The "Fair" Triathlon: Equating Standard Deviations Using Bayesian Nonlinear Models

Gilbert W. Fellingham
gwf@byu.edu

C. Shane Reese
reese@stat.byu.edu

See next page for additional authors

Follow this and additional works at: https://scholarsarchive.byu.edu/facpub

Part of the Statistics and Probability Commons

Original Publication Citation

BYU ScholarsArchive Citation
https://scholarsarchive.byu.edu/facpub/327
The “Fair” Triathlon: Equating Standard Deviations Using Bayesian Nonlinear Models

S. McKay Curtis∗  Gilbert W. Fellingham†
C. Shane Reese‡

∗North Carolina State University, s.mckay.curtis@gmail.com
†Brigham Young University, gwf@byu.edu
‡Brigham Young University, reese@stat.byu.edu

Copyright ©2006 The Berkeley Electronic Press. All rights reserved.
The “Fair” Triathlon: Equating Standard Deviations Using Bayesian Nonlinear Models

S. McKay Curtis, Gilbert W. Fellingham, and C. Shane Reese

Abstract

The Ironman triathlon was created in 1978 by combining events with the longest distances for races then contested in Hawaii in swimming, cycling, and running. The Half Ironman triathlon was formed using half the distances of each of the events in the Ironman. The Olympic distance triathlon was created by combining events with the longest distances for races sanctioned by the major federations for swimming, cycling, and running. The relative importance of each event in overall race outcome was not given consideration when determining the distances of each of the races in modern triathlons. Thus, there is a general belief among triathletes that the swimming portion of the standard-distance triathlons is underweighted. We present a nonlinear Bayesian model for triathlon finishing times that models time and standard deviation of time as a function of distance. We use this model to create “fair” triathlons by equating the standard deviations of the times taken to complete the swimming, cycling, and running events. Thus, in these “fair” triathlons, a one standard deviation improvement in any event has an equivalent impact on overall race time. We conclude that a ratio of roughly 1:4:17 for swim distance to run distance to bike distance generates appropriate distances for a “fair” triathlon. So, for example, the Olympic triathlon swim distance should be increased from 1.5 km to 2.5 km to more fairly value each discipline in the race.

KEYWORDS: Three-parameter lognormal distribution, flexible tail thickness, growth models
1. Introduction

According to popular myth, the first Ironman triathlon was conceived in a bar in Hawaii. Chatting over a few beers, Naval Commander John Collins debated with several others the following question: “Who are the toughest athletes—swimmers, cyclists, or runners.” Collins suggested that to answer the question athletes from each discipline should compete in an event that consists of all three disciplines run back-to-back. Collins proposed a race comprised of the longest events in swimming, cycling, and running then contested in Hawaii, USA—the 2.4 mile (3.9 km) Waikiki Rough Water Swim, the 112 mile (180 km) Around Oahu Bike Ride, and the 26.2 mile (42 km) Honolulu Marathon. (For simplicity, all distances will be written in metric from now on.) So, in 1978, Collins and a little over a dozen others participated in the first Ironman triathlon. (For Collins’ comments and clarifications on the genesis of the Ironman triathlon see http://vnews.ironmanlive.com/vnews//1043361628/?keywords=John:collins)

The popularity of the Ironman triathlon increased dramatically after the second Hawaii Ironman when Sports Illustrated ran a story on the 1979 winner Tom Warren and the American Broadcasting Company (ABC) began coverage of the Hawaii triathlon in 1980. Soon, other shorter-distance triathlons emerged. The Half-Ironman triathlon was formed by taking half the distances of each event in the Ironman triathlon (1.9 km swim, 90 km bike, and a 21 km run). The International Triathlon Union (ITU) formed the Olympic triathlon by combining the longest event distances used in races sanctioned by each of the major federations for swimming (the International Federation for Swimming), cycling (International Cycling Union), and track (International Association of Athletics Federations). (See http://www.triathlon.org/tv/FAQs.htm) The Olympic Triathlon consists of a 1.5 km swim, a 40 km bike, and a 10 km run.

Today, the three major triathlon types are the Olympic, Half-Ironman, and Ironman triathlons. However, triathletes can compete in a plethora of other triathlons of different distances. Sprint triathlons are composed of events of shorter distances than the Olympic triathlon. Triathletes also compete in Double and Triple Ironman triathlons, which—as their names imply—are twice and three times the distance of the Ironman triathlon.

As the history of the modern triathlon indicates, little consideration was given to the relative difficulty of each event when forming the various triathlons. In fact, empirical evidence indicates that the swimming portion of the modern triathlon is extremely underweighted. Dengel et al. (1989),

Published by The Berkeley Electronic Press, 2006
in a study of eleven male triathletes, report that, “swimming time was not significantly related to overall triathlon time.” Currently, the ratio of swim:run:bike distances is 1:11:48 in the Ironman (and races that are multiples of the Ironman) and 1:7:27 in the Olympic Triathlon.

Wainer and DeVeaux (1994) make a persuasive argument that in a fair triathlon each segment should take roughly the same time. Using world records to estimate times, they come up with an optimal ratio of 1:3.5:8 for the swim:run:bike distances.

We define the fair triathlon as one in which a one standard deviation improvement in time in any of the individual triathlon components results in the same improvement in overall triathlon time. In this paper we present a Bayesian model for finishing times in each of the components of the modern triathlon. The model is fit three different times - one for swim times, one for run times, and one for bicycling times. The three models are then used to find the distances of each event that equate standard deviations. Segment distances that result in distributions with equal standard deviations are used to construct triathlons of different distances, and a general ratio of event distances is proposed to be used in the construction of fair triathlons.

Bayesian methods provide an ideal framework for creating a fair triathlon. As mentioned previously, the second moment is of primary interest in this study. The Bayesian formulation allows us to estimate the entire posterior distribution of finish times. Using this formulation we can ”match” posterior distributions and equate second moments so that the variability is comparable across the three different segments of a triathlon.

2. Data Description
The data used are a collection of finishing times from 55 different triathlons found on the world wide web. We consider triathlons from 1996 to 2003. None of the triathlons included in the data set allowed drafting in the bicycling portion. Table 1 contains information on the length of each triathlon, the number of triathlons, and the number triathletes for each triathlon type in the data set. By far, the most numerous triathlons in the data set are the three standard distance triathlons—Olympic, Half Ironman, and Ironman. However, the shorter-distance sprint triathlons and the longer-distance double and triple Ironman triathlons allow prediction of standard deviations at shorter or longer distances when constructing the fair triathlon.
Table 1

Types of triathlons in the dataset and their respective distances in kilometers for the swim, bike, and run. Triathletes is the number of triathletes in the data set who recorded a finishing time in that type of triathlon.

<table>
<thead>
<tr>
<th>Type</th>
<th>Swim (km)</th>
<th>Bike (km)</th>
<th>Run (km)</th>
<th>Number of Triathlons</th>
<th>Triathletes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sprint 1</td>
<td>0.40</td>
<td>16.1</td>
<td>4.8</td>
<td>1</td>
<td>269</td>
</tr>
<tr>
<td>Sprint 2</td>
<td>0.50</td>
<td>19.0</td>
<td>5.0</td>
<td>2</td>
<td>1,246</td>
</tr>
<tr>
<td>Sprint 3</td>
<td>0.80</td>
<td>32.2</td>
<td>10.0</td>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>Sprint 4</td>
<td>0.75</td>
<td>20.0</td>
<td>5.0</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>Sprint 5</td>
<td>0.40</td>
<td>15.0</td>
<td>5.0</td>
<td>1</td>
<td>116</td>
</tr>
<tr>
<td>Olympic</td>
<td>1.50</td>
<td>40.0</td>
<td>10.0</td>
<td>13</td>
<td>5,686</td>
</tr>
<tr>
<td>Half Ironman</td>
<td>1.93</td>
<td>90.1</td>
<td>21.1</td>
<td>10</td>
<td>5,168</td>
</tr>
<tr>
<td>Ironman</td>
<td>3.90</td>
<td>180.2</td>
<td>42.2</td>
<td>15</td>
<td>13,427</td>
</tr>
<tr>
<td>Double Ironman</td>
<td>7.72</td>
<td>360.5</td>
<td>84.3</td>
<td>5</td>
<td>53</td>
</tr>
<tr>
<td>Triple Ironman</td>
<td>11.6</td>
<td>540.7</td>
<td>126.5</td>
<td>5</td>
<td>30</td>
</tr>
</tbody>
</table>

3. The Model

This section describes the model used to fit the triathlon finishing times. Each event in a triathlon is modeled independently; therefore, the model described in this section is used to fit the data three different times—one to the swim times, once to the bike times, and once to the run times.

3.1 Sampling Density Model

The sampling density model is the population density from which the observations in the data are drawn. Thus, the sampling density should “match” the data in terms of its general shape. The shape of the distribution is discussed in section 3.1.1, and the parameters of the distribution—which must be modeled according to distance—are discussed in sections 3.1.2 and 3.1.3.

3.1.1 General Shape

The shape of the sampling density should approximately match the shape of the data. Figure 1 is a plot of finishing times for swimming, cycling, and running from three different triathlon types—the Olympic, Half Ironman, and Ironman. As Figure 1 shows, finishing times in all three events and across triathlons of different lengths are clearly non-normal. In all cases, the data are right skewed and, in some cases, have
Figure 1. Histograms of the finishing times for each of the three events in the Olympic, Half Ironman, and Ironman triathlons in the data set. Each histogram is fit with a density smoother to highlight the shape of the data.
fairly thick tails. Thus, the sampling density for the model must be skewed and have flexibility in its tail thickness.

The three-parameter lognormal distribution satisfies both concerns. The lognormal distribution is right skewed, and the three-parameter lognormal distribution has an extra parameter $\alpha$ (sometimes called the “peakedness parameter,” see Gajewski et al., 2004) that allows for flexibility in the specification of the tail.

Figure 2 is a plot of three three-parameter lognormal distributions. Each distribution in the figure has $E[X] = 40$; but each distribution has a different value of the peakedness parameter $\alpha$. The three-parameter lognormal distribution with peakedness of zero is simply the two-parameter lognormal distribution. The figure shows that, as the peakedness parameter increases, more of the density shifts into the tails of the distribution, and vice-versa.

The density of the three-parameter lognormal is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma^2(x+\alpha)}e^{-\frac{1}{2\sigma^2}[(\ln(x+\alpha)-\mu)^2]},$$

where parameter $\alpha$ in (1) is the peakedness parameter, $\mu$ is the mean of the distribution of $\ln(X + \alpha)$, and $\sigma^2$ is the variance of the distribution of $\ln(X + \alpha)$. The expected value and the variance of the three-parameter lognormal distribution are

$$E(X) = e^{\mu+\sigma^2/2} - \alpha$$
$$V(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$$

Our proposal uses the square root of equation (3) to construct the fair triathlon. With the model for $\mu$ discussed in Section 3.1.2, equation (3) becomes a function of distance and can then be used to find distances in each event that give the same standard deviation.

3.1.2 Central Moment Formulation If $X$ follows a lognormal distribution, then the parameter $\mu$ in the lognormal density is the mean of $\ln(X)$. If $X$ follows a three-parameter lognormal distribution, then the parameter $\mu$ is the mean of $\ln(X + \alpha)$, where $\alpha$ is the peakedness parameter. Thus, any model for $\mu$ should closely fit the logarithm of the data. Figure 3 shows plots of the log of finishing times for the three triathlon events versus distance. The plots show a definite nonlinear trend in the means for swimming, cycling, and running. Thus, any model for $\mu$ should reflect the
Figure 2. The three parameter lognormal distribution with three different values for the peakedness parameter $\alpha$ and $E[X] = 40$.

Figure 3. Plots of the logarithm of event times versus distance. Scales on the x and y axes are different for each plot.
nonlinear relationship between the log of finishing time and distance.

In order to formulate a nonlinear model for these data we can consider the (log of) finishing times as the "growth" variable and distance as the time variable. Schnute (1981) presents a general class of growth models. A special case of these growth models is used for modeling $\mu$. This model is

$$\mu = \tau_1 + (\tau_2 - \tau_1) \frac{1 - e^{-\beta(d-\delta_1)}}{1 - e^{-\beta(\delta_2-\delta_1)}}, \quad (4)$$

where $\tau_1$ is the log of the sum of the peakedness parameter $\alpha$ and the time it takes to complete an event (either swim, cycle, or run) of distance $\delta_1$. The log of the sum of the peakedness parameter $\alpha$ and the time it takes to complete an event of distance $\delta_2$, is $\tau_2$. The growth rate of finishing times according to distance is $\beta$, and $d$ is the actual distance of the event.

The model in equation (4) is particularly nice for Bayesian modeling because of the interpretability of the parameters $\tau_1$ and $\tau_2$. The parameters $\tau_1$ and $\tau_2$ are the log of the sum of $\alpha$ and finishing times for distances of $\delta_1$ and $\delta_2$. This interpretability makes specifying prior distributions more manageable.

The value of $\beta$ determines the shape of the growth curve. Positive values of $\beta$ result in growth curves that are concave down. Negative values of $\beta$ result in growth curves that are concave up. Therefore, beliefs about the shape of the growth in finishing times can be incorporated into the prior distributions for $\beta$ as shown in Figure 4.

Although not as easily interpretable as $\tau_1$ and $\tau_2$, the parameter $\beta$ has some nice properties that are useful in prior elicitation. Larger values of $\beta$ give longer finishing times for shorter distances than smaller values of $\beta$. So, for example, $\beta$ for swim times as a function of distance should be much larger than $\beta$ for cycling times.

### 3.1.3 Higher Order Moment Formulations

If $X$ follows a three-parameter lognormal distribution, the parameter $\sigma^2$ is the variance of $\ln(X+\alpha)$. Thus, any model for $\sigma^2$ must approximate the spread of the log of the data. In Figure 3, the spread of the log finishing times at each distance is approximately equal. Therefore, $\sigma^2$ is assumed to be the same for all distances within an event (swimming, cycling, or running).

### 3.1.4 The Likelihood

The likelihood for the data is now determined by equations (1) and (4), where $\mu$ in (1) is written as a function of the distance $d$ of the event according to the growth curve $\tau_1 + (\tau_2 - \tau_1) \frac{1 - e^{-\beta(d-\delta_1)}}{1 - e^{-\beta(\delta_2-\delta_1)}}.$
Figure 4. The value of $\beta$ determines the shape of the growth curve. Positive values of $\beta$ result in growth curves that are concave down. Negative values of $\beta$ result in growth curves that are concave up. (In both plots, $\tau_1 = 4.73$, $\delta_1 = 24.8$, $\delta_2 = 112$. For the solid line $\tau_2 = 6.08$ and $\beta = 0.008$. For the dashed line $\tau_2 = 5.4$ and $\beta = -0.008$.)
The quantities $\delta_1$ and $\delta_2$ are not parameters but are constants that are chosen to be the distances that correspond respectively to parameters $\tau_1$ and $\tau_2$. For each triathlon segment, (i.e. swim, run, bike), $\delta_1$ and $\delta_2$ are the distances corresponding to an Olympic triathlon and an Ironman triathlon, respectively. For example, in the swim component, $\delta_1 = 1.5$ km, and $\delta_2 = 3.9$ km.

3.2 Prior Distribution Model

The likelihood for one segment contains five parameters, thus prior distributions must be specified for $\tau_1$, $\tau_2$, $\beta$, $\alpha$, and $\sigma^2$ for the three different events in a triathlon. The choices for the distributional form and parameter values of each prior distribution are explained in this section.

The parameters $\tau_1$ and $\tau_2$ are defined as the natural log of the sum of $\alpha$ and finishing times for particular events of distances $\delta_1$ and $\delta_2$. Theoretically, $\tau_1$ and $\tau_2$ can be negative for certain values of finishing times and $\alpha$ (i.e. if the finishing time and $\alpha$ sum to a number between zero and one). However, in each of the triathlon components, the values of $\delta_1$ and $\delta_2$ correspond to the Olympic triathlon distances and the Ironman triathlon distances, respectively. Finishing times at these distances always exceed one minute. Therefore, the quantities $\tau_1$ and $\tau_2$ are always positive quantities. The gamma distribution is defined for positive real numbers, and the gamma distribution is also flexible in its shape—it can be symmetric or skewed. Therefore, we use gamma prior distributions for $\tau_1$ and $\tau_2$ in each event.

The peakedness parameter $\alpha$ and $\sigma^2$ in the three-parameter lognormal distribution are both defined as positive quantities. Again, the gamma distribution is defined for positive real numbers and offers flexibility in its shape. Thus, as with $\tau_1$ and $\tau_2$, we use gamma prior distributions for $\alpha$ and $\sigma^2$.

The parameter $\beta$ represents the “growth rate” of finishing times. Theoretically, $\beta$ can be either positive or negative, which suggests using a prior distribution defined on the entire real line. Also, a priori, there is no reason to believe that $\beta$ is more likely to take on larger values than smaller values (or smaller values than larger values), which suggests using a symmetric prior distribution. Because of these considerations, a normal prior distribution is used for $\beta$. 
Table 2
Table of parameter values for prior distributions for all three portions of a triathlon.

<table>
<thead>
<tr>
<th>Prior Parameter</th>
<th>Swim</th>
<th>Cycle</th>
<th>Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{\tau_1}$</td>
<td>1156.0</td>
<td>529.0</td>
<td>430.0</td>
</tr>
<tr>
<td>$b_{\tau_1}$</td>
<td>340.0</td>
<td>115.0</td>
<td>104.0</td>
</tr>
<tr>
<td>$a_{\tau_2}$</td>
<td>900.0</td>
<td>1600.0</td>
<td>529.0</td>
</tr>
<tr>
<td>$b_{\tau_2}$</td>
<td>200.0</td>
<td>267.0</td>
<td>92.0</td>
</tr>
<tr>
<td>$a_\alpha$</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>$b_\alpha$</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>$a_\sigma$</td>
<td>9.0</td>
<td>1.7</td>
<td>4.0</td>
</tr>
<tr>
<td>$b_\sigma$</td>
<td>300.0</td>
<td>130.0</td>
<td>200.0</td>
</tr>
<tr>
<td>$m_\beta$</td>
<td>0.6</td>
<td>0.1</td>
<td>0.05</td>
</tr>
<tr>
<td>$s^2_\beta$</td>
<td>0.04</td>
<td>0.25</td>
<td>$2.25 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

The prior distributions used for the parameters are therefore

\[
\begin{align*}
\pi(\tau_1) & \propto \tau_1^{a_{\tau_1}-1} e^{-\tau_1 b_{\tau_1}} \quad \text{(gamma)}, \\
\pi(\tau_2) & \propto \tau_2^{a_{\tau_2}-1} e^{-\tau_2 b_{\tau_2}} \quad \text{(gamma)}, \\
\pi(\beta)  & \propto \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(\beta - m)^2}{2s^2}} \quad \text{(normal)}, \\
\pi(\sigma^2) & \propto (\sigma^2)^{a_\sigma-1} e^{-\sigma^2 b_\sigma} \quad \text{(gamma)}, \quad \text{and} \\
\pi(\alpha)  & \propto \alpha^{a_\alpha-1} e^{-\alpha b_\alpha} \quad \text{(gamma)},
\end{align*}
\]

where the values for the prior parameters are listed in Table 2.

The values in Table 2 were chosen using moment matching. In other words, values for the mean and standard deviation of the prior distributions were chosen first. Then the mean and standard deviation were used to solve for the actual parameters of the prior distributions.

Values for the prior distribution of $\tau_1$ were chosen based on beliefs about an average triathlete. Because $\tau_1$ is the log of the sum of $\alpha$ and the time it takes to finish a race of distance $\delta_1$, we chose our prior for $\tau_1$ based on the log of the time we believed it would take an average triathlete to swim, bike, or run a distance of $\delta_1$, where values for $\delta_1$ for each component are the Olympic triathlon distances—1.5 km swim, 40.0 km bike, and a 10.0 km run. By taking the log of various plausible values for finishing times, we obtained an approximate mean and variance for the prior distributions of $\tau_1$ in each of the triathlon components. For example, based on our prior knowledge, we felt an average triathlete might take as little as 20 minutes or as long as 40
Table 3

Table of means and standard deviations for prior distributions of each parameter in the model. Distances $\delta_1$ and $\delta_2$ (in miles) are also listed for the parameters $\tau_1$ and $\tau_2$.

<table>
<thead>
<tr>
<th>Prior Parameter</th>
<th>Swim</th>
<th>Bike</th>
<th>Run</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\delta_1$= 1.5</td>
<td>$\delta_1$= 40.0</td>
<td>$\delta_1$= 10.0</td>
</tr>
<tr>
<td>$\tau_1$</td>
<td>3.4 0.1</td>
<td>4.6 0.2</td>
<td>4.15 0.2</td>
</tr>
<tr>
<td></td>
<td>$\delta_2$= 3.9</td>
<td>$\delta_2$= 180.2</td>
<td>$\delta_2$= 42.2</td>
</tr>
<tr>
<td>$\tau_2$</td>
<td>4.5 0.15</td>
<td>4.6 0.2</td>
<td>5.75 0.25</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>20.0 10.0</td>
<td>20.0 10.0</td>
<td>20.0 10.0</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.03 0.01</td>
<td>0.013 0.01</td>
<td>0.02 0.01</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.60 0.20</td>
<td>0.02 0.01</td>
<td>0.03 0.015</td>
</tr>
</tbody>
</table>

minutes to swim 1500 meters in an Olympic triathlon. Taking the log of 20 and 40 minutes suggests a prior distribution for $\tau_1$ with a mean of 3.4 and a standard deviation of 0.01. The mean and the variance for the distributions of $\tau_1$ in the cycling and running events were obtained in a similar fashion, and parameters for the gamma prior distributions were solved using these means and standard deviations. Values for the parameters in the prior distributions for $\tau_2$ were chosen similarly.

The parameter $\sigma^2$ is the variance of the distribution of $\ln(X+\alpha)$, which can be thought of more simply as the variance of $\ln(X)$. The parameter values for the prior distributions of $\sigma^2$ are based on beliefs about the maximum and minimum time it would take a triathlete to swim, bike, or run a specified distance. The distances used for this were 1.6 km for the swim, 40.2 km for the bike, and 9.7 km for the run (which are the approximate distances of an Olympic triathlon). By using the range of the log of the proposed maximum finishing times and the log of the proposed minimum finishing times, we were able to obtain an estimates of the variance of $\ln(X)$ in each triathlon component by dividing the range by six and squaring the result. For example, in the swim event, an extremely fast swimmer might swim 1.6 km in 15 minutes, and an extremely slow swimmer might take as long as 55 minutes. Or, alternatively, a fast swimmer might take 25 minutes to swim 1.6 km, and a slow swimmer might take 45 minutes to swim 1.6 km. Based on the difference between the log of 15 and the log of 55 divided by 6 and the difference between the log of 25 and the log of 45 divided by six, the prior distribution for $\sigma^2$ in the swim event was centered.
Table 4
Possible values for $\beta$ and their corresponding finishing times for a swim of 1.6 km and 2.4 km. The means of the prior distributions for $\tau_1$ and $\tau_2$ were used in equation (4) along with the values of $\beta$ in the table to calculate the predicted finishing times.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>Finish Time for 1.6 km</th>
<th>Finish Time for 2.4 km</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>32</td>
<td>47</td>
<td>15</td>
</tr>
<tr>
<td>0.3</td>
<td>32</td>
<td>49</td>
<td>17</td>
</tr>
<tr>
<td>0.5</td>
<td>32</td>
<td>51</td>
<td>18</td>
</tr>
<tr>
<td>0.7</td>
<td>33</td>
<td>53</td>
<td>20</td>
</tr>
<tr>
<td>0.9</td>
<td>33</td>
<td>55</td>
<td>22</td>
</tr>
<tr>
<td>1.1</td>
<td>33</td>
<td>57</td>
<td>24</td>
</tr>
<tr>
<td>1.3</td>
<td>34</td>
<td>59</td>
<td>25</td>
</tr>
</tbody>
</table>

at 0.03 with a standard deviation of 0.01.

The parameter $\beta$ is the “growth” rate of the finishing times. To choose a prior distribution for $\beta$ in each of the three triathlon components, we calculated predicted finishing times for several different values of $\beta$ at two different distances. We looked at the change in the finishing time between the two distances to determine what values of $\beta$ gave reasonable changes in finishing times for the corresponding change in distance. For example, Table 4 contains predicted finishing times for a 1.6 km and a 2.4 km swim. These finishing times were computed using equation (4) with the means of the prior distributions for $\tau_1$ and $\tau_2$ and several values of $\beta$. For $\beta = 0.1$, the increase in time between a 1.6 km swim and a 2.4 km swim is 15 minutes. This implies that the average triathlete would finish the third 0.8 km of a 2.4 km swim at a pace slightly faster than the pace of the first 0.8 km. For $\beta = 1.3$, the increase in time between a 1.6 km swim and 2.4 km swim is 25 minutes. This implies that the average triathlete would finish the third 0.8 km of a 2.4 km race at a pace nearly twice as slow as the pace of the first 0.8 km. Both of the above scenarios are unlikely; thus, a normal prior distribution with a mean of 0.6 and a standard deviation of 0.2 was chosen for $\beta$ in the swim model. This prior distribution places most of the density over the more plausible intermediary values of $\beta$.

A priori, we knew very little about the parameter $\alpha$. However, we believed the parameter to be nonzero in each event. We chose a value for the mean of the prior distribution based on several plots of the three-parameter lognormal distribution (see Figure 2) and chose a large variance for each
distribution. A priori, we had no reason to believe that the peakedness parameter in each event should be different from the peakedness parameter in another event. Thus, the prior distribution for $\alpha$ in each event has a mean of 20 and a variance of 100.

### 4. Results

The joint posterior distribution for all parameters in the model is intractable. Therefore, an MCMC approach was used to draw samples from the posterior distribution. Specifically, Gibbs sampling (Geman and Geman, 1984) was used with nested Metropolis-Hastings algorithms (Metropolis et al., 1953 and Hastings, 1970) to generate draws from the posterior distributions of each parameter. We used the `gibbsit` software (Raferty and Lewis, 1996) to check for convergence. After thinning the chains for each parameter, the `gibbsit` output indicated that our chains had converged.

Table 5 contains a summary of the posterior distributions for each pa-
rameter in each model. As expected, the posterior mean for $\beta$ in the swim is greater than the posterior mean for $\beta$ in the run, and the posterior mean for $\beta$ in the run is greater than the posterior mean for $\beta$ in the bike. Also, the posterior distributions for the other parameters yielded sensible results according to their parameter interpretations. For instance, the posterior mean for $\tau_1$ in the swim component is 3.381. Exponentiating this result and subtracting the posterior mean for $\alpha$ gives a finishing time of 26 minutes. This finishing time is a reasonable finishing time for a swim distance of 1.5 km. Likewise, the posterior means for $\tau_1$ in the bike and the run yield estimated finishing times of 67 minutes for a bike distance of 40 km and 43 minutes for a run distance of 10 km.

Figure 5 is a plot that checks the fit of the model in each triathlon component. As mentioned in Section 3, when $X$ has a three-parameter lognormal distribution, $\ln(X+\alpha)$ is distributed normally with mean $\mu$ and variance $\sigma^2$. The model for triathlon finishing times models $\mu$ with equation (4). Figure 5, then, is a plot of the log of the sum of the posterior mean of $\alpha$ and the finishing times in each component. The line on the plot is equation (4) with the posterior means for each of the parameters—$\tau_1$, $\tau_2$, $\beta$. The figure shows that our model for the mean seems reasonable in all three triathlon events.

Figure 6 provides another graphical model check. As mentioned in Section 3.1.2, the distribution of $\ln(X+\alpha)$ is normal in a three-parameter lognormal distribution. Thus Figure 6 contains normal probability plots for the residuals in our model. Because $\ln(X+\alpha)$ follows a normal distribution when $X$ is distributed as a three-parameter lognormal random variable, the residuals are calculated as

$$\ln(X+\hat{\alpha}) - \hat{\mu},$$

(5)

where $\hat{\alpha}$ is the posterior mean for $\alpha$, $\hat{\mu} = \hat{\tau}_1 + (\hat{\tau}_2 - \hat{\tau}_1) \frac{1-e^{-\hat{\beta}(d-\delta_1)}}{1-e^{-\hat{\beta}(\delta_2-\delta_1)}}$, and $\hat{\tau}_1$, $\hat{\tau}_2$, and $\hat{\beta}$ are the posterior means for $\tau_1$, $\tau_2$, and $\beta$.

5. The Fair Triathlon

The standard deviation of the three-parameter lognormal distribution is

$$\sqrt{e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)}.$$  

(6)

Notice that equation (6) is a function of the parameter $\mu$. In the model for each triathlon component, the parameter $\mu$ is a function of distance. Thus,
Figure 5. Plots of the log of the sum of the posterior mean of $\alpha$ plus the finishing time for each triathlon component. The line on the plot is equation (4) with the posterior means of $\tau_1$, $\tau_2$, and $\beta$. $\otimes$ represents the mean of finishing times for each event.
Figure 6. Probability plots of the residuals for each event. The residuals were calculated as the difference between the log of the sum of finishing time and the posterior mean for alpha and the predicted value for $\mu$ at a given distance.
in the triathlon model, the standard deviation is a function of the distance. Substituting equation (4) for $\mu$ into equation (6) gives

$$\sqrt{e^{2 \tau_1 + \tau_2 - \tau_1} \frac{1 - e^{-\beta(d - \delta_1)}}{1 - e^{-\beta(d - \delta_2)}} + \sigma^2 (e^{\sigma^2} - 1)}.$$ (7)

Thus equation (7) can be used along with parameter estimates (posterior means) to find the distances in each triathlon component that give the same standard deviation.

Table 6 contains several fair triathlons of different lengths. Each triathlon uses the run triathlon component as the base distance. In other words, the triathlons were computed by using a run distance to solve for the standard deviation. Then, the distances for the swim and bicycle components were solved for that particular standard deviation. The run distances in Table 6 correspond to distances of 10, 15, 20, and 25 kilometers.

The first triathlon in Table 6 uses the same running distance as the current Olympic Triathlon and, therefore, can be thought of as the fair Olympic triathlon. The swim distance is approximately 60% longer and the bike distance is about 28% shorter in the fair version of the Olympic triathlon than in the current Olympic triathlon.

If a distance of 21 km (the current run distance for the Half Ironman triathlon) is used as the base run distance, then the swim and bike distances are 5.5 km and 104.2 km respectively. The fair distance for the swim in a Half Ironman is 183% longer and the bike distance is about 16% longer than the current distances for the swim and the bike in the Half Ironman.

If a distance of 42.2 km (the current run distance for an Ironman triathlon) is used as the base run distance, then the model calculates a
swim distance much longer than the longest swim distance in the data set. This distance is extremely unreliable because of the lack of data at the longer distances. However, Table 6 also lists the ratios of the bike distances to the swim distances and the run distances to swim distances for the triathlons in the table. These ratios suggest an approximate rule of thumb for constructing the fair triathlon. This rule of thumb is the ratio 1:4:17 of swim to run to bike. If the rule of thumb is used, then a fair Ironman triathlon is a 10.6 km swim, a 179.3 km bike, and a 42.2 km run. The swim distance is 175% longer than the swim distance in the current Ironman, while the bike distance is virtually the same.

The difficulty in calculating the fair Ironman underscores possible unreliability in constructing a fair triathlon with the current data. The only data on swimming distances longer than the 3.9 km swim in the Ironman come from Double and Triple Ironman triathlons. For obvious reasons, not many triathletes compete in Double and Triple Ironman triathlons, which makes obtaining large amounts of data at these distances impossible. And those triathletes who do participate in Double and Triple triathlons cannot be considered “typical” triathletes by any means. Thus predictions for fair triathlons that require swim distances longer than the current distances in the data set are unreliable, and predictions for swim distances longer than 3.9 km may be influenced by the types of triathletes that choose to compete in Double and Triple Ironman triathlons.

The self-selection bias in the Double and Triple Ironman triathlons may well extend to the shorter distances as well. As Wainer and DeVeaux (1994) point out, undervaluing the swimming portion of any triathlon would predispose better swimmers to not compete. They make the case that data from all triathlons must be flawed because of the bias against swimmers. For this reason, they based their model on world record performances. Their ratios of 1:3.5:8 value swimming even more than our ratios of 1:4:17, quite possibly because of the self-selection bias inherent in triathlon data.

However, despite the lack of data at longer swim distances, and the self-selection bias imposed by using data where better swimmers quite possibly do not compete, the data clearly show that the swimming portion of the triathlon is underweighted. Our model gives a longer distance for the swim portion of the triathlon, which will clearly improve the fairness of the triathlon regardless of the problems with the data.
6. Conclusion

In summary, the major triathlons of today are severely imbalanced in the relative importance they place on each event. Strong swimmers are at a disadvantage, not because of any lack of athleticism, but because of the improper relative event distances of the major triathlons. We have presented a model for the finishing times in each event of a triathlon, and we have used this model to create the fair triathlon. Based on the results of our model, we suggest a simple ratio for constructing fair triathlons—1:4:17 for swim to run to bike. Only by using fair triathlons can we ever hope to answer John Collins’ question, “Who are the better athletes?”

References


