4**
- **Aho - Chapter 3.3-3.6**

Poisson Distribution

The Poisson is a ***discrete*** distribution that describes the probability of observing a certain number of events in a given interval of time (or space), if the events occur ***randomly*** and ***independently*** of each other. It is used to model stochastic processes, such as shot noise or scattered objects.

The Poisson is actually a **limiting case of the binomial** when n is large and p is small, and it is particularly useful for ***rare*** events. We will talk about this more below.

Examples in ***time***:

- The number of gamma rays detected in a scintillation counter over time (the number of particles emitted over time due to exponential decay)
- The number of mutations that accumulate per generation in biological systems
- The number of accidents in New York City per day
- The number of coffee orders per minute at the Matto coffee shop on Mercer

Examples in ***space***:

- The number of mutations per kilobase after treatment with a mutagen (e.g. EMS)
- Distribution of bacterial colonies on an evenly spread agar plate, or of bacteriophage plaques plated at low multiplicity
- Distribution of gas molecules in a closed container
- The number of typos per page in a book
- Distribution of chocolate chips in cookie dough

Any process that gives rise to events that are randomly distributed in space or time, where events happen at a constant rate and the number of occurrences in non-overlapping intervals is statistically independent, is called a **Poisson process**.

Visualization

The Poisson has a single parameter, λ, which is called the ***rate constant***. Lambda is the ***expected number of events per unit time*** t. The figure and code below show how the shape of the PDF and CDF change with increasing λ.

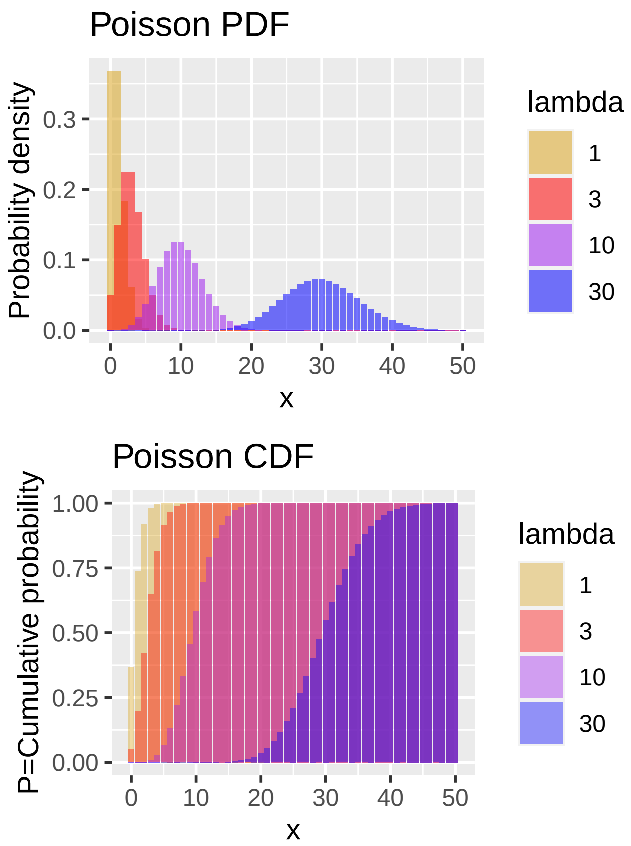
```r
x=c(0:50)
pois.dist = data.frame(x = x,
                       l1 = dpois(x, 1),
                       l3 = dpois(x, 3),
                       l10 = dpois(x, 10),
                       l30 = dpois(x, 30))
pois.dist.long = gather(pois.dist, "lambda", "density", 2:5)
pois.dist.long = pois.dist.long %>% mutate(lambda = substring(lambda,2))
pois.dist.long = pois.dist.long %>% mutate(lambda = factor(lambda, c(1,3,10,30)))

p.pdf = ggplot(pois.dist.long, aes(x=x)) +
  geom_bar(aes(y=density, fill=lambda), alpha=0.54,
           stat="identity", position="identity") +
  scale_fill_manual(values = c("goldenrod","red","purple","blue")) +
  ylab("Probability density") +
  ggtitle("Poisson PDF") +
  coord_fixed(ratio = 100)
```

```
pois.cum = data.frame(x = x,
                      l1 = ppois(x, 1),
                      l3 = ppois(x, 3),
                      l10 = ppois(x, 10),
                      l30 = ppois(x, 30))
pois.cum.long = gather(pois.cum, "lambda", "density", 2:5)
pois.cum.long = pois.cum.long %>% mutate(lambda = substring(lambda,2))
pois.cum.long = pois.cum.long %>% mutate(lambda = factor(lambda, c(1,3,10,30)))

p.cdf = ggplot(pois.cum.long, aes(x=x)) +
  geom_bar(aes(y=density, fill=lambda), alpha=0.4,
           stat="identity", position="identity") +
  scale_fill_manual(values = c("goldenrod","red","purple","blue")) +
  ylab("P=Cumulative probability") +
  ggtitle("Poisson CDF") +
  coord_fixed(ratio = 40)

ggarrange(p.pdf, p.cdf, nrow=2, ncol=1)
```



The PDF shows the ***probability seeing x events in one unit of time***. When events are rare, the rate constant λ is small and it is very unlikely that we will see a lot of events per unit time, so the distribution is squished to the left (a.k.a. “right-skewed”). When λ is large, that means that events are much more frequent, so it’s much more likely we will see lots of events per unit time.

We can see from the PDF that the distribution starts to look a lot like a normal distribution as λ gets bigger.

Poisson PDF

If X follows a Poisson distribution, we write $X \sim POI(\lambda)$. The formula for the Poisson **PDF** is:

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where X is a Poisson random variable and λ is the ***rate*** at which events occur in one unit of time (or space). The single parameter λ represents both the expected value (mean) and the variance (σ^2), so it determines both the ***location*** and ***scale*** of the distribution:

$$E(X) = Var(X) = \lambda$$

Alternative formula

In the above equation λ is the ***expected number of events per unit of measure*** t, where $t = 1$. Since this is just the ***mean*** number of events per t units of time ($\lambda = \mu/t$), we can easily substitute $\mu = \lambda t$ in the above equation:

$$P(X = x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} = \frac{e^{-\mu}\mu^x}{x!}$$

The Poisson is often expressed using this alternative notation. When $t = 1$, this boils down to the first form of the equation, with just λ instead of μ.

If we are interested in the number of events that occur within a larger interval than one unit, for example the number of traffic accidents per week vs. per day, then the expression for the PDF in terms of μ is what we want. If λ is defined as accidents/day, then we would set $t = 7$, making $\mu = \lambda * t = \lambda * 7$.

Alternatively, we could redefine λ in terms of accidents/week, which would make it 7 times as large as the daily rate, and then use $t = 1$ week. These end up exactly the same because we are simply adjusting a scaling factor, so the formula still works out.

The important thing is that we want to define our problem in terms of the units and intervals that we are interested in.

CDF

The CDF gives the ***total probability***, which is area under the PDF. Since the Poisson is a discrete distribution, the cumulative probability for x is just the sum of the PDF from zero to some value of x:

$$P(X \leq x) = \sum_{x=0}^{k} \frac{e^{-\lambda t}(\lambda t)^x}{x!} = e^{-\mu}\sum_{x=0}^{5} \frac{\mu^x}{x!}$$

Assumptions

Like any other distribution, the Poisson makes certain assumptions about the nature of the data:

- The number of observed events $x \in \{0, 1, 2, ...\}$ is independent in any interval.
- Events occur at a constant (random) rate, $\lambda > 0$, over defined intervals of time (or space).
- The probability of observing two or more events in the same interval will approach 0 as intervals become smaller.

Example: Single-nucleotide variants

The number of **single-nucleotide variants (SNVs)** between *C. elegans* strains from Bristol, UK (N2) and Hawaii (CB4856) is **1.5 per 1000 nt**. We assume that the number of SNVs in any random 1kb interval across the genome is *independent* of the number in any other interval.

Crosses between these two lines are often used to map mutations that give rise to specific phenotypes (either natural variants, or in genetic screens of EMS-mutagenized animals). These are identified by selecting for the phenotype among the F2 progeny, and then finding regions that are homozygous for one allele or the other.

Let's say we would like to ask, ***What is the probability of observing zero SNVs in a random 5kb interval?*** The Poisson equation for this question would be:

$$P(X = 0) = \frac{e^{-\mu}\mu^{x}}{x!} = \frac{e^{-\mu}\mu^{0}}{0!} = e^{-\mu} = e^{-\lambda t}$$

So what is $\mu = \lambda t$? Given that we see ***1.5 SNVs per kilobase, on average***, we can define the problem using a unit of 1nt or a unit of 1kb. It doesn't really matter, we just need to makes sure that we scale everything properly.

For a "unit" size of 1nt, $\lambda = 1.5/1000 = 0.0015$ and $t = 5000$, whereas for a unit of 1000nt, $\lambda = 1.5$ and $t = 5$. Either way, μ works out to be the same: $\mu = \lambda t = (0.0015) * 5000 = 1.5 * 5 = 7.5$.

Now we can answer our question:

$$P(X = 0) = e^{-\lambda t} = e^{-7.5} = 5.53 * 10^{-4}$$

Poisson functions in R

In R, the `pois` family of functions assumes that the units have already been scaled so that the interval of interest is unit size ($t = 1$). In other words, just think of the `lambda` parameter as representing $\mu = \lambda t$, where $t = 1$.

That means that if you want to find the number of mutations per 5kb, you need to scale `lambda` to $1.5 * 5 = 7.5$. In other words, there is no way to specify $\lambda = 1.5$ and $t = 5$.

```
# the rate is given as lambda = 7.5 per unit t, where t = 5kb
dpois(0,lambda=7.5)
```

```
## [1] 0.0005530844
```

```
# Here the PDF and the CDF are the same since we are asking for only one value.
ppois(0,lambda=7.5)
```

```
## [1] 0.0005530844
```

Now let's say we want to ask,

What is the probability of observing 5 or fewer mutations in a random selection of 10000nt?

This is a question of the total probability $P(X \leq 5)$, for which we need the CDF. Since we are given a multiple of our original units (1kb), we can use the formula with $\mu = \lambda t$ to get the expected number of mutations per 10kb. And since this is a discrete distribution, we just sum up the individual probabilities from $x = 0$ to $x = 5$:

$$P(X \leq 5) = \sum_{x=0}^{5} P(X = x) = e^{-\mu} \sum_{x=0}^{5} \frac{\mu^x}{x!}$$

⇒ ***How would we compute this in R?***

Answer

Since R assumes that we have already scaled λ according to the interval of interest, for any question about the interval of 1kb, we will use $\lambda = 1.5$, and for any question about an interval larger than 1kb we will use $\mu = \lambda t$, where t is some multiple of the given unit measure.

Here our unit interval is 10kb instead of 1kb, so we need to use $\mu = 1.5 * 10 = 15$. This is the value we plug in for the `lambda` parameter.

- Since this is a discrete distribution, one way to do this would be to use the PDF (density function) to sum the probabilities up to 5:

```
# sum of probabilities for X = {0,1,2,3,4,5}
sum(dpois(0:5,lambda=15))
```

```
## [1] 0.002792429
```

- Equivalently, we could use the CDF (the total probability function) to compute this value:

```
# mu = lambda * t = 1.5 * 10 = 15
ppois(5,lambda=15) # now plug in mu instead of the original value of lambda
```

```
## [1] 0.002792429
```

Technical note Since this is a **discrete** distribution, x must be an integer. So for finding the probability of 5 SNVs per 10kb, we **cannot** just divide λ by 10 and ask for $Pr[0 \leq X \leq 0.5]$ over 1000 nt. Instead, we have to factor in the interval 10kb by using the formula $\mu = \lambda t = 1.5 * 10$, so we can get $Pr[0 \leq X \leq 5]$ in this interval.

Also note:

- Calling `rpois()` with a non-integer value will just return the number of values corresponding to `int(n)`.
- Calling `dpois()` with a non-integer value will give an error.
- R implements `ppois()` using the Gamma function, which is continuous, so it will give a result if you call it with a non-integer value, but this is not technically valid.
- The `qpois()` function, of course, takes continuous values for a cumulative probability between 0 and 1.

```
rpois(0:5.25,lambda=1) # uses 0:5
## [1] 1 2 0 0 1 1
dpois(0.5,lambda=1) # raises an error
## Warning in dpois(0.5, lambda = 1): non-integer x = 0.500000
## [1] 0
ppois(0.5,lambda=1) # returns a value, but not "technically" valid
## [1] 0.3678794
#                      for a discrete distribution
qpois(seq(from=0, to=1, by=0.1),lambda=1) # takes any probability value(s)
##  [1]   0   0   0   0   1   1   1   1   2   2 Inf
                                          # between 0-1
```

Exponential Distribution

The exponential distribution is a ***continuous*** distribution that, like the Poisson distribution, also relates to a **Poisson process**.

Whereas the ***Poisson*** describes the **number of events** that occur in a given period of time, the ***exponential*** describes **the length of time elapsed** between consecutive events. You can also think of this as the ***waiting time* until the first observed event**.

Examples:

- Radioactive decay (time to the first emission, given a particular rate of decay)
- Type II survivorship (where mortality rates are about the same at any age)
- Battery life (for a given failure rate)

PDF and CDF

The exponential distribution describes the density of waiting times ("failures") before the next Poisson event. For a random variable that follows this distribution, we write $T \sim EXP(\lambda)$. The **PDF** is:

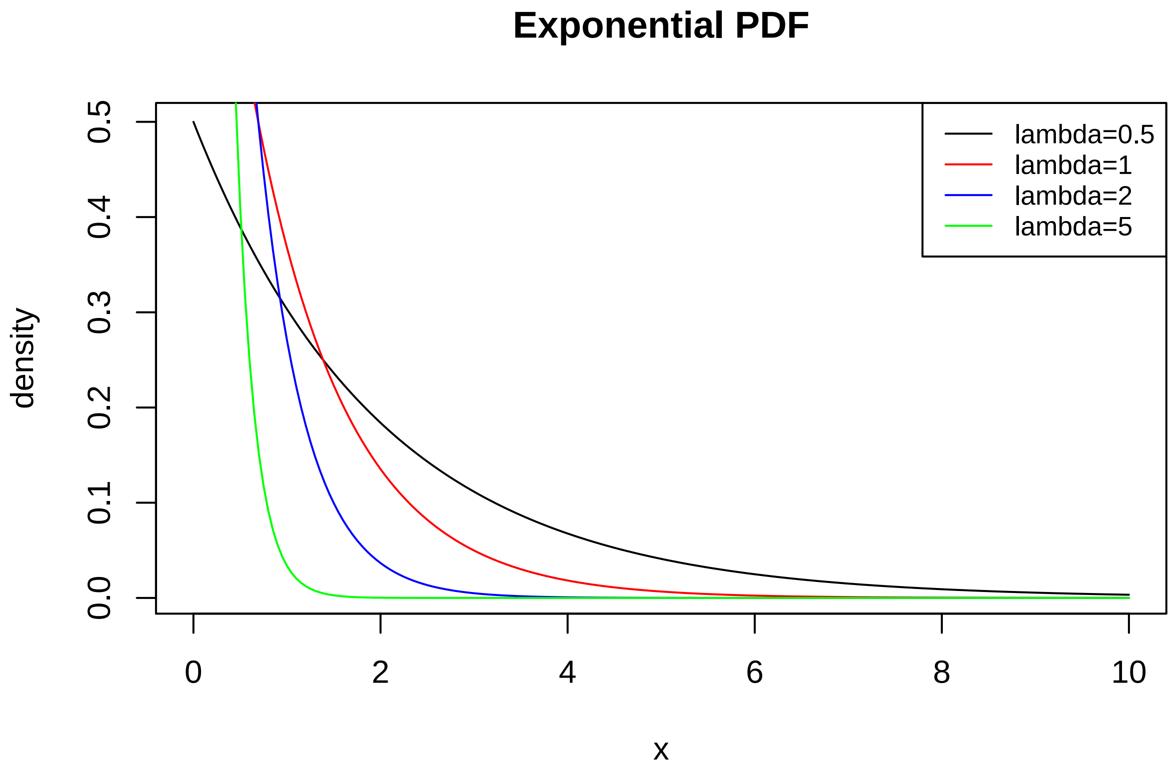
$$f(t) = \lambda e^{-t\lambda}$$

where $\lambda > 0$ and $t \geq 0$.

The expected value $E(T)$ (the mean) and standard deviation are both $1/\lambda$:

$$\mu = \sigma = 1/\lambda \;\; ; \;\; \sigma^2 = (1/\lambda)^2 = 1/\lambda^2$$

```
interval.exp<-seq(0, 10, 0.05)
plot(interval.exp,dexp(interval.exp,rate=0.5),type = "l", xlab="x",ylab="density", main="Exponential PDF
lines(interval.exp,dexp(interval.exp,rate=1),col="red")
lines(interval.exp,dexp(interval.exp,rate=2),col="blue")
lines(interval.exp,dexp(interval.exp,rate=5),col="green")
legend("topright", legend=c("lambda=0.5","lambda=1","lambda=2","lambda=5"),
       col=c("black"," red", "blue", "green"), lty = 1, cex=0.8)
```



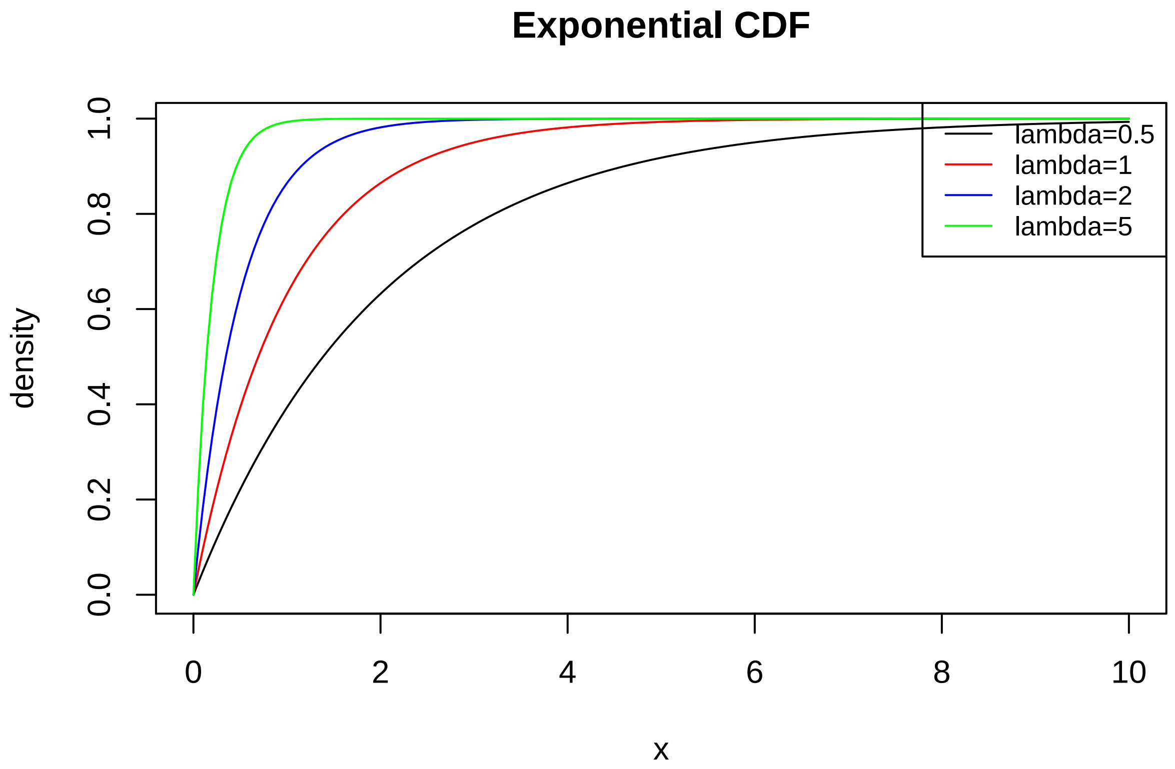
The **CDF** has a closed form that does not require integration:

$$F(t) = \int_0^\infty \lambda e^{-\lambda t} dt = 1 - e^{-t\lambda}$$

If we change the name of our random variable to x, then we have the probability that a Poisson event occurs within a timespan of zero to x:

$$Pr[X \leq x] = 1 - e^{-x\lambda}$$

```
interval.exp<-seq(0, 10, 0.05)
plot(interval.exp,pexp(interval.exp,rate=0.5),type = "l", xlab="x",ylab="density", main="Exponential CD
lines(interval.exp,pexp(interval.exp,rate=1),col="red")
lines(interval.exp,pexp(interval.exp,rate=2),col="blue")
lines(interval.exp,pexp(interval.exp,rate=5),col="green")
legend("topright", legend=c("lambda=0.5","lambda=1","lambda=2","lambda=5"),
       col=c("black"," red", "blue", "green"), lty = 1, cex=0.8)
```



The probability that a Poisson event happens ***within the interval*** between x_1 and x_2 is:

$$Pr[x_1 < X \leq x_2] = F(x_2) - F(x_1) = (1 - e^{-x_2\lambda}) - (1 - e^{-x_1\lambda})$$

Survivorship

The **survivorship**, or the length of time ***before*** something happens, is the same as probability that the time to a Poisson event is ***greater*** than x:

$$Pr[X > x] = 1 - (1 - e^{-x\lambda}) = e^{-x\lambda}$$

Example: SNVs

Going back to our example of SNVs in *C. elegans*, we had a rate of 1.5 mutations per 1000nt, and we used the Poisson to find the probability of finding exactly 0 SNVs in 1kb, or up to 5 mutations in 10kb. To ask about x *number of events* in an interval of 10kb nt, we used $\mu = \lambda t = 1.5 * 10 = 15$.

Here, our random variable x represents an **interval** of *time* or *space* (vs. the number of events per interval of time). These questions are really just the flip side of the same coin!

Note: *For the Exponential function family* `exp()` *in R, λ is called the* ***rate****. Here the rate of SNV's is 1.5 per kb, so we use* `rate=1.5`*, where 1kb is "1" unit. Alternatively, we can use* `rate=0.0015`*, where the unit is 1nt.*

⇒ ***What is the average number of nt within which we would expect to find one SNV?***

Answer

```
# EXP(X) = 1 / rate
1/1.5         # this is the number of kb since unit size is 1kb
```

```
## [1] 0.6666667
```

```
(1/1.5)*1000  # we need to multiply by 1000 to get result in nt
```

```
## [1] 666.6667
```

```
1/(1.5/1000)  # or convert to 0.0015 per nt unit
```

```
## [1] 666.6667
```

⇒ ***What is the probability of an SNV occurring somewhere within 1000 nt? (I.e. what is the probability that a span of 1kb will contain at least one SNV?)***

Answer

$$Pr[X \leq 1000] = 1 - e^{-x\lambda} = 1 - e^{-1000*1.5} = 1 - e^{-1.5} = 0.777$$

```
pexp(1, 1.5, lower.tail=T)           # P of finding at least 1 per 1000nt unit
## [1] 0.7768698
pexp(1000,rate=0.0015, lower.tail=T) # same: P from 0-1000 for unit=1nt
## [1] 0.7768698
```

Notice that asking for the probability of finding ***at least one*** SNV per 1000nt using the Exponential, or asking the probability of ***NOT finding 0*** in 1000nt using the Poisson, are two ways of asking the same question:

```
# P of NOT finding 0 (different way to say the same thing)
ppois(0, 1.5, lower.tail=F)
```

```
## [1] 0.7768698
```

⇒ ***What is the probability that a span of 1000nt (1kb) will contain NO SNV's? (I.e. what is the probability of NOT finding a single SNV in 1000nt?)***

Answer

This is the same as asking for the survivorship, or the ***upper-tail*** probability of finding at least one within 1000 nt.

$$P(X > 1000) = 1 - P(X \leq 1000) = e^{-x\lambda} = e^{-1.5} = 0.223$$

```r
# P of finding a span of 1000 bases that contains no SNV
# (i.e., P of not finding an SNV within 1kb)

# these are all equivalent
exp(-1.5)
## [1] 0.2231302
pexp(1, rate=1.5, lower.tail=F)  # 1 unit at rate = 1.5/unit
## [1] 0.2231302
pexp(1000, 0.0015, lower.tail=F) # same: 1000 units at rate = 0.0015/unit
## [1] 0.2231302
1-pexp(1000,rate=0.0015)
## [1] 0.2231302
```

The survivorship is the same as $P(X = 0)$ for the Poisson distribution with $\lambda = 1.5$ and $t = 1$:

```r
ppois(0, lambda=1.5, lower.tail=T)
```

```
## [1] 0.2231302
```

```r
# but this is not the same, because this distribution has a different shape!
ppois(0,0.0015)
```

```
## [1] 0.9985011
```

Relationships between distributions

Poisson and Exponential distributions

The Poisson is a discrete distribution describing the ***number of events per unit time***, and the exponential is a continuous distribution describing the ***length of time between events***. For a Poisson process, then, if events happen at a rate of λ per unit time on average, an average of λt events will occur per t unit of time.

Classic example: Radioactive decay The decay of radioactive isotopes is a Poisson process, since radioactive isotopes have a constant half-life, and each successive emission is independent of the last one. The Poisson and exponential are complementary ways of describing this process.

- The slower the decay rate (longer half-life), the larger the time interval between observed events (exponential), and the fewer events observed per unit time (Poisson).
- The probability of time elapsed between decay events follows a continuous exponential distribution.
- The probability of detecting a certain number of decay events per unit time with a scintillation counter follows a discrete Poisson distribution.

Survival function The ***survival function*** is the complement of the CDF of a particular distribution. If the CDF is $F(X) = P(X \leq x)$, then the **survivorship** is $P(X > x) = 1 - F(X)$.

For a Poisson distribution $P(X = x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$, $P(X = 0) = e^{-\lambda t}$ is the probability of *no* events within t units of time. This is the same as the survivorship, or the probability that the time T to the *first occurrence* exceeds t:

$$P(T > t) = P(X = 0|\mu = \lambda t) = e^{-\lambda t}$$

You can see that this is just one minus the exponential CDF, which gives the total probability that an event *does* occur within an interval of time t:

$$P(T \leq t) = 1 - P(X = 0|\mu = \lambda t) = 1 - e^{-\lambda t}$$

For example, the ***exponential survivorship*** function describes situations where the *probability of mortality is the same for all individuals in a population*, since the rate of events is always the same for a Poisson process. In other words, the chance of dying is independent of age! This is not true for mammals, but it is true of some birds, rodents, lizards, and sea animals.

Poisson approximation of the Binomial distribution

The Poisson is the **limiting** case of a Binomial distribution for **rare** events. That is, a Poisson with parameter$\lambda = np$ approximates the Binomial when n is large and p is small.

For example, let's consider hits on a website. This could be modeled as a Binomial distribution with a Bernoulli trial every minute for an hour, with $n = 60$. If $p = 0.1$, this would be the same as a Poisson distribution with 6 events per hour, i.e. $\lambda = 6$.

The Poisson has some advantages over the Binomial:

- The Poisson can contain more than one event per unit interval (whereas the binomial only handles one, since it describes Bernoulli trials, which only have a binary outcome)
- It can be easier to work with the Poisson since it has only one parameter instead of two.

Starting with the Binomial probability, it can be shown that the limit of $P(x)$ as n goes to infinity and p goes to 0 is the equation for the Poisson distribution. Setting $p = \frac{\lambda}{n}$, the Binomial formula may be written in terms of λ as:

$$\begin{aligned}
P(x) &= \lim_{n\to\infty} \binom{n}{k} (p)^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \binom{n}{k} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^n (1 - \frac{\lambda}{n})^{-k} \\
&= 1 * \frac{\lambda^k}{k!} e^{-\lambda} * 1 \\
&= \frac{e^{-\lambda}\lambda^k}{k!}
\end{aligned}$$

All the terms including n in a factorial expression cancel out to 1 as n goes to infinity, and the last term also goes to 1 in the limit. So why does the second to last term equal $e^{-\lambda}$? The definition of e is:

$$e = \lim_{x\to\infty} \left(1 + \frac{1}{x}\right)^n$$

If we set $x = -\frac{n}{\lambda}$, then we find that:

$$\lim_{x\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = \lim_{x\to\infty} \left(1 + \frac{1}{x}\right)^{-\lambda x} = e^{-\lambda}$$

Normal approximation of the Poisson and Binomial distributions

Like the Binomial distribution for large n, the Poisson is well approximated by the normal distribution when λ is large (and the sampling size is very large). This can be seen from the graph above for the Poisson distribution with $\lambda = 30$.

Poisson, Exponential, and Gamma distributions

We will not go over the Gamma distribution here, but it is commonly encountered in Bayesian probability applications, as a congugate prior distribution for a number of likelihood functions.

Conceptually, you can think of the exponential function as the **waiting time** until the **first** Poisson-distributed event, and Gamma as the waiting time until the nth Poisson-distributed event (or the nth change in a Poisson process). So, *the exponential distribution is a special case of the Gamma distribution* [in which the “shape” parameter equals 1].