2011-2015 Boston Marathon Results Analysis

Peter Chng

More boring running stuff!

Introduction

- Scraped, cleaned/normalized data and put into a single CSV file
  - http://boston.r.mikatiming.de/{year}
- Available at: https://github.com/pchng/boston-marathon-2011-2015
- Structure of data:

```
# Example of what the data looks like:
results[results.name_location.str.contains('Chng')]
```

	year	place	place_gender	place_division	name_location	country	bib	gender	age_group	half_split	finish_net	finish_gun
1548	2011	2478	2288	1549	Chng, Peter (CAN)	CAN	NaN	M	18-39	01:32:34	03:07:20	03:10:37
25486	2012	3052	2648	1608	Chng, Peter (CAN)	CAN	3179	M	18-39	01:39:51	03:29:51	03:31:43
48147	2013	5708	4919	2715	Chng, Peter (CAN)	CAN	2284	M	18-39	01:38:48	03:19:25	03:20:53
64975	2014	3075	2881	1963	Chng, Peter (CAN)	CAN	2664	M	18-39	01:29:17	03:04:11	03:05:22
97055	2015	3329	3125	2238	Chng, Peter (CAN)	CAN	2168	M	18-39	01:30:15	03:03:34	03:04:52

Sample size and demographics

- Normally consistent but has varied from ~17K to ~31K in recent years
- Qualification times and impact on demographics
  - BQ times change impact?
  - M/F ratio roughly inline with other big marathon races

Differences across years

- Differences attributable to two main properties
  1. Underlying ability/fitness of the runners
  2. Environmental conditions
- Difficult to determine (1), except to say that tougher BQ standards have almost certainly made the field faster
- (2) is directly observable and can have a much larger impact
  - Heat and humidity generally bad
  - {Tail, Head}wind is generally {good, bad}

Differences across years

- Compare 2011 (13 C) vs 2012 (31 C):
  - Median finishing time 26 mins slower (4:11 vs. 3:45)
    - Going from 4:11 to 3:45 through a fitness gain is a huge improvement
  - Much higher standard deviation
- Effect of heat:
  - Slower runners slowed more
  - Standard deviation increases more than median
  - Right tail gets drawn out more; positive skew
- Challenging assumptions
  - 2015 considered to be tough because of miserable conditions: heavy wind, rain, cold
  - Aggregate stats show this to be the fastest year from 2011-2015

Distribution of finishing times

- Almost always a big drop-off after 4 hrs
  - Less frequent: After 3-hrs
  - More evidence of running to a goal, rather than “all out”
- Distribution of half-split times is more “smooth”
  - bin[i] <= bin[i+1], i < mode position
  - bin[i] >= bin[i+1], i >= mode position

Success rate of a sub-3?

- How to determine:
  1. Make some assumptions (possibly incorrect!)
  2. Calculate success ratio based on these assumptions
  3. Make a graph
- Correlated with the “toughness” of the environmental conditions; varies widely

M18-39 age group

- Behaviour has changed with the changing BQ standards
- Aggregate behaviour: Now aiming for sub-3:00 instead of sub-3:10
- Much more competitive: Bad news for me!

Negative split

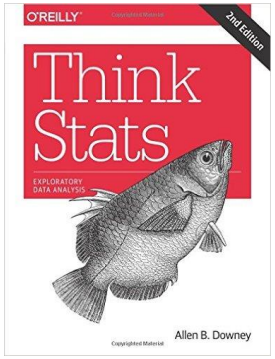
- Very few negative split runners in any given year
- Again, correlated with environmental conditions, mainly temperature
- No male/female differences
- Doesn’t get much better if relax the definition of “negative split”

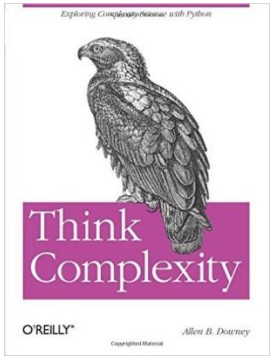
Normality of finishing times?

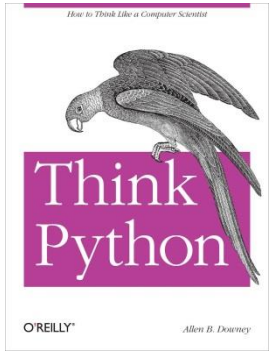
- If normal model is good, then mean and standard deviation completely describe distribution
- Some studies suggested the normal model is a good model for marathon finishing times
  - They used statistical tests for normality
- I could not reproduce with my data
  - Just looking at the distributions, they don't look normal
  - Needed someone with more insight

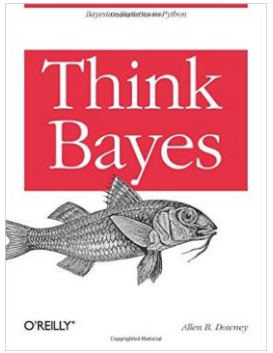
Normality of finishing times?

- Dr. Allen Downey: CS Professor, runner, wrote some books









Normality of finishing times?

- Ask the right question:
  - Is data normally distributed? Most likely no
    - Statistical tests for normality test the likelihood of a sample being drawn from a normal distribution
    - Real world-data truncated/bounded
  - Instead, check if the normal model is a good model for your data
    - Slight difference
    - Just do visual comparison: CDFs, Q-Q plots

Normality of finishing times?

- Result: Found normal model mostly deviated
  - Still unable to explain differences, but at least have pretty graphs
  - No sub-group by {year, gender} was modeled well by a normal model; even specific age groups (M18-39) were not represented well by a normal model
  - Normal model might be “good enough” since the cost of being incorrect is pretty much zero (Not so in other situations)
- Reasons? Can only speculate:
  - People run to goal times
  - When things go bad, they really go bad (positive/right skew)