Original Article

DOI: http://dx.doi.org/10.16926/par.2017.05.07

Statistical Description of Arrival Sequences in Amateur Long-Distance Races

Authors' Contribution:

A – Study Design
B – Data Collection
C – Statistical Analysis
D – Manuscript Preparation
E – Funds Collection

Beata Białek

Department of Physics, Inha University, Korea

Abstract

In this paper, we analyze data sets, collected from the public web sites, containing the finishers arrival times in five different half marathons. We focus on finding patterns in the arrival times distributions as well as on studying the differences between the female and male runners' arrival times distributions. The main statistical tools used are Analysis of Variance and Kolmogorov-Smirnov test for identifying underlying distributions and for comparing distributions of two samples. The obtained results indicate that the dynamics of groups of female runners is different than that of male runners. Also, a meaningful factor for the shape of the distribution is time limit set by the races organizers.

Keywords: *running, long-distance run, amateur races*

Address for correspondence:
Beata Bialek, Department of Physics, Inha University, Korea, email: b.bialek@yahoo.com

Recevied: 10.01.2017; Accepted: 20.01.2016; Published online: 5.04.2017

INTRODUCTION

Running is one of the most popular physical activities. Amateur runners challenge themselves taking part in long distance races organized all over the world. The growing popularity of running is reflected in the growing number of organized events and the increasing number of participants.

Long distance races include those that cover distances no less than 5 km. Probably the most popular among amateur athletes are 10 km races. Running longer distances, such as half marathons (21.0975 km) and marathons (42.195 km), demands more serious systematic training. Judging from the number of participants, long-distance running events are dynamic mass gatherings. As such, they have been analyzed as stochastic processes with attention being paid to congestion-diffusion effects [1], appearance of critical phenomena in the dynamic process [2], crowding being the effect of both competition and cooperation between the runners [3], or correlations in arrival sequences [4]. As far as the above research topics are concerned, the focus was on marathon races, especially those very popular, attended by tens of thousands of participants. It has been found that certain regularities exist in athletic races. For example, Garcia-Manso et al. showed that speed of athletes as a function of time in middle and long distance races (from 1500 m to marathon) can be modelled by power laws [5].

Inspired by the above-mentioned research, we asked ourselves the following questions: (1) is the distribution of athletes' arrival times to the finish line independent on the race distance? (2) is there any difference between arrival times distributions between races held in different continents? (3) is the distribution of arrival times of female athletes the same as that of male athletes?

Answering questions (1) may give more insight on the dynamics of long-distance races, which may be of value for the races organizers who can estimate the race time limits, the amount of water or snack supply on the routes, or the logistic aspects of the finish line surroundings. Finding a positive answer to question (2) may be interesting from sociological, leisure and recreation, and marketing point of views. Finally, answering question (3), specifically if it is found that there is a difference between male and female arrival times distribution, may trigger some interest for further research in this topic.

METHOD

Data

In this paper, we present analysis of finishing times recorded in five small size half marathon (21.095 km) races. The races were held in the United States, Poland, South Korea, and Australia in last two years. Two of the races were organized exclusively for female runners. The number of participants in the races varied from less than 1000 to more than 3000, as shown in Table 1. The data for the analysis were extracted from public domain world-wide-web sites. Some of the website give information about the athletes age, but some do not. Therefore, we focused on the racing times and we did not analyze the age structure within the groups. Information about the participants' gender was always available, however, and that let us distinguish between the races dominated by male runners and those in which the finish times distribution for the whole marathon pack is affected by the presence of a large proportion of female runners.

Table 1. List of analyzed races together with the number of finishers

	Race	Abbreviation	Location	Time Limit	Number of Finishers		
					Women	Men	total
1	Silesia Half Marathon	Sillesia	Poland	3 h	403 23%	1346 77%	1749
2	Seoul Open Race Half Marathon	Seoul	South Korea	3 h	188 15%	1054 85%	1242
3	Mercedes Benz Half Marathon	Merc-Benz	United States (Alabama)	4 h	1841 56%	1419 44%	3260
4	Nike Women's Half Marathon	NikeW	Australia	3 h	1566 100%	6 0.0%	1572
5	Bellin Half Marathon	BellinW	United States (Wisconsin)	4 h	897 99%	11 1.0%	908

Statistical method

One way analysis of variance (ANOVA) was used in order to compare the mean arrival times for each group of runners we were interested in. Tukey's method was applied to compare the means pairwise. The optimal distribution for the data was searched for based on the probability plots and goodness-of-fit tests. Kolmogorov-Smirnov test of normality was used in order to determine whether or not the distribution of the arrival times in studied races was normal. The same test was also used to compare the distributions of the arrival times for various groups of runners. Finally, simple linear regression method was used to find the relationship between arrival times of finishers from the race main pack (for explanation, see the Discussion section below) and their rank.

RESULTS

For the purpose of this work, we used raw data available at the following websites:

- https://silesiamarathon.pl/strefa-zawodnika/wyniki/wyniki-2016-ok?runid=62 (Silesia Half Marathon, 2016);
- http://www.seoul-race.co.kr/record/index03.php (Seoul Open Race Half Marathon, 2016);
  http://www.besttimescct.com/results/Mercedes16_Half_Overall.HTML (Mercedes Benz Half Marathon, 2016);
- http://fairfax.tiktok.biz/list/nikesydney/2016/21km/ (Nike Women's Half Marathon, Sydney, 2016);
- http://www.onlineraceresults.com/race/view_race.php?race_id=48813#racetop (Bellin Women's Half Marathon, 2016).

All records were considered in the analyses but those who did not finish the race.

The Results of Exploratory Data Analysis

Selecting the races held on different continents for the analysis was not expected to lead to large differences in the distributions of the finish times. As the exploratory data analysis shows, however, there are differences in basic statistical characteristics between the races. Some of the basic statistical data analysis results are summarized graphically in figure 1 in the form of box plots. Every box plot contains a rectangle whose vertical size indicates the range of the arrivals times for the middle half of all participants; the minimum and the maximum values are represented by the lines extended from the boxes; extreme values that are considered "unreasonable" from statistical point of view (outliers) are indicated by asterisks; the lines

dividing the boxes indicate the medium values of the arrivals times. From the plot in Figure 1., we can read the following information:

- The races held in Korea (Seoul) and Poland (Silesia) are characterized by nearly a symmetrical distribution of the arrival times
- The distributions of the arrival times in the races dominated by female participants are skewed to the right, i.e., there are many more "slow" runners as compared with "fast" runners
- The range of the arrival times is smaller in the races held in Poland, Korea, and Australia (NikeW race) as compared with the races organized in the United States
- The minimum arrival times are similar in all the races
- There are numerous outliers when it comes to the late arrival times

Putting together information from Table 1 and Figure 1, we can presume that the proportion of female to male runners may be an important factor for the arrival time distributions in half-marathon races. Another factor may be the time limit set by the organizers for the participants to arrive to the finish line.

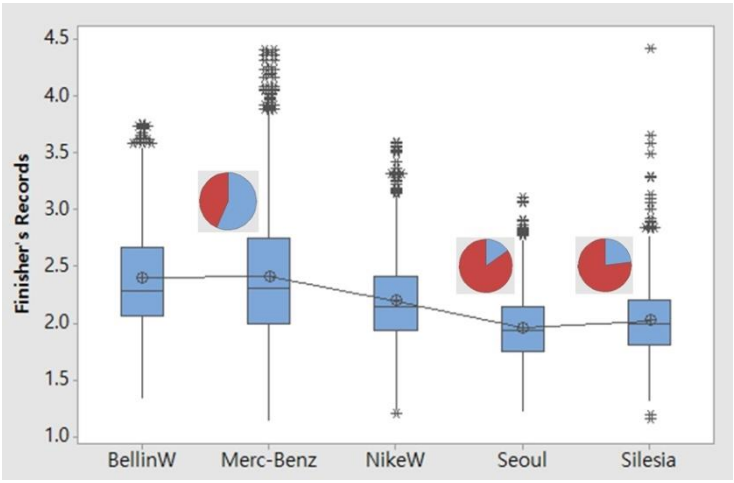

Figure 1. Box plots of the analyzed finishers' arrival times grouped by races. Three pie charts visualize proportions of female participants (blue color) in the three races. Finisher's Records are in the units of hours. See text for detail explanation of the plot.

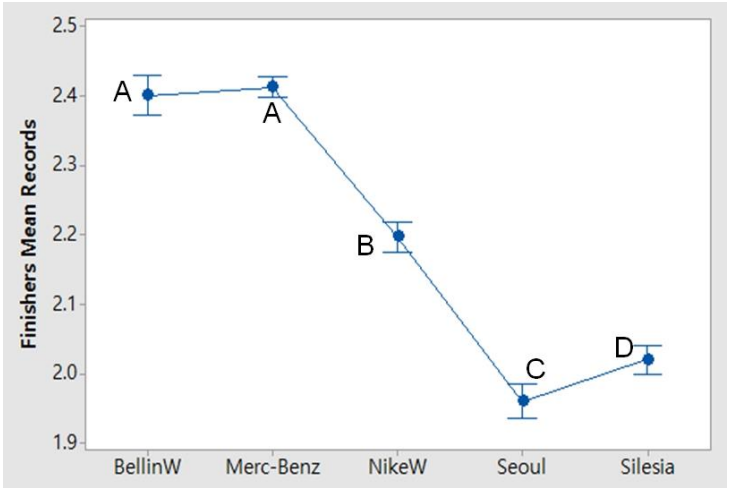

Figure 2. Comparison of finishers' mean arrival times with 95% confidence intervals shown. Different letters mean that the means differ significantly.

Large differences between the mean arrival times are seen in Figure 2, where the results of the means comparison with the use of ANOVA method are shown. The same letters next to the mean values indicate that the means are not statistically different, which is only in the case of two women's half marathons. Again, the races with a small proportion of female runners are characterized with much shorter mean arrival times.

Arrival times distributions

We first investigated the dependence of arrival times as a function of position (rank) of the athletes. In order to compare qualitative properties of the functions characterizing races with various numbers of participants, we projected the rankings in each race on [0 1] interval dividing the true ranking by the number of participants in the race. In Figure 3, the arrival times as functions of the normalized rankings are shown.

The recorded arrival times in the races with large ratio of female to male runners or in those organized exclusively for women, even after being logarithmically transformed did not appear to be normally distributed. In Figure 4 we present the histograms of arrival times distributions together with the normal probability plots for the data logarithmically transformed for two races: Seoul and BellinW. These two sets of graphs illustrate the two cases described above. Figure 4 contains also the probability plots associated with the data.

In Figure 5 a comparison of the eCDFs obtained for the two women's-only half marathons are shown. The two distributions cannot be considered the same, which was confirmed by the result of the Kolmogorov-Smirnov 2-sample test with the calculated test statistics greater than the critical value expected for the two samples having the same distribution.

In order to investigate whether the time limits set by the organizers may be one of the factors determining the arrival times distribution, we plotted eCDFs for the races with different proportion of female participants and we grouped the plots by the time limit set.The plots of eCDFs are shown in Figure 6.

In Figure 7, we show the cumulative percentage of finishers counted every 90 seconds from the time the winner crossed the finish line in the two mixed-gender races, i.e., Seoul and Merc-Benz, with the smallest and the largest proportion of female runners, respectively. The plots of the cumulative percentages for all runners (blue), female athletes (red), and male athletes (green) are shown. Since in the Seoul race there were only 15% female participants, the cumulative percentage function for all runners is basically defined by the group characteristics of male runners (Fig. 5b). In the Mercedes-Benz race, the proportion of women participants was almost 57%, and therefore the "All" curve falls close to the middle between the curves obtained for the two groups of runners.

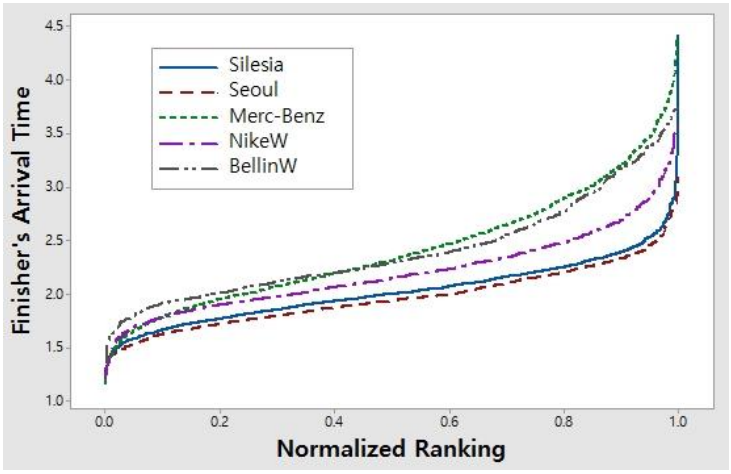

Figure 3. Finishers' arrivals times as functions of normalized rankings. Finisher's Arrival Times are in the units of hours.

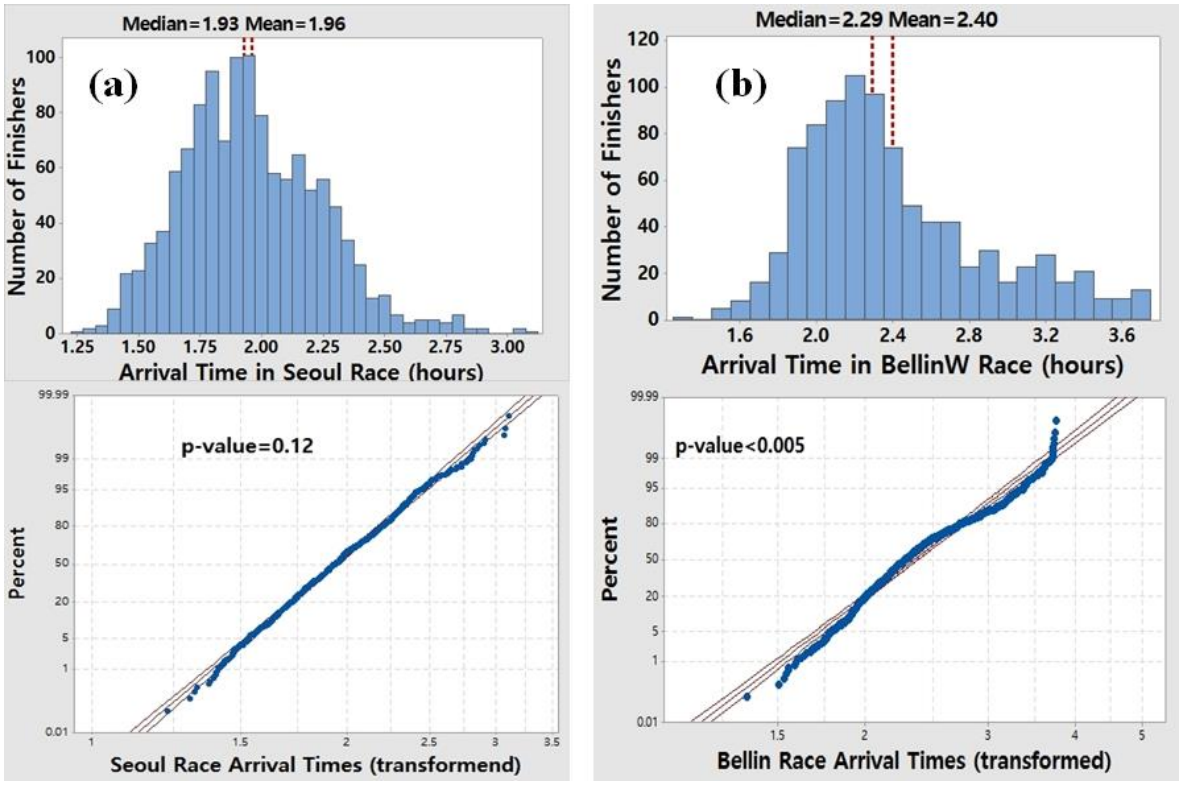

Figure 4. Histograms of the arrival times (top) together with the probability plots for the data logarithmically transformed (bottom) for (a) Seoul Open Race with 15% female runners and (b) Bellin Women's race with 99% of female runners.

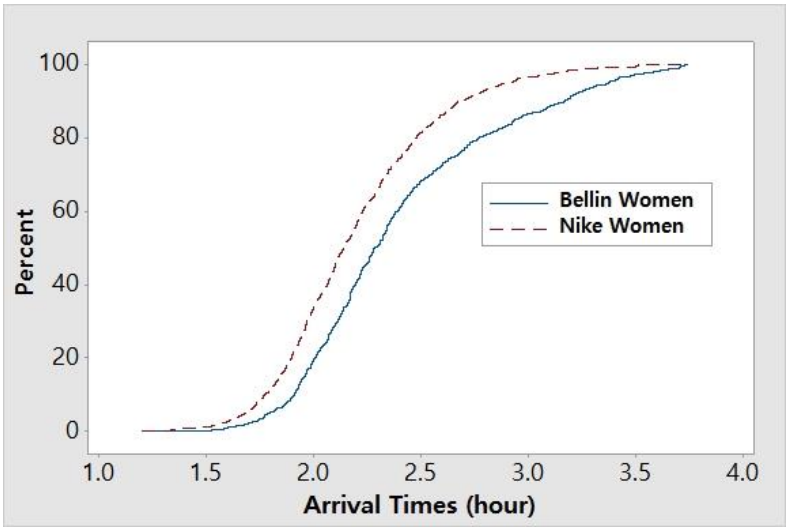

Figure 5. Empirical cumulative distribution functions plotted for the two women's-only half marathons.

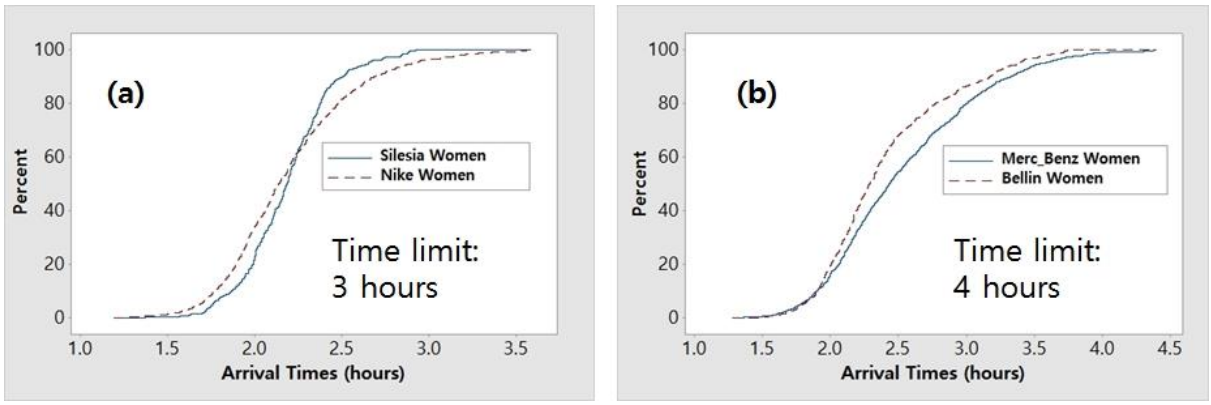

Figure 6. Empirical cumulative distribution functions plotted for the results recorded for (a) women athletes in Silesia half marathon together with women runners in Nike race, and (b) women athletes in Merc-Benz half marathon together with women runners in Bellin race.

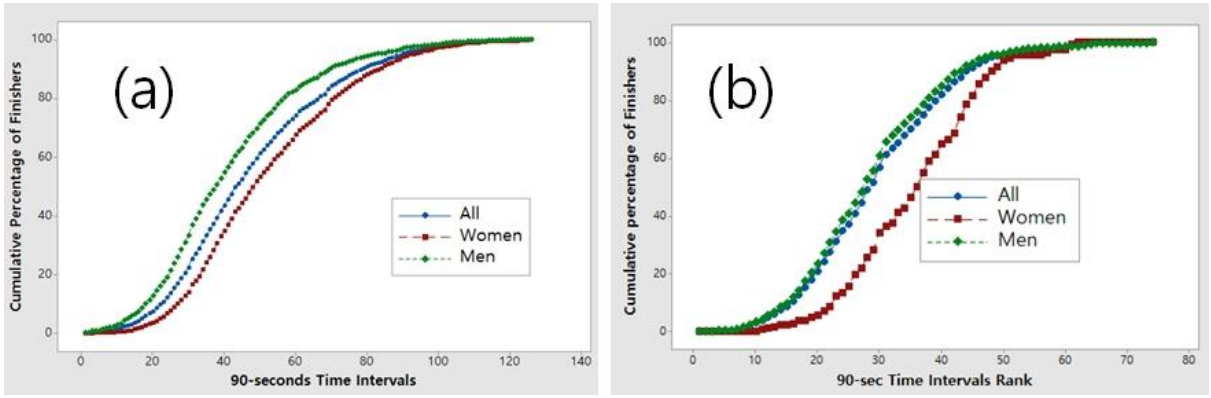

Figure 7. Cumulative percentage of finishers counted every 90 seconds from the time the first athletes arrived in (a) Merc-Benz half marathon and (b) Seoul half marathon.

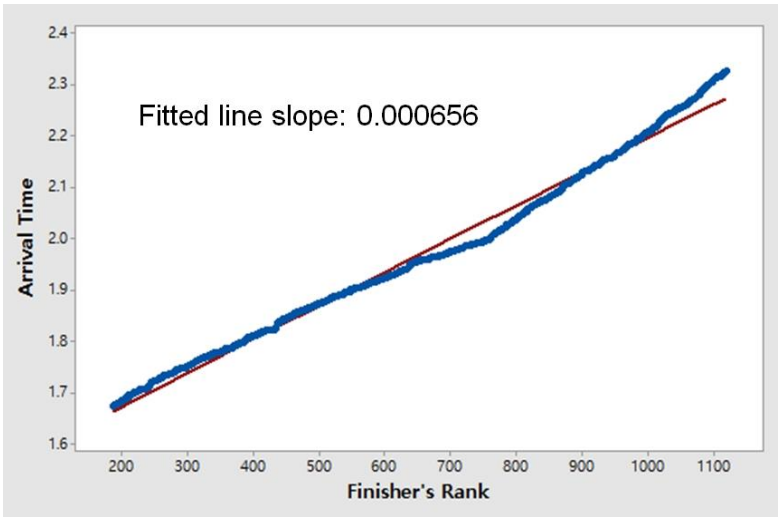

Figure 8. Seoul half marathon main race pack arrival times versus finisher's rank together with a fitted regression line.

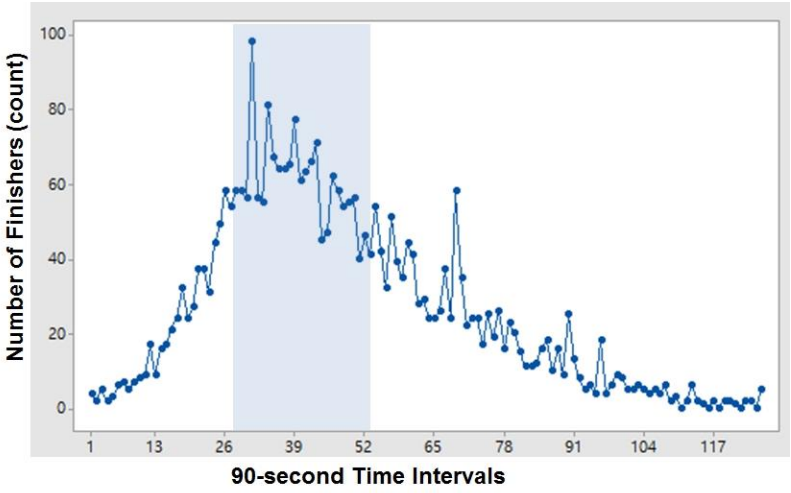

Figure 9. Number of runners arriving to the finish line at 90-seconds intervals as calculated for Mercedes-Benz half marathon. Shaded area indicates the main race pack.

The plots shown in Figure 7 are helpful in finding what percentage of runners consist so called main race pack, i.e., the volume of participants whose arrival times depends linearly on the ranking. We estimate that in the races with a small number of female participants 15% of athletes belonged to a high-performance group (lower tail in the plots) and 10% of runners belong to slow-pace group (upper tail). In the races dominated by female athletes, the slow-pace group is larger, and the non-linear behavior in the group dynamics is observed earlier; we estimate that not more than 60% of all runners arrive to the finish line in the steady rate on average.

In Figure 8, the linear dependence of the main race pack arrival times versus finishers rank in Seoul half marathon is shown together with the calculated regression line. The calculated regression line slope was 0.000656 (in hours), which means that when the main race pack is observed, on average one runner arrives to the finish line every 2.4 seconds.

Similar slope calculated for other races main packs were: 0.000456 hour or 1.64 second for Silesia race, 0.000389 hour (1.43 second) for Mercedes-Benz race, 0.000587 hour (2.11 second) for Nike-Women race, and 0.001446 hour (5.20 second) for Bellin-Women race.

It must be stressed that the discussed linear dependence of the arrival times observed for the main race pack is averaged over the whole group. In fact, the arrivals of the runners to the finish line is much more complex, dynamic behavior. Figure 9 illustrates the process. We counted the number of athletes coming to the finish line in regular time intervals of 90 seconds. As shown in Figure 9, the number of counts fluctuates indicating clustering of runners during the race, which is mirrored at the finish line. Such clustering was discussed regarding marathons [2]. In our opinion, it may be the result of cooperation between the runners, especially as the main race pack is concerned. Clustering in the group of fast runners (left tail of the plot in Figure 9) although not so obvious, is still visible. It may be the effect of competition between the runners. On the other hand, clustering in the group of slower runners and walkers (right tail of the plot) may be the effect of social behavior of the participants. Similar results were obtained for all the analyzed races.

DISCUSSION

The plots shown in Figure 3 have very similar shape. For each race, we found a curvilinear dependence between the arrival times and the finishers' rankings. At the minimum times limits, the curves are concave, while at the maximum time limits they are convex.

Between the inflection points, the dependence between the arrival time and the position at the finish line is nearly linear in every race. This characteristic overlaps with the one found in analysis of big size full marathon races [2]. The linear parts of the plots obtained for the half marathons analyzed in this work are not all parallel to each other: only the Silesia and the Seoul half marathons have almost identical arrival time – ranking characteristics; the long upper tail observed in the Silesia characteristic indicates presence of a severe outlier. The two races had common features: they had time limits set to 3 hours and the proportion of women participating in the races was very small compared with the other races. All curves are between the curves visualizing Seoul half marathon (the smallest slope in the linear part) and Mercedes-Benz half marathon (the largest slope). The shapes imply also that the distributions are skewed to the right and may not be normal. Indeed, the goodness of fit tests carried out for the sets of arrival times in all the five races confirmed that the hypotheses that the distributions are normal must be rejected.

Logarithmical transformation often used to normalize right-skewed data was not always a hundred percent effective in normalizing the data. We applied the Kolmogorov-Smirnov goodness of fit test to the data transformed in such a way. As it is seen from Figure 4, only the most symmetrically distributed data recorded in Seoul and Silesia races, in which the percentage of female runners was very small, appeared to be normal after the transformation. Based on the probability plots also shown in Figure 4, we can see that in the race dominated by men only few points fall slightly outside of the 95% confidence interval for lognormal distribution, while in the women's-only race majority of the data, even after transformation, depart from the normal distribution. Similar departure from normality was found not only for the data collected in the Nike Women's half marathon, but also for the data recorded in Mercedes-Benz race, in which 56% of runners were women. That leads us to another question to ask: Does the distribution of arrival times of female runners always differ from that of male runners ?

To find an answer to that question we plotted and compared empirical cumulative distribution functions (eCDFs) for the races of interest. The result, shown in Figure 5, was unexpected, as we predicted that the distributions of the arrival times in one-gender races will be similar. There is, however, a factor that might be responsible for the observed differences: the time limits were set differently in the two races; they were 3 and 4 hours in Nike and Bellin half marathons, respectively.

The assumption that the time limit to finish the race may be an important factor in shaping the distribution of the results was checked by comparing eCDFs for the following pairs of races: Nike Women and Silesia half marathons with the time limits set to 3 hours, and Bellin Women and Mercedes-Benz half marathons with time limit set to 4 hours. Only female runners' arrival times were considered as far as the mixed races are concerned.

CONCLUSIONS

We analyzed half marathons finishers arrival times in order to find whether or not there existed some differences in the arrival times distributions because of the races locations and proportion of female to male runners. Using simple tools, we found that in all races the pattern of arrival times versus finishers positions is similar. Specifically, there is always a small group of fast runners, a big group of finishers that may be called main race pack, and a group of slower runners and walkers which size is dependent of race specifics. On the other hand, there are significant differences between the distributions of the arrival time of groups of female and male athletes. Comparison of such gender-specific distributions between various races indicate that the shape of the distribution may be strongly dependent on the time limit set by the races organizers.

We found that the mean arrival time, so also the mean speed of the half marathon packs, depends on the ratio of female to male participants. Interestingly, proportions of female runners is much higher in the races organized in the United States than in the races organized in Poland or South Korea. It may be interesting to extend this work in order to find if this is true for all running events in Europe and Asia, or if it is specific for the two countries.

The fact that female runners move slower as a group than male runners is not surprising, but it projects on different distributions of the finishers at the finish line. Race organizers may be interested in finding more specifics about such distributions in order to organize the events in a way that meets all participants' needs.

The plots shown in Figure 6. Confirm our assumption about the importance of the time limit set by organizers on the distribution of the arrival times of female runners. Even though the Kolmogorov-Smirnov tests values are still greater than the critical value, it is visible that the differences between the distributions are smaller if the races with the same time limits are grouped together. This may lead us to a conclusion that the dynamic of female runners as a group is in a larger degree determined by the time limit announced by the organizers than by the fact that the race is for the females only.

ACKOWLEDGEMENTS

This work was in a large degree prepared for the Capstone Project in the Master of Applied Statistics program of Pennsylvania State University. The author is extremely grateful to Dr. Xiaoyue Niu, as well as to Anthony Carra and Nkiruka Atuegwu for critical reviews and valuable discussions regarding the work.

REFERENCES

1. Rodriguez E, Espinosa-Paredes G, Alvarez-Ramirez J. Convection–diffusion effects in marathon race dynamics. Physica A 2014, 393: 498–507. http://dx.doi.org/10.1016/j.physa.2013.09.051
2. Alvarez-Ramirez J, Rodriguez E. Scaling properties of marathon races. Physica A 2006, 365: 509–520. doi: 10.1016/j.physa.2005.09.066
3. Sabhapandit S, Majumdar S N, Redner S. Crowding at the front of marathon packs. J. Stat. Mech. Theor. Exp. 1008, 3: L03001. http://dx.doi.org/10.1088/1742-5468/2008/03/L03001
4. Alvarez-Ramirez J, Rodriguez E, Dagdug L. Time-correlations in marathon arrival sequences. Physica A 2007, 380: 447-454. http://dx.doi.org/10.1016/j.physa.2007.03.008
5. Garcia-Manso J M, Martin-Gonzales J M, Davila N, Ariazza E. Middle and long distance athletics races viewed from the perspective of complexity. J. Theor. Biol. 2005, 233: 191-198. doi: 10.1016/j.jtbi.2004.10.014

Cite this article as:
Białek B. Statistical Description of Arrival Sequences in Amateur Long-Distance Races, Phys Activ Rev 2017, 5: 44-53