The Minnesota Journal of Undergraduate Mathematics

The Reality of Olympic Sprinting

Patrick Gravelle

Queen's University

The Minnesota Journal of Undergraduate Mathematics

Volume 3 Number 1 (2017-2018)

Sponsored by
School of Mathematics
University of Minnesota
Minneapolis, MN 55455

The Minnesota Journal of Undergraduate Mathematics

Volume 3 Number 1 (2017-2018)

The Reality of Olympic Sprinting

Patrick Gravelle

Queen's University

Abstract. Every four years, elite athletes gather at the Summer Olympic Games, aiming to be crowned as the world's best. Audiences are spoiled with fantastic performances from competitors such as Michael Phelps and Usain Bolt, and watch each event to see not only who will win a gold medal, but whether any world records will be set in the process. But how frequent are those records? This paper aims to answer that question for the track sprinting events. First we construct a cumulative distribution function to model the probability that an Olympic sprinter will be able to race within a certain range of times. After estimating parameter values for the model, we use techniques from probability theory to determine the likelihood than an elite athlete will be able to break a current world record.

1. Introduction

The Olympic Games have been the venue for numerous world record achievements in track sprinting races. The world watches these elite sprinters, hoping to see an athlete sprint faster than any man or women in history. The current world record times remain in the bottom corner of the television screen, tantalizing audiences. This begs the question of whether our expectations of world records is too high. The goal of this paper is to determine the probability that a fan at the Olympics will witness a single sprinting world record, multiple world records, or no record at all.

The models constructed for this paper only apply to elite Olympic sprinters, as those are the individuals that will break the existing world records and who have done so in the past. In the context of this paper, sprinting events are defined to be the 100 meter and 200 meter dashes for both men and women, yielding four separate events. The data used [4] were collected prior to the 2016 Rio Summer Olympic Games and include times for each of those events. These data sets are in descending order, from the fastest recorded times in an official event to the slowest. Each set for the men's sprints contain nearly 3000 data entries, whereas the data sprints comprises nearly 2000 entries.

2. Distribution Function

To interpret the data from [4], it is necessary to create a model that to determine the probability than an event will occur. One way this can be accomplished is through construction of a cumulative distribution function (CDF), which is a function of ordered cumulative event probabilities for a particular data set. This is used because it allows for the ease of calculation for the probability that an event will occur (e.g. a sprinter whose time is the fastest in history).

To construct a CDF, we choose a lower bound y_0, which represents an impossibly fast time, and an upper bound y_1, which is so high that we are certain the sprinters will complete the race within that time. The interval $[y_0, y_1]$ of finishing times between is then scaled to $[0,1]$, yielding the following CDF:

$$F(y) = \begin{cases} 0 & \text{if } y \leq y_0 \\ \left(\frac{y-y_0}{y_1-y_0}\right)^{\alpha} & \text{if } y_0 < y < y_1 \\ 1 & \text{if } y \geq y_1 \end{cases} \tag{1}$$

where α is a shape parameter which alters the shape of the distribution. This CDF will be used to determine the probability that a sprinter will finish a race within a certain time. The probability of a sprinter finishing in a time equal to or less than y_0 is 0, but by time y_1 the sprinter will have finished with probability 1. Equation (1) will allow us to compute the probability of the sprinter finishing with a time less than or equal to y for $y_0 < y < y_1$, which then enables us to calculate the probability of world records being set.

3. Estimating Alpha

The CDF in (1) has a single parameter, α. The effect of changing α on the shape of the distribution will be more thoroughly understood once the relationship between y_0, y_1, and α is established. First, one must estimate the value of α, which may be done through the method of moments. This method begins by taking the derivative of the CDF to obtain the probability density function, or PDF. The PDF is used to calculate the expected value of the distribution, $E(X)$, as follows:

For a continuous random variable X with CDF $F(x)$, the expected value of X is defined as

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

where $F'(x) = f(x)$, and μ can be estimated by $\bar{x}$ which is the mean time of the data set.

Taking the derivative of $F(x)$:

$$\begin{aligned}\frac{d}{dx}F(x) &= \frac{d}{dx}\left(\frac{x-y_0}{y_1-y_0}\right)^{\alpha}\\ f(x) &= \alpha\left(\frac{x-y_0}{y_1-y_0}\right)^{\alpha-1}\cdot\frac{1}{y_1-y_0}\end{aligned}$$

Substituting $f(x)$ into $E(X)$ and integrating with respect to x:

$$\begin{aligned}\mu = E(X) &= \int_{y_0}^{y_1} x\alpha(\frac{x-y_0}{y_1-y_0})^{\alpha-1}\frac{1}{y_1-y_0}dx\\ &= \int_{y_0}^{y_1}\frac{\alpha}{(y_1-y_0)^{\alpha}}(x-y_0+y_0)(x-y_0)^{\alpha-1}dx\\ &= \int_{y_0}^{y_1}\frac{\alpha}{(y_1-y_0)^{\alpha}}\Big((x-y_0)^{\alpha}+y_0(x-y_0)^{\alpha-1}\Big)dx\\ &= \frac{\alpha}{(y_1-y_0)^{\alpha}}\int_{y_0}^{y_1}(x-y_0)^{\alpha}dx+\frac{\alpha y_0}{(y_1-y_0)^{\alpha}}\int_{y_0}^{y_1}(x-y_0)^{\alpha-1}dx\\ &= \frac{\alpha}{(y_1-y_0)^{\alpha}}\int_{y_0}^{y_1}(x-y_0)^{\alpha}d(x-y_0)+\frac{\alpha y_0}{(y_1-y_0)^{\alpha}}\int_{y_0}^{y_1}(x-y_0)^{\alpha-1}d(x-y_0)\\ &= \frac{\alpha}{(y_1-y_0)^{\alpha}}\frac{(x-y_0)^{\alpha+1}}{\alpha+1}\Bigg|_{y_0}^{y_1}+\frac{\alpha y_0}{(y_1-y_0)^{\alpha}}\frac{(x-y_0)^{\alpha}}{\alpha}\Bigg|_{y_0}^{y_1}\\ &= \frac{\alpha}{\alpha+1}(y_1-y_0)+y_0\\ &= (1-\frac{1}{\alpha+1})(y_1-y_0)+y_0\\ &= y_1-\frac{1}{\alpha+1}(y_1-y_0)\end{aligned}$$

Rearranging for α yields

$$\begin{aligned}\frac{1}{\alpha+1}(y_1-y_0) &= y_1-E(x)\\ \alpha+1 &= \frac{y_1-y_0}{y_1-E(x)}\\ \alpha &= \frac{y_1-y_0}{y_1-E(x)}-1.\end{aligned}$$

Finally, using $\bar{x}$ as an estimate for $\mu = E(X)$ gives our estimate for the parameter α:

$$\hat{\alpha} = \frac{y_1-y_0}{y_1-\bar{x}}-1. \tag{2}$$

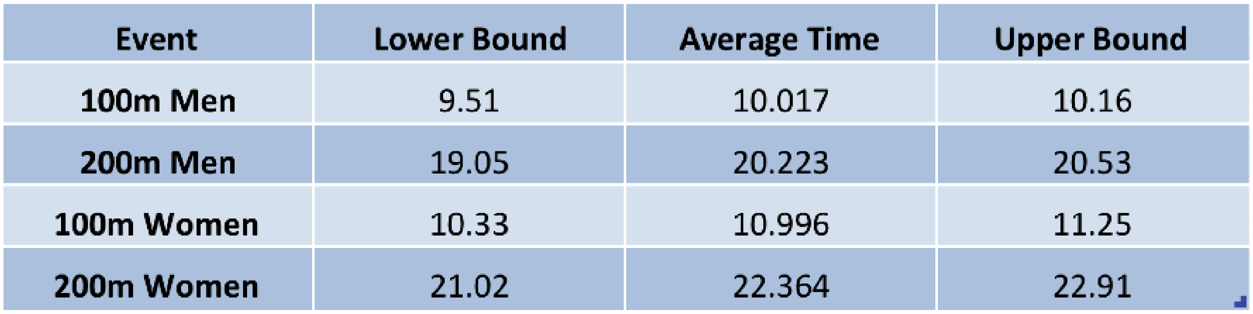
Thus from (2), the estimate of $\hat{\alpha}$ is dependent upon the bounds of the distribution and the average time from the data set. Since $\bar{x}$ is a constant value calculated from each sprinting event's data set, to change $\hat{\alpha}$ one must change the bounds of the distribution, ie. y_0 and y_1. Thus, following the determination of these bounds, a sensitivity analysis testing a range of different values for y_0 and y_1 will establish an understanding of how these different values affect $\hat{\alpha}$, and the individual and cumulative event probabilities of a world record being broken. Moreover, this will allow for the bounds to accurately represent past data in the likelihood of a world record being broken.

4. Obtaining Unknown Values: y_0, y_1, and $\bar{x}$

Before calculating the probability that an elite Olympic sprinter will break the world record in a particular race, we must choose bounds y_0 and y_1 for the event. For y_0, we used the theoretical fastest times for each race, calculated in [1]. For the Men's 100m Dash, $y_0 = 9.51$, which is 0.07 less than the current men's world record. For the Women's 100m Dash, $y_0 = 10.33$, a -0.16 difference from the current women's world record. Taking twice these differentials and adding them to their respective sex's 200m world record time yields the remaining two lower bounds. This is done because the world record for both the Men's and Women's 200m Dash are approximately twice that of their respective 100m races, yielding $y_0 = 19.05$ for the men and $y_0 = 21.02$ for the women.

To determine y_1, take the positive differential between the world record and y_0, and add this number to the last distinct time in the main list from [4]. For example, the Men's 100m Dash times from [4] range from 9.58 to 10.09. Taking positive 0.07 added to the last distinct time of 10.09 yields an upper bound of 10.16.

Thus, the last remaining unknown value required to estimate $\hat{\alpha}$ is $\bar{x}$. To obtain this value for each event, take the sum of all times recorded for the specific race from [4] and divide by the total number of race times listed. This process is repeated for each of the three remaining events.



Event	Lower Bound	Average Time	Upper Bound
100m Men	9.51	10.017	10.16
200m Men	19.05	20.223	20.53
100m Women	10.33	10.996	11.25
200m Women	21.02	22.364	22.91

Table 1. The unknown values used to calculate specific event probabilities.

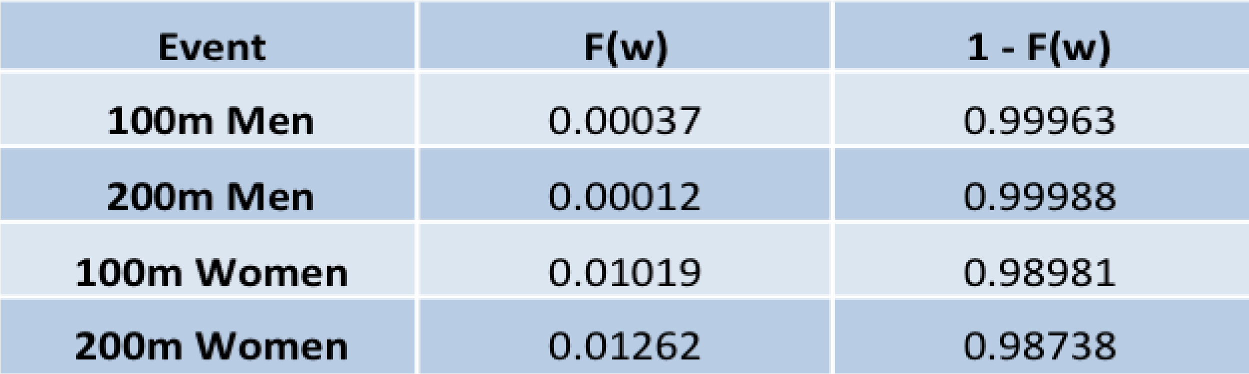
5. Individual Event Probabilities

To compute the probabilities for each individual event, one must first substitute the obtained values for y_0, y_1, and $\bar{x}$ into (2) and solve, to estimate for $\hat{\alpha}$. Subsequently, y_0, y_1, and $\hat{\alpha}$ are substituted into (1). By replacing y with the current world record w, of the event corresponding to the values of y_0, y_1, and $\bar{x}$ used, the general formula is

$$F(w) = \left(\frac{w - w_0}{w_1 - w_0}\right)^{\hat{\alpha}} \tag{3}$$

where $w_0 = y_0$ and $w_1 = y_1$.

Solving (3) for each event yields the probability that an elite Olympic sprinter will break the world record of that specific event. The results are presented in Table 2.



Event	F(w)	1 - F(w)
100m Men	0.00037	0.99963
200m Men	0.00012	0.99988
100m Women	0.01019	0.98981
200m Women	0.01262	0.98738

Table 2. The probabilities of a new world record by an elite Olympic sprinter, $F(w)$, in each specific event, along with the probability of no new world record in that same event, $1 - F(w)$.

6. Sensitivity Analysis

Although the results of section 4 provide reasonable estimates for y_0 and y_1 because the conclusions by [1] are considered to be the fastest possible times for each race, an analysis is in order to test how different bound values affect the results.

The values to be chosen for y_0 and y_1 are those which best emulate real-world results of the probability that a world record will be broken. These results can be computed through [4] by determining the number of times a world record was broken, and dividing this total by the cumulative number of data points listed in the data set. This analysis ensures that the bounds of each model properly match current results, and can be considered accurate in predicting the likelihood of future world records.

As mentioned in section 3, the mean time for each data set from [4] does not change without additional race times; thus, changes of $\hat{\alpha}$ are dependent upon y_0 and y_1 only. One must then consider each possible scenario in which the values of y_0 and y_1, either increase, decrease, or remain the same, resulting in 9 possible combinations of changes. Taking the initial values for y_0 and y_1 as the reference case, there are 8 general cases to compare against the reference for each race. The process for each of these comparisons occurs is as follows:

a.: Choose an $\epsilon > 0$ that will be used as the difference margin for each comparison (ie. $\epsilon = 0.05$).

b.: Determine $w_0 = y_0 \pm \epsilon$ and $w_1 = y_1 \pm \epsilon$, for each sprinting event. This yields the values necessary to compute each general case: y_0 stays the same, decreases, or increases, and separately y_1 stays the same, decreases, or increases.

c.: Determine the α_0 corresponding to each general case, and note whether it is greater or less than the base case $\hat{\alpha}$.

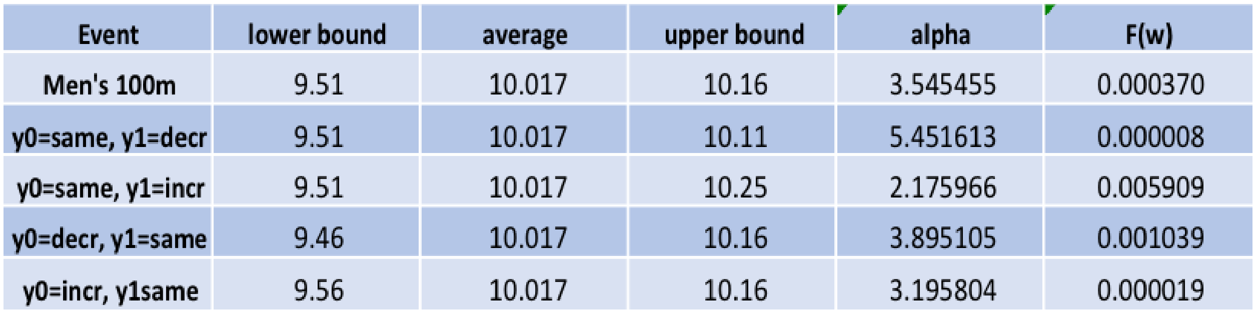
d.: Compute (3) using α_0 and the corresponding w_0 and w_1. Identify how this new α_0 affects the probability of a world record being achieved.

As observed from (3), $\left(\frac{w-w_0}{w_1-w_0}\right) < 1$, for all $w_0 < w < w_1$ and w_0, w_1 positive real numbers, resulting in the expectation that increasing $\hat{\alpha}$ would yield a smaller $F(w)$ and decreasing $\hat{\alpha}$ would yield a greater $F(w)$. However, this is not always true because the values of y_0 and y_1 are changing along with $\hat{\alpha}$. This is shown empirically in Table 3 and graphically in Figures 1 and 2 through a subset of comparisons computed for the Men's 100m Dash.

Following the computation of all 8 general cases for each sprinting event, it was determined that values of $\hat{\alpha}$, both above and below the baseline $\hat{\alpha}$, produce world record probabilities greater and less than the baseline value. This holds true for each event and thus, it cannot be concluded that specific values of $\hat{\alpha}$ alone, yield greater or lower world record probabilities.

Additionally, setting $\epsilon = 0.05$ and $\epsilon = 0.10$ for the 100m and 200m events respectively, resulted in a range of probabilities from $1.58 \times 10^{-6}\%$ and 2.5%, with much of the lower values originating from the men's events and the higher values from the women's.

However, in each of the events, there resulted a general case that matched the world record probability of the real-world data, as shown in Table 4. This allowed for a model comparison between the original model from section 4 (denoted Model 1) and the near perfect model (denoted Model 2), in which the world record probabilities from each model are evaluated empirically against the real-world results. The model that produced the probabilities closest to that of the past data was chosen to be the best predictive model.



Event	lower bound	average	upper bound	alpha	F(w)
Men's 100m	9.51	10.017	10.16	3.545455	0.000370
y0=same, y1=decr	9.51	10.017	10.11	5.451613	0.000008
y0=same, y1=incr	9.51	10.017	10.25	2.175966	0.005909
y0=decr, y1=same	9.46	10.017	10.16	3.895105	0.001039
y0=incr, y1same	9.56	10.017	10.16	3.195804	0.000019

Table 3. The baseline case and 4 of the 8 general cases are shown for the Men's 100m sprint. Changes made to the bounds are specified in the first column where ϵ is an arbitrary value that is added or subtracted from the bounds ($\epsilon = 0.05$).

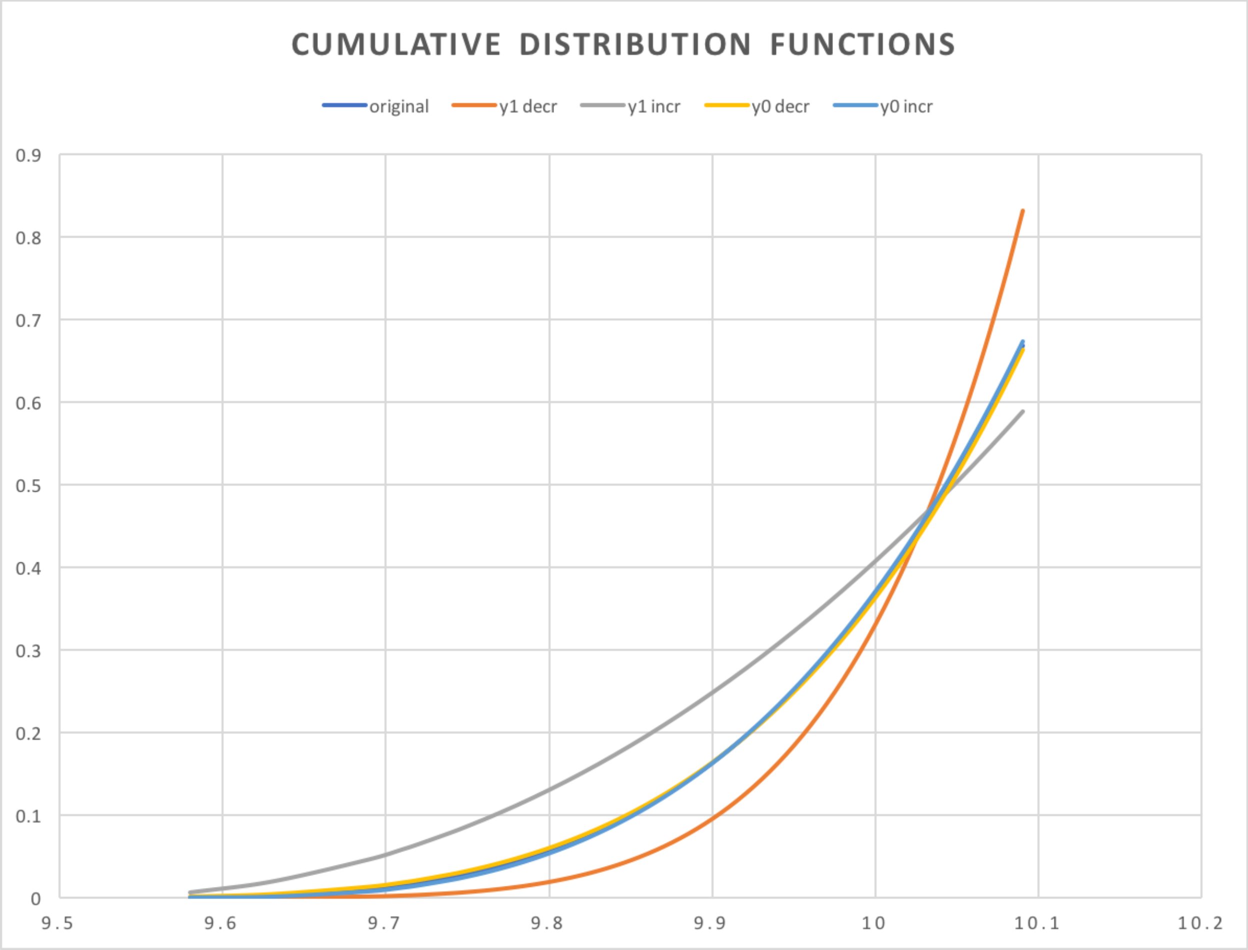


Figure 1. The effects of the sensitivity analysis on the cumulative distribution function. The cases described in Table 3 are shown. As the dataset does not contain values near the upper bound, the distributions presented will progress towards a probability of 1 as the race times increase, reaching this value in correspondence with the data in Table 3.

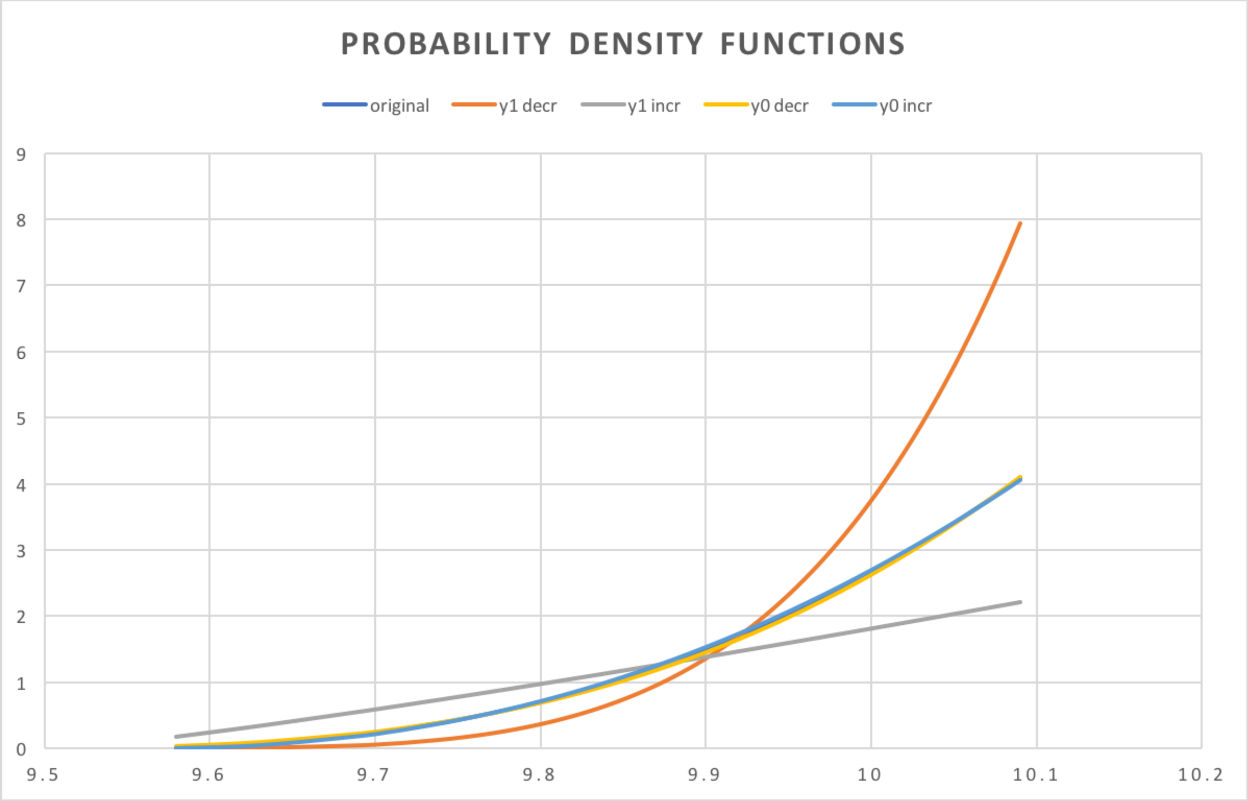
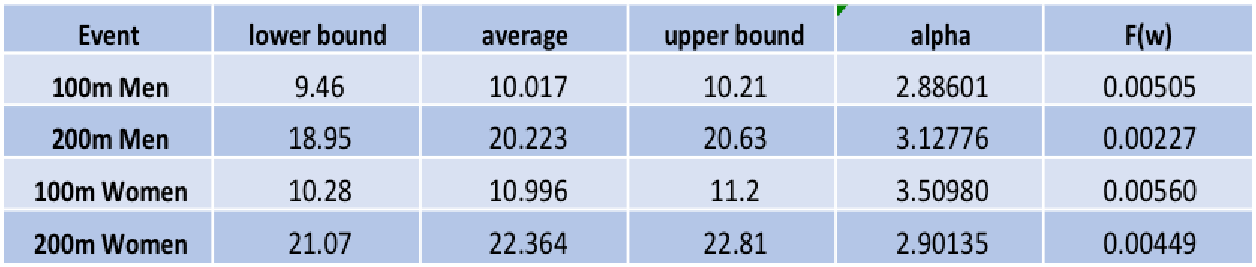


Figure 2. The effects of the sensitivity analysis on the probability density function. Cases described in Table 3 are shown.



Event	lower bound	average	upper bound	alpha	F(w)
100m Men	9.46	10.017	10.21	2.88601	0.00505
200m Men	18.95	20.223	20.63	3.12776	0.00227
100m Women	10.28	10.996	11.2	3.50980	0.00560
200m Women	21.07	22.364	22.81	2.90135	0.00449

Table 4. The values for Model 2 yield near identical world record probabilities when compared with the real world data from [4].

7. Cumulative Probabilities

Using each event's world record probability, it is now possible to determine the probability that a world record will be broken in zero, one, or any combination of the events. These probabilities are obtained using the Law of Total Probability and the binomial coefficient, $\binom{n}{Y}$, where n represents the number of events (ie. $n = 4$), and Y represents the number of world records observed, for $Y = 0, 1, 2, 3, 4$. This accounts for each scenario in which Y world records are broken.

7.1. **Probability Formulae.** Determining the formulae for the cumulative probabilities allowed for a model comparison of the final results for Model 1 and Model 2 against the real-world data.

Denote $F(w)_i = P_i$ where each $F(w)_i$ corresponds to an event for $i = 1, 2, 3, 4$ (where $1 =$ 100m Men, $2 =$ 200m Men, $3 =$ 100m Women, $4 =$ 200m Women).

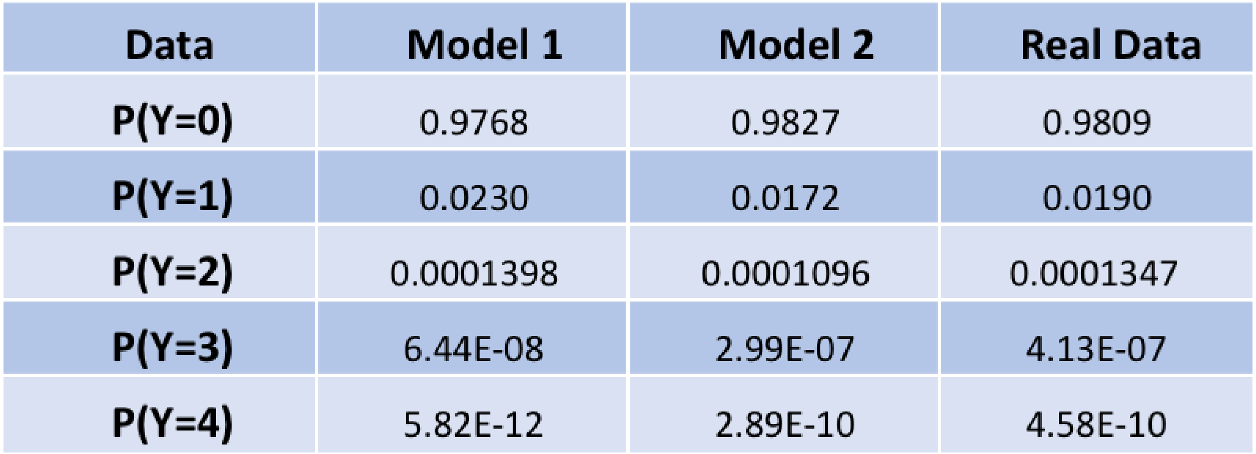
7.2. **Results.** Using the formulae from Subsection 7.1, the cumulative probabilities are obtained for both models and the real-world data from [4]. See Table 4.

Thus, one can conclude that Model 2 represents a reasonably accurate prediction of future world record probabilities, as it is within 0.5% of the real world data for the probability of no world record being broken and is subsequently closer for each of the remaining probabilities.

7.3. **Meaning Behind the Numbers.** These probabilities indicate the challenge, even for the elite Olympic sprinters, in achieving a world record. As the calculations show, there is approximately a 1.9±0.4% chance that one may observe a single world record in any of the four events, and an incredibly small 0.01% chance of witnessing two of the four world records broken. The probabilities for three and four world records being broken simultaneously are virtually negligible. Perhaps the most glaring result is the probability that one will not witness a single new world record in these events is approximately 98±0.5%.

World Records	Equation	Conditions
0	$P(Y=0)=\prod_{s=1}^{4}(1-P_s)$	where $s \in \mathbb{Z}$
1	$P(Y=1)=\sum_{i=1}^{4}P_i\prod_{s=1}^{4}(1-P_s)$	where $i,s \in \mathbb{Z}$ and $i \neq s$
2	$P(Y=2)=\sum_{i=1}^{4}P_iP_j\prod_{s=1}^{4}(1-P_s)$	where $i,j,s \in \mathbb{Z}$ and $i \neq j \neq s$
3	$P(Y=3)=\sum_{i=1}^{4}P_iP_jP_k\prod_{s=1}^{4}(1-P_s)$	where $i,j,k,s \in \mathbb{Z}$ and $i \neq j \neq k \neq s$
4	$P(Y=4)=\prod_{s=1}^{4}P_s$	where $s \in \mathbb{Z}$

Table 5. The equations used to determine the probability that a world record will be observed Y times.



Data	Model 1	Model 2	Real Data
P(Y=0)	0.9768	0.9827	0.9809
P(Y=1)	0.0230	0.0172	0.0190
P(Y=2)	0.0001398	0.0001096	0.0001347
P(Y=3)	6.44E-08	2.99E-07	4.13E-07
P(Y=4)	5.82E-12	2.89E-10	4.58E-10

Table 6. Cumulative probabilities for Model 1, Model 2, and the real-world data.

8. Conclusion

Many studies and research projects have been conducted in an effort to determine the fastest possible race time, or the probability of breaking a world record, especially for the 100-meter dash. Even with different approaches being applied to these tasks, the conclusions often identify that breaking a world record in a sprinting event is a tremendously difficult feat. Indeed, some world records can remain unbroken for decades. Although it is nearly a universal desire to witness an athlete run faster than any human prior to his or her time, it rarely happens. This rarity is what makes a new world record so special, for if its occurrence was routine, these events would have reduced excitement and anticipation. Thus, during the next Summer Olympics, if one or more new world records are observed, appreciate the truly rare moment. However, in the most likely, yet disappointing circumstance that no new world records are set, it is what the world should have been expecting.

9. Nota Bene

The data from [4] used in this paper preceded the 2016 Rio Summer Olympic Games. It is of interest to note that there were no new world records broken in either of the Men's or Women's 100 and 200 meter races at the 2016 Summer Games.

This model does not take into consideration that athletic achievement is improving over time. The model gives equal weight to race times achieved from the past as it does current race times, thus resulting in potentially lower world record probabilities. A future study or improvement to this modelling method would be to systematically account for the aging effect of past race times.

10. Acknowledgements

Many thanks go to the help from Dr. Andrew P. Dean for supervising and guiding this research project, as well as Dr. Deli Li for helping with the major statistical areas of this paper. This paper was inspired from the work of [5].

References

[1] Einmahl, John, H.J., and Smeets, Sander, G.W.R., Ultimate 100-m world records through extreme-value theory, *Statistica Neerlandica* **65** (2011).

[2] B.R., Ranking sports' popularity 'And the silver goes to...', *The Economist* (2011).

[3] Wood, Robert, The Most Popular Summer Olympic Sport, *Topend Sports* (2015).

[4] Larsson, Peter, Track & Field All-time Performances, *http://www.alltime-athletics.com* (2016).

[5] Noubary, Reza, What is the Speed Limit for Men's 100 Meter Dash, *Mathematics and Sports* **43** (2010).

Student biography

Patrick Gravelle: (patrick.gravelle@queensu.ca) Patrick is currently a senior at Queen?s University in Kingston, Ontario, majoring in statistics. Beginning in fall 2018, he will commence his studies in a Master's of Science Biostatistics program.