hot hands in hockey
are they real? does it matter?

Namita Nandakumar
Wharton School, University of Pennsylvania
@nnstats
Let’s talk about streaks.

2017-18 Flyers’ record as coin flips:
first 16 games: $HTHTHHTHTTHTTHTTH$
next 16 games: $TTTTTTTTTTTTTHHHHHHH$
the rest: $THTTTHTHHHHTHHTTTTHHHH \ldots$

This seems like a *ridiculous* coin.

~ *The Athletic Philadelphia*, 11/27/2017
Back in January, I simulated sequences of wins and losses for a team with the Flyers’ record and showed that it was pretty unlikely for those teams to ever reach a 10-game losing streak.
But I’m zeroing in on a small, particularly interesting part of a larger sequence.

(Another small, interesting part of a larger sequence? The Penguins have lost 3 games in a row.)

The point is we need something a bit more comprehensive.
What do I want from a measure of streakiness?

- (This may differ from what you want from a measure of streakiness.)

- Takes the whole sequence into account.
  
  *Sure, the Flyers lost 10 games in a row, but what about the rest of it?*

- Puts the streaks in context.
  
  *It’s easy to go on a losing streak if you’re the 2017-18 Arizona Coyotes.*

- Imposes as few assumptions as possible.
  
  *I am often wrong about things.*

- Is easily computed, summarized, and conveyed.
  
  *I want to quickly look at all teams and all goal scorers.*

- Works for binary incidence data.
  
  *Did a team win or not? Did a player score or not?*
What do I mean by streakiness/clumpiness?

- Things are more “clumped” together than they would be if events in a sequence were randomly distributed, and this could be due to:

  - **Sequential dependence**: The next game’s outcome is impacted by whether you won the last one, or two, or three, and so forth.

  - **Non-stationarity**: There is a non-constant probability of success and the team/player goes through relatively good and bad periods.

- **Note**: I’m not at any point making any statements about the underlying quality of any team or player.

  *Essentially, I’m treating observed outcomes as fixed, and then seeing if they’re streakier than what we’d expect from flipping a (not necessarily 50/50) coin.*
The #AdvancedStat

- “New Measures of Clumpiness for Incidence Data” addresses major problems with existing hot hand measures.

- In particular, Zhang et al. find that the normalized entropy of inter-event times is a robust measure of clumpiness that minimizes misclassification error compared to other metrics.

- **Inter-event time**: # of time periods between wins or goals or whatever.

- Entropy is a measure borrowed from information theory that is related to disorderliness and uncertainty.

- More importantly, it has many desirable properties when utilized with inter-event times to evaluate clumpiness.
How does it work? Let’s look at the Flyers again.

- Calculate inter-event times $x_i$:
  
  \[
  \begin{array}{c}
  H & T & H & T & H & T & H & \ldots \\
  \end{array}
  \]
  
  \[
  x_i = 1 & 2 & 2 & 1 & 2 \\
  62 & 62 & 62 & 62 & 62 \\
  \]

- Divide by the length of the sequence + 1 to normalize.
  
  \[
  x_i = 1 & 2 & 2 & 1 & 2 & \ldots \\
  62 & 62 & 62 & 62 & 62 \\
  \]

- Multiply each $x_i$ by the log of itself, sum it all up, divide by log(number of successes + 1), add 1 to keep things positive.

  \[
  \ldots H_p = 0.08558024 \quad \text{(as of 2/22/2018)}
  \]

- We know that higher values correspond to more streakiness, but is that high or low?
Simulate sequences to contextualize results.

- I simulated 10,000 sequences of wins and losses from a 32-29 team, then calculated the normalized entropy for each sequence.

- It’s clear that the Flyers’ actual results were quite a bit streakier than most simulated sequences of their record.

- In fact, they were in the 99th percentile of streakiness.
Why do we need to do the simulate?

- You might think that if one team has a higher raw entropy value than another, they’re streakier.

- But the range of reasonable values depends on the length of the sequence and number of successes.

- We’d naturally expect worse teams to have larger spacings between wins.
What about other streakiness metrics?

- **Runs Test:**
  Count the number of runs of heads or tails within a sequence.

  \[
  H \ T \ H \ T \ HH \ T \ HT \ H \ TT \ H \ T \ TTTTTTTTTTT \ HHHH \ ... 
  \]

  No statistically significant evidence of “streakiness” at a 5% level for the 2017-18 Flyers.

- This disagrees with our normalized entropy test, which would classify the Flyers as streaky at a 5% level of significance.

- **Who’s right?** We can’t ever be sure, but...

- Normalized entropy more accurately classified computer-generated streaky data as streaky as compared to the runs test and others (Zhang et al.).
Streakiness of NHL Teams

- Streakiness percentile represents % of randomly generated sequences from a team’s win-loss record that they are streakier than.

  Higher = more streaky.

- NHL teams have not been too streaky this season... except for the Flyers.
Streakiest Teams Since 2009-10

- Columbus Blue Jackets, 2014-15
- Philadelphia Flyers, 2017-18

Un-Streakiest Teams Since 2009-10

- Dallas Stars, 2009-10
- Philadelphia Flyers, 2011-12

Season summaries via Hockey Reference.
Streakiness of NHL Player Scoring

- Blake Wheeler and Nick Backstrom are usually kind of streaky.
- Tomas Tatar and Joe Pavelski are usually kind of un-streaky.
- The vast majority of players with 15+ goals in 41+ games played in each of the last 5 seasons have had relatively streaky and un-streaky seasons.
Repeatability: Between Seasons
Repeatability: Within Season

Teams

Streakiness Percentile: Second Half

Players

Streakiness Percentile: First Half

$r = -0.077$

$r = 0.067$
Season Streakiness vs. Success

(This is what we’d expect due to having contextualized streakiness based on number of successes and failures.)
I could have just shown you this:
Is streakiness real?

- Yes, in the sense that there are certainly sequences in recent NHL history that appear “streaky.”
- Which could be due to chance, but also due to external factors (ex. CBJ injury issues in 2014-15).
- This is a pretty simple statement but it does matter!
- A frequent claim related to hot hands is that we are spotting patterns where there are none.

Is streakiness repeatable?

- No.

Does streakiness matter?

- No.
Thank you for listening!

Special thanks to:
- Prof. Shane Jensen (Wharton Statistics) for guidance
- Hockey Reference for data

Code/slides/data/etc. will be tweeted out @nnstats.
Appendix: Normalized Entropy Calculation (Source)

1. Convert the individual-level transaction data into incidence/binary data, if necessary.
2. Compute the Inter-event times (IETs).
   \[ x_i = \begin{cases} 
   t_1, & \text{if } i = 1, \\
   t_i - t_{i-1}, & \text{if } i = 2, \ldots, n, \\
   N + 1 - t_n, & \text{if } i = n + 1,
   \end{cases} \]

   where \( t_i \) and \( x_i \) are the \( i_{th} \) occurrence of event time and IETs, respectively, and \( n \) and \( N \) represent the number of visits and total time intervals.
3. Rescale the IETs. Divide IETs by \( N + 1 \).
4. Compute the normalized entropy-like

\[
H_p: 1 + \frac{\sum_{i=1}^{n+1} \log(x_i) \cdot x_i}{\log(n + 1)}.
\]
Appendix: Runs Test Calculation (Source)

H₀: the sequence was produced in a random manner
Hₐ: the sequence was not produced in a random manner

Test Statistic:

\[ Z = \frac{R - \bar{R}}{s_R} \]

where \( R \) is the observed number of runs, \( \bar{R} \), is the expected number of runs, and \( s_R \) is the standard deviation of the number of runs. The values of \( \bar{R} \) and \( s_R \) are computed as follows:

\[ \bar{R} = \frac{2n_1 n_2}{n_1 + n_2} + 1 \]

\[ s_R^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} \]

with \( n_1 \) and \( n_2 \) denoting the number of positive and negative values in the series.

Significance Level: \( \alpha \)

Critical Region: The runs test rejects the null hypothesis if

\[ |Z| > Z_{1-\alpha/2} \]