

How do we calculate the probability of an extreme event?

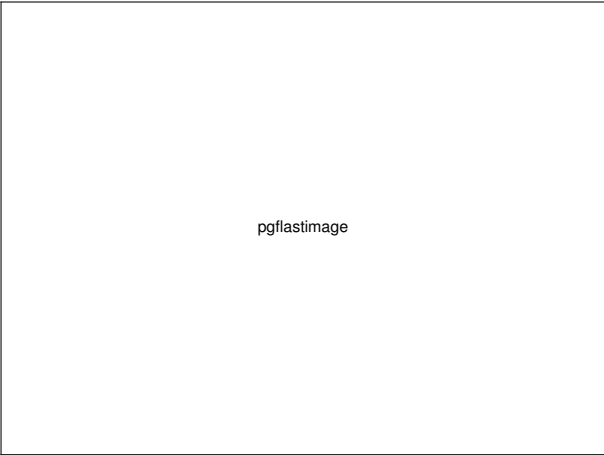
Peter Taylor

The University of Melbourne

July 27, 2012

Floods in Toowoomba

Tuesday, January 11, 2011.



pgflastimage

Floods in Victoria



Saturday, January 15, 2011.

A problem to think about

Suppose that you are the engineer for the local government authority. The authority has asked you to build a levee to protect the town from flooding.

Your task is to design it. In particular, you need to decide how high it should be. Considerations are:

- If there is a flood higher than the levee, it will 'overtop' and inundate the town.
- The higher the levee is, the more expensive it will be to build. There could be other downsides in terms of amenity if the levee is built too high.

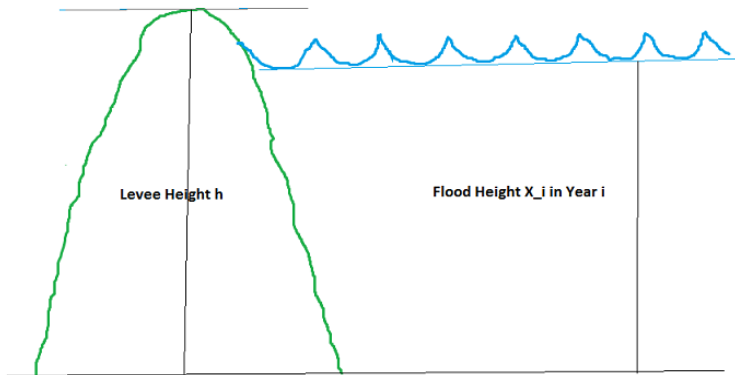
A simple model

Typically you will have some historical data about flood levels.

Year	Maximum level
1841	5.3m
1842	6.5m
1843	5.6m
⋮	⋮

However, the time series of data may not contain any instance of a flood as extreme as the one that we want to protect against.

A simple model



Once we build the levee, its height h is fixed. However the flood height at any given time is a **random variable**. Let X_i be the maximum flood height in year i .

A simple model

Assume that the levee has a design lifetime of T years. In order to inform our decision about how high to build the levee, we need to work out the probability that a flood will overtop the levee at some time in the next T years if we build it at height h .

It is actually easier to work out the probability that a flood will **not** overtop the levee. Mathematically, we can write this as

$$P(M_T \leq h)$$

where

$$M_T = \max_{i=1, \dots, T} X_i.$$

A simple model

Straightaway we have

$$P(M_T \leq h) = P(X_i \leq h \text{ for all } i = 1, \dots, T).$$

We can rewrite this as

$$P(M_T \leq h) = P(X_1 \leq h, X_2 \leq h, \dots, X_T \leq h).$$

If we now assume that the maximum flood levels in each year are independent (this could be a dubious assumption), then we can write this as

$$P(M_T \leq h) = P(X_1 \leq h)P(X_2 \leq h) \dots P(X_T \leq h).$$

A simple model

So

$$P(M_T \leq h) = \rho^T$$

where

$$\rho = P(X_i \leq h).$$

Since it is a probability, ρ is a number between 0 and 1. If $\rho = 1$ (what does this mean physically), then $\rho^T = 1$ for all T .

On the other hand if ρ is any other number between 0 and 1, ρ^T will approach zero as the design lifetime T gets large.

A simple model

We would like to derive an expression for ρ^T where T is reasonably large, say $T = 50$ or $T = 100$.

The problem is that we have no direct way of estimating from the data what the value of ρ is.

Remember that, just because we have never seen a flood of height h , it does not mean that it can't happen.

An aside

To illustrate how statisticians think about such problems, I'm going to consider a related problem, which you may be familiar with. This is to calculate

$$P(AV_T < h)$$

where

$$AV_T = (X_1 + X_2 + \cdots + X_T)/T$$

is the average (rather than the maximum) flood height over our planning period T .

An aside

One of the most well-known results in statistics, the **Central Limit Theorem** tells us that, provided the mean and variance of the X_i exist, we can find scaling factors a_T and shift factors b_T such that

$$P(AV_T \leq a_T h + b_T) \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^h e^{-y^2/2} dy$$

as T gets big.

The expression on the right is the distribution function for the **Standard Normal Distribution**.

Two reasons to believe in God

- The limiting distribution is normal, whatever the distribution of the X_j , provided that it has a mean and variance.
- The scaling and shift factors depend on the distribution of the X_j only through some simple measures.

Specifically, the factor b_T is the **mean** μ of the random flood heights X_j , which we can estimate from our data by putting it equal to the sample average.

The factor a_T is σ/\sqrt{T} , where σ^2 is the **variance** of the X_j , which we can also estimate from our data.

The Central Limit Theorem

The upshot is that

$$P(AV_T \leq \tilde{\sigma}h/\sqrt{T} + \tilde{\mu}) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^h e^{-y^2/2} dy$$

where $\tilde{\mu}$ and $\tilde{\sigma}$ are our estimates of μ and σ respectively.

This is the same as saying

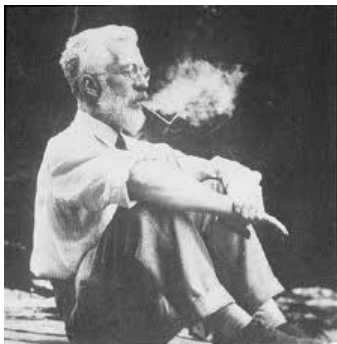
$$P(AV_T \leq h) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(h-\tilde{\mu})\sqrt{T}/\tilde{\sigma}} e^{-y^2/2} dy.$$

Back to our problem

Our approach to making statements about probabilities involving the maximum is essentially the same: we just need to change the details concerning the limiting distribution and the estimation of the scaling and shift parameters.

The analogue of the Central Limit Theorem that applies to maxima is the **Three Types Theorem**, which was originally stated by **Fisher and Tippett (1928)**, and later proved rigorously by **Gnedenko (1943)**.

Some personal information

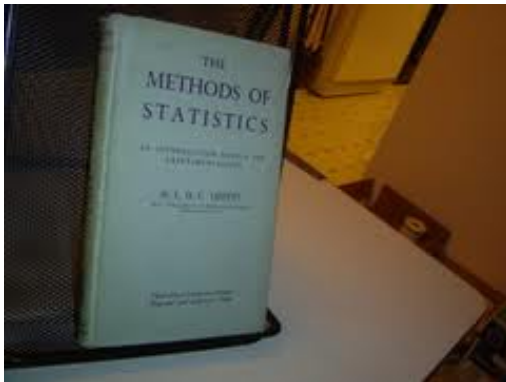


R. A. Fisher (1890-1962). A genuine great in two fields of science.

Hald said 'a genius who almost single-handedly created the foundations of statistical science'.

Dawkins said 'the greatest biologist since Darwin'.

Some personal information



L. H. C. Tippett (1902-1985). An English statistician. Spent his entire career, 1925 to 1965, on the staff of the Shirley Institute, Manchester. Mostly known for his work with the textile industry.

Some personal information



B. V. Gnedenko (1912-1995).

A leading member of the Russian school of probability theory and statistics.

He also worked on applications of statistics to reliability and quality control in manufacturing.

The Three Types Theorem

Says that **if** we can find a sequence of scaling constants a_T and shifts b_T such that

$$P(M_T \leq a_T h + b_T) \rightarrow F(h)$$

as T gets big, then (up to scale and shift) the limit function $F(h)$ must take one of three forms.

$$F(h) = \exp(-e^{-h}) \quad \text{The Gumbel Distribution}$$

$$F(h) = \begin{cases} 0, & x < 0 \\ \exp(-x^{-\alpha}) & x \geq 0 \end{cases} \quad \text{The Fréchet Distribution}$$

$$F(h) = \begin{cases} \exp(-|x|^\alpha) & x < 0 \\ 1 & x \geq 0. \end{cases} \quad \text{The Weibull Distribution}$$

where $\alpha > 0$.

Some personal information



E.J. Gumbel (1891-1966). A German mathematician, who was a prominent anti-Nazi intellectual. Fled to France in 1932 and to the US in 1940.

Some personal information



M.R. Frechét (1878-1973). Made major contributions to the topology of point sets and introduced the concept of a metric spaces.

Some personal information



Walodi Weibull 1887-1979

Photo by Sam C. Saunders

E.H.W. Weibull (1887-1979). A Swedish engineer, who published many papers on strength of materials, fatigue, rupture in solids, bearings, and the Weibull distribution.

A simple model

The problem of making predictions about the maximum of the X_i is a bit more complicated than the problem of making predictions regarding the average of the X_i , but it can be done along similar lines.

- First, we have to work out which of the limiting distributions is the appropriate one, and if it is Frechét or Weibull, we need to estimate the parameter α .
- Second, we have to work out the scaling constants a_T and the shifts b_T .
- Finally, provided that T is large, we can use the Three Types Theorem to calculate the probabilities.

A simple model

The first part is not too bad. The limiting distribution depends on the probabilities $P(X_1 \leq h)$ in a relatively systematic way.

- If there is a fixed maximum level for the X_i , then the limiting distribution is Weibull. The value of α depends on the behaviour of $P(X_1 \leq h)$ for h just less than the maximum level.
- If the probabilities $P(X_1 \leq h)$ have what is called a **heavy tail**, then the limiting distribution is Fréchet. We can get the value of α from the nature of the heavy tail.
- In all other cases (which include most of the standard probability distributions in undergraduate courses), the limiting distribution is Gumbel.

A simple model

We still have the problem of estimating the scaling constants a_T and the shifts b_T . Unfortunately, unlike the Central Limit Theorem, these depend on more complicated probabilistic properties of the X_i than just the mean and variance.

This means that we do have to put some effort into modelling the X_i . There are techniques, such as the **Peaks over thresholds** method for doing this. Basically we need to know more about the distribution $P(X_i \leq h)$ than when we were deriving the distribution of the average.

I found some packages on the web that claim they can do it.

A simple model

Some examples of a_T and b_T are

- If $P(X_i \leq h) = 1 - e^{-h}$ (Exponential), then $a_T = 1$ and $b_T = \log T$.
- If $P(X_i \leq h) = 1 - h^{-\alpha}$ (Pareto), then $a_T = T^{1/\alpha}$ and $b_T = 0$.
- If there is a ω , such that

$$P(X_i \leq h) \begin{cases} \approx 1 - (\omega - h)^\alpha & h \text{ close to } \omega \\ = 1 & h \geq \omega \end{cases}$$

(Bounded Support), then $a_T = T^{-1/\alpha}$ and $b_T = \omega$.

Conclusion

We derive the appropriate form of F and estimate a_T and b_T .
Then

$$P(M_T \leq \tilde{a}_t h + b_t) \approx F(h).$$

This is the same as saying

$$P(M_T \leq h) \approx F((h - b_T)/a_T).$$

Conclusion

So we could choose the height h

- so that the probability

$$1 - F((h - b_T)/a_T)$$

of overtopping during the planning horizon T is sufficiently small, say less than ϵ , or

- so that the trade-off between

$$1 - F((h - b_T)/a_T)$$

and the cost $C(h)$ of building the levee to height h is acceptable.