Forecasting extreme outcomes

This document explores and develops methods for forecasting extreme outcomes, such as the maximum of a sample of n independent and identically distributed random variables. I was inspired to write this by Jaime Sevilla's recent post with research ideas in forecasting and, in particular, his suggestion to write an accessible introduction to the Fisher–Tippett–Gnedenko Theorem.

I'm very grateful to Jaime Sevilla for proposing this idea and for providing great feedback on a draft of this document.

Summary

The Fisher–Tippett–Gnedenko Theorem is similar to a central limit theorem, but for the *maximum* of random variables. Whereas central limit theorems tell us about what happens on average, the Fisher–Tippett–Gnedenko Theorem tells us what happens in extreme cases. This makes it especially useful in risk management, when we need to pay particular attention to worst case outcomes. It could be a useful tool for forecasting tail events.

This document introduces the theorem, describes the limiting probability distribution and provides a couple of examples to illustrate the use (and misuse!) of the Fisher–Tippett–Gnedenko Theorem for forecasting. In the process, I introduce a tool that computes the distribution of the maximum n iid random variables that follow a normal distribution centrally but with an (optional) right Pareto tail.

Summary:

- The Fisher–Tippett–Gnedenko Theorem says (roughly) that if the maximum of n iid random variables—which is itself a random variable—converges as n grows to infinity, then it must converge to a generalised extreme value (GEV) distribution
- Use cases:
    - When we have lots of data, we should try to fit our data to a GEV distribution since this is the distribution that the maximum should converge to (if it converges)
    - When we have subjective judgements about the distribution of the maximum (e.g. a 90% credible interval and median forecast), we can use these to determine parameters of a GEV distribution that fits these judgements
    - When we know or have subjective judgements about the distribution of the random variables we're maximising over, the theorem can help us determine the distribution of the maximum of n such random variables

    for large *n* – but this can give very bad results when our assumptions / judgements are wrong
- Limitations:
    - To get accurate forecasts about the maximum of *n* random variables based on the distribution of the underlying random variables, we need accurate judgements about the right tail of the underlying random variables because the maximum will very likely be drawn from the tail, especially as *n* gets large
    - Even for data that is very well described by a normal distribution for typical values, normality can break down at the tails and this can greatly affect the resulting forecasts
        - I use the example of human height: naively assuming normality underestimates how extreme the tallest and shortest humans are because height is “only” normally distributed up to 2-3 standard deviations around the mean
    - Modelling the tail separately (even with quite a crude model) can improve forecasts
- This simple tool might be good enough for forecasting purposes in many cases
    - It assumes that the underlying r.v.s are iid and normally distributed up to *k* standard deviations above the mean and that there is a Pareto tail beyond this point
    - Inputs:
        - 90% CI for the underlying r.v.s
        - *n* (the number of samples of the underlying random variables)
        - *k* (the number of SDs above the mean at which the Pareto tail starts); set this high if you don’t want a Pareto tail
    - Output: cumulative distribution function, approximate probability density function and approximate expectation of the maximum of *n* samples of the underlying random variables
- Request for feedback: I’m not an experienced forecaster and I don’t know what kind of information and tools would be most useful for forecasters. Let me know how this kind of work could be extended or adapted to be more useful!

I expect the time-poor reader to get most of the value from this document by reading the informal statement of the Fisher–Tippett–Gnedenko Theorem, the overview of the generalised extreme value distribution, and the shortest and tallest people in the world example, and then maybe making a copy and playing around with the tool for forecasting the maximum of *n* random variables that follow normal distributions with Pareto tails (consulting this as needed).

Informal statement of the theorem

Let $X_1, X_2, \cdots, X_n$ be iid random variables and let $M_n = \max\{X_1, X_2, \cdots, X_n\}$ be the maximum of these random variables. If M_n converges to a non-degenerate random variable as n grows to infinity, then that limiting random variable must follow a generalised extreme value distribution.[1]

Note that the theorem says that *if* the maximum converges, *then* it must converge to a generalised extreme value distribution. It does not say that the maximum necessarily converges. Contrast this with the CLT, which ensures convergence to a normal distribution, so long as the mean and variance are finite. There are additional conditions that ensure convergence, but I won't go into them here except to say that it seems to me that the maximum of most commonly used distributions will converge.[2]

Most of the theory is framed in terms of the maximum of random variables but can be adapted to apply to minima too. For example, I estimate the height of both the tallest and shortest people in the world.

Generalised extreme value distribution

Overview

There are three types of GEV distribution: Gumbel, Fréchet and (reversed) Weibull distribution. They differ primarily along the following dimensions: (i) the heaviness of their tails; (ii) the types of underlying distributions that generate them; (iii) their support – that is, the values to which they assign positive (i.e. non-zero) probability.

The right tail of the underlying random variables determines the right tail of the limiting maximum random variable since the maximum of a large number of random variables will likely be drawn from the right tail. Exponentially decreasing tails (e.g. normal, exponential distributions) produce a Gumbel distribution, which also has exponentially decreasing tails. An important exception is the lognormal distribution, which has a heavier than exponential tail, but also generates a Gumbel distribution.[3] Bounded distributions (e.g. uniform, beta distributions) produce a Weibull distribution, which has an upper bound – this makes sense since the maximum can't

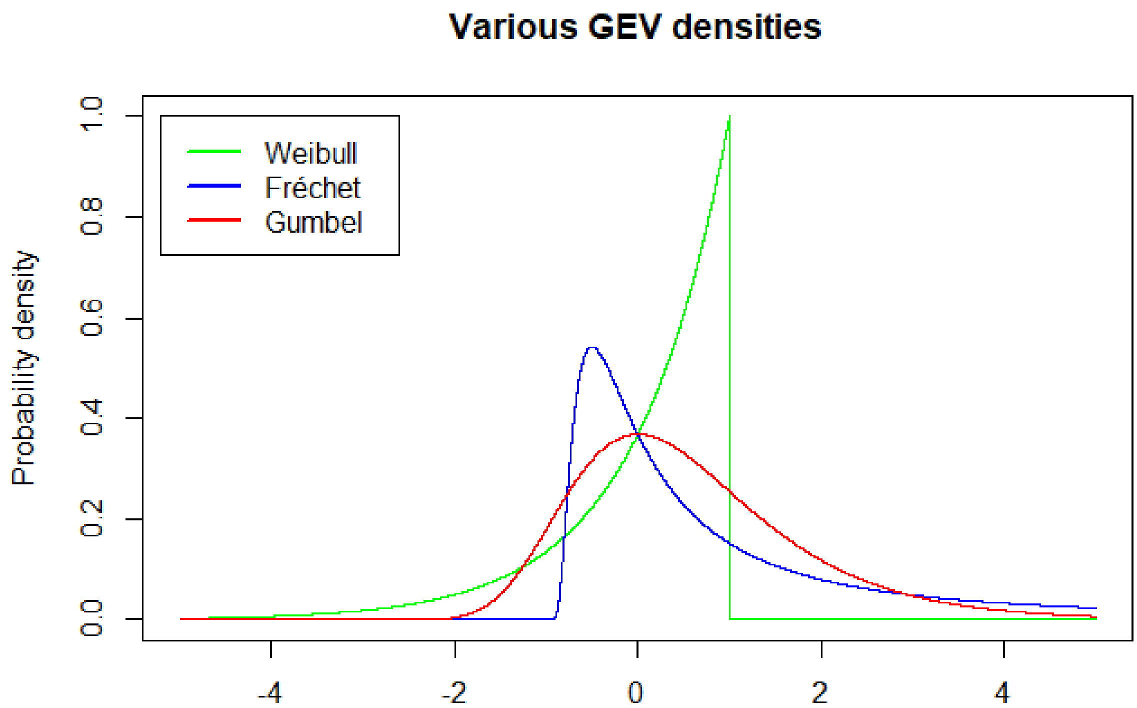
[1] Technically, the theorem applies to M_n after "suitable normalisation". The precise details of normalisation are subtle, potentially confusing to unfamiliar readers and not that practically important, so I've relegated them to a more technical appendix.

[2] One caveat: after suitable normalisation. See the appendix for more on this.

[3] I've seen this in a few sources, e.g. pg. 59 Kotz, S., & Nadarajah, S. (2000). *Extreme value distributions: theory and applications*. However, when I ran simulations with lognormal random variables, I actually got a Fréchet distribution (with a small but significantly positive shape parameter). It's possible I didn't run enough simulations to get convergence to Gumbel distribution (lognormal r.v.s can be very slow to converge due to heavier than exponential tails).

be greater than the upper bound of the underlying random variables. Distributions with polynomially decreasing tails (e.g. Pareto, Cauchy distributions) produce a Fréchet distribution, which has a heavy right tail and a lower bound (though the lower bound isn't very important in practice).

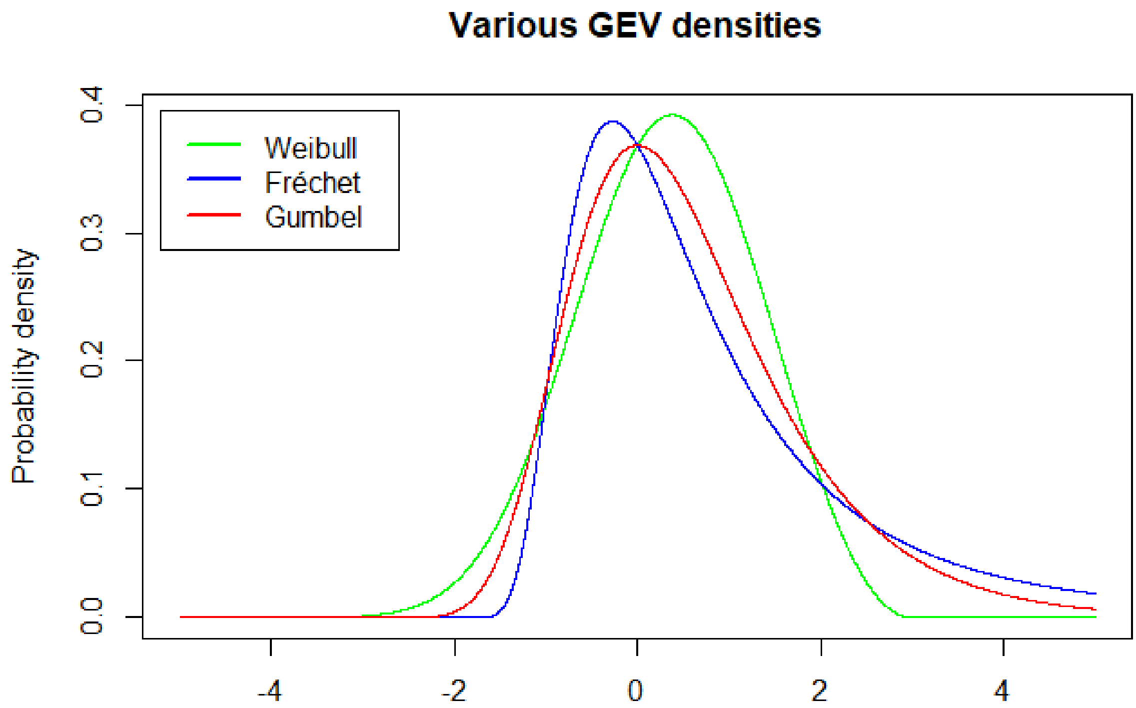
Here are some example plots of GEV distributions. The first image illustrates three GEV densities, one of each type. The Gumbel distribution is the standard Gumbel distribution but the Weibull and Fréchet distributions are relatively extreme: this Weibull distribution has significant probability density near the upper bound and this Fréchet distribution has a sufficiently heavy tail for the mean to be infinite. I'll describe the parameters of these distributions precisely in a later section.[4]



The second image again shows example densities of each type of GEV distribution. The same Gumbel distribution is displayed but the Weibull and Fréchet distributions are much less extreme.[5]

[4] For completeness, here are the parameters of these distributions. All three have $\mu = 0$ and $\sigma = 1$. The Weibull, Gumbel and Fréchet distributions have $\xi = -1, 0, 1$, respectively. See GEV parameters for an explanation of these parameters.

[5] Here, we have again $\mu = 0$ and $\sigma = 1$ for all three distributions. The Weibull, Gumbel and Fréchet distributions have $\xi = -1/3, 0, 1/3$, respectively.



The following table summarises some of the key features of the three types of GEV distribution.

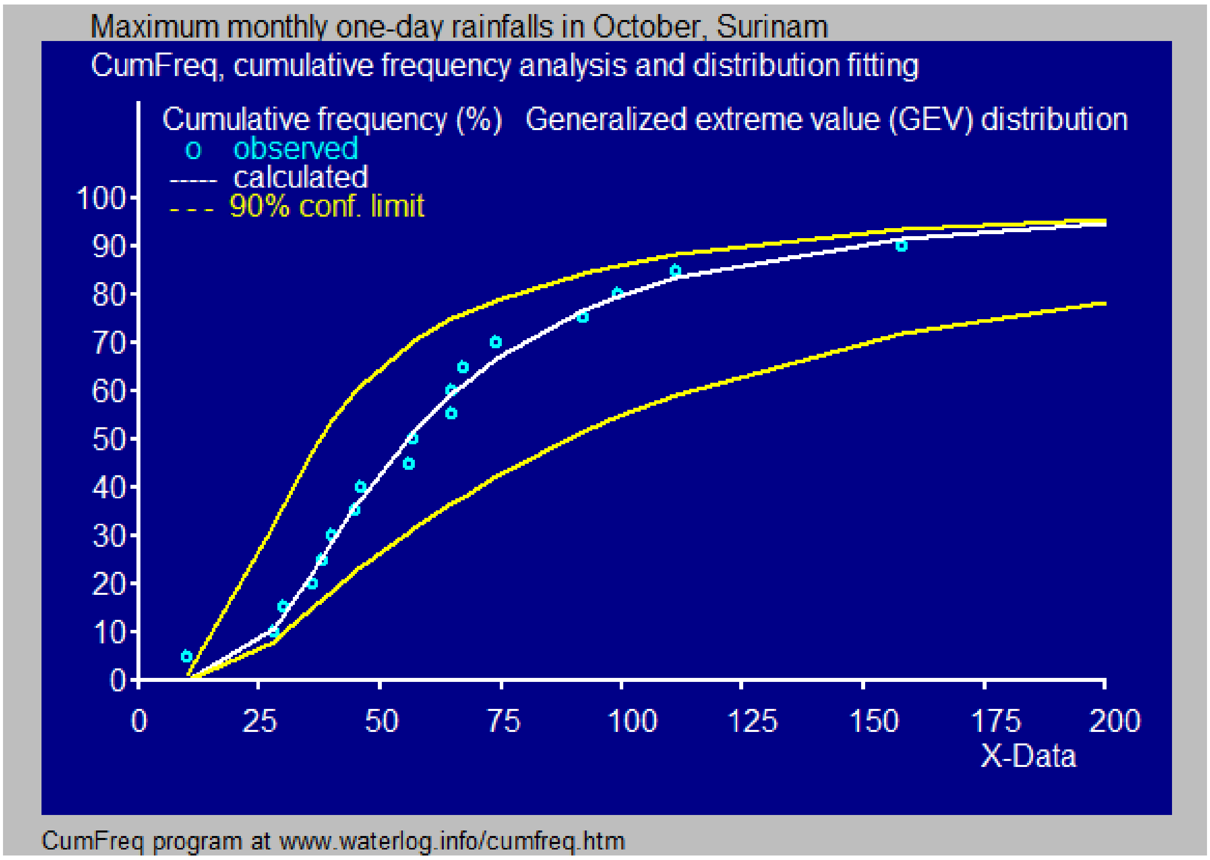
	Gumbel	(Reversed) Weibull	Fréchet
Support and tails	Unbounded support with relatively thin tails that decay exponentially	Bounded from above	Bounded from below with heavy right tail
Generated by	Distributions with exponentially decreasing tails, e.g. normal, exponential, gamma. Also lognormal distribution.	Distributions that are bounded from above, e.g. uniform, beta.	Distributions with polynomially decreasing tails, e.g. Cauchy, Pareto, Student's t

A note to avoid unnecessary confusion: the Weibull distribution that arises as a generalised extreme value distribution is a little different to the Weibull distribution found elsewhere, e.g. on Wikipedia. In other contexts, the Weibull distribution has a lower bound rather than an upper bound (it's used to model minima) and is parameterised slightly differently. For this reason, the Weibull distribution discussed here, as a type of generalised extreme value distribution, is sometimes called a *reversed* Weibull distribution.

Applications

The GEV distribution has been applied in several engineering and weather forecasting contexts, as well as to model tail risks in insurance and finance.[6] For example, given data on the annual highest tide, one can estimate the distribution of the highest tide over 100 years (or more) in order to determine how high to build flood defences. The GEV distribution has found similar applications with extreme temperature (for both hottest and coldest temperatures), maximum wind speeds, minimum and maximum rainfall (droughts and floods), maximum earthquake magnitude.

In most real-world applications, a GEV distribution is fitted to data of maxima or minima. As a concrete example, this image, taken from waterlog.info (via Wikipedia), shows a GEV distribution fitted to maximum monthly one-day rainfalls in October in Surinam:



[6] See also applications of the Gumbel, Weibull and Fréchet distributions.

GEV parameters

The generalised extreme value distribution (GEV) is characterised by a cumulative distribution function (cdf) of the form $e^{-t(x)}$, where $t(x) = \{ \begin{smallmatrix} (1+\xi(x-\mu)/\sigma)^{-1/\xi} & if\ \xi\neq0 \\ e^{-(x-\mu)/\sigma} & if\ \xi=0 \end{smallmatrix}$. The precise form of the cdf isn't too important for forecasting applications but it's helpful to understand what the distribution's parameters are. The distribution has three parameters: (i) a location parameter, μ, which, roughly speaking, controls where the centre of the distribution is located; (ii) a scale parameter, σ, which, roughly, controls how much the distribution is stretched out; (iii) a shape parameter ξ, which, roughly, controls the tail behaviour of the distribution.

Greater ξ corresponds to a heavier right tail. We have a Weibull distribution when $\xi < 0$, Gumbel when $\xi = 0$, and Fréchet when $\xi > 0$. When $\xi \geq 1$, the right tail is sufficiently heavy that the mean of the distribution is infinite and when $\xi \geq \frac{1}{2}$, the variance is infinite.

The Gumbel distribution is the limiting distribution of the Fréchet or Weibull distributions as $\xi \to 0$. This is illustrated by the density plots above. All plotted distributions have $\mu = 0$ and $\sigma = 1$ so that we can focus on the tails and the differences between the three types of GEV distribution, which are primarily determined by ξ. The Weibull, Gumbel and Fréchet distributions in the first image have $\xi =- 1, 0, 1$, respectively. In the second image, the Weibull, Gumbel and Fréchet distributions have $\xi =- 1/3, 0, 1/3$, respectively. We see that the first three distributions look very different but the latter three distributions look quite similar. Moving ξ closer to 0 moves the Weibull and Fréchet distributions closer to a Gumbel distribution. Although each type can look quite different, this is one of the reasons it makes sense to unify them as a single distribution – the GEV distribution.

(Warning: Don't confuse μ and σ for the mean and standard deviation – the mean and standard deviation of GEVs (when finite) can be given in terms of $\mu,\ \sigma,\ \xi$ but are not equal to μ and σ, respectively. The normal distribution is peculiar in having its mean and standard deviation equal to its location and scale parameters.)

Forecasting examples

When forecasting the maximum of several iid random variables, the Fisher–Tippett–Gnedenko Theorem suggests that a GEV distribution could be a good fit. Which type is best will depend on the distribution of the underlying random variables. In sufficiently data-rich contexts, one can fit a GEV distribution to data to determine the parameter values. This often isn't feasible for important forecasting questions though. The following examples illustrate the use (and misuse!) of the Fisher–Tippett–Gnedenko Theorem for forecasting.

Subjective forecasting

In many forecasting contexts, we don't have rich data and we don't have a precise sense of the distribution of the underlying random variables. We might still make use of a GEV distribution though to forecast the maximum given subjective estimates, e.g. for the median and 0.05 and 0.95 quantiles.

The quantile function for a GEV distribution (with $\xi \neq 0$) is $Q(p) = \mu + \frac{\sigma}{\xi}((-\log p)^{-\xi} - 1)$. So if a forecaster thinks the median of M_n is about 3, and the 0.05 and 0.95 quantiles (i.e. 5th and 95th percentile outcomes of M_n) are about 1 and 10 respectively, this gives:

$$1 = \mu + \frac{\sigma}{\xi}((-\log 0.05)^{-\xi} - 1)$$
$$3 = \mu + \frac{\sigma}{\xi}((-\log 0.5)^{-\xi} - 1)$$
$$10 = \mu + \frac{\sigma}{\xi}((-\log 0.95)^{-\xi} - 1)$$

We have three equations and three unknown parameters, μ, σ, ξ, so we can solve these equations numerically to find the parameters (e.g. using the nleqslv function in R). In this case, this yields $\mu \approx 2.41$, $\sigma \approx 1.52$, $\xi \approx 0.32$. The resulting expected value is $E(M_n) = \mu + \sigma(\Gamma(1 - \xi) - 1)/\xi \approx 3.99$, where Γ is the Gamma function.

If a forecaster has more subjective quantile estimates, they can write down more quantile equations and find parameter values that minimise (e.g.) the sum of the squares of the differences. More heuristically, with an understanding of the GEV parameters and some qualitative judgements about what the distribution of the maximum should look like, you might be able to find a good enough fit by playing around with different parameter values, e.g. using the gevd functions in R.

The shortest and tallest people in the world

Often, we won't have a good sense of the distribution of the maximum a priori but are able to make judgements about the distribution of the underlying random variables and we can use these to make judgements about the maximum. Here, I use human height as an example of how to do this *and how it can go very wrong*.

Human height is approximately normally distributed, so the distribution of the tallest shortest people in the world should be related to a Gumbel distribution. I will focus on the shortest adult woman and the tallest adult man across North America, Europe, East Asia as Our World in Data and my regional adult population data source have good data only for these regions. I'll refer to these regions as "NAEEA" for short. According to Our World in Data, across NAEEA, female height follows a normal distribution with mean 164.7 cm and standard deviation 7.1 cm and male height follows a normal distribution with mean 178.4 cm and standard deviation 7.6 cm. So in this region, female height has cdf $\Phi(\frac{x-164.7}{7.1})$ and quantile function $164.7 + 7.1\Phi^{-1}(p)$ and male height has cdf $\Phi(\frac{x-178.4}{7.6})$ and quantile function $178.4 + 7.6\Phi^{-1}(p)$, where Φ is the standard normal cdf.

The total adult population in NAEEA is about 2 billion, so there are about 1 billion adult men and 1 billion adult women across these regions.

In general, if underlying data have cdf F and are of a distribution type that generates a Gumbel distribution, and we have n samples, then the maximum of the n samples (for large n) will be approximately Gumbel-distributed with $\mu = F^{-1}(1 - \frac{1}{n})$ and $\sigma = F^{-1}(1 - \frac{1}{ne}) - F^{-1}(1 - \frac{1}{n})$, where F^{-1} is the quantile function (inverse of the cdf).[7]

Then the height of the tallest man should be approximately Gumbel-distributed with $\mu \approx 178.4 + 7.6\Phi^{-1}(1 - \frac{1}{1\times10^9}) \approx 224$ cm (or about 7 feet and 4 inches) and

$$\sigma = 178.4 + 7.6\Phi^{-1}(1 - \frac{1}{e\times10^9}) - 178.4 - 7.6\Phi^{-1}(1 - \frac{1}{1\times10^9})$$

$$= 7.6(\Phi^{-1}(1 - \frac{1}{e\times10^9}) - \Phi^{-1}(1 - \frac{1}{1\times10^9})) \approx 1.22.$$

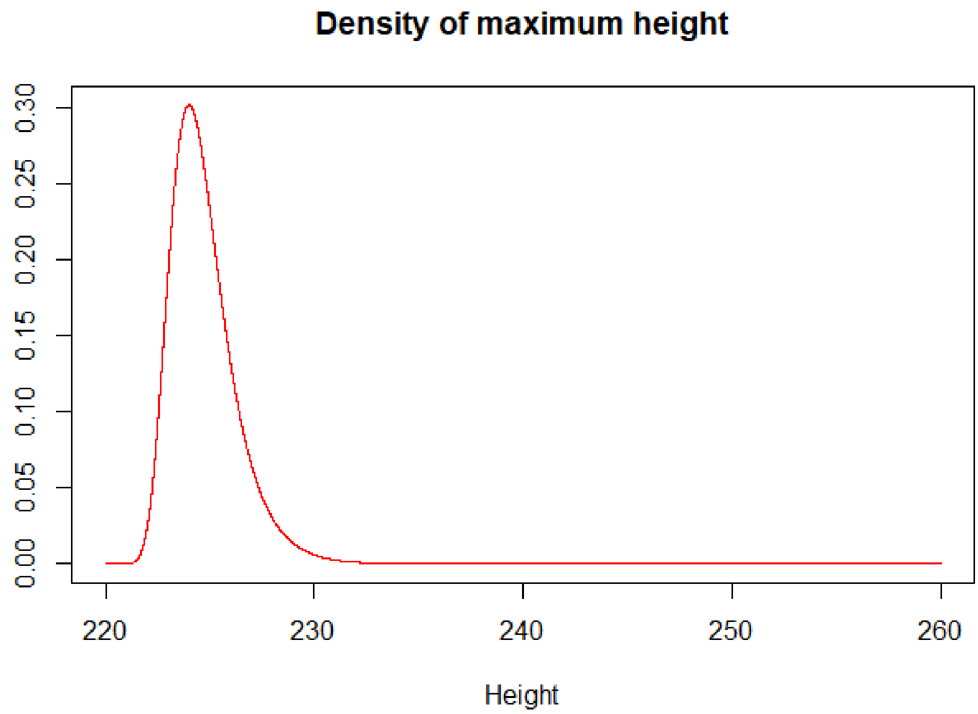
The mean of a Gumbel distribution is $\mu + \sigma\gamma$, where $\gamma \approx 0.577$ is the Euler-Mascheroni constant and the quantile function is $Q(p) = \mu - \sigma\log(-\log p)$ (where $\log$ is the natural logarithm). This gives an expected value of $224 + 1.22 \times 0.577 \approx 225$ cm, or about 7 feet and 4.5 inches, about 6 standard deviations above the population mean. The 90% confidence interval is about $[223, 228]$.

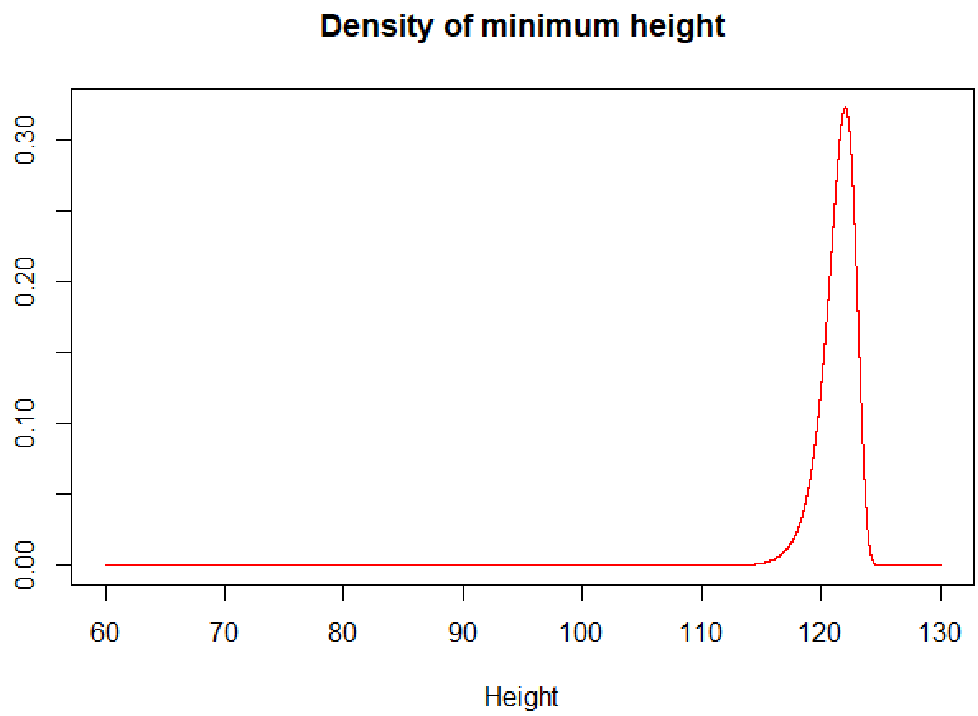
The computations for the shortest woman are similar. Notice that the Fisher–Tippett–Gnedenko Theorem applies to maxima rather than minima, but we can just consider negative heights instead – the maximum of negative heights will be the negative of the minimum height (e.g. $\max\{-1, -2, -4\} = -1 = -\min\{1, 2, 4\}$). So the negative height of the shortest woman is approximately Gumbel-distributed with $\mu = -164.7 + 7.1\Phi^{-1}(1 - \frac{1}{1\times10^9}) \approx -122$ cm and $\sigma \approx 1.14$. This gives an expected minimum height of about 121 cm (about 4 feet), which is about 6 standard

[7] This is described here for normal random variables and for exponential random variables here but I think it generalises to other distributions that generate Gumbel distributions.

deviations below the population mean. The 90% confidence interval is about $[119, 123]$.

Here are the densities for the maximum and minimum height:





Reality check: how tall are the tallest and shortest people in the world? It turns out they're more extreme than the above calculations predict. The tallest living person in NAEEA is Sun Mingming, who is 236 cm (7 ft 9), which is about 7.5 standard deviations above the mean and 1.5 standard deviations above the expected maximum computed above. As very tall people go, globally and historically, Sun Mingming isn't that tall. Sultan Kösen, though not in the regions considered, is 251 cm (8 ft 3 in), which is 9.5 standard deviations above the mean, and there are plenty more very tall people living today and in the past. Some care is needed in considering cases like this: if we expand the region considered to include the whole world (and/or no longer living people), then we're increasing the sample size and a larger sample will have a larger maximum because there are more chances to have large outliers. Additionally, the data for NAEEA might not be representative of the rest of the world. Conservatively increasing the sample size to 4 billion to account for the global living male population and assuming (likely incorrectly) that male height follows the same distribution globally as it does in NAEEA, the results don't change drastically though: the expected maximum is about 226 cm (just 1 cm taller). Overall, the predicted maximum height seems too conservative and potentially much too conservative.

For context, the probability of a draw from a normal distribution being 8 or more standard deviations away from the mean for is about 10^{-15}, so even with 10 billion (i.e. 10^{10} draws), it's very unlikely that a normal distribution could produce such outliers.

The other end is even more extreme. I couldn't easily find the height of the shortest living woman in NAEEA but there are strong reasons to think that the above prediction is much too conservative. Bridgette Jordan, an American who died in 2019, was 69 cm (2 ft 3 in), which is about 13 standard deviations below the mean and 7 standard deviations below the expected minimum computed above. Jyoti Amge, from India, is the shortest living woman at 63 cm (2 ft ¾ in), about 14 standard deviations below the mean. There are plenty more very short people living today.

There's a really important lesson here: you have to take the tails into account carefully. A normal distribution describes height for most of the population very well but it breaks down at the tails and in estimating extreme outcomes, the tails are really important because the tails determine how likely extreme outcomes are. In this case, the tails are heavier than those of a normal distribution, so the Fisher–Tippett–Gnedenko Theorem applied to normal random variables produces a maximum and a minimum with insufficiently heavy tails.

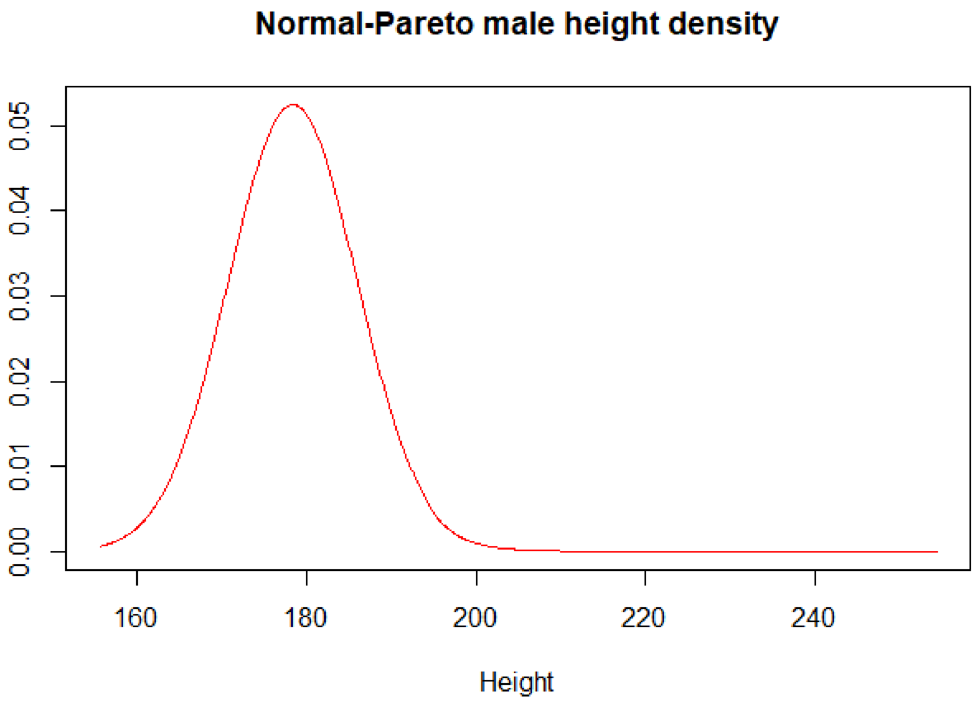
We can do better by making the tails heavier. Here's a pretty crude attempt to do that which gets pretty good results in this case. Instead of supposing height is normally distributed, let's assume it's only normally distributed within some range of the mean and then has Pareto tails. Here, I will assume that height is normally distributed within 2 standard deviations of the mean, i.e. up to about 194 cm for men and down to 151 cm for women. I selected this value because of a vague comment in this blog post.[8] (If you have better data, let me know.)

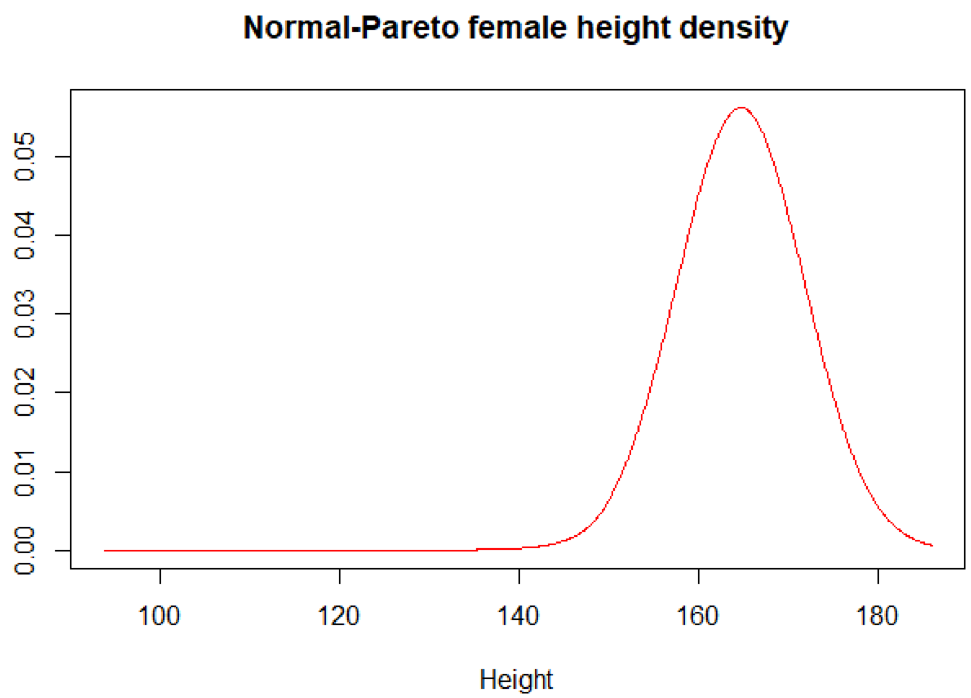
We have to do some work to get reasonable-ish parameters for the tails and then some more to compute the resulting distribution of the maximum. The main idea is that by assuming both the cdf and pdf of the underlying random variables are continuous (at the point at which the tail begins), we pin down precisely what the parameters of the Pareto tail need to be. This gives us the cdf of the underlying random variables and we can use this to work out the cdf, pdf, and quantile function of the maximum. The full derivations are in the appendix.

The resulting 90% CI for maximum height in 1 billion men is $[252, 269]$ and the mean is about 259 cm. The result is probably too high: the tallest man in the world is 251 cm, just below our 90% CI for the tallest man in NAEEA and the tallest man in NAEEA is only 236 cm. The result for the shortest woman is still not extreme enough though: the 90% CI is about $[80, 96]$ with a mean of about 90 cm (original estimate: mean: 121; 90% CI: [119,123]). Recall that the shortest woman in the world is 63 cm and the shortest woman in NAEEA was, until 2019, 69 cm. This is indicative of there being more very short people than very tall people, so that normality breaks down closer to the mean on the left tail than the right tail and/or the left tail decays even more slowly than the right tail.

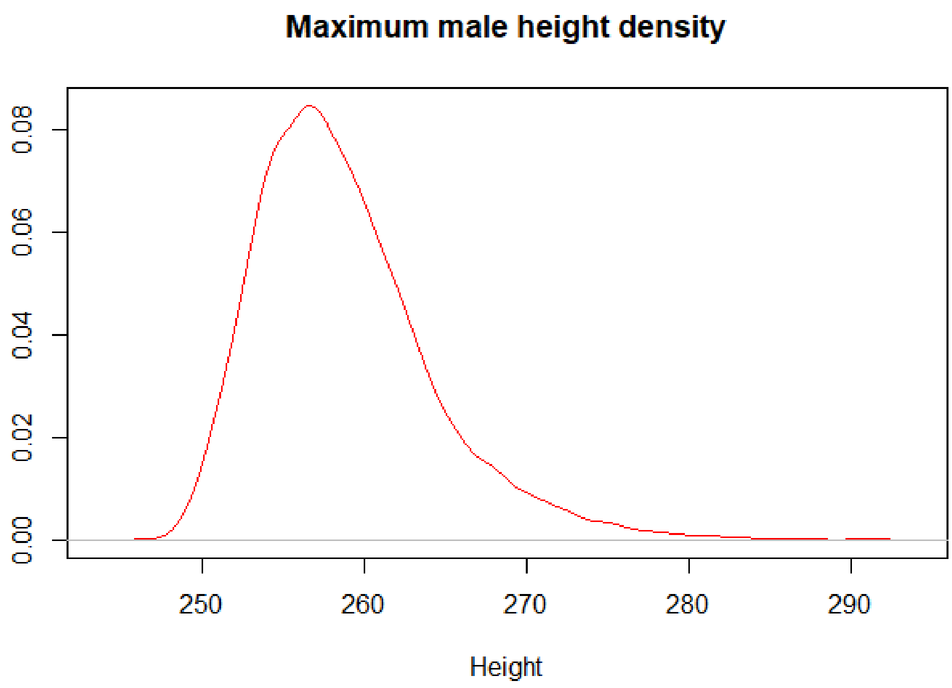
Here are the resulting densities for male height, female height, and maximum male height, and minimum female height:

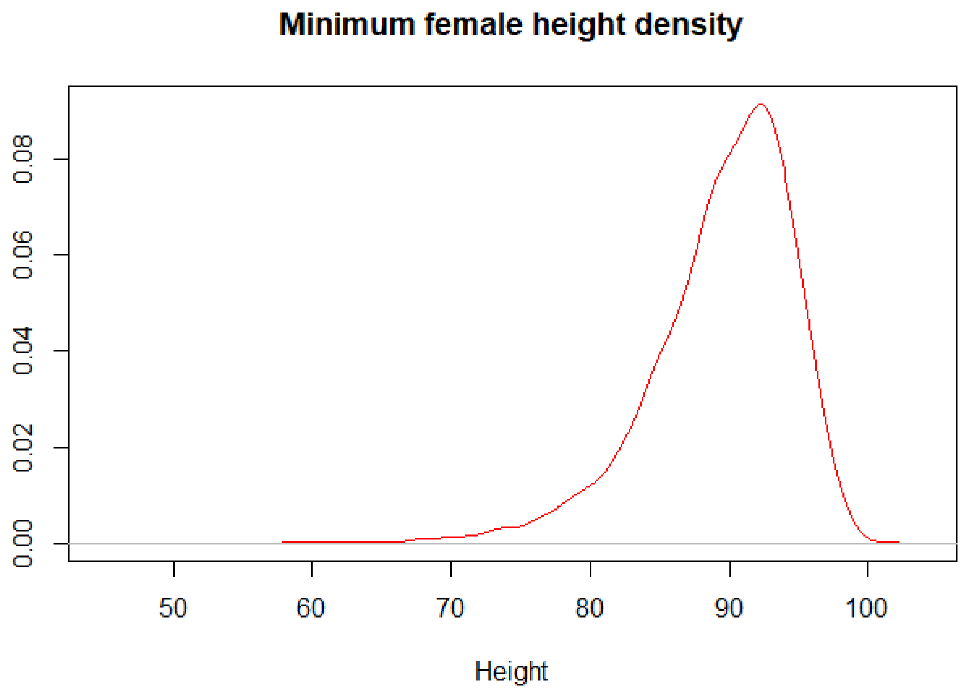
[8] "The normal distribution describes heights remarkably well near the mean, even a couple standard deviations on either side of the mean."





Notice that these densities don't look *that* different to the normal densities. But the resulting densities for maximum and minimum heights look very different to the previous densities for maximum and minimum heights.





These aren't perfectly smooth because they're plotted from simulations (described in the appendix), but don't worry about that – there's nothing weird going on here.

The results depend on the choice of where the Pareto tail starts (which I selected as 2 standard deviations from the mean, after very little thought or research). The following table shows how sensitive the results for the tallest man are to this choice. $\bar{x}$ is the point at which the Pareto tail starts, given in terms of the mean height (178.4

cm for men and 164.7 cm for women) and some multiple of standard deviations of height (7.6 cm for men and 7.1 cm for women). For greater values of $\bar{x}$, the Pareto tail starts later, and the resulting distribution of the tallest male height is less extreme and the resulting distribution of shortest female height is less extreme.

$\bar{x}$	**Expected height of tallest man (cm)**	**90% CI**	$P(X_i \geq \bar{x})$
$\mu + 1.5\sigma$	279	$[269, 293]$	6.7%
$\mu + 2\sigma$	259	$[252, 269]$	2.3%
$\mu + 2.5\sigma$	246	$[241, 254]$	0.62%
$\mu + 3\sigma$	238	$[234, 245]$	0.13%
$\mu + 4\sigma$	229	$[226, 234]$	0.003%
$\mu + 5\sigma$	226	$[223, 229]$	0.0000294%

$\bar{x}$	**Expected height of shortest woman (cm)**	**90% CI**	$P(X_i \leq \bar{x})$
$\mu - 1.5\sigma$	71	$[57, 80]$	6.7%
$\mu - 2\sigma$	90	$[80, 96]$	2.3%
$\mu - 2.5\sigma$	101	$[94, 107]$	0.62%
$\mu - 3\sigma$	109	$[103, 113]$	0.13%
$\mu - 4\sigma$	117	$[113, 120]$	0.003%
$\mu - 5\sigma$	120	$[117, 123]$	0.0000294%

We see that the results are quite sensitive to the choice of where the Pareto tail starts, mainly up to 3 or 4 standard deviations away from the mean. By 5 standard deviations above the mean, the results are almost identical to the original prediction (for men: (mean: 225, 90% CI: [223,228]); for women: (mean: 121, 90% CI: [119,123])). I imagine choosing where the Pareto tail should start will often be challenging in practice. To get a sense of what a reasonable choice for the start of the Pareto tail might be, the right hand column shows the probability of a randomly selected individual lying in the Pareto tail, i.e. the proportion of the population that lies in the Pareto tail.

Forecasting tool for the maximum of *n* iid random variables that follow normal distributions with optional Pareto tails

This google sheet can compute the cdf and quantile function of the maximum of *n* iid random variables that follow normal distributions up to some point $\bar{x}$ with a Pareto tail beyond this point. It also estimates the pdf and expected value of the maximum (though I'd recommend other methods for high stakes decisions).

As inputs, the sheet requires:

- 90% CI for the underlying r.v.s (0.05 and 0.95 quantiles)
- The number of samples of the underlying random variables, *n*
- The number of SDs above the mean at which the Pareto tail starts (i.e. *k* such that $\bar{x} = \mu + k\sigma$

The Pareto tail is essentially optional: set *k* to be very high for a negligible Pareto tail (even $k = 8$ is usually enough).

The sheet calculates the mean and standard deviation of a normal distribution that fits the (symmetric) 90% CI provided by the user. Note that this is uniquely determined: there is precisely one normal distribution that fits a given CI. It then calculates the Pareto parameters that produce a Pareto tail that starts at $\bar{x}$ and is continuous with the normal cdf and pdf. This is also uniquely determined: there is precisely one Pareto tail that starts at a given $\bar{x}$ and is continuous with a given normal cdf and pdf at that point.

The normal and Pareto parameters determine the cdf and quantile function of the underlying random variables we're maximising over. The sheet computes the cdf and quantile function of the maximum of *n* such random variables using the cdf of the underlying random variables, and it uses these to approximate the density and expected value of the maximum. These estimates are crude and may not be reliable. However, in all the computations I carried out in the process of writing this document, the expected value of the maximum estimated by the sheet was equal to the actual expected value computed more robustly in R, when rounded to the nearest integer (i.e. the nearest cm in the height example).[9]

More details of the derivations are in the notes in the sheet and in the appendix.

[9] The beauty of Google sheets is in its simplicity, user-friendliness and ease of shareability. One can do much more in R or other programming languages though, so I'd recommend using one of those if more sophisticated or robust analysis is required.

Conclusion

Forecasting extreme outcomes is important because we often want to ensure robustness to extreme outcomes, not just typical or likely outcomes (e.g. we want our flood defences to be tall enough to protect us from the worst floods of their lifetime, not just most typical floods). The Fisher–Tippett–Gnedenko Theorem provides some guidance for forecasting the maximum of a large number of random variables by telling us that this will converge to a generalised extreme value distribution, if it does indeed converge. This gives forecasters a good first place to look when forecasting extreme events, especially in data rich contexts.

However, forecasting extreme outcomes is really hard because results are very sensitive to the tails. The height example showed that naively assuming that human height is normally distributed and applying the Fisher–Tippett–Gnedenko Theorem yields too conservative results. And height is often used as a prime example of something that's normally distributed! Adding Pareto tails might be a useful tool for accounting for heavier than normal tails, e.g. via this spreadsheet. The most challenging part of using this method is deciding where the Pareto tail should start (i.e. where normality breaks down), and I imagine this could be quite difficult to pin down accurately and precisely without lots of data.

Finally, I'm not an experienced forecaster and I don't know what kind of information and tools would be most useful for forecasters. Let me know how this kind of work could be extended or adapted to be more useful!

Appendix

This is more technical.

Pareto tails

We wish to replace the right tail of a normal distribution with a Pareto tail. We can work out reasonable-ish parameters for the tail from the cdf of the Pareto distribution: $H(x) = 1 - (\frac{x_m}{x})^{\alpha}$, where x_m and α are the parameters to be determined. In order to determine x_m and α, we need to set two constraints: one per parameter. There are two very natural constraints: (1) the Pareto cdf should equal the normal cdf at the point at which the Pareto tail begins (i.e. the cdf is continuous); (2) the Pareto pdf should equal the normal pdf at the point at which the Pareto tail begins (i.e. the pdf is continuous). Once we impose these constraints, x_m and α will be uniquely determined.

Let Φ and $\phi = \Phi'$ denote the standard normal cdf and pdf, respectively. Given a normal distribution with mean μ and standard deviation σ, we impose the following constraints:

$$H(\bar{x}) = 1 - (\frac{x_m}{\bar{x}})^{\alpha} = \Phi(\frac{\bar{x}-\mu}{\sigma})$$

$$H'(\bar{x}) = \frac{\alpha x_m^{\alpha}}{\bar{x}^{\alpha+1}} = \phi(\frac{\bar{x}-\mu}{\sigma})$$

Solving these for x_m and α yields:

$$\alpha = \bar{x}\phi(\frac{\bar{x}-\mu}{\sigma})/(1 - \Phi(\frac{\bar{x}-\mu}{\sigma}))$$

$$x_m = \bar{x}(1 - \Phi(\frac{\bar{x}-\mu}{\sigma}))^{1/\alpha} = \bar{x}(1 - \Phi(\frac{\bar{x}-\mu}{\sigma}))^{(1-\Phi(\frac{\bar{x}-\mu}{\sigma}))/(\bar{x}\phi(\frac{\bar{x}-\mu}{\sigma}))}$$

In the human height example, with $\bar{x} = \mu + 2\sigma = 193.6$, this yields the following parameters for male height:

$$\alpha = 193.6\phi(\frac{193.6-178.4}{7.6})/(1 - \Phi(\frac{193.6-178.4}{7.6})) \approx 60.5$$

$$x_m = 193.6(1 - \Phi(\frac{193.6-178.4}{7.6}))^{1/60.5} \approx 182$$

Male height has cdf F such that $F(x)$ is equal to the normal cdf for $x < \bar{x}$ and is equal to the Pareto cdf for $x \geq \bar{x}$.

Female height is similar, except the Pareto tail is on the left (and is therefore a reflected version of a right Pareto tail). There are multiple ways to adapt the method for a right Pareto tail to a left Pareto tail. I think the most straightforward in this case is to give female height a right tail and run the same calculations as above, and then reflect everything about the centre of the normal distribution, so that the Pareto tail is on the left.[10] For female height, we have $\bar{x} = \mu + 2\sigma = 178.9$, which yields $\alpha \approx 60$ and $x_m \approx 179$. Reflecting the resulting cdf about the centre, $\mu = 164.7$, we have that female height is normally distributed down to $\underline{x} = \mu - 2\sigma = 2\mu - \bar{x} = 150.5$ and has cdf G such that $G(x)$ is equal to the normal for $x > \underline{x}$ and cdf equal to $(\frac{x_m}{2\mu - x})^\alpha$ for $x \leq \underline{x}$.

Maximum and minimum heights of Normal-Pareto random variables

We now have the cdfs of male and female height and we can use these to compute the cdfs, pdfs and quantile functions of maximum and minimum height. Denote the male height cdf by F with pdf $f = F'$. The maximum of n iid samples has cdf

$$F_n(x) = P(\{X_1 \leq x\} \cap \cdots \cap \{X_n \leq x\}) = \prod_{i=1}^{n} P(\{X_i \leq x\}) = F(x)^n$$

The first equality follows from the fact that if the maximum of n random variables is less than x, then each of the n random variables must be less than x. The second equality follows from the independence of the samples and the third equality from the fact that they're identically distributed.

By the chain rule, the maximum male height has pdf $f_n(x) = nf(x)F(x)^{n-1}$. The quantile function is the inverse of the cdf: $Q_n(p) = F^{-1}(p^{1/n})$, where F^{-1} is the quantile function for height, and is equal to the normal quantile function or Pareto quantile function at $p^{1/n}$, depending on whether $p^{1/n}$ is in the Pareto tail or not, i.e. whether $p^{1/n} \geq F(\bar{x})$.

90% CIs (and other CIs) can be computed directly from the quantile functions. For example $Q_{10^9}(0.05) = F^{-1}(0.05^{10^{-9}}) = x_m(1 - 0.05^{10^{-9}})^{-1/\alpha} \approx 252$ since the Pareto quantile function is $Q_{Pareto}(p) = x_m(1 - p)^{-1/\alpha}$ and $0.05^{10^{-9}}$ is sufficiently high that it

[10] One could instead put the tail on the left from the start, which would require slightly different equations for the Pareto tail and to determine the parameters of the Pareto tail. The end result is the same.

corresponds to a point in the Pareto tail (as opposed to the normal body of the distribution), i.e. $F(\bar{x}) < 0.05^{10^{-9}}$. That is, the lower bound of the 90% CI is 252.

Similarly, with female height following cdf G and pdf $g = G'$, the minimum of n iid samples has cdf

$$G_n(x) = P(\min\{X_1, \cdots, X_n\} \leq x\}) = 1 - P(\min\{X_1, \cdots, X_n\} > x) = 1 - (1 - G(x))^n$$

The minimum female height has pdf $g_n(x) = ng(x)(1 - G(x))^{n-1}$. However, in practice, it's often easier just to work with the maximum and then reflect the results in the centre of the normal body at the end (i.e. map x to $2\mu - x$).

The means of the maximum and minimum height can be computed from the pdfs $f_n(x) = nf(x)F(x)^{n-1}$ and $g_n(x) = ng(x)(1 - G(x))^{n-1}$ by computing the relevant expectation integrals (perhaps analytically—I haven't tried—or definitely numerically). This method is a little fiddly though since the F, f, G and g have normal components and Pareto components. Instead, I estimated the means via simulations. The idea is to take a large sample of random variables with cdf F_n (i.e. a sample of maxima) and this sample will approximately follow the right distribution. In particular, its mean should be close to the mean of the maximum male height. You can also estimate quantiles this way by computing quantiles of the random sample (quantiles estimated this way were very close to the exact quantiles computed as above).

In general, to get a random sample following cdf L, you can take a random sample following a uniform distribution on $[0, 1]$ (e.g. with the runif function in R) and then apply the quantile function L^{-1} to the sample. For the height example, I took a uniform random sample of size 10,000. I then fed this into the quantile function for the maximum height of 1 billion men, i.e. $Q_{10^9}(p) = F^{-1}(p^{10^{-9}})$. With extremely high probability, we don't need to worry about the normal component of F because $p^{10^{-9}}$ is sufficiently high that it will lie in the Pareto tail, even for small values of p. So, after checking that the smallest draw from the sample will lie in the Pareto tail (it essentially always will in this example) we can raise each element of the uniform random sample to the power 10^{-9}, pass these through the Pareto quantile function and this gives us a large sample of maximum heights. The reported means were computed by taking the mean of such samples. The plotted densities of maximum and minimum heights are density plots of such samples.

Formal statement of the Fisher–Tippett–Gnedenko Theorem

Let $X_1, X_2, \cdots, X_n$ be iid random variables with cdf F and let $M_n = \max\{X_1, X_2, \cdots, X_n\}$. If there exist two sequences of real numbers a_n, b_n with $a_n > 0$ and a non-degenerate cdf G, such that $\lim_{n\to\infty} P(\frac{M_n - b_n}{a_n} \leq x) = G(x)$ for all x at which G is continuous, then G follows a generalised extreme value distribution. That is, if there are suitable normalising sequences a_n, b_n such that $\frac{M_n - b_n}{a_n}$ converges in distribution to some non-degenerate random variable, then that random variable must follow a generalised extreme value distribution.

Normalisation

This section is intended to make the formal statement clearer for the curious but unfamiliar reader. I don't think it's practically important.

The sequences a_n and b_n and the term $\frac{M_n - b_n}{a_n}$ perhaps look daunting but they're just a technical requirement to ensure convergence. They're exactly analogous to the term $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$ in the classical CLT with normalising sequences $c_n = \sigma/\sqrt{n}$ and $d_n = \mu$ (in this formulation, d_n is independent of n).

The unnormalised sample mean converges to the degenerate random variable that always takes the value of the population mean but normalisation of the sample mean prevents it converging to this degenerate random variable. Similarly, without normalisation, the maximum, M_n will converge to the degenerate random variable that only ever takes the value of the population maximum or to infinity. (More precisely, M_n converges in probability to $\sup\{x : P(X_i \leq x) < 1\}$, which may be infinite.)

In many applications, we want to know about the distribution of the sample mean or maximum, not just the numbers they converge to, and this requires non-zero variance. For the sample mean, we obtain that by multiplying the by $\sqrt{n}$: $Var(\sqrt{n}\bar{X}_n) = nVar(\bar{X}_n) = \sigma^2$. The factor $\sqrt{n}$ is just large enough to make the variance non-zero as the sample size grows and just small enough to prevent the variance blowing up to infinity. The sequence a_n plays this role for the sample maximum.

Less formally, if $\frac{M_n - b_n}{a_n}$ converges in distribution to $GEV(0, 1, \xi)$, then we can say that M_n is approximately GEV distributed with scale parameter a_n, location parameter b_n, and shape parameter ξ, for large n. This follows from the fact that if $X \sim GEV(\mu, \sigma, \xi)$, then $mX + b \sim GEV(m\mu + b, m\sigma, \xi)$ (described here and readily derived from the GEV cdf).