Order Statistics

This is a preview for a group project.
Contact Joseph Marks for a full version.

Supervisor: Robin Henderson

2019/20

Abstract

In this report, we discuss the topic of order statistics. Opening with a motivating example, we consider the distribution of order statistics in two special cases and then in the general case, with histograms simulating different cases from the uniform distribution.

We discuss the expected values of order statistics, deriving and applying the formula for the expected value of the i-th order statistic.

We consider the two extreme values of order statistics, a subtopic known as extreme value theory. This leads to discussion of the generalised extreme value distribution, used to model extrema, and applications thereof. The three special cases of this distribution are also described.

We further investigate extremes: analysing rainfall in the City of Durham over the past 140 years and producing accompanying graphical plots. A comparison is made between Durham and Stornoway in Scotland, of maximum and expected rainfall for both locations.

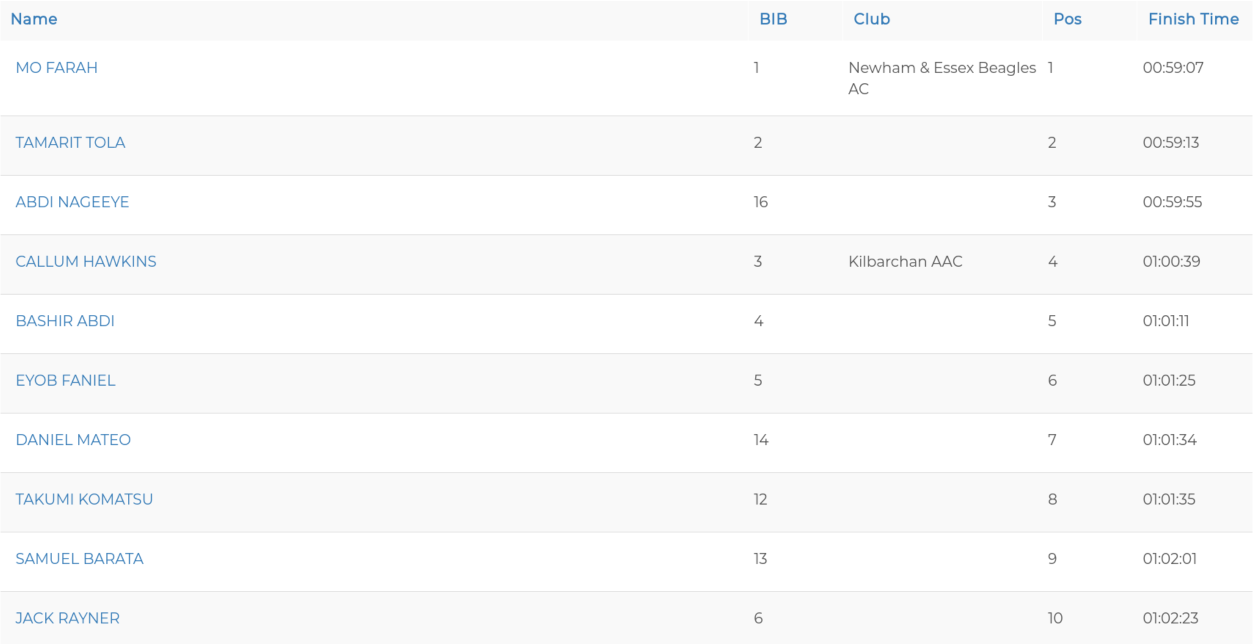
Our motivating example is linked to the generalised extreme value distribution and we present our findings with figures before carrying out a comparative analysis between different events.

This report concludes by summarising the topics covered, discussing our assumptions and positing some further analysis that could be carried out.

1 Introduction

1.1 Motivating Example

A real-world case of order statistics can be seen in Figure 1, showing the times of a number of athletes from the Great North Run in 2019. Each competitor has their individual time, and these have been ordered from quickest to slowest, with the smallest number as the first statistic and each consecutive entry being larger. One question a statistician may ask is: could one construct a model which can make predictions about the times of the first or last runner?



Name	BIB	Club	Pos	Finish Time
MO FARAH	1	Newham & Essex Beagles AC	1	00:59:07
TAMARIT TOLA	2		2	00:59:13
ABDI NAGEEYE	16		3	00:59:55
CALLUM HAWKINS	3	Kilbarchan AAC	4	01:00:39
BASHIR ABDI	4		5	01:01:11
EYOB FANIEL	5		6	01:01:25
DANIEL MATEO	14		7	01:01:34
TAKUMI KOMATSU	12		8	01:01:35
SAMUEL BARATA	13		9	01:02:01
JACK RAYNER	6		10	01:02:23

Figure 1: Great North Run 2019 Elite Men's results[1]

Order statistics are also useful for the detection of outliers. A test can determine whether an observation is discordant with a model, or if another model needs to be made that is a better fit.

1.2 Notation and Definitions

Given some random variables $X_1, X_2, \ldots, X_n$, which are independently and identically distributed, we use the mapping $X_i \mapsto X_{(i)}$ such that the following inequality holds:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

We define $X_{(i)}$ as being the i-th order statistic, where $X_{(1)}$ is defined as $\min\{X_1, X_2, \ldots, X_n\}$ and $X_{(n)}$ as $\max\{X_1, X_2, \ldots, X_n\}$.

In the motivating example from Figure 1, X_i would represent the time of the runner with the i-th bib number, and $X_{(i)}$ the runner finishing in the i-th position.

Throughout this report we will use $F(x)$ and $f(x)$ to denote the cumulative distribution function (cdf) and probability density functions (pdf), respectively, of X_i. Additionally, all values are given to four decimal places of accuracy unless otherwise stated.

2 Probability Distributions

2.1 Special Cases: First & Last Order Statistic

To derive the cdf of the first order statistic, $F_{X_{(1)}}(x)$, we first consider its definition:

$$\begin{aligned} F_{X_{(1)}}(x) &= \Pr\left(X_{(1)} \leq x\right) \\ &= 1 - \Pr\left(X_{(1)} > x\right). \end{aligned}$$

By definition, if $X_{(1)}$ is greater than some arbitrary x, then so are all other X_i, therefore:

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - \Pr\left(X_1 > x, X_2 > x, \ldots, X_n > x\right) \\ &= 1 - \Big([1 - F_{X_1}(x)]\,[1 - F_{X_2}(x)] \ldots [1 - F_{X_n}(x)] \Big) \\ &= 1 - \prod_{i=1}^{n} [1 - F_{X_i}(x)]. \end{aligned}$$

As each X_i is independently and identically distributed, we let each $F_{X_i}(x) = F(x)$, hence:

$$F_{X_{(1)}}(x) = 1 - [1 - F(x)]^n. \tag{1}$$

Consequently, to derive the probability density function of the first order statistic, $f_{X_{(1)}}(x)$, we differentiate $F_{X_{(1)}}(x)$ as follows:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\left[F_{X_{(1)}}\right] &= \frac{\mathrm{d}}{\mathrm{d}x}\left[1 - [1 - F(x)]^n\right] \\ &= -n\,[1 - F(x)]^{n-1}\,\frac{\mathrm{d}}{\mathrm{d}x}\left[1 - F(x)\right] \\ &= n\,[1 - F(x)]^{n-1}\,f(x). \end{aligned}$$

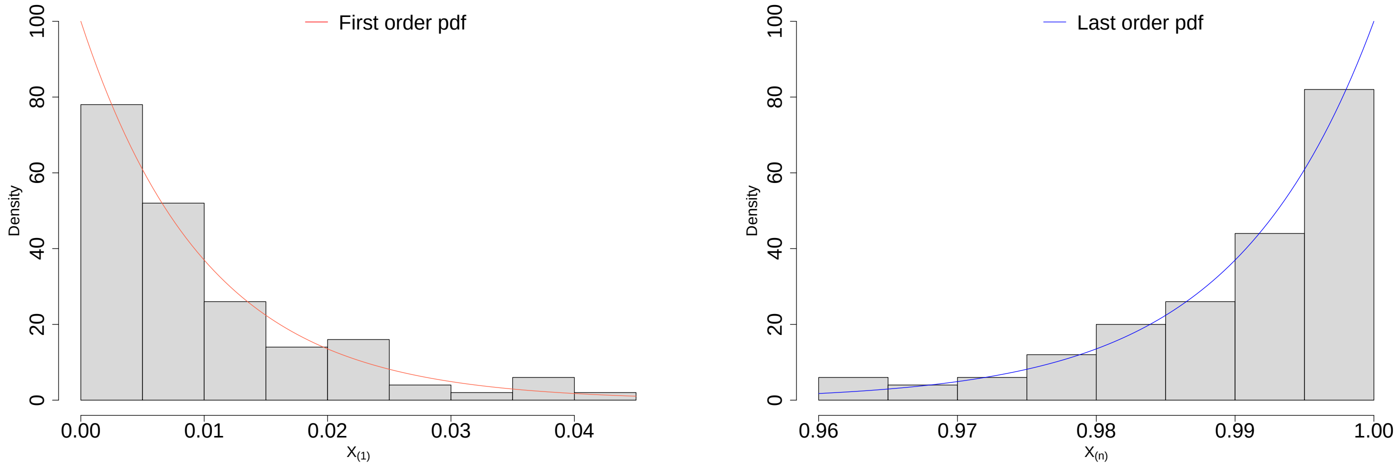
Our final result is:

$$f_{X_{(1)}}(x) = n\,[1 - F(x)]^{n-1}\,f(x). \tag{2}$$

A similar argument (see appendix) can be made for the last order statistic cdf and pdf. In the next section we will verify the following results:

$$F_{X_{(n)}}(x) = [F(x)]^n, \tag{3}$$

$$f_{X_{(n)}}(x) = n\,[F(x)]^{n-1}\,f(x). \tag{4}$$



(a) First order statistic distribution.

(b) Last order statistic distribution.

Figure 2: Histogram of simulated final order statistics where $X_i \sim \mathrm{Unif}(0, 1)$, $n = 100$.