06051540-MATH70076-assessment-1

MSc in Statistics 2025/26, Imperial College London

Justin Upson

Question 1

For $\xi \neq 0$:

We know from the cumulative distribution function that:

$$F(x; \sigma, \xi, u) = 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)_+^{-1/\xi} \tag{1}$$

Rewriting to make x the subject:

$$\begin{aligned}
\left(1 + \frac{\xi(x-u)}{\sigma}\right)_+^{-1/\xi} &= (1 - F) \qquad (2)\\
1 &= (1 - F)\left(1 + \frac{\xi(x-u)}{\sigma}\right)_+^{1/\xi} \qquad (3)\\
(1 - F)^{-1} &= \left(1 + \frac{\xi(x-u)}{\sigma}\right)_+^{1/\xi} \qquad (4)\\
& \qquad (5)
\end{aligned}$$

For $\xi > 0$, we see that $\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{1/\xi} > 0$, so we get:

$$\begin{aligned}
(1 - F)^{-\xi} &= 1 + \frac{\xi(x-u)}{\sigma} \qquad (6)\\
(1 - F)^{-\xi} - 1 &= \frac{\xi(x-u)}{\sigma} \qquad (7)\\
\left((1 - F)^{-\xi} - 1\right) \times \frac{\sigma}{\xi} &= x - u \qquad (8)\\
u + \left((1 - F)^{-\xi} - 1\right) \times \frac{\sigma}{\xi} &= x \qquad (9)\\
& \qquad (10)
\end{aligned}$$

So $F_X^{-1} = u + \Big((1-x)^{-\xi} - 1\Big) \times \frac{\sigma}{\xi}$ But given our inputs of $F_X^{-1}(x)$ vary between 0 and 1, we can write that:

$$F_X^{-1}(x) = u + \Big((x)^{-\xi} - 1\Big) \times \frac{\sigma}{\xi} \tag{11}$$

For $\xi < 0$, we know that we have quickly decaying tails with finite upper endpoint. With $x > u$, this finite endpoint is met when

$$\begin{align}
1 + \frac{\xi(x-u)}{\sigma} &= 0 \tag{12}\\
\frac{\xi(x-u)}{\sigma} &= -1 \tag{13}\\
x - u &= -\frac{\sigma}{\xi} \tag{14}\\
x &= u - \frac{\sigma}{\xi} \tag{15}\\
& \tag{16}
\end{align}$$

So for $x > u - \frac{\sigma}{\xi}$, produced by the inverse function above we discard the values of x.

For $\xi = 0$:

$$\begin{align}
F = 1 - exp\Big(-\frac{x-u}{\sigma}\Big) \tag{17}\\
exp\Big(-\frac{x-u}{\sigma}\Big) = 1 - F \tag{18}\\
-\frac{x-u}{\sigma} = ln(1-F) \tag{19}\\
-x + u = \sigma ln(1-F) \tag{20}\\
u - \sigma ln(1-F) = x \tag{21}\\
\tag{22}
\end{align}$$

So $F_X^{-1} = u - \sigma ln(1-x)$

But given our inputs of $F_X^{-1}(x)$ vary between 0 and 1, we can write that:

$$F_X^{-1}(x) = u - \sigma ln(x) \tag{23}$$

Question 2a

Defining the quantile function, and using the basis of the cdf from question 1, we get that:

```r
qgpd <- function (p, sigma=1, xi=0, u=0){
  if (p<0||p>1){
    return(warning("NaNs produced - p must be between 0 and 1"))
  }
  else if (sigma<=0){
    return(warning("NaNs produced - sigma must be greater than 0"))
  }
  else if (xi != 0){
    return(u + ((1-p)^(-xi)-1) * sigma / xi)
  } else {
    return(u - sigma * log(1-p))
  }
}
```

By default, the expected inputs for the function are sigma=1, xi=0, u=0. The code also prevents inputs where p values are less than 0, where p values exceed 1, or where sigma is less than or equal to 0.

The expected output is a real number greater than u that is unbounded if xi is greater than or equal to 0. The expected output is less than u - sigma/xi if xi is less than 0.

Regarding the behaviours of the quantile function: The larger the value of xi, the slower the tail decays. In this for xi >= 0 the functions output approaches infinity as p -> 1. For xi < 0, the functions output has a maximum at u - sigma/xi (when p -> 1). For larger values of sigma, the slower the tail decays.

Question 2b

```r
qgpd(0.5,2,-0.4,1.5)
```

```
[1] 2.710709
```

```r
qgpd(0.75,2,-0.4,1.5)
```

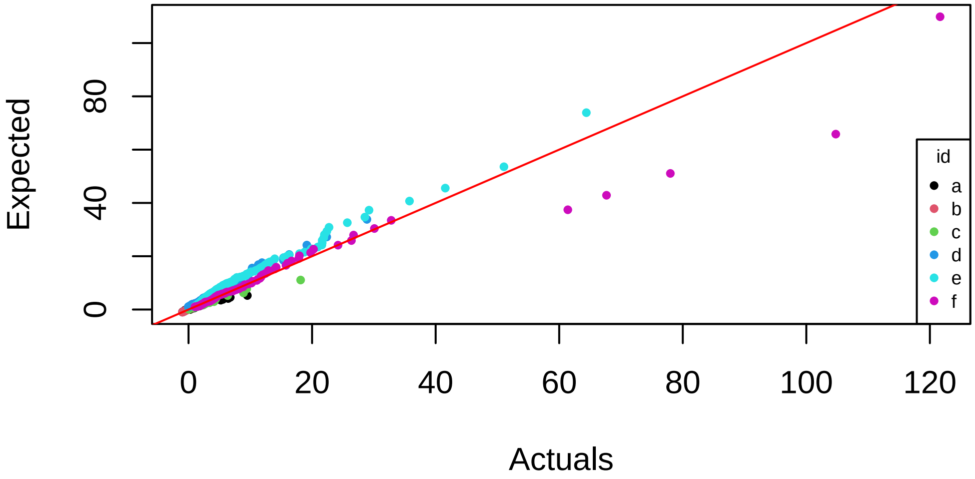
```
[1] 3.628254
```

```r
qgpd(0.99,2,-0.4,1.5)
```

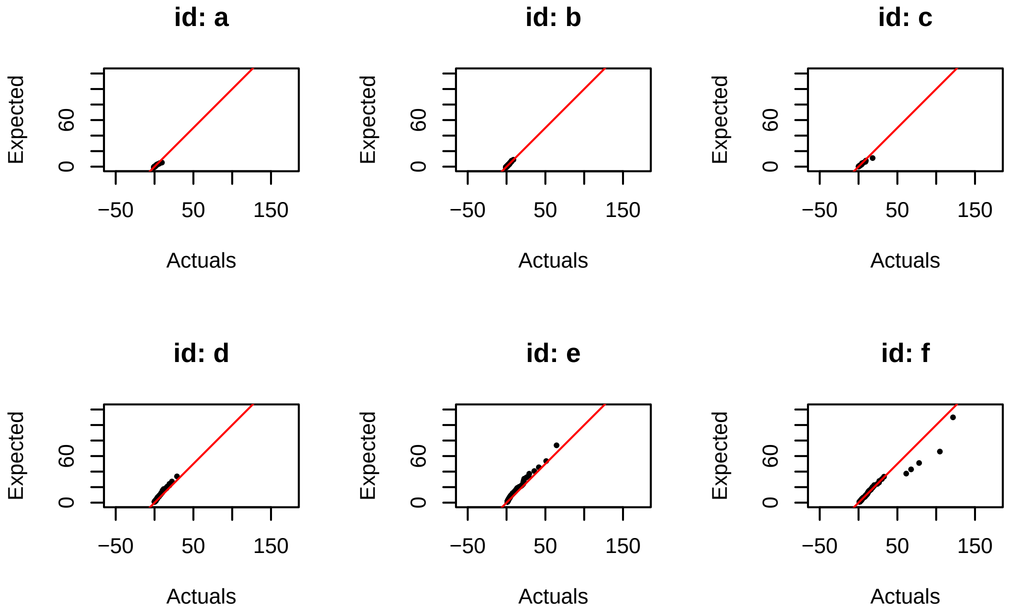
```
[1] 5.707553
```

Question 3

Graphing our actual vs expected split by id, we see that:

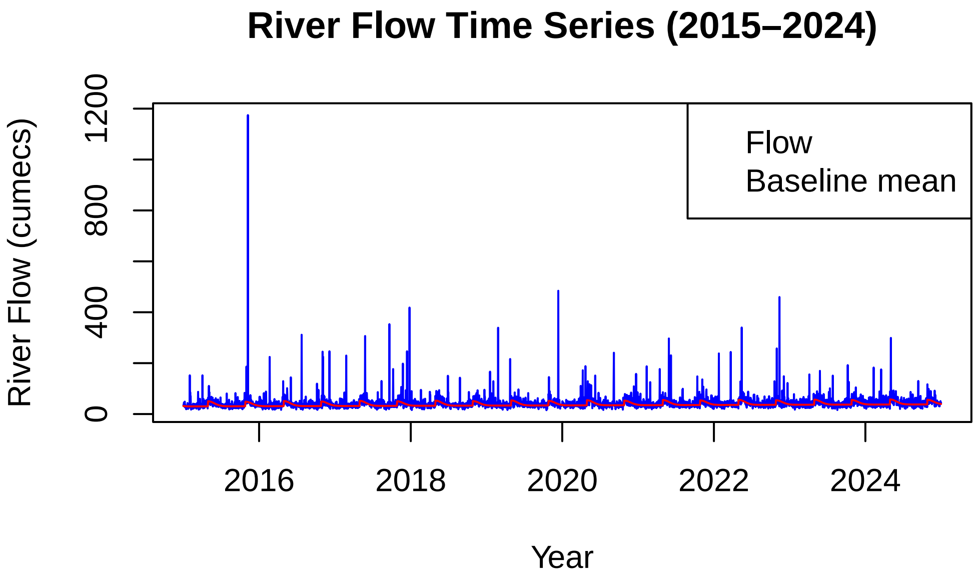


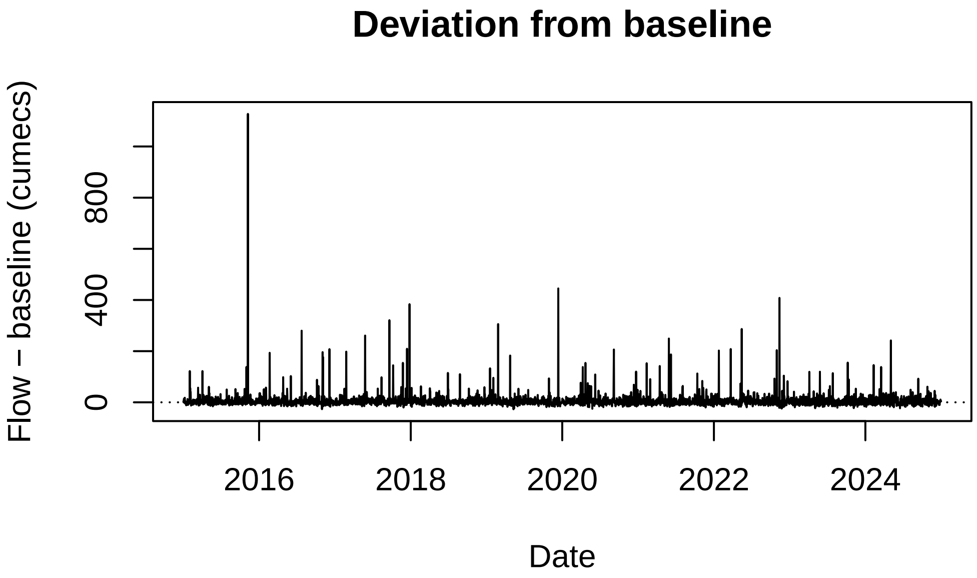
From the figure above, we see some interesting results – although most actuals seem in line with the expecteds (being close to the figures red line), this is not always the case. Ultimately, however, the different ids need to be separated so as to provide a more granular analysis of the distributions:



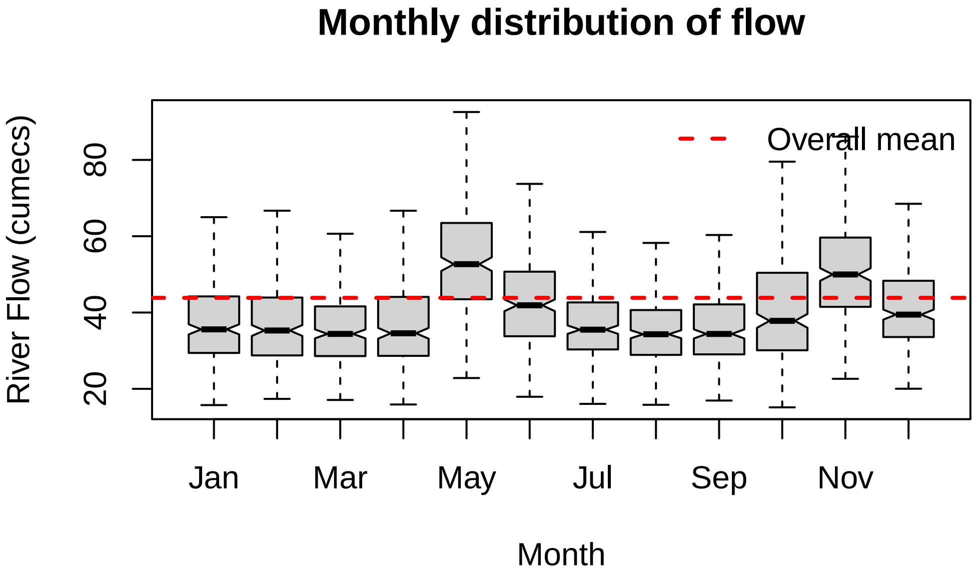
From these smaller figures, we can much more readily see which of the actual distributions align with the expected distributions. We can therefore see that:

Question 4

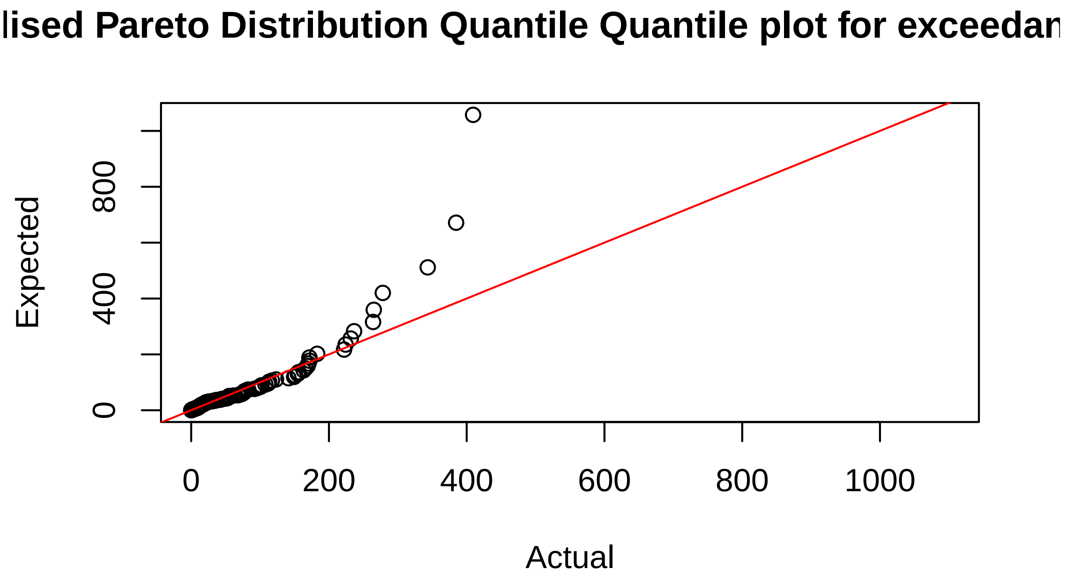




Determining whether the distribution of river flows is constant over time, we can see from the above that an assumption of constant flow is flawed. This is because we see that the speed of the river's flow varies tremendously - at numerous (albeit brief) intervals over the ten year period, the speed of the river is many multiples of the mean river speed. This is something we can demonstrably see when comparing the mean red dashed line with the time series plot above.



By using a box and whisker plot which considers each month, we can see how seasonal changes cause profound effects in the speed of the river's flow. The month of May, for example has an interquartile range in river flow speeds that exceed the entire interquartile range in river flow speeds from the month of March.



To determine whether it is appropriate to model large data flows with the stated GPD, I constructed a quantile quantile plot contrasting the actual dataset, and the expected data. In this, the expected data was calculated using the inverse cumulative distribution.

From the graph, we can see that it is not appropriate to model large data flows with the GPD provided in the question. This is because we see a significant deviation of the plotted points from the $y = x$ line - the expected values don't align with the actual values when dealing with the largest river flows.