Lecture 5: Estimating the Probability of Extreme Events

8 February 2019

Overview

A key topic in risk analysis is estimating the probability of extreme and unlikely events occurring. Examples:

- What is the probability that a financial institution will lose more than £1bn in a day, based on the previous returns of a portfolio it owns?
- What is the probability of a terrorist attack more deadly than the 9/11 New York event occurring in the next 20 years?
- What is the probability of a magnitude 8 or greater earthquake occurring in a particular geographical region?

We discussed the first example last week in the context of value-at-risk analysis. Today we will focus on the second, and look at terrorism data.

Setting

- The techniques used come from the area of **extreme value statistics** which is a branch of statistics that studies the probability of large/extreme events occurring.
- We will follow a methodology similar to that used in a recently published academic paper (available from the course moodle page).
- As a small difference, their paper used frequentist methodology while we work in a broadly Bayesian setting.

Data

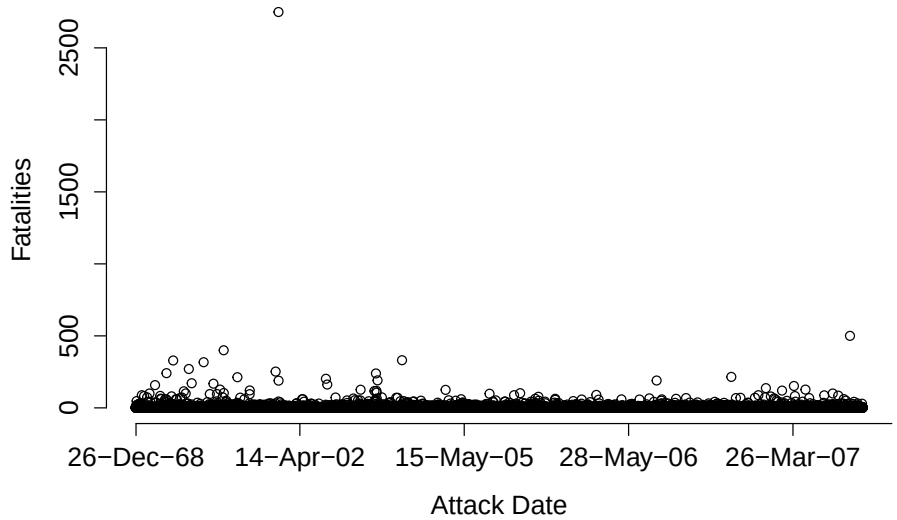
- We look at data from the RAND-MIPT terrorism database which contains a list of all terrorist attacks to occur worldwide between 1967-2007 (RAND is a famous policy think-tank)
- The data is publicly available, and can be downloaded from: www.rand.org/nsrd/projects/terrorism-incidents/download.html
- Each attack record contains the following fields – Date, City, Country, Perpetrator, Weapon, Injuries, Fatalities, Description
- There are 13,274 deadly attacks in total over this period (i.e. attacks with more than 1 fatality).

Sample Data

- We will focus only on the Fatalities column.
- We are interested in modelling the number of people killed in a typical terrorist attack.
- The purpose is to make statistical predictions about the probability of large terrorist attacks happening in the future.

Visualization

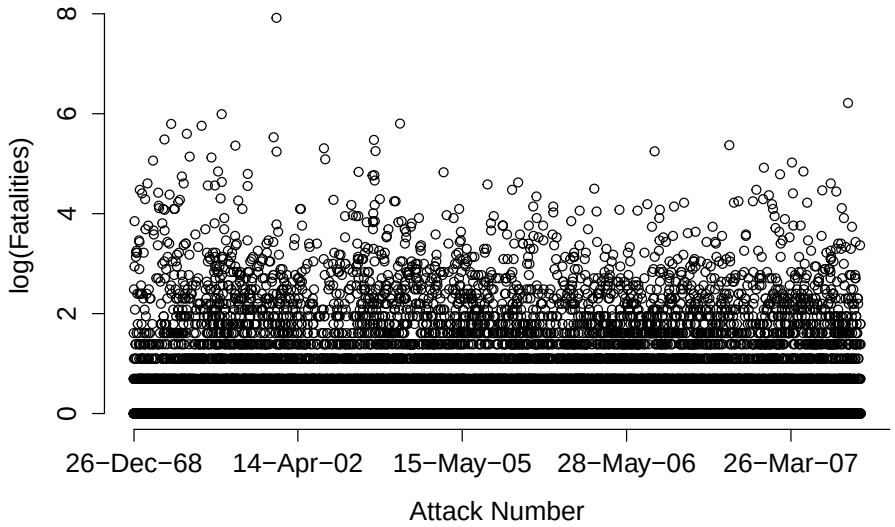
Whenever you work with data in any situation, the first thing to do is to **plot the data** to get an idea what it looks like.



The frequency of terrorist attacks seems to have increased drastically after 9/11. However the presence of outliers means its hard to interpret this plot since everything is so squashed together.

Visualization

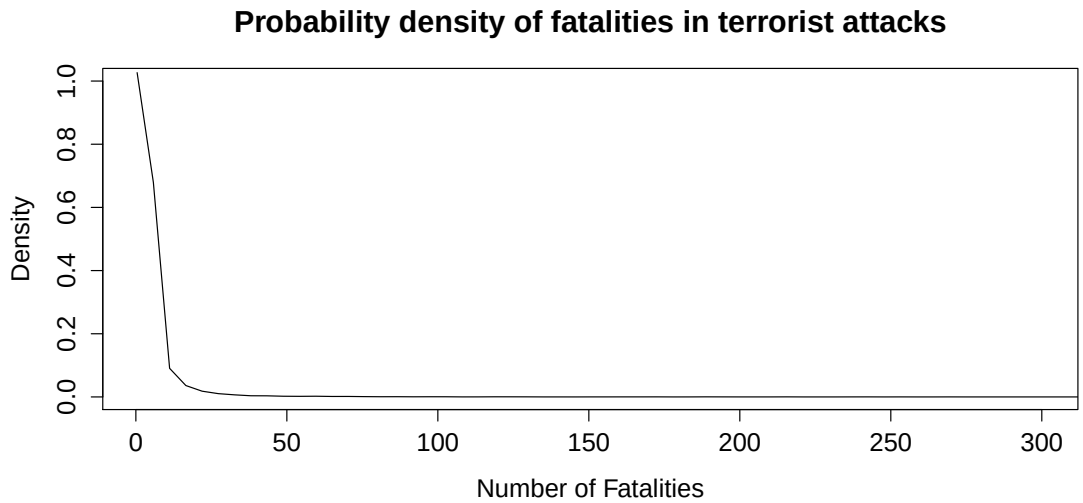
To fix this, we can take the logarithm of the fatalities.



Although the frequency of attacks increased after 9/11, there is no obvious change in the number of fatalities. The data looks stationary (i.e. not changing over time)

Visualization

We can also plot the density of the fatalities data:



The vast majority of attacks (7,965 out of 15,532) have only a single fatality, and 93.6% of attacks kill fewer than 10 people

First Steps

We can also compute some summary statistics:

- The mean (average) number of deaths per attack is 4.14
- The median number of deaths per attack is 1
- The modal number of deaths per attack is 1
- The standard deviation of the number of deaths per attack is 25.11

In summary most attacks do not kill many people. But there are several attacks that do cause high number of fatalities – for example, 13 of the attacks killed more than 200 people.

So deadly attacks are rare, and there are not many of them. Extreme value statistics is concerned with estimating the probabilities of rare events happening.

The Goal..

- From a statistical perspective, our goal is to try and build a model that can predict the probability of deadly terrorist attacks occurring.
- We want to answer questions like "Given that a terrorist attack occurs, what is the probability of 20 or more people dying?" or "What is the probability of an attack more deadly than 9/11 occurring in the next 20 years?"
- Of course, the techniques we will use are not restricted to just modelling terrorist attacks - they can broadly be applied to any situation where we are concerned with estimating the probability of certain events happening (the Value-at-Risk analysis from last week was similar)

Statistical Analysis

- We will (for now) treat the fatality count of each attack as independent of all other attacks. This is broadly justified based on the previous plots of the data where there was no obvious serial correlation.

- Suppose we have n attacks (here $n = 13,274$). For $1 \leq i \leq n$, let Y_i denote the number of deaths in the i^{th} attack. We assume each Y_i is independently drawn from some distribution $p(y|\theta)$:

$$Y_i \sim p(y|\theta), \quad 1 \leq i \leq n,$$

where $p(y|\theta)$ specifies some probability distribution (e.g. Exponential) and θ is an unknown vector of parameters.

y

Basic Model - Exponential Distribution

- For the first half of this lecture we will assume that the fatalities in terrorist attacks follow an Exponential distribution $p(y|\lambda) = \lambda e^{-\lambda y}$ (here $\theta = \lambda$).
- The Exponential distribution is a standard model for phenomena which seem to decay exponentially, such as the empirical density plot shown a few slides ago. We will investigate how reasonable this assumption is in the second half of the lecture!
- Note: the Exponential distribution assumes the Y_i variables are continuous. But here they are discrete (number of deaths). However continuous distributions tend to be easy to work with, and modelling discrete data as if it were continuous is usually fine, as long as the numerical values cover a wide range (they are between 1 and 2,749 here)

Fatalities in a Single Attack

- First consider the question: "What is the probability of D or more people dying in a single terrorist attack?"
- As a point of notation, we will use Y_i to denote the fatalities in our data set (i.e. in the 13,274 attacks that have already happened), and $\tilde{Y}$ to denote the fatalities in a future attack that we are predicting (where we are assuming that $\tilde{Y}$ has the same distribution as Y_i, i.e. $\tilde{Y} \sim p(y|\theta)$).
- In other words, we are using the observed data $Y_1, \ldots, Y_n$ to make predictions about future attacks $\tilde{Y}$. So "What is the probability of D or more people dying in a single terrorist attack?" is equivalent to computing $p(\tilde{Y} \geq D)$

Fatalities in a Single Attack

- If the distribution $p(y|\theta)$ is given, and θ **is also known** (i.e. for now we are treating it as known rather than estimating it), then the probability of a particular attack killing more than D people is given by integrating over the tail of the probability distribution

$$p(\tilde{Y} \geq D) = \int_D^\infty p(y|\theta)dy = 1 - F_Y(D|\theta)$$

where $F_Y(D|\theta) = p(Y \leq D|\theta)$ is the distribution function.

- Note: remember we are using a continuous distribution to model Y here, so $p(Y \leq D|\theta) = p(Y < D|\theta)$.

Example

- We model the fatalities using an Exponential distribution. The average number of fatalities is 4.17, so the maximum likelihood estimate of λ is $\hat{\lambda} = 1/4.17 = 0.24$.
- Assume for now that **this is the true value of** λ (i.e. that there is no estimation uncertainty). The probability of more than 10 people dying in a particular attack is then:

$$1 - \int_0^{10} 0.24 e^{-0.24y} dy = 0.091$$

- So there is a 0.9% chance of 10 or more people dying in a single attack.

Example

Similarly for $D \in \{10, 20, 50, 100, 200, 500, 2749\}$ the probabilities of D or more people dying in a single attack are:

D	Probability
10	9.09×10^{-2}
20	8.28×10^{-3}
30	7.53×10^{-4}
50	6.23×10^{-6}
100	3.89×10^{-11}
200	≈ 0
500	≈ 0
2749	≈ 0

So under this model the probability of over 100 people dying is extremely small – around 0.000000004%

The probability of 200 or more deaths is so small it is essentially zero.

Distribution of Sample Maximum

- So far we have considered how to compute the probability of **one particular attack** killing D or more people.
- However, there may be many attacks occurring over a given time period.
- Suppose there are m attacks during some period. We now want to ask about the probability of **at least one** of these m attacks killing D or more people. Hence, if the m attacks $\tilde{Y}_1, \ldots, \tilde{Y}_m$ occur what is the probability that at least one has a value greater than D?.
- More formally, let $\tilde{Y}_1, \ldots, \tilde{Y}_m$ be independent draws from $p(y|\theta)$. What is the distribution of the sample maximum $M = \max_i \tilde{Y}_i$?

Distribution of Sample Maximum

- Again we assume for now that θ is known exactly, not estimated.

- For any value D, the probability that the maximum M is less than D is equal to the probability that all the $\tilde{Y}_i$ variables are less than D:

$$\begin{aligned} p(M \leq D) = p(\max_i \tilde{Y}_i \leq D) &= \prod_{i=1}^{m} p(\tilde{Y}_i < D) \\ &= p(Y \leq D)^m \\ &= F_Y(D)^m \end{aligned} \tag{1}$$

So:

$$p(M \geq D) = 1 - p(Y \leq D)^m = 1 - F_Y(D)^m$$

Distribution of Sample Maximum - Example

- Based on the terrorism dataset, there have been approximately 2000 terrorist attacks each year, from the year 2002 onwards.
- **Question**: if the death count in terrorist attacks is known to follow an $Exponential(0.24)$ distribution and 2000 attacks occur in a given year, what is the probability that at least one will kill 30 or more people?
- **Answer**: We need the probability that the sample maximum is equal to or greater than 30. By the previous slide, the distribution of the sample maximum M is:

$$p(M \geq 30) = 1 - p(Y \leq D)^m = 1 - \left[\int_0^{30} \hat{\lambda} e^{-\lambda y} dy\right]^{2000} = 0.78$$

- So there is a 78% chance of at least one attack killing 30 or more people

Distribution of Sample Maximum - Example

Similarly, if 2000 attacks occur in a year, then the probability of seeing at least one that kills D or more people is:

D	Probability
10	≈ 1
20	≈ 1
30	0.78
50	0.01
100	7.78×10^{-8}
200	≈ 0
500	≈ 0
2749	≈ 0

Incorporating Uncertainty

- Of course in practice λ is not known, and must be estimated.
- In the last few slides, we simply assumed that λ was equal to the maximum likelihood estimate and plugged in this number.
- As we have discussed, this is typically bad practice since it ignores the uncertainty in λ.

- To incorporate this uncertainty, we can estimate λ using Bayesian inference and then perform analysis using the posterior distribution $p(\lambda|\mathbf{y})$.

Incorporating Uncertainty

- One of the nice aspects of Bayesian statistics is that it is very easy to make predictions about future events – we simply replace any unknown quantities with their posterior distribution, and average over this (by integrating).
- Suppose $p(\lambda|\mathbf{y})$ is our posterior estimate of λ after observing some data $\mathbf{y} = \{y_1, \ldots, y_n\}$. We predict future values $\tilde{Y}$ using:

$$p(\tilde{Y} \geq D|\mathbf{y}) = \int p(\tilde{Y} \geq D|\lambda)p(\lambda|y)d\lambda = \int (1 - F_Y(D|\lambda))p(\lambda|\mathbf{y})d\lambda$$

- Remember our notation: $\mathbf{y} = (y_1, \ldots, y_n)'$ denotes the 13,274 attacks we have observed.
- $\tilde{Y}$ denotes a future attack we are predicting, which has the same distribution as Y_i)

Incorporating Uncertainty

- In our Exponential example we need to learn λ using Bayesian inference.
- The conjugate prior is Gamma(α, β) where we choose α and β to reflect prior beliefs about λ:

$$p(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

- The posterior distribution in this case is (Exercise 1, Question 5):

$$p(\lambda|y) = Gamma(\alpha + n, \beta + \sum_{i} y_i)$$

Incorporating Uncertainty

- The cumulative distribution function for the Exponential distribution is $p(Y \leq D) = 1 - e^{-\lambda D}$

$$p(\tilde{Y} \geq D) = \int (1 - F_Y(D|\lambda)) p(\lambda|Y) d\lambda = \int (e^{-\lambda D}) \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} \lambda^{\tilde{\alpha}-1} e^{-\tilde{\beta}\lambda} d\lambda$$

  where $\tilde{\alpha} = \alpha + n$ and $\tilde{\beta} = \beta + \sum_i y_i$.
- Simplifying gives:

$$p(\tilde{Y} \geq D) = \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} \int \lambda^{\tilde{\alpha}-1} e^{-\lambda(\tilde{\beta}+D)} d\lambda$$

- As always, when working with conjugate priors we evaluate this integral by recognising it has the same form as the prior distribution (Gamma in this case).

Incorporating Uncertainty

- We define $\hat{\beta} = \tilde{\beta} + D$ and part under the integral sign becomes:

$$\int \lambda^{\tilde{\alpha}-1} e^{-\lambda\hat{\beta}} d\lambda = \frac{\Gamma(\tilde{\alpha})}{\hat{\beta}^{\tilde{\alpha}}}$$

- Thus the whole thing becomes:

$$p(\tilde{Y} \geq D) = \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} \frac{\Gamma(\tilde{\alpha})}{\hat{\beta}^{\tilde{\alpha}}}$$

- Substituting everything back in gives:

$$p(\tilde{Y} \geq D) = \frac{(\beta + S)^{\alpha+n}}{(\beta + S + D)^{\alpha+n}}, \quad S = \sum_{i=1}^{n} y_i$$

where α and β are the parameters of the prior.

Incorporating Uncertainty

- As always, if we have strong prior knowledge about λ, then we choose an informative prior that reflects this.
- Otherwise, we choose an uninformative prior which reflects our ignorance.
- A fairly common choice of uninformative prior for the Exponential distribution parameter is Gamma$(0.01, 0.01)$, i.e. $\alpha = 0.01$ and $\beta = 0.01$.

Incorporating Uncertainty

- Using this prior, we compute the probability of more than D people dying in a single attack for various values of D, and compare this to our previous findings when we assumed that λ was equal to the maximum likelihood estimate:

D	Probability assuming $\lambda = \hat{\lambda}$	Probability With Estimated λ
10	9.09×10^{-2}	9.10×10^{-2}
20	8.28×10^{-3}	8.29×10^{-3}
30	7.53×10^{-4}	7.55×10^{-4}
50	6.23×10^{-6}	6.27×10^{-6}
100	3.89×10^{-11}	3.97×10^{-11}
200	≈ 0	≈ 0
500	≈ 0	≈ 0
2749	≈ 0	≈ 0

Incorporating Uncertainty

- In this particular case, incorporating uncertainty about λ makes very little impact on the final results.
- The main reason for this is that we used an uninformative prior, and also had a large number of observations (13,274) so the effect of the prior distribution is minimal. Recall:

$$p(\tilde{Y} \geq D) = \frac{(\beta + S)^{\alpha+n}}{(\beta + S + D)^{\alpha+n}}, \qquad S = \sum_{i=1}^{n} y_i$$

- Since S is very large when there are a lot of observations, it will dominate this term and the prior term β will hence have little effect unless it is very large (in the uninformative case it was 0.01).

Summary

- We have learned how to compute the probability of extreme events occurring under the Exponential probability model, both when λ is known, and when it has been estimated.
- However there is a problem – we are finding that the probability of terrorist attacks with 200 or more deaths occurring is so low that it is almost impossible.
- But the historical record shows that these attacks do happen – specifically in our 13,274 attacks, there are 13 that have 200 or more deaths.

How can this be explained?

The Role of Assumptions

- The key issue here is that we **assumed** the Exponential distribution was a good model for this data. But the fact that we are assigning very small probabilities to events that regularly happen suggests that this may not be the case.

- This is a key lesson: while the choice of distribution $p(y|\theta)$ for the likelihood is important in all statistical modelling, it is **especially important** when it comes to analysing the probability of large (extreme) events occurring.

- If we model $p(y|\theta)$ using a different distribution than the Exponential, we might end up with very different results. Hence we must be **very** careful when it comes to choosing our statistical model.

The Role of Assumptions - Exponential Decay

- Recall the form of the Exponential distribution:

$$p(y|\lambda) = \lambda e^{-\lambda y}$$

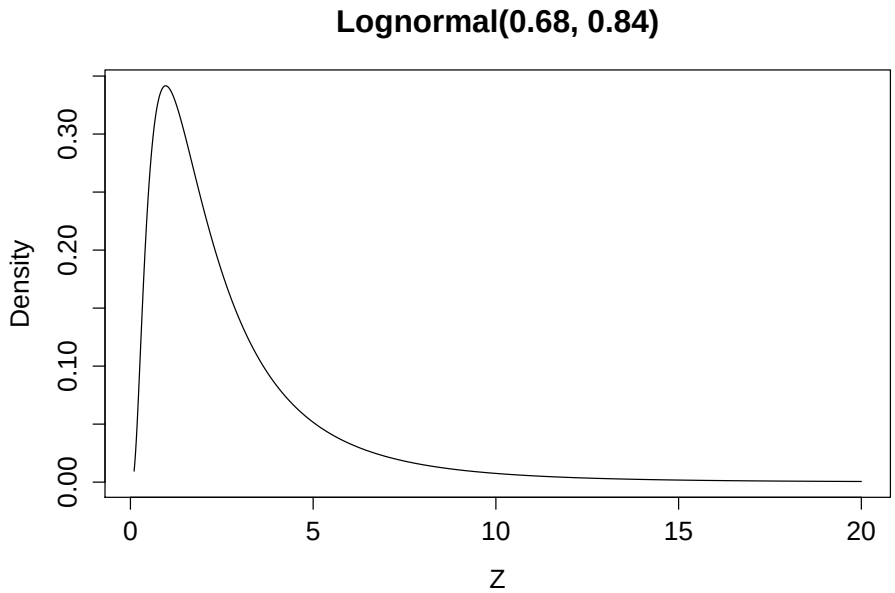
- This function decays exponentially in y – which means that it decays very fast as y becomes larger.
- This means that under the Exponential model, the probability of large values occurring is very small.
- This is not a realistic model in many situations where outliers often occur!

Lognormal Distribtuion

- An alternative distribution which is often used to model the probability of extreme events occurring is the Lognormal distribution.
- A random variable Z has a $Lognormal(LN)$ distribution with parameters μ and σ^2 if $\log(Z)$ has a $N(\mu, \sigma^2)$ distribution.
- In other words, if $Z \sim LN(\mu, \sigma^2)$ then $Z = e^X$, where X has a $N(\mu, \sigma^2)$ distribution.
- Like the Exponential distribution (and unlike the Normal), the Lognormal distribution only puts non-zero probability on the interval $(0, \infty)$ – i.e. draws from this distribution are always positive real numbers.

Lognormal Distribtuion

- The Lognormal distribution can replicate behaviour that looks like Exponential decay. E.g. if $\mu = 0.68$ and $\sigma^2 = 0.84$ this distribution looks like:



- The fact this has two parameters (μ, σ^2) while the Exponential only has one (λ) means that it can better fit the tails of empirical data.

Lognormal Distribution

- We can easily derive the functional form of the Lognormal distribution using the standard univariate transformation theorem.
- Let $Z \sim Lognormal(\mu, \sigma^2)$ then $Z = e^X$ where $X \sim \mathrm{N}(\mu, \sigma^2)$, so:

$$p(Z \leq z|\mu, \sigma^2) = p(e^X \leq z) = p(X \leq \log(z)) = \Phi_{\mu,\sigma^2}(\log(z))$$

where Φ_{μ,σ^2} denotes the Normal cumulative probability distribution:

$$p(z|\mu, \sigma^2) = \frac{d}{dz} p(Z \leq z) = \frac{d}{dz} \Phi_{\mu,\sigma^2}(\log(z)) = \frac{1}{z} \phi_{\mu,\sigma^2}(\log(z))$$

where ϕ_{μ,σ^2} is the Normal density, so:

$$p(z|\mu, \sigma^2) = \frac{1}{z} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(\log(z)-\mu)^2}{2\sigma^2}\right)}$$

Parameter Estimation

- To estimate the parameters of the Lognormal distribution, we essentially just take the log of the observations and treat them as if they were Normal. If our observations are $Y_1, \ldots, Y_n$ then the maximum likelihood estimates are:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \log(y_i)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\log(y_i) - \hat{\mu})^2$$

- Bayesian inference for these parameters is done similarly – just take the log of the observations and then use the methodology from last week.

Predictions Under The Lognormal Model

- If we fit the Lognormal distribution to the terrorism data, the MLE of the parameters is $\hat{\mu} = 0.68$ and $\hat{\sigma}^2 = 0.84$.
- We can then compute the probability of large terrorist attacks occurring under this model.
- If we again treat $\hat{\mu}$ and $\hat{\sigma}^2$ as being equal to the true parameter values (ignoring estimation uncertainty) like we initially did for the Exponential distribution, we can find (e.g.) the probability of 10 people dying in a single attack in the same way as before:

$$p(\tilde{Y} \geq 10) = \int_{10}^{\infty} p(y|\theta)dy$$

where we now use a Lognormal model for $p(y|\theta)$.

Comparison - Single Attack

For $D \in \{10, 20, 30, 50, 100, 200, 500\}$ we can hence compare the probability of D or more fatalities in a single attack under both the Exponential and Lognormal models:

D	Exponential Model	Lognormal Model
10	9.09×10^{-2}	3.80×10^{-2}
20	8.28×10^{-3}	5.71×10^{-3}
30	7.53×10^{-4}	1.48×10^{-3}
50	6.23×10^{-6}	2.08×10^{-4}
100	3.89×10^{-11}	9.11×10^{-6}
200	0	2.30×10^{-7}
500	0	7.63×10^{-10}
2749	0	1.33×10^{-15}

Comparison - Multiple Attacks

Similarly, if 2000 attacks occur in a year, then the probability of seeing at least one that kills D or more people is:

D	Exponential	Lognormal
10	≈ 1	≈ 1
20	≈ 1	≈ 1
30	0.78	0.95
50	0.01	0.34
100	7.78×10^{-8}	0.01
200	≈ 0	4.61×10^{-4}
500	≈ 0	1.53×10^{-6}
2749	≈ 0	2.67×10^{-11}

Again, large attacks are much more likely under the Lognormal model. This highlights how much the choice of likelihood function can affect the results – particularly when D is large.

Summary

- So, what have we learned?
- When doing statistical analysis, choosing the right model (Exponential vs Lognormal etc) is always important. But it is extremely important when measuring the probability of **rare** events happening.
- Since we don't have many observations in the tails of the distribution (only 13 of the 13,274 events had more than 200 fatalities), it will be parametric form of the distribution which carries most weight in prediction. As such, we have to get this right!
- Of course, the question is now – how do we know which distribution to use? How do we know whether the Exponential or Lognormal distribution gives a better fit to the data? This leads us to the topic of **model selection**

Model Selection

Model Selection

- Model selection is the task of choosing which of two or more different probability models (usually likelihood functions) are better suited to modelling a particular data set.
- For example, we here want to choose between the Exponential and Lognormal models.
- Model selection is one of the most controversial areas of statistics – if you ask a group of statisticians the best way to do it, you will get several different answers.

Bayesian Model Selection

In theory, Bayesian model selection is simple and logical. Suppose we have K different models $M_1, \ldots, M_K$. In our case $K = 2$, and the models are:

- $M_1 : p(\mathbf{y}|\theta)$ should be modelled using an Exponential distribution
- $M_2 : p(\mathbf{y}|\theta)$ should be modelled using a Lognormal distribution

The Bayesian approach involves computing the posterior distribution of both models $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$. These respectively correspond to the belief we have about Models 1 and 2 being correct after seeing the data.

We then go with the most probable model – e.g. use an Exponential distribution if $p(M_1|\mathbf{y}) > p(M_2|\mathbf{y})$

Bayesian Model Selection

We can compute these posterior distributions using Bayes theorem.
For Model 1:

$$p(M_1|\mathbf{y}) = \frac{p(\mathbf{y}|M_1)p(M_1)}{p(\mathbf{y})}$$

Here $p(M_1)$ is the prior belief we have the Model 1 is correct before seeing the data, and $p(\mathbf{y}|M_1)$ is the **marginal likelihood** of the data Y under Model i.
Similarly for Model 2:

$$p(M_2|\mathbf{y}) = \frac{p(\mathbf{y}|M_2)p(M_2)}{p(\mathbf{y})}$$

Bayesian Model Selection

- Note that $p(\mathbf{y})$ occurs in both formula and does not depend on the model. Since it is common to both, we can simply ignore it.
- We will also usually assume that the prior $p(M_i)$ on each model is equal – i.e. we do not assume any model is more likely than the others.
- As such, the only terms that matter are the $p(\mathbf{y}|M_i)$ terms. We will choose the model for which $p(\mathbf{y}|M_i)$ is largest.

Bayesian Model Selection - Marginal Likelihood

- Let θ_i denote the vector of unknown parameters that occurs in Model i. In our case, $\theta_1 = \{\lambda\}$ and $\theta_2 = \{\mu, \sigma^2\}$ corresponding to the parameters of the Exponential and Lognormal distributions respectively.
- Then:

$$p(\mathbf{y}|M_i) = \int p(\mathbf{y}|\theta_i)p(\theta_i)d\theta_i$$

- Problem: In practice these integrals tend to be very difficult to calculate outside simple situations.

Bayesian Model Selection

Although this approach is simple in theory, in practice computing these integrals can be difficult. In this course we will hence use an approximation to $p(M_i|\mathbf{y})$ which is very easy and quick to calculate.

This approximation is called the **Bayesian Information Criterion** (BIC). It states that as long as we have "enough" observations $\mathbf{y}$, then:

$$\log p(\mathbf{y}|M_i) \approx \log p(\mathbf{y}|\hat{\theta}_i) - 0.5 k_i \log(n)$$

where:

- n is the number of observations ($13,274$ in our case)
- $\hat{\theta}_i$ is the maximum likelihood estimate of θ_i in Model i
- k_i is the number of parameters in Model i. Since the Exponential distribution has a single parameter λ, we have $k_1 = 1$. Similarly $k_2 = 2$ since the Lognormal distribution has 2 parameters.

Bayesian Model Selection - Example

- Let's see how this works in practice using the terrorism data. First, we consider model M_1 (Exponential distribution). As discussed earlier, the MLE of λ is $\hat{\lambda} = 0.24$.

- The BIC approximation is hence:

$$\log p(\mathbf{y}|M_1) \approx \sum_{i=1}^{13,274} \left[\log\left(\hat{\lambda} e^{\hat{\lambda} y_i}\right)\right] - 0.5 \times \log(13,274) = -32243.51$$

Bayesian Model Selection - Example

- Similarly for the Lognormal distribution we found that the MLEs were $\hat{\mu} = 0.68$ and $\hat{\sigma}^2 = 0.84$. So:

$$\log p(\mathbf{y}|M_2) \approx -26687.5$$

- This is substantially higher than the BIC for the Exponential model, meaning that the Lognormal gives a much better fit to this data.

- Therefore, model M_2 is preferred.

Bayesian Model Selection - Summary and Warning

- This method of doing Bayesian model selection is very general – in practice if we have multiple models to choose between, we simply compute/approximate $p(\mathbf{y}|M_i)$ for each, and choose the model for which this is largest.
- We will not prove why the BIC approximation works, but in short, the proof treats model selection as a decision theory problem where there is a 0-1 loss function.
- Recall from Lecture 2 that under a 0-1 loss function, the posterior is best summarised using its mode.
- The BIC approximation should hence **not be used in situations where we have very strong prior beliefs, or only a small number of observations**.

Some Extreme Value Theory

A Problem Still Remains...

In the paper that this lecture is loosely based, it is claimed that the probability of an event as deadly as 9/11 (2,749 deaths) occurring over their 40 year period is as high as 0.35.

But we found that the probability of more than a few hundred deaths occurring was tiny.

Model Selection is Hard!

So far we have only shown that the Lognormal distribution is preferable to the Exponential distribution for this data. But this does not mean it is the best (or even an appropriate) model!

Normally when doing applied statistics we can look at plots of the data, come up with some reasonable choices of probability model, and then do model selection. And this would be fine if we are only interested in $p(\tilde{Y} \geq D)$ when D is small.

However when D is large (e.g. $D = 2,749$), even small misspecifications of the probability model will lead our predictions to be out by many orders of magnitudes. So just assuming that the data is either Exponential or Lognormal is problematic. Fortunately there is a mathematical result that can help us out...

Pickands-Balkema-de Haan Theorem

We now come to what is essentially the "Fundamental Theorem of Extreme Value Statistics". It is comparable to the Central Limit Theorem in the sense of being extremely general, and very powerful.

The PBH theorem essentially says that if we go far enough into the extreme tails (i.e. when D is large), almost **every** probability distribution eventually starts to look "the same"

(This is comparable to the Central Limit Theorem, which tells us that if we add up enough independent and identically distributed random variables, the sum will always have the same distributional form [Normal] regardless of the distribution of the variables)

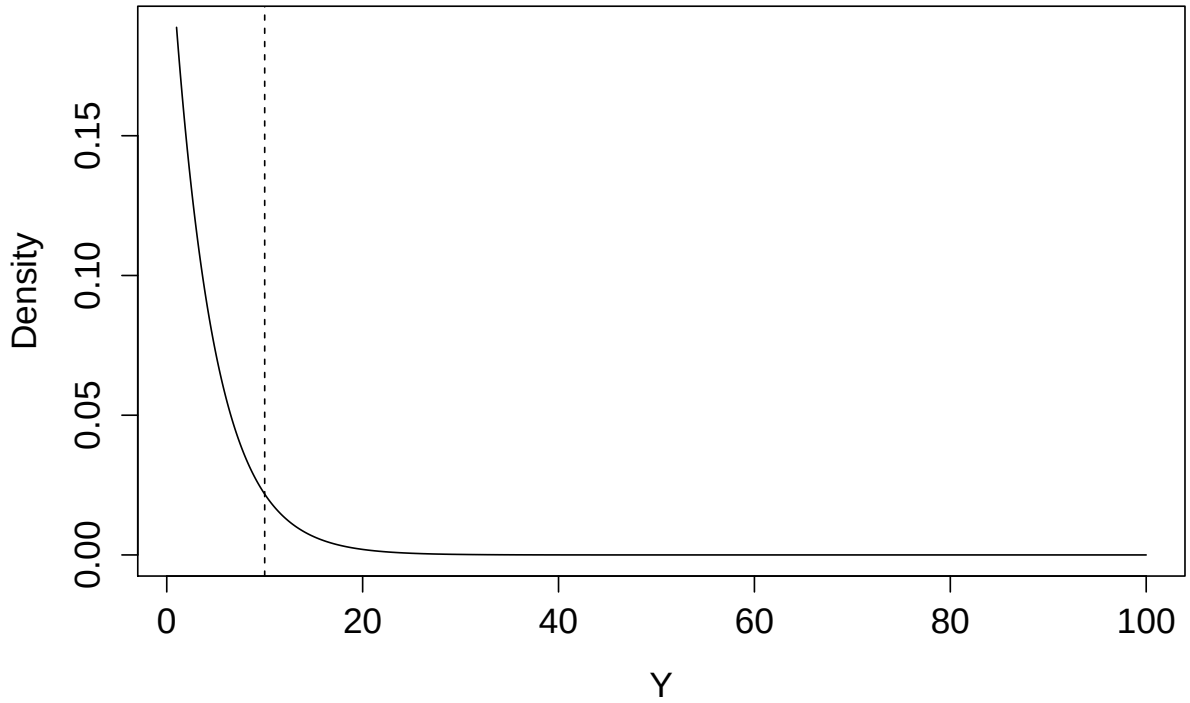
Pickands-Balkema-de Haan Theorem

More formally, suppose we have observations $Y_1, \ldots, Y_n$ from **any** probability distribution, and we are interested in computing $p(Y \geq D)$. We do not want to assume any particular function form for $p(y|\theta)$ (e.g. Exponential or Lognormal).

Let u denote some threshold value. The basic idea is that to avoid having to make assumptions about the form of $p(y|\theta)$, we can instead only model the distribution of the Y_i values **above this threshold** rather than the whole distribution.

Pickands-Balkema-de Haan Theorem

So we only attempt to model the tail distribution to the right of u (the dotted line), while ignoring the rest of it:



Pickands-Balkema-de Haan Theorem

Define $F_u(D-u) = P(Y-u \leq D-u | Y \geq u)$ to be the "conditional exceedence distribution" – it models the distribution of Y **conditional on Y being greater than u**. It is hence a model for the tail part of the distribution only.

The PBH theorem states that as $u \to \infty$ then $F_u(D-u)$ converges to the Generalised Pareto Distribution **regardless of the true distribution of** Y. Hence for essentially any distribution (Exponential, Lognormal, etc) if we go far enough into the tails, we can model it using the Generalised Pareto Distribution (GPD)

As such, to estimate the probability of extreme events occurring when we don't have a good choice for $p(y|\theta)$, we can choose to instead model only the distribution tails, using the GPD. As long as we choose u to be large enough, this will be a good model.

GPD distribution

Aside from the threshold choice u, the Generalised Pareto Distribution has two parameters k and σ. Its distribution function is:

$$p(Y \leq y|u, k, \sigma) = 1 - (1 - k\frac{y-u}{\sigma})^{(1/k)}$$

and the density function is:

$$p(y|u, k, \sigma) = \frac{1}{\sigma}(1 - k\frac{y-u}{\sigma})^{(1/k-1)}$$

It is important to remember that we can only use this model when $Y \geq u$!

GPD distribution - Parameter Estimation

Estimating the parameters (k, σ) can be difficult - there is no conjugate prior distribution

Point estimates can be derived using the method of moments, and these tend to be reasonably close to the MLE.
Suppose that out of the n observations y_i, there are m of them greater than u. Denote these by $y'_1, \ldots, y'_m$. The estimators are:

$$\hat{\sigma} = 1/2((\bar{y'}/s)^2 + 1)\bar{y'}$$

$$\hat{k} = 1/2((\bar{y'}/s)^2 - 1)$$

where $\bar{y'}$ and s^2 denote the sample mean and variance of the transformed observations $z_1 = y'_1 - u, \ldots, z_m = y'_m - u$:

$$\bar{y'} = \frac{1}{m}\sum_{i=1}^{m} z_i, \qquad s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(z_i - \bar{y'})^2$$

GPD distribution - Fitting to Our Data

Let's fit the GPD distribution to the terrorism data. The first step is choosing the threshold u – i.e. how to we determine the point in the tails above which the GPD gives a good approximation?

This is a difficult question with no single answer! A heuristic which is often used is to choose u to be the 95^{th} percentile of the observed data (i.e. if we have n observations, then choose u to be the $0.95 \times n^{th}$ smallest observation, which is the $13274 * 0.95 = 12609^{th}$ smallest observation in our case.

In the paper this lecture is based on they chose $u = 10$, so we will also use this.

GPD distribution - Fitting to Our Data

We discard all values of Y_i that are smaller than 10. This leaves 986 attacks. Using only these attacks, we compute the estimates:

$$\hat{\sigma} = 8.24$$

$$\hat{k} = -0.60$$

Now suppose we want to find the probability of D or more people dying in a single terrorist attack under the GPD. The calculation is slightly different to before – the GPD only models the tail behaviour, so we need an additional term which reflects the probability of an observation being in the tail (i.e. being greater than u).

GPD distribution - Making Predictions

By the theorem of total probability we have:

$$p(Y \geq D) = p(Y \geq D|Y \geq u)p(Y \geq u) + p(Y \geq D|Y < u)p(Y < u)$$

The term $p(Y \geq D|Y < u)$ is equal to zero when $D > u$. We can estimate $p(Y \geq u)$ by the empirically observed frequency of attacks that killed at least u people. When $u = 10$ there are 986 such attacks, so $p(Y \geq u) = 986/13274 = 0.074$

Finally $p(Y \geq D|Y \geq u)$ comes from then cumulative distribution function of the GPD distribution.

Comparison - Single Attack

For $D \in \{10, 20, 30, 50, 100, 200, 500, 2749\}$ we can hence compare the probability of D or more people dying in a single attack under the Exponential, Lognormal, and GPD models:

D	Exponential Model	Lognormal Model	GPD Model
10	9.09×10^{-2}	3.80×10^{-2}	0.07
20	8.28×10^{-3}	5.71×10^{-3}	0.03
30	7.53×10^{-4}	1.48×10^{-3}	0.02
50	6.23×10^{-6}	2.08×10^{-4}	0.007
100	3.89×10^{-11}	9.11×10^{-6}	0.002
200	0	2.30×10^{-7}	0.0008
500	0	7.63×10^{-10}	0.0002
2749	0	1.33×10^{-15}	0.00001

This results in a much higher probability that extreme events will occur than under the previous models!

Comparison - Multiple Attacks

Similarly, if 2000 attacks occur in a year, then the probability of seeing at least one that kills D or more people is:

D	Exponential	Lognormal	GPD
10	≈ 1	≈ 1	≈ 1
20	≈ 1	≈ 1	≈ 1
30	0.78	0.95	≈ 1
50	0.01	0.34	≈ 1
100	7.78×10^{-8}	0.01	≈ 1
200	≈ 0	4.61×10^{-4}	0.80
500	≈ 0	1.53×10^{-6}	0.30
2749	≈ 0	2.67×10^{-11}	0.02

Comparison - Multiple Attacks

Now that we have fit a more principled tail model (rather than simply guessing Exponential or Lognormal) we see the probabilities of extreme events occurring is much higher!

Under the GPD model we find that there is a 2% probability that an attack at least as deadly 9/11 (which killed 2,749 people) will happen in any given year. This is very high!

Doing a similar calculation, the probability of such an attack occurring in the next decade is around 19%

This assumes that the typical rate at which terrorist attacks occur remains at around 2000 attacks per year (recall that this has only been the case since the year 2002 - previously there were far fewer attacks each year). In this weeks exercise sheet we discuss whether this assumption is reasonable!

Summary

The purpose of this lecture was to show how to estimate the probability of extreme events occurring using a real dataset.

Working with real data is difficult because we have to choose which probability model to use, which is something of an art.

The choice of model becomes more and more important when the events we are trying to estimate become more and more extreme (i.e. as D becomes larger).

Summary

One of the important messages from this lecture is that the probabilities we estimate are **very** sensitive to the modelling assumptions.

As such you should be cautious whenever you read a newspaper article about a scientific study which claims to have estimated the probability of some catastrophic event occurring - it will usually be the case that they would have reached a very different answer if they had made small changes in their model!

This is one of the difficult areas of applied statistics for this reason.