MTH1004 Summative Coursework for Year 1, Term 2

Report 1

Introduction

Taking the daily rainfall totals over 40 years which exceeded 25mm, I will determine a suitable distribution, assessing the level of error/ uncertainty for each estimation. Finally, I will compute the 10-year return level on this distribution, that is exceeded on average every 10 years.

Modelling with Exponential and Gamma Distributions

While there is a significant positive skew in the amount of excess rainfall over 25mm that occurs, $Figure\ 1$ shows no trend in the data across the days this excess rainfall is encountered, so it can be reasonably modelled as realisations of independent and identically distributed random variables.

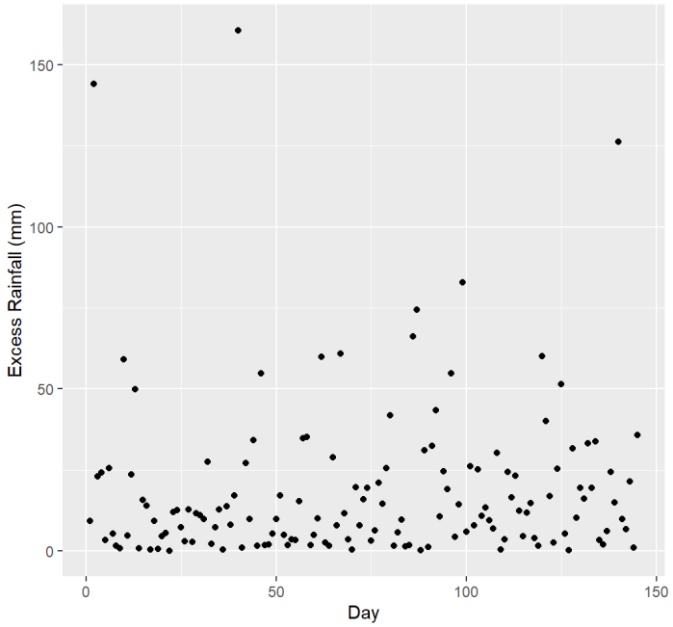

Figure 1: Scatter plot of those days with excess rainfall over 25mm, occurring over 40 years.

Since the datapoints are all positive real numbers, it is fitting to assume an Exponential or Gamma Distribution where the range space is $y > 0$. Attempting to model them as Exponential Distributions using the method of moments for parameter estimation:

$$Let\ Y_1 \dots Y_{145}\ be\ iid\ Exp(\lambda)$$

$$Given\ that\ E(Y) = \frac{1}{\lambda}, we\ can\ estimate: \bar{y} = \frac{1}{\hat{\lambda}}, so\ \hat{\lambda} = \frac{1}{\bar{y}}$$

$$For\ the\ sample\ data\ y_1 \dots y_{145};\ \bar{y} = 18.9mm\ (3.s.f)$$

$$Y \sim Exp\left(\frac{1}{18.9}\right)\ or\ Y \sim Exp(0.0529)\ (3.s.f)$$

Alternatively modelling them as Gamma Distribution yields the following:

$$Let\ Y_1 \dots Y_{145}\ be\ iid\ Ga(\alpha, \beta)$$

$$known\ point\ estimation\ is\ given\ by\colon\ \hat{\alpha} = \frac{\bar{y}^2}{s^2} and\ \hat{\beta} = \frac{\bar{y}}{s^2}$$

$$For\ the\ sample\ data\ y_1 \dots y_{145}, \bar{y} = 18.9mm\ and\ s = 24.8mm\ (3.s.f), so$$

$$Y \sim Ga\left(\frac{18.9^2}{24.8^2}, \frac{18.9}{24.8^2}\right)\ or\ Y \sim Ga(0.579, 0.0307)\ (3.s.f)$$

Comparing both distributions superimposed on the data ($Figure\ 2$) shows Exponential Distribution (Black) with similar fit for the middle 50% of the data but poorly skewed compared with Gamma Distribution (Red) which is more tightly fitting over the whole data range. This can be clearly seen in the q-q plot below ($Figure\ 3$) which demonstrates a more extreme skew for Exponential Distribution at the upper tail, indicating an overly light right-hand tail.

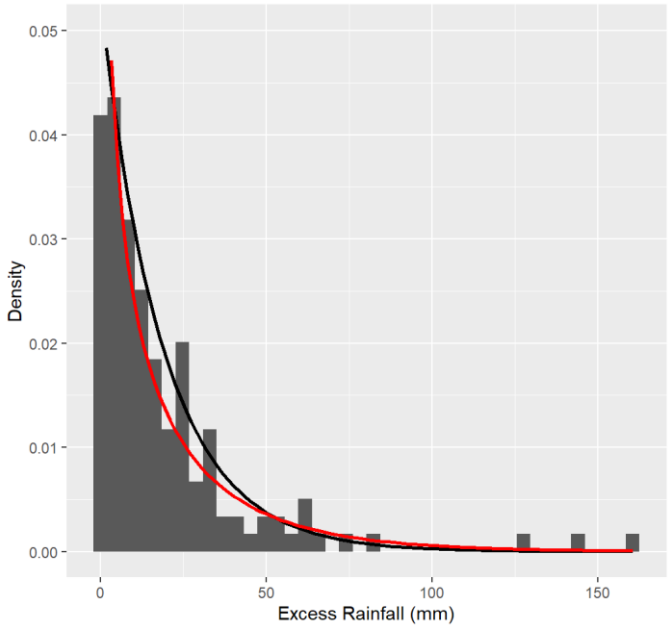

Figure 2: Density Histogram of the sample data superimposed with Exp (black) and Ga (red) Distributions.

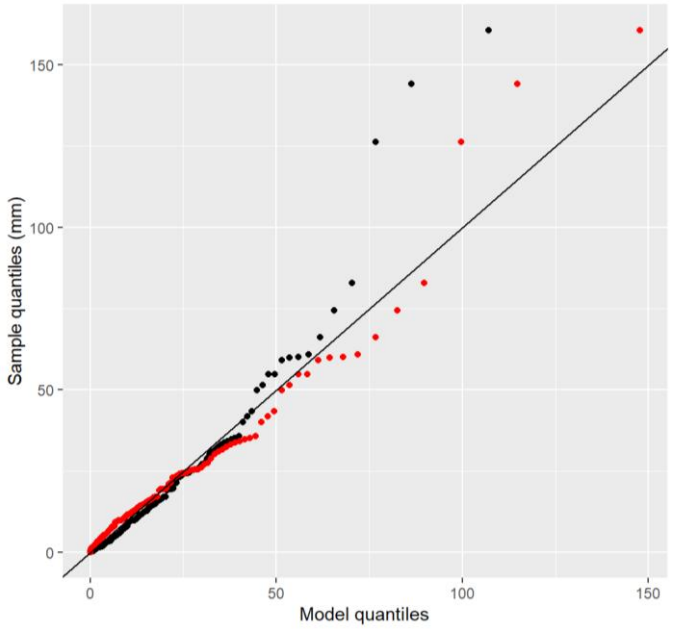

Figure 3: q-q plot showing Exp (black) and Ga (red) model structures compared with the sample data.

Then, to validate using the Gamma Distribution, we can find the standard errors in its parameter (point) estimations by simulation, visualised below ($Figure\ 4$):

$$standard\ error\ in\ \alpha = 0.108\ (3.s.f)$$

$$standard\ error\ in\ \beta = 0.00685\ (3.s.f)$$

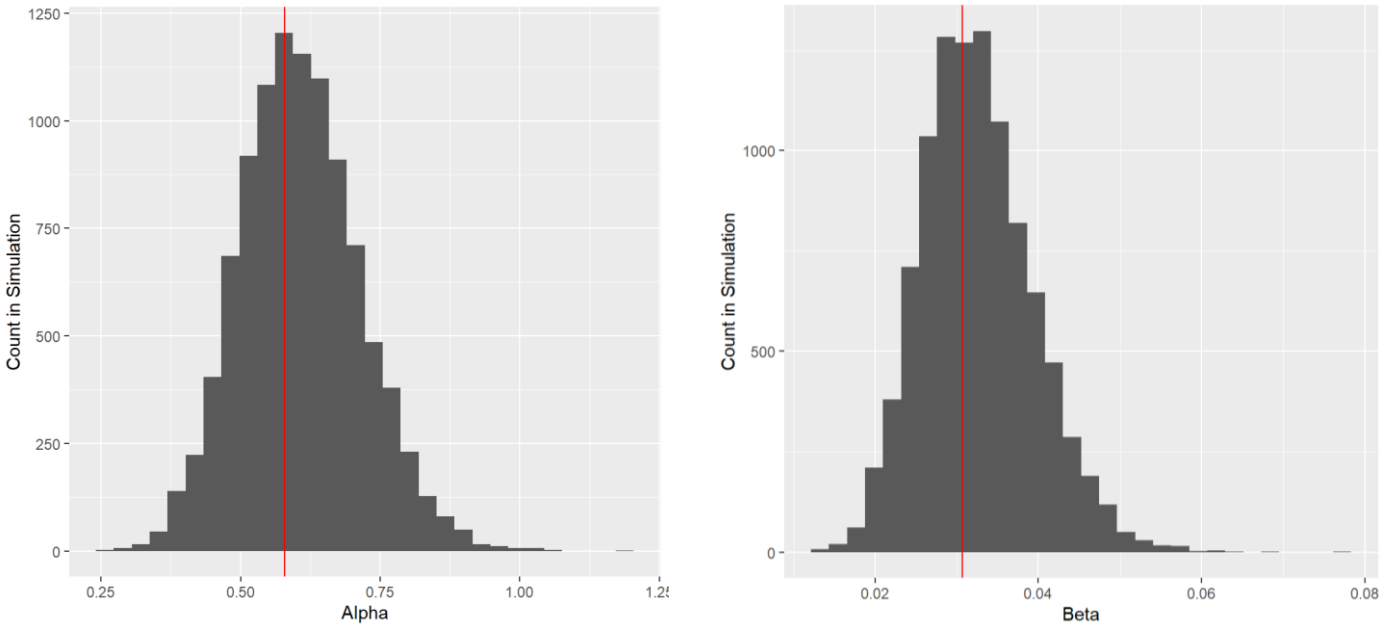

Figure 4: Two Histograms showing the how α and β vary over the simulation and a visualisation of the error given compared with the estimation (red).

The Gamma Distribution may be a reasonably accurate model to assume for these independent and identically distributed random variables with high confidence in β, in particular, compared with reasonable doubt in the relatively high error in α. This is reflected in the previously pointed feature of the Gamma Distribution being well spread (β) but tightly fitted/ sharply peaked to the left (α).

Modelling with given M Distribution

Now given model $M(\sigma, \gamma)$, the same process can be done to determine its feasibility to model Excess Rainfall occurring over these 40 years as follows:

$$Let\ Y_1 \dots Y_{145}\ be\ iid\ M(\sigma, \gamma)$$

$$Given\ that\ E(Y) = \frac{\sigma}{1-\gamma} = \bar{y}\ and\ Var(Y) = \frac{\sigma^2}{\{(1-\gamma)^2(1-2\gamma)^2\}} = s^2,$$

$$by\ substituting\ \sigma = \bar{y}(1-\hat{\gamma}), we\ can\ reduce\ to: 1 - 2\hat{\gamma} = \frac{\bar{y}^2}{s^2}\ so\ \hat{\gamma} = \frac{1}{2}(1 - \frac{y^2}{s^2})$$

$$and: \hat{\sigma} = \bar{y}\left(1 - \frac{1}{2} + \frac{\bar{y}^2}{2s^2}\right) = \frac{1}{2}\bar{y}(1 + \frac{\bar{y}^2}{s^2})$$

$$For\ the\ sample\ data\ y_1 \dots y_{145}, \bar{y} = 18.9mm\ and\ s = 24.8mm\ (3.s.f), so$$

$$Y \sim M(14.9, 0.210)\ (3.s.f)$$

This M Distribution fits the data similarly well to the previous estimates ($Figure\ 5$), however a noticible improvement can be seen in the q-q plot below ($Figure\ 6$) indicating much more balanced skew which is relatively more accurate despite naturally light right-hand tail where any continuous distribution would show imbalance for data that "fades out" as $y \to \infty$, due to real world limitations, put simply in this case – there can only be so much rain in one day.

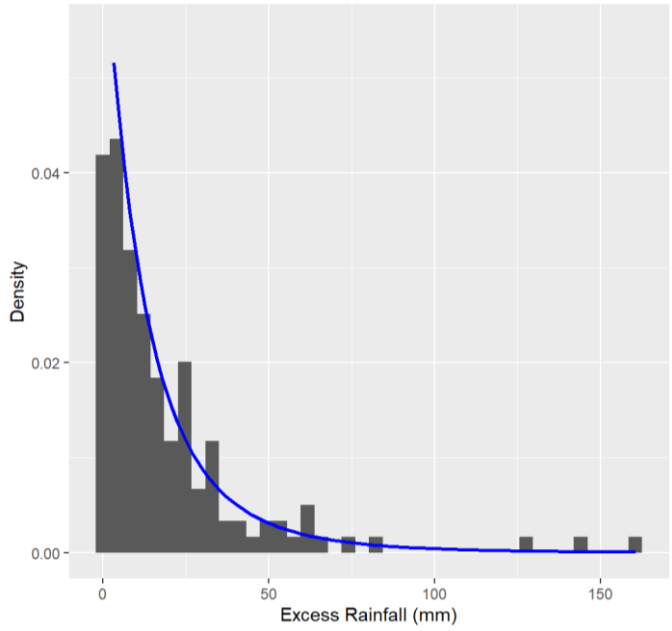

Figure 5: Density Histogram of the sample data superimposed with $M(\sigma, \gamma)$ (blue) Distribution.

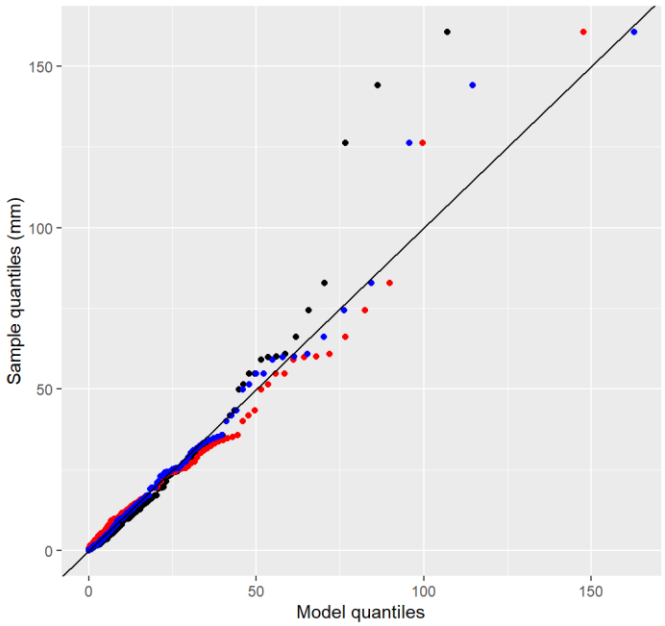

Figure 6: q-q plot showing Exp (black), Ga (red), M (blue) model structures compared with the sample data.

Then, to validate using the M Distribution, we can find the standard errors in its parameter (point) estimations by simulation, again visualised below ($Figure\ 7$):

$$standard\ error\ in\ \sigma = 1.89\ (3.s.f)$$

$$standard\ error\ in\ \gamma = 0.0900\ (3.s.f)$$

Not only does the error in σ indicate major inaccuracy, the histograms in $Figure\ 7$ show an offset in both parameters suggesting an inadequate model for this particular use case, therefore I will use the Gamma Distribution determined previously to model the excess rainfall amounts as realisations of independent and identically distributed random variables.

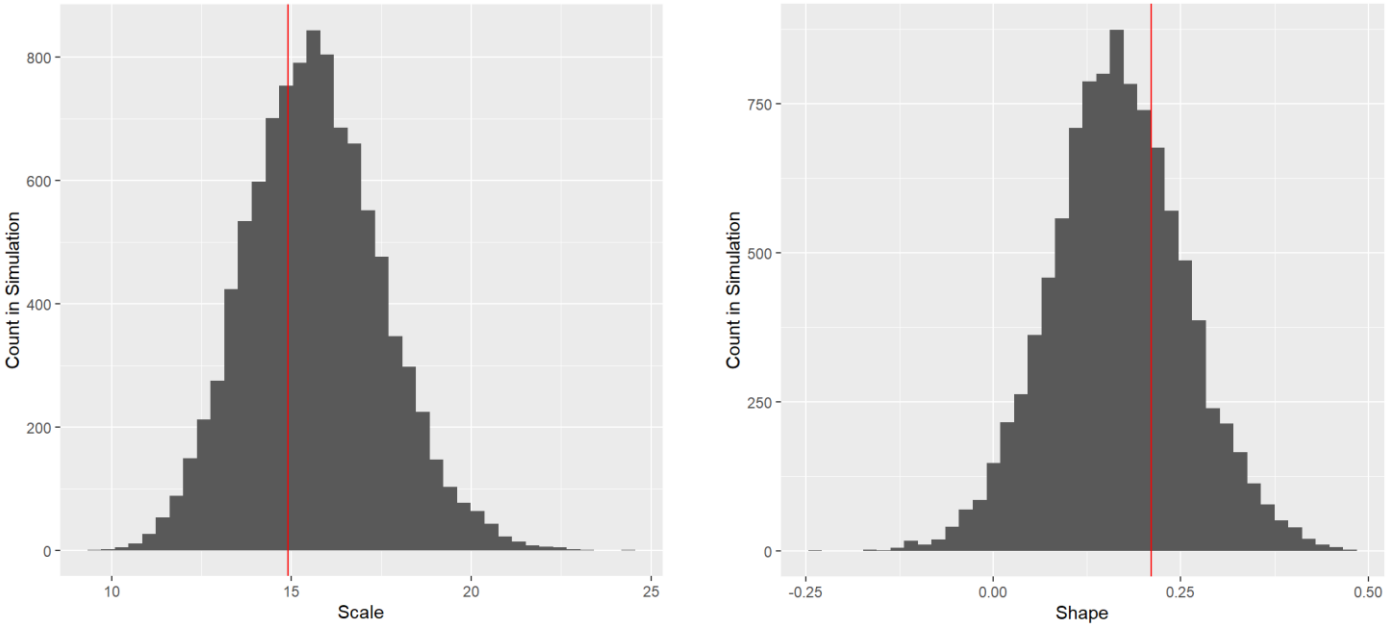

Figure 7: Two Histograms showing the how σ(scale) and γ(shape) vary over the simulation and the reliability of the a visualisation of the error given compared with the estimation (red).

Finding the m-year Return Level

To find the m-year return level x, first we need the expected proportion of the sample data which should be found to exceed the level $x - 25$, given by:

$$P(y > x - 25) = \frac{1}{365mp}$$

Where $m = 10$ in this case, and $p = 0.01$, so:

$$Let\ k = x - 25$$

$$P(y > k) = \frac{1}{36.5} = 1 - P(y > k)$$

$$ie.\ 1 - F(k) = \frac{1}{36.5}$$

However we are not given, nor is it easy to find, the cdf of the Gamma Distribution, we can however manipulate the expression into a question of quantiles:

$$F(k) = 1 - \frac{1}{36.5} = \frac{71}{73}$$

So at "$q_{\frac{71}{73}}"th$ will be the data value (k) where a proportion of $\frac{71}{73}$% of the sample' s range falls below it ($F(k)$) in other words, the value at which only a proportion of $\frac{1}{36.5}$ of the data exceeds k.

This gives:

$$k = 86.1mm\ (3.s.f)$$

$$x = k + 25 = 111mm\ (3.s.f)$$

Summary

I assessed the feasibility of modelling the daily excess rainfall amounts as realisations of independent and identically distributed random variable with Exp and Gamma Distributions, of which I then determined Gamma to be more reliable, even compared to the given model M using method of moment estimations and validating these by computing the standard error on each estimation.

I was able to then calculate the 10-year return level x = 111mm (3.s.f) by taking a unique approach using quantiles.

Report 2

Introduction

By analysing the results of a yearlong study on the effects of a new antibiotic, I will assess whether it is more or less effective than an existing drug, take into account which hospital the patient attended and evaluate the overall study.

General effectiveness of the New Antibiotic

Ignoring which hospital each patient visited, we can interpret every patient's results as a independent and identically distributed Bernoulli Trials with some constant probability which is the same for all patients taking the new antibiotic. This is working under the assumption that every patient has the same "medical features" and therefore are all equal members of the target population, in other words we expect them all to react the same to the drug (constant p). This allows us to model the sample of 98 patients as a Binomial Distribution such that:

$$Y \sim B(n = 98, p)$$

$$Let\ p = the\ probability\ that\ any\ randomly\ chosen\ patient\ is\ treated\ successfully$$

$$within\ 2\ weeks\ in\ the\ (target)\ population$$

$$using\ point\ estimation: p = sample\ proportion\ treated\ successfully = \frac{72}{98}$$

$$p = 0.735\ (3.s.f)$$

Due to large $n > 50$ and large $p > 0.1$, this Binomial Model can approximate to a Normal Distribution, where:

$$\mu = np = 72$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{19.1}\ (3.s.f)$$

$$Y \dot{\sim} N(72, 19.1)\quad (3.s.f)$$

This allows us to easily find the equal-tailed confidence interval at 95%, for example, via standardization:

$$\bar{y} \pm \frac{z_p S}{\sqrt{n}}\ \ for\ p = 0.025\ and\ s = sample\ sd = 19.1\ (3.s.f)$$

Now, number of successfully treated patients in sample of 98 is, on average:

$$(71.1, 72.9)\ patients\ where\ n = 98$$

As a percentage this interval can be more effectively compared:

$$(72.6, 74.4)\%\ of\ all\ patients\ recieving\ the\ new\ antibiotic$$

This is notably better than the existing antibiotic with 70% success rate, however we have not yet questioned the assumption that "both hospitals are the same", ie. all the patients are equal members of the target population.

Considering Difference between Hospital A and B

The same analysis as above can be carried out, all the relevant datapoints below in $Figure\ 8$.

Sample	n	p	$\bar{y}$	s	interval	as %
Hospital A patients	55.0	0.836	46.0	7.53	(44.0, 48.0)	(80.0, 87.3)
Hospital B patients	43.0	0.605	26.0	10.3	(22.9, 29.1)	(53.3, 67.6)

Figure 8: Table showing figures from confidence interval calculation per hospital, all in 3.s.f

Despite being a strong contender in the general case, the new antibiotic behaves very differently when looking specifically at each hospital. In Hospital A, the antibiotic is being given to a widely varied group of people in terms of "medical features", a very strong proportion of which ≈80-87% generally are successfully treated within 2 weeks. This is significantly better than the existing antibiotic option which generally clears infections for 70% within 2 weeks.

On the other hand, in Hospital B a far smaller proportion of patients are successfully treated within 2 weeks with this new antibiotic and more importantly even less than the existing drug. This suggests that these patients, being at a specialized Hospital, share some "medical feature" making the new antibiotic comparatively weaker/ less reliable despite being in the target population (those infected with the specific bacterial infection).

Evaluating the Study

Evaluating the study further, the poor results found at Hospital B are likely to be the result of confounding and the organisers should implement strategies to avoid this occurring. For example, matching pairs of people with similar backgrounds, health conditions, personal features (age, sex etc) across those in and not in Hospital B will allow the researchers to avoid known confounders.

To tackle confounders which are not well known or cannot be determined easily, such as hidden genes/ unknown medical history, the researchers should implement randomisation to randomly determine which individual of each pair is in control vs. intervention group.

The study does not describe any control group since the assumption is that the drug works and the test is just to learn how effective it is compared to other solutions, however a control is necessary to isolate the effect of taking the antibiotic. Alternatively, they should consider a whider more varied sample to test on, potentially taking a stratified sample across different types of hospitals/ patients/ "medical features".

Summary

I determined a Binomial Distribution for the sample of patient results, which due to CLT I was able approximate as a Normal Distribution. This allowed for easy computation of the confidence interval on my point estimate for population mean from the sample, showing that in the general case, the new antibiotic was more effective than the existing option. However, looking into each Hospital individually make it clear that while the drug was highly effective in general, a subset of the target population concentrated at Hospital B may suffer from a confounder hindering the new antibiotic.