Statistics and Application 统计学与应用, 2017, 6(5), 526-532
Published Online December 2017 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2017.65059

Hans汉斯

A Probability Distribution Problem in Sports Games

Xingyu Chen[1], Xiaowei Cai[2]

[1]Jiangsu Tianyi High School, Wuxi Jiangsu
[2]National Supercomputing Center in Wuxi, Wuxi Jiangsu
Email: xingyu_chen.alice@hotmail.com

Received: Dec. 2nd, 2017; accepted: Dec. 18th, 2017; published: Dec. 25th, 2017

Abstract

In mathematical modelling problems relating to sports games, the running or swimming time of athletes is usually assumed to follow a normal distribution. This paper statistically and theoretically analyzes the data collected from some real Marathon and Triathlon games, and shows that in a sports event the athletes' average speed has a general normal distribution, whereas the time taken by the athletes arriving at a fixed point (e.g., the destination) in the race course does not follow such type of distribution, but rather a parameter-varying right-skewed normal-like distribution. Moreover, the mathematical formula for the probability density function of the athletes' race time is derived, and the results obtained may also be applied to the corresponding mathematical modelling problems of similar sports games or other application areas.

Keywords

Mathematical Modelling, Normal Distribution, Probability, Sports Games, Marathon

体育比赛中的一类概率分布问题

陈星宇[1]，蔡晓伟[2]

[1]江苏省天一中学，江苏 无锡
[2]国家超级计算无锡中心，江苏 无锡
Email: xingyu_chen.alice@hotmail.com

收稿日期：2017年12月2日；录用日期：2017年12月18日；发布日期：2017年12月25日

摘　要

在有关体育比赛的数学建模问题中，运动员的跑步、游泳等比赛时间通常被假设为服从正态分布。本文

文章引用: 陈星宇, 蔡晓伟. 体育比赛中的一类概率分布问题[J]. 统计学与应用, 2017, 6(5): 526-532.
DOI: 10.12677/sa.2017.65059

通过对若干马拉松及铁人三项比赛成绩进行统计和理论分析后发现：运动员的平均速度服从一般的正态分布，但是运动员到达赛程中某一固定点(比如终点)所用的时间并不服从正态分布，而是服从一个变参数的、右偏的类正态分布。文中推导了时间概率分布密度函数的数学表达式，该结论可应用于类似比赛或其它应用领域的相应数学建模问题。

关键词

数学建模，正态分布，概率，体育比赛，马拉松比赛

1. 引言

体育运动是现代人类生活中的重要组成部分，体育比赛的激烈性和挑战性也常常吸引人们积极参与。为了提高体育比赛成绩，发掘人体潜能，科学家会运用统计学、数学、物理学等各种学科知识对体育运动展开科学研究，从中也产生了众多有意义的数学问题。

大部分体育比赛参与的运动员都不会太多，但是马拉松和铁人三项等体育运动是特例，参与者可达成千上万，业余者和专业运动员可同台竞技，因此这两项运动特别能吸引广大群众积极参加。正因为马拉松比赛中的运动员人数众多，所以运动员的运动速度或比赛时间可当作随机变量来处理，然后就可以采用统计方法来研究马拉松比赛中的问题。统计分析的基础是要知道随机变量的概率分布，通常由实验数据或经验假设给出。由于正态分布是常见的经典分布之一，因此正态分布假设在体育运动统计分析中得到广泛应用[1] [2] [3]。

由于正态分布比较常见，因此当通过初步统计发现实验数据的统计图形近似于钟形对称时，就会轻易地认为相应的随机变量服从正态分布，但是真实情况未必如此。本文以文献[3]的研究作为基础，对马拉松运动员的速度和比赛时间的概率分布开展了研究。

文献[3]研究了马拉松比赛途径上设置水站的优化问题。所谓水站优化问题，是对不同水站要储备多少饮用水进行估算，使得不同时间到达水站的运动员都能拿到合适的饮用水。显然，要解决水站优化问题，首先就必须估算出不同时间段到达某个水站的人数，然后才能估算出水站应该储备多少饮用水。这就涉及到到达时间的概率分布问题。文献[3]的方法是，将运动员通过某一水站的时间作为独立同分布随机变量，通过每隔 15 分钟统计到达水站的运动员人数，绘出直方图(见图 1，该图给出了 18 km 位置和 31 km 位置两个水站的统计数据直方图)，并由此认为根据中心极限定理，流量(某一时间段内跑过该站的总人数)应该服从正态分布，见式(1)。

$$f(t)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] \tag{1}$$

然后文献[3]在假设所有运动员的速度是匀速的条件下，得到了跑过位于x位置水站的人员流量密度函数：

$$f(t)=\frac{K(x)}{\sigma(x)\sqrt{2\pi}}\exp\left[-\frac{(t-x/v)^2}{2\sigma^2(x)}\right] \tag{2}$$

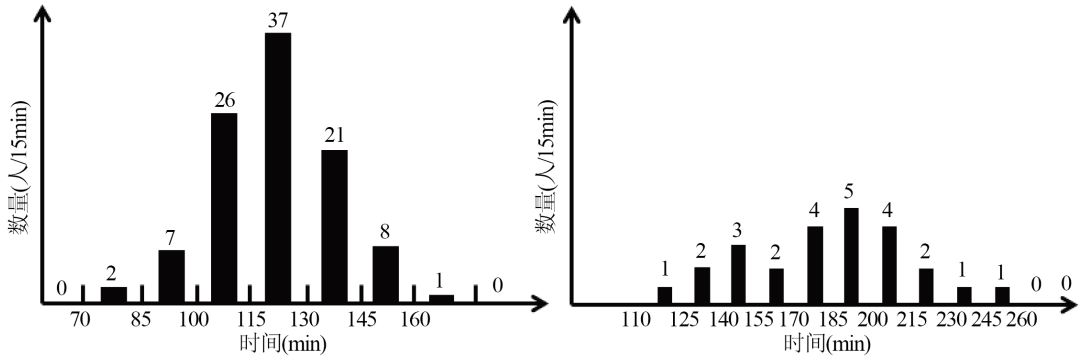


Figure 1. Histogram of flows for 18 km and 31 km (cited from [3])
图 1. 18 km 和 31 km 流量图(引自文献[3])

其中 $K(x)$ 为跑过 x 水站的总人数， v 为平均速度。

在式(1)和(2)基础上，文献[3]建立了水站优化配置的数学模型并用计算机求得了优化解。

然而，本文通过简单的定性分析发现，在运动员速度 v 为匀速的条件下，跑过 x 水站的人员概率密度函数不太可能是正态分布。由于运动员的速度是有上下限的，设运动员的最大可能速度为 $v_{\max}$ ，则运动员的平均速度在理论上的取值范围为 $v\in(0,v_{\max})$ ，从而运动员从出发点 0 跑到 x 水站的时间为

$$t=\frac{x}{v} \tag{3}$$

因此时间 t 的取值范围为 $t\in(x/v_{\max},+\infty)$ 。

从上述取值范围就可看出，时间t在左边是有限值，而在右边可取到无穷，两边并不对称，据此可推测时间t的概率密度函数 $f(t)$ 必然是非对称图形(见图 2)，因此 $f(t)$ 不太可能是正态分布(因为正态分布是对称的)，也就是说，用式(1)作为时间t的概率密度函数是不准确的。

文献[3]出现这种假设的根本原因在于统计数据时忽略了速度比较慢的选手人数，即忽略了统计图形右边部分拖延到无穷但是数值较小的部分(见图 2 右边虚线的右侧部分)，于是认为剩下的是对称的正态分布图形。

虽然 $f(t)$ 不可能是对称的正态分布，但注意到运动员的速度 v 也是一个随机变量，其取值范围为 $v\in(0,v_{\max})$ ，上下限都是有限值，因此以速度 v 为统计标准的概率密度函数 $f(v)$ 倒是有可能相对中间值对称，即速度 v 可能服从正态分布。为了验证这一点，通过从网上收集真实马拉松比赛的数据，对上述猜测进行了分析与检验。

2. 比赛数据的统计分析

为了验证上节的猜想，本文从中国田径协会官方网站[4]上获取了 2013 年重庆国际马拉松赛成绩表中给出的数据，考虑到运动员参加并完成半程马拉松更具有随机性(能跑完全程马拉松的运动员大都进行过专门训练)，因此我们选择其中的男子半程马拉松的成绩来进行统计，这样在去掉其中的专业运动员人数后，剩余总人数为 K = 1521 人。对通过 x = 5 km 和 x = 10 km 水站的时间和速度的频率分布分别进行了统计，其中速度 $v=x/t$ 可由式(3)计算得到。利用 Matlab 中的函数 hist 绘制出相应的频率直方图，分别见图 3 和图 4。其中图 3(a)的横坐标表示时间，纵坐标表示频率，即某段时间段内到达的人数 n 与总人数 K 之比 $f(t)\approx n/K$ ， 图 3(b)及后面各图中横纵坐标的含义与此类似。

从图 3 和图 4 均可看出，反映时间分布的(a)图确实类似于图 2 中的右偏图形，而反映速度分布的(b)图更近似于对称的正态分布图形。

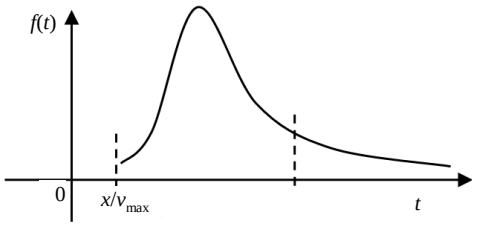
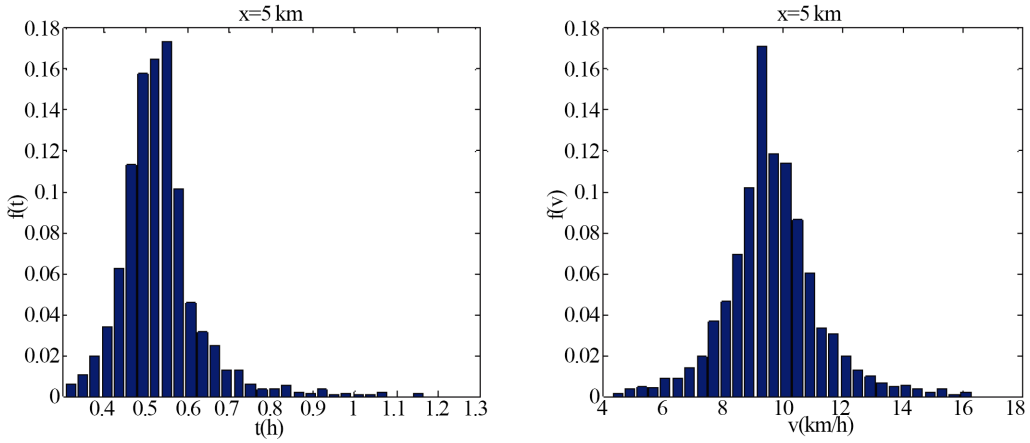


Figure 2. Diagrammatic sketch of the probability density function $f(t)$

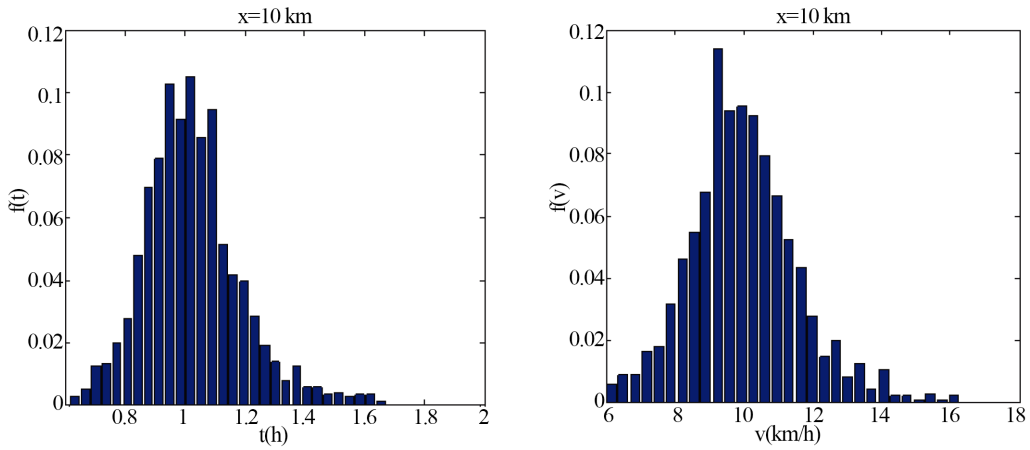
图 2. 概率密度函数 $f(t)$ 的示意图



(a) 时间分布直方图　　(b) 速度分布直方图

Figure 3. Frequency histogram for x = 5 km water station

图 3. x = 5 km 水站处的频率直方图

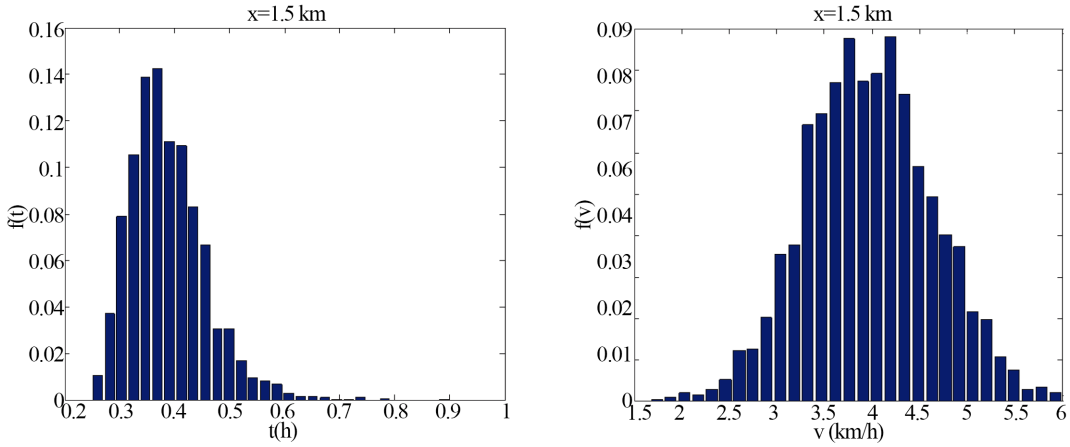


(a) 时间分布直方图　　(b) 速度分布直方图

Figure 4. Frequency histogram for x = 10 km water station

图 4. x = 10 km 水站处的频率直方图

为了进一步验证上述结论，本文采用了美国数学及其应用联合会网站[5]上 2016 年美国高中生数学建模竞赛(HiMCM)的 Problem A 提供的近期某铁人三项比赛的成绩表，选择其中的 MOpen 选手的游泳比赛成绩进行统计(总人数为 2147 人)，得到的时间分布和速度分布直方图见图 5。从图 5 可得出类似的结果，即速度分布更接近对称的正态分布图形，而时间分布是右偏的图形。



(a) 时间分布直方图　　(b) 速度分布直方图

Figure 5. Frequency histogram of the swimming results
图 5. 游泳比赛成绩的频率直方图

综上所述，本文认为用正态分布来描述运动员的速度分布更为合理，进而可由速度分布推导出相应的时间分布。

3. 速度的概率密度函数

根据上一节的分析，设运动员的速度分布为正态分布，即有

$$f(v)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(v-\mu)^2}{2\sigma^2}\right]=N(\mu,\sigma^2) \tag{4}$$

其中期望值 μ 和标准差 σ 可分别按下述式(5)和式(6)来估计

$$\mu=\bar{v}=\frac{v_1+v_2+,\cdots,+v_K}{K} \tag{5}$$

$$\sigma=std(v)=\sqrt{\frac{(v_1-\bar{v})^2+(v_2-\bar{v})^2+,\cdots,+(v_K-\bar{v})^2}{K-1}} \tag{6}$$

以前述马拉松赛到达=10 km 水站的速度分布(见图 4)和铁人三项中游泳比赛的速度分布(见图 5)为例，发现它们分别与 $N(1.99,2.42)$ 和 $N(3.97,0.45)$ 的正态分布符合得很好，见图 6。

4. 时间的概率密度函数

根据前面的讨论，时间与速度存在关系式 $t=x/v$，因此时间 t 的概率密度函数可以从速度 v 的概率密度函数公式(4)推导出来。这里需要利用如下的随机变量的函数的分布定理[6]：

定理：设 X 是连续随机变量，具有概率密度函数 $f_X(x), x\in(-\infty,+\infty)$。$Y=g(X)$ 是另一随机变量。若 $y=g(x)$ 严格单调，其反函 $h(y)$ 数有连续导函数，则 $Y=g(X)$ 的概率密度函数为

$$f_Y(y)=\begin{cases} f_X[h(y)]|h'(y)|, & y\in(a,b) \\ 0, & \text{其他} \end{cases} \tag{7}$$

其中 $a=\min\{g(-\infty),g(+\infty)\}$，$b=\max\{g(-\infty),g(+\infty)\}$。

在本文所讨论的问题中，随机变量 $t=x/v$，因此由上述定理及式(4)可得时间 t 的概率密度函数为

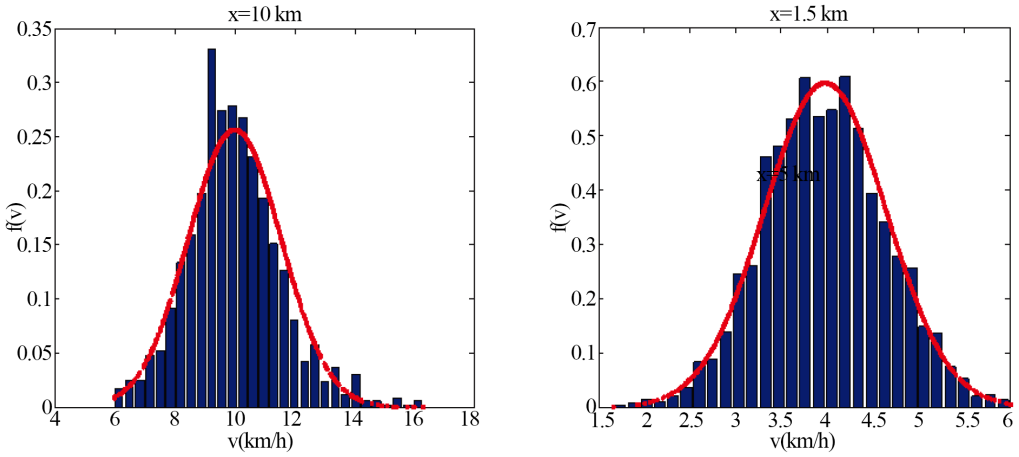
$$f(t)=\frac{x/(\mu t)}{(t\sigma/\mu)\sqrt{2\pi}}\exp\left[-\frac{(t-x/\mu)^2}{2(t\sigma/\mu)^2}\right]=x/(\mu t)\cdot N\left(x/\mu,(t\sigma/\mu)^2\right) \tag{8}$$

其中的参数 μ 和 σ 可分别按式(5)和(6)估算。

时间 t 的概率密度函数式 $f(t)$ 在形式上可写成一个变参数的类正态分布函数 $N\left(x/\mu,(t\sigma/\mu)^2\right)$ 与 $x/(\mu t)$ 的乘积，但它却不是常规的正态分布函数，其图形是不对称的。由式(8)及图 6 中马拉松赛和游泳比赛的速度分布，分别计算出相应的时间分布并与图 4 和图 5 的时间分布频率直方图中的数据进行对比，相应结果见图 7。从图中可看出，运动员比赛时间的统计数据与式(8)的计算结果非常吻合。

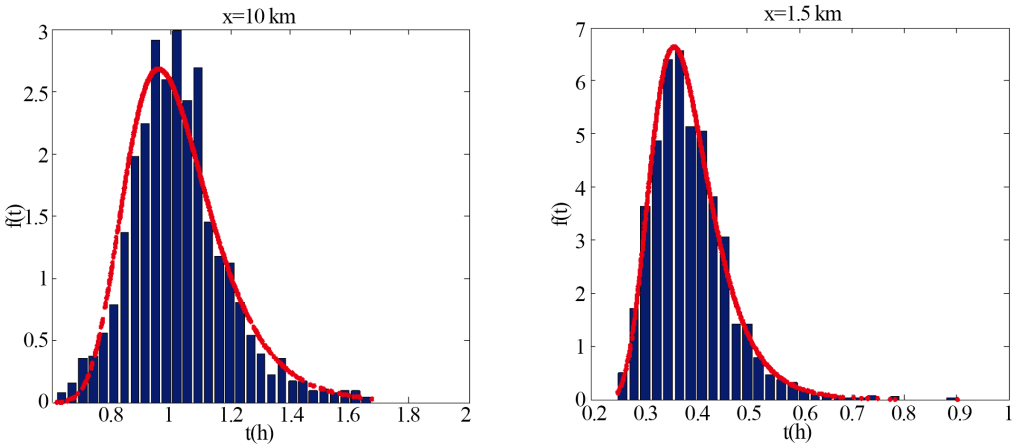
5. 结论

本文基于若干马拉松赛及铁人三项比赛运动员的成绩表，对相应赛事中运动员的速度分布以及到达赛程中固定点的时间分布进行了分析，从数据统计和理论建模两方面，均证明了运动员的速度分布接近于正态分布，但到达固定点(比如终点)的时间分布并不是正态分布，而是一种变参数的右偏的类正态分布



(a) 马拉松赛　(b) 游泳比赛

Figure 6. Comparison of normal probability density function and frequency histogram of speed
图 6. 速度的正态分布密度函数与频率直方图的比较



(a) 马拉松赛　(b) 游泳比赛

Figure 7. Comparison of normal-like probability density function and frequency histogram of time
图 7. 时间的类正态分布密度函数与频率直方图的比较

分布。本文推导了这种时间分布的概率密度函数，给出了其理论表达式。需要注意的是，即使在大样本条件下，时间的非对称分布也不可能趋向正态分布，这与大数定律或中心极限定理无关。该结论可用于其他类似运动比赛中的数学建模或管理优化，也可用于交通流的研究。

参考文献 (References)

[1] 田利军. 探析正态分布理论在体育运动统计分析中的作用[J]. 吉林师范大学学报(自然科学版), 2012(3): 148-151.

[2] 许莉, 黄宗文. 体育统计常用指标的概率分布初探[J]. 体育科技, 1999, 20(2): 53-56.

[3] 李栋, 杨明进. 马拉松服务方案优化研究[J]. 体育科学, 2008, 28(4): 30-38.

[4] 中国田径协会官方网站. 2013 重庆国际马拉松赛成绩表[EB/OL]. http://www.athletics.org.cn/competition/results/2013-07-14/414937.html

[5] Consortium for Mathematics and Its Applications (COMAP). HiMCM 2016 Problems. http://www.comap.com/highschool/contests/himcm/2016problems.html

[6] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程[M]. 北京: 高等教育出版社, 2014.