1 Poisson Models

The HURDAT data consists of 6-hourly updates of the status of tropical cyclones.

There are various ways of analyzing the data based on what the goals of the analysis are.

1.1 Poisson Distribution for Counts

We are interested in modeling the total number of storms per year and investigating the clustering of storms over short time periods.

Let's start with models for the total number of storms.

It is natural to aggregate the data into a format that has a row for each year, and a total number of storms of each type, as in

year	tropical cyclones	tropical storms	hurricanes	landfalling hurricanes
⋮	⋮	⋮	⋮	⋮
2019	20	16	6	2
2020	31	29	14	11
2021	21	20	7	4
2022	16	14	9	5
2023	21	19	7	2
2024	18	18	11	8

When doing this, be careful to include years that have zero storms.

Notation for the data:

$i = 1, \ldots, n$ for the row of data (loosely referred to as "year")

y_{i1} = number of tropical cyclones in year i

y_{i2} = number of tropical storms in year i

y_{i3} = number of hurricanes in year i

y_{i4} = number of landfalling hurricanes in year i

The Poisson model is a reasonable starting point because our data are counts ranging from possibly zero with no upper bound.

We are specifically interested in overdispersion relative to the Poisson distribution, so here is the Poisson model for y_{ik}:

$$Y_{ik} \overset{ind}{\sim} \text{Poisson}(\lambda_k)$$

We are saying here that the distribution of storms of type k (TC, TS, H, LH) is Poisson with expected value λ_k, which allows for a different mean for each type of storm.

The poisson distribution has the following mass function:

$$P(Y_{ik} = x) = \frac{\lambda_k^x e^{-\lambda_k}}{x!}, \quad \text{for } x = 0, 1, 2, \ldots \tag{1}$$

The expectation and variance of the Poisson are

$$E(Y_{ik}) = \lambda_k, \quad \text{Var}(Y_{ik}) = \lambda_k$$

See if you can work this out, for practice.

We are also saying that the number of storms is independent across years, which means that knowledge of the number of storms in one year does not impact our beliefs about the number of storms occurring in any other year.

Suppose we have some other model for y_{ik}:

$$Y_{ik} \overset{ind}{\sim} F(\theta_k)$$

We say that distribution F is overdispersed relative to the poisson if

$$\text{Var}(Y_{ik})/E(Y_{ik}) > 1$$

under distribution F.

This is what you're testing for in your assignment. You may elect to build a specific alternative model, use a simulation-based procedure, or some other method to conduct the test.

1.2 Poisson Process Models for Event Time Data

The next part of the assignment is about clustering. Just as overdispersion is defined relative to the Poisson distribution, clustering can be defined relative to a distribution.

A natural choice is the Poisson process, which is a model for the times at which events occur.

This is a bit trickier to define, so let's focus on hurricanes.

Let e_i be the time that the ith hurricane forms, in units of days since Jan 1, 1851 (or some other appropriate time unit). This can be defined as the first time at which the hurricane is classified as a hurricane.

This gives us a sequence of event times:

$$e_1, e_2, \ldots, e_n$$

where n is the total number of hurricanes.

How do we relate these times to a Poisson distribution? Well, if we have these times $e_1, \ldots, e_n$, we can count up the number of events that occur in any specified time interval. Define $y([t_1, t_2])$ to be the number of events that occur in the interval $[t_1, t_2]$.

For example, you can recover the counts for each year by defining intervals that start on January 1 and end on December 31 of that year:

$y_{13} = y([1851.00, 1852.00])$

If you know the event times, you can clearly calculate y for any interval. Conversely, if you know the count y for any interval, you can recover the event times to arbitrary accuracy by considering a sequence of very small intervals.

The Poisson process model is the random variable version of y, defined as follows:

$$Y([t_1, t_2]) \sim \text{Poisson}(\lambda(t_2 - t_1)) \quad \text{for } t_2 > t_1$$
$$Y([t_1, t_2]) \perp\!\!\!\perp Y([t_3, t_4]) \quad \text{for non-overlapping} \quad [t_1, t_2], [t_3, t_4]$$

Basically, the model says that the counts have a Poisson distribution with mean proportional to the length of the interval, and the number of counts in two non-overlapping intervals are independent.

Because of the independence assumption, we can view the Poisson process as a model for events that are independent of one another. If we know how many events occurred last week, that has no bearing on the number of events we expect to see this week.

What does clustering mean?

How might you look for evidence of clustering relative to the Poisson process?

1.3 Waiting time distribution

If we pick a time, say t_0, we can define the time we wait until the next event occurs, T, and work out its probability distribution. Clearly $T > 0$ and is a continuous variable.

$$P(T > t) = P(Y([t_0, t_0 + t]) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$
$$P(T < t) = 1 - e^{-\lambda t},$$

which gives us the CDF. We take a derivative to get the pdf:

$$f(t) = \lambda e^{-\lambda t} \quad \text{for} \quad t > 0,$$

which you should recognize as the exponential distribution with rate λ, which has expected value $1/\lambda$. This makes sense because if you expect λ events per unit time, you should expect to wait $1/\lambda$ time units between events.

1.4 Two independent Poisson processes

Suppose we have two independent poisson processes Y_1 and Y_2 with rates λ_1 and λ_2.

Pick a time t_0. What's the chance that the next event comes from Y_1 instead of Y_2?

Define the waiting times T_1 and T_2. These are clearly independent because the two processes are independent.

How do we calculate the probability? Consider a very fine sequence of wait times $0, \Delta, 2\Delta, \ldots$. The event that $T_1 < T_2$ is roughly

$$\bigcup_{j=0}^{\infty} \{j\Delta < T_1 < (j+1)\Delta \quad \text{and} \quad T_2 > (j+1)\Delta\}$$

These are disjoint events, so the probability is

$$P(T_1 < T_2) = \sum_{j=0}^{\infty} P(j\Delta < T_1 < (j+1)\Delta \text{ and } T_2 > (j+1)\Delta\} \tag{2}$$

$$\approx \sum_{j=0}^{\infty} f_1(j\Delta)\Delta e^{-\lambda_2 j\Delta} \tag{3}$$

Taking the limit as $\Delta \to 0$ gives the integral

$$\begin{aligned} P(T_1 < T_2) &= \int_{t=0}^{\infty} f_1(t)e^{-\lambda_2 t}dt \\ &= \int_{t=0}^{\infty} \lambda_1 e^{-\lambda_1 t}e^{-\lambda_2 t}dt \\ &= \lambda_1 \int_{t=0}^{\infty} e^{-(\lambda_1+\lambda_2)t}dt \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}\int_{t=0}^{\infty}(\lambda_1+\lambda_2)e^{-(\lambda_1+\lambda_2)t}dt \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2} \end{aligned}$$

because the last integral is the pdf for an exponential variable with rate $\lambda_1 + \lambda_2$, which must integrate to 1.

If $\lambda_1 = \lambda_2$, you get 1/2 which makes perfect sense if they have the same rate.

1.5 Heterogeneous Poisson Processes

You will definitely be able to find clustering in the times of storms relative to the Poisson process defined in the previous section, also known as a homogeneous Poisson process. That's because there is a "hurricane season" where storms are more likely to occur in the months June to November, with maximum activity in early September.

This is a known type of clustering and not the type we are interested in.

We are really interested in clustering, or lack of clustering, relative to a heterogeneous Poisson process, defined as follows. Let $\lambda(t) > 0$ be a rate that depends on time. Define the process Y as follows:

$$\begin{aligned} &Y([t_1, t_2]) \sim \text{Poisson}\Big(\int_{t_1}^{t_2} \lambda(t)dt\Big) \\ &Y([t_1, t_2]) \perp\!\!\!\perp Y([t_3, t_4]) \quad \text{for non-overlapping} \quad [t_1, t_2], [t_3, t_4] \end{aligned}$$

You can see that you get back the original homogeneous poisson process if $\lambda(t)$ is constantly λ, since $\int_{t_1}^{t_2} \lambda(t)dt = (t_2 - t_1)\lambda$ in that case.

I really want you to look for clustering or lack of clustering relative to a heterogeneous Poisson process.

1.6 Simulating from a Poisson Process

Depending on how you do the analysis, you might find it useful to simulate from a Poisson process.

There is more than one way to do it, but you'll want to define the interval over which you are simulating events. Let's call it $[0, t_{max}]$

To simulate the times at which events occur, you can simply simulate the waiting periods between successive events, which is possible for the homoegeneous Poisson process, since it follows an exponential distribution. Here is a routine:

set $E_0 = 0$

set $S = 0$

set $J = 0$

while $S < t_{max}$

1. Simulate wait time $T_{J+1} \sim$ Exponential(λ)
2. set $S = E_J + T_{J+1}$
3. if $S < t_{max}$
   (a) set $E_{J+1} = E_J + T_{J+1}$
   (b) set $J = J + 1$

Finally, your set of event times is $E_1, \ldots, E_J$

Here is another way:

Simulate the total number of events $J \sim$ Poisson(λt_{max})

Simulate the J event times $E_1, \ldots, E_J$ independently and uniformly over $[0, t_{max}]$

This method is considerably simpler. Why does it follow a poisson process?

Consider an interval $[t_1, t_2]$ with $0 < t_1 < t_2 < t_{max}$.

Under this simulation process,

$$Y([t_1, t_2]) = \sum_{j=1}^{J} \mathbf{1}\{t_1 < E_j < t_2\}$$

This is a little tricky to evaluate because the number of terms in the sum is a random variable J.

We can still do it though if we work a little bit. Here's one way to think about the count $Y([t_1, t_2])$. We first simulate a poisson random variable J, and then we loop over the J events, each time deciding whether or not to count it as part of $Y([t_1, t_2])$ by flipping a coin with probability of heads $(t_2 - t_1)/(t_{max} - 0)$.

So it's like a binomial with a random number of trials that needs to be accounted for. This is also called a "thinned" poisson random variable.

$$P(Y[t_1,t_2]=x)=\sum_{j=x}^{\infty}P(J=j)P(\text{Binomial}(j,(t_2-t_1)/t_{max})=x)$$

$$P(Y[t_1,t_2]=x)=\sum_{j=x}^{\infty}\frac{\lambda^j e^{-\lambda}}{j!}\frac{j!}{x!(j-x)!}p^x(1-p)^{j-x}$$

where $p=(t_2-t_1)/t_{max}$

Let $k=j-x$. Then we have $j=k+x$ and

$$\begin{aligned}P(Y[t_1,t_2]=x)&=\sum_{k=0}^{\infty}\frac{\lambda^{k+x}e^{-\lambda}}{(k+x)!}\frac{(k+x)!}{x!(k)!}p^x(1-p)^k\\&=\frac{\lambda^x p^x e^{-\lambda}}{x!}\sum_{k=0}^{\infty}\frac{\lambda^k(1-p)^k}{k!}\\&=\frac{\lambda^x p^x e^{-\lambda}}{x!}e^{\lambda(1-p)}\\&=\frac{(\lambda p)^x e^{-\lambda p}}{x!}\end{aligned}$$

It remains to be shown that if $t_1<t_2<t_3<t_4$, then $Y([t_1,t_2])$ and $Y([t_3,t_4])$ are independent.

We can simulate from an inhomogeneous Poisson process in a similar way:

Simulate J events from a Poisson distribution with mean $\int_0^{t_{max}}\lambda(t)dt$.

Simulate the J times $E_1,\ldots,E_J$ from a continuous probability distribution proportional to $\lambda(t)$.

Showing that $Y([t_1,t_2])$ is Poisson with mean $\int_{t_1}^{t_2}\lambda(t)$ follows the exact same proof as above.

1.7 Simulating

To simulate from an inhomogeneous Poisson process, we need to sampling from a distribution proportional to $\lambda(t)$. If $\lambda(t)$ is proportional to a Gaussian, or an exponential, a Gamma distribution, then we can use built in R functions to do the sampling.

If we can easily calculate the cdf $\Lambda(t)=\int_0^t\lambda(s)ds$ and its inverse $\Lambda^{-1}(q)$, then we can simulate J uniforms $U_1,\ldots,U_J$ and then event times $E_j=\Lambda^{-1}(U_j)$.

You could actually do this using the empirical quantiles of the data.

If you can't do any of that, you could also use rejection sampling. Here's how it works.

We want to sample a random variable from a distribution proportional to $\lambda(t)$, and repeat this J times independently. Here's how to do it once:

Let λ_{max} be the largest value of $\lambda(t)$ on $[0,t_{max}]$. Simulate a uniform random variable D on $[0,t_{max}]$. Simulate a uniform random variable U on $[0,1]$. If $U<\lambda(D)/\lambda_{max}$, then keep D. Otherwise, reject it and repeat the process until you accept it.