Streaky Good Models

Incorporating Short-Run Trends into NBA Predictions

Harry Jain
Advisors: Jim Glenn
&& Kevin O’Neill

ROADMAP

Background and Motivation

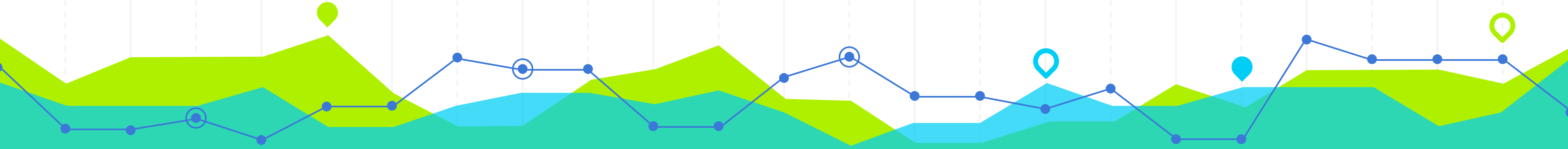
Project Outline and Goals

The Hot Hand Fallacy

Measuring Streakiness

Simulations and Results

Future Directions and Applications to Modeling



Background and Motivation

1

Why are we interested in studying streaks in sports?

The streak has become my identity; it’s who I’ve become.

- *Cal Ripken, Jr.*

Characteristics of NBA Data

- “NBA data is subject to **relatively little randomness**…” (FiveThirtyEight)
  - Small rotations with regular minutes distribution
  - Large amount of games
    - 82 regular season games
    - Up to 28 postseason games

Typical Predictive Models

Elo Ratings

- Provides a **mathematical metric for calculating win probability** that is **updated** from each game result

$$E_A = \frac{1}{10^{\frac{-(R_A - R_B + Adj)}{400}} + 1}$$

$$R'_A = R_A + K(S_A - E_A)$$

- Used by FiveThirtyEight

Adjusted Point Differentials

- Predicts **average margin of victory**
- **Bayesian framework** based on priors like offensive and defensive efficiency, schedule strength, preseason expectations, etc.
- Used in ESPN’s BPI and Basketball Reference’s SRS

Machine Learning

- Can predict either game results (**classification**) or victory margins (**regression**)
- Generally **supervised learning**, e.g. neural networks, logistic regressions, or random forests using season data
- Used in research projects and newer models

Mathematics of the Elo Model

At its core, the Elo model assumes that performance of each player in each game is a normally distributed random variable, essentially making it a normal or logistic regression model with the properties below.

- **Independent variables:** team ratings, represented by:

$$\begin{aligned} R_A &= \text{rating for Team A} \\ R_B &= \text{rating for Team B} \end{aligned}$$

- **Dependent variables:** game results, represented as:

$$S_A = \begin{cases} 1 & \text{Team A win} \\ 0 & \text{o.w.} \end{cases}$$

- **Probabilistic model:** the probability of Team A beating Team B is:

$$P(\text{Team A win}) = E_A = \sigma(r_{A,B}) \;\; \text{with} \;\; \sigma(r) = \frac{1}{10^{-\frac{r}{s}}+1} \;\; \text{and} \;\; r_{A,B} = R_A - R_B$$

$$P(\text{Team B win}) = \sigma(-r_{A,B}) = 1 - \sigma(r_{A,B})$$

Mathematics of the Elo Model (cont.)

Now, updating the model is equivalent to **minimizing the log loss of the prediction**, represented by:

$$\ell = \begin{cases} -\log \sigma(r_{A,B}) & S_A = 1 \\ -\log \sigma(-r_{A,B}) & S_A = 0 \end{cases}$$

which we can minimize with **stochastic gradient descent**

$$R'_A = R_A - \eta \frac{d\ell}{dR_A} = \begin{cases} R_A + \eta \frac{d}{dR_A} \log \sigma(r_{A,B}) & S_A = 1 \\ R_A - \eta \frac{d}{dR_A} \log \sigma(-r_{A,B}) & S_A = 0 \end{cases}$$

resulting in our **update rule**

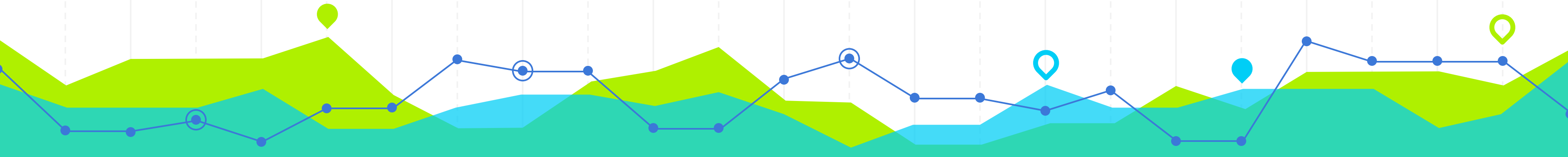
$$R'_A = \begin{cases} R_A + K\sigma(-r_{A,B}) & S_A = 1 \\ R_A - K\sigma(r_{A,B}) & S_A = 0 \end{cases} \text{ with } K = \frac{\eta \log 10}{s}$$

or, more compactly

$$R'_A = R_A + K(S_A - E_A)$$

Motivation

- While these aforementioned frameworks are effective at predicting games, they are **slow to react to team trends**
  - E.g. injuries, changes in team chemistry, etc.
- One reason is because they are based on the **assumption of normal, logistic, or binomial data**
  - Supposes that the expected performances of a team changes *slowly* over time
- Yet streaks are seemingly predictive of future results, both empirically and theoretically



2

Project Outline and Goals

How can we use streaks to improve game predictions?

Outline of Project

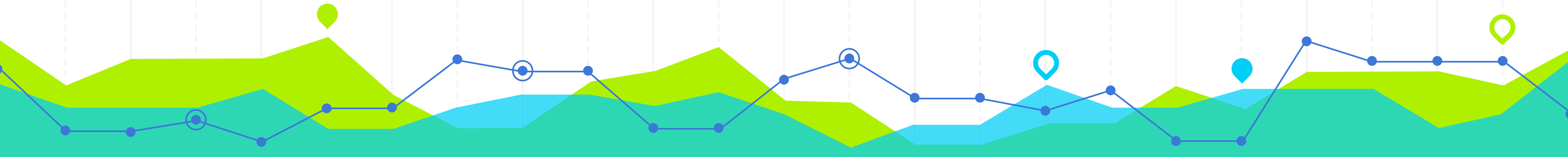
Goal: Increase the sensitivity of NBA models so that they better incorporate short-run performance, aka “streakiness”

Justification: Prove that the “Hot Hand” does exist and significantly impacts “success” probability for NBA teams

Method: Determine aptly powerful and unbiased measures of streakiness

Application: Utilize these measures to update the Elo ratings of NBA teams more dynamically by adjusting *K* parameter of the update equation

$$R'_A = R_A + K(S_A - E_A)$$



3

The Hot Hand Fallacy

Why are previous rejections of the “Hot Hand” statistically flawed?

Binomial Representation of NBA Season

Null Hypothesis: Teams perform relatively uniformly across a season, winning each game with probability equal to their winning percentage. Thus, we can represent the games as a **sequence of Bernoulli trials for *n* random variables**

$$\mathbf{W} = \{W_1, \dots W_n\} = \{W_i\}_{i=1}^n$$

Equivalent to a **Binomial Distribution**

$$B(n, p)$$

with the following properties

- **Parameters:** win probability *p* for each of the *n* **independent** games
- **Mean:** $E[\mathbf{W}] = np$
- **Variance:** $\mathrm{Var}[\mathbf{W}] = np(1-p)$
- **PMF:** $P(\mathbf{W} = k) = \binom{n}{k} p^k (1-p)^{n-k}$

The Hot Hand Fallacy

The idea of the “Hot Hand Fallacy” was popularized by Gilovich, Vallone, and Tversky (**GVT**) in their 1985 paper “The Hot Hand in Basketball: On the Misperception of Random Sequences.”

- **Goal:** Determine whether empirical and experimental data on shooting streaks differ from binomial expectations.
- **Method:** Calculate the percentage of makes after streaks of three makes and misses, and complete a paired t-test
- **Results:** 49% shooting percentage on “Hot Streaks” and 45% on “Cold Streaks”
- **Conclusion:** Using a paired t-test, this difference is not significant and the hot hand is “a powerful and widely shared cognitive illusion”

Theorem: Streak Selection Bias

Let $\mathbf{W} = \{W_i\}_{i=1}^{n}, n \geq 3$ be a sequence of independent Bernoulli trials, each with equal probability of success $0 < p < 1$.

Then, say $S_k(\mathbf{W})$ represents the subset of trials that immediately follow k consecutive successes, i.e. a "run" of length k, i.e.

$$S_k(\mathbf{W}) := \{i : \prod_{j=i-k}^{i-1} X_j = 1\} \subseteq \{k+1, \ldots, n\}$$

Furthermore, define $\hat{P}_k(\mathbf{B})$ to be the proportion of successes of trials in $S_k(\mathbf{W})$, i.e.

$$\hat{P}_k(\mathbf{W}) := \frac{\sum_{i \in S_k(\mathbf{W})} W_i}{|S_k(\mathbf{W})|}$$

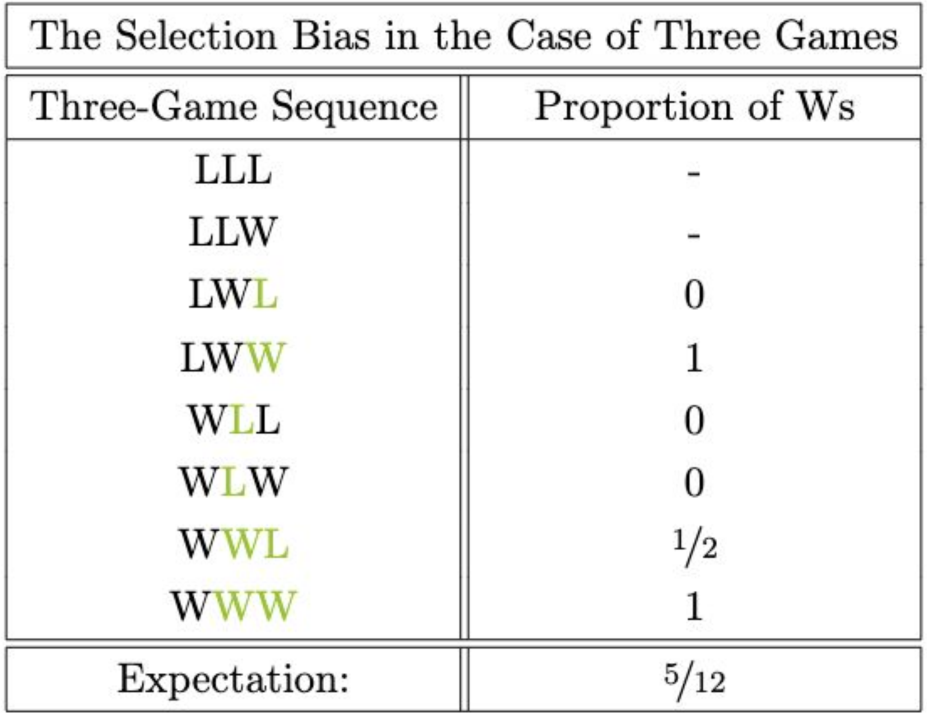
In this setting, $\hat{P}_k(\mathbf{W})$ is a biased estimator of the probability that an independent trial $W_t = 1$ given that the preceding k $(1 \leq k \leq n-2)$ trials were 1, i.e.

$$P(X_t = 1 \mid \prod_{j=t-k}^{t-1} W_j = 1) = p$$

In fact, we can say

$$E[\hat{P}_k(\mathbf{W}) \mid S_k(\mathbf{W}) \neq \emptyset] < p$$

Intuitively, this means that we expect a lower probability of success for this biased sampling technique.



The Selection Bias in the Case of Three Games	
Three-Game Sequence	Proportion of Ws
LLL	-
LLW	-
LWL	0
LWW	1
WLL	0
WLW	0
WWL	$^1/_2$
WWW	1
Expectation:	$^5/_{12}$

A simple example of the selection bias for a *n = 3* games and “streaks” of length *k = 1*

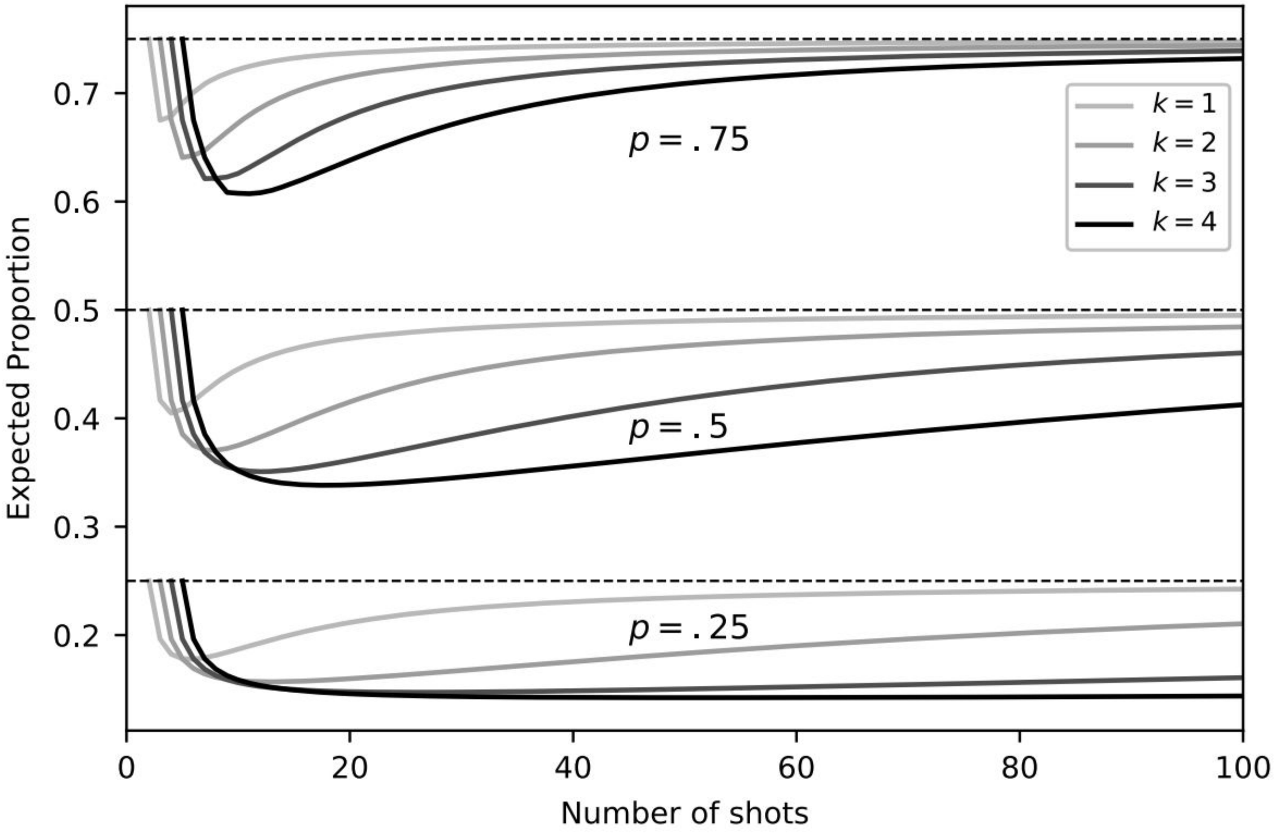


FIGURE 1.—The expected value of the proportion of successes on trials that immediately follow k consecutive successes, $\hat{P}_k(\mathbf{X})$, as a function of the total number of trials n, for different values of k and probabilities of success p, using the formula provided in Supplemental Material Appendix E.1 (Miller and Sanjurjo (2018)).

Proof Sketch

First, we convert our expected value into a conditional probability where τ is a random trial drawn uniformly from all trials following a win streak of length *k*

$$E[\hat{P}_k(\mathbf{W}) \mid S_k(\mathbf{W}) \neq \emptyset] = P(W_\tau = 1 \mid S_k(\mathbf{W}) \neq \emptyset)$$

which we can then break down into a sum over all possible selections $\tau = t$

$$P(W_\tau = 1 \mid S_k(\mathbf{W}) \neq \emptyset) = \sum_{t=k+1}^{n} P(W_t = 1 \mid \tau = t, \prod_{i=t-k}^{t-1} W_i = 1) P(\tau = t \mid \prod_{i=t-k}^{t-1} W_i = 1)$$

Here, the rightmost probabilities sum to 1, so we just need to show the left is less than *p* in all iterations of the sum. Using Bayes's rule, we can get

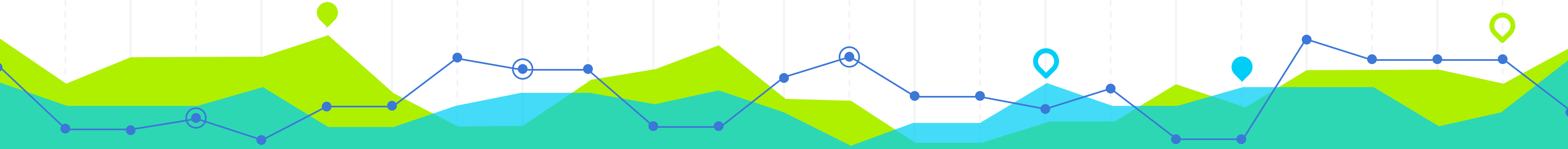
$$\frac{P(W_t{=}1 | \tau{=}t, \prod_{i=t-k}^{t-1} W_i{=}1)}{1{-}P(W_t{=}1 | \tau{=}t, \prod_{i=t-k}^{t-1} W_i{=}1)} = \frac{P(\tau{=}t | W_t{=}1, \prod_{i=t-k}^{t-1} W_i{=}1)}{P(\tau{=}t | W_t{=}0, \prod_{i=t-k}^{t-1} W_i{=}1)} \times \frac{p}{1{-}p}$$

so the left is less than *p* as long as

$$P(\tau = t \mid W_t = 1, \prod_{i=t-k}^{t-1} W_i = 1) < P(\tau = t \mid W_t = 0, \prod_{i=t-k}^{t-1} W_i = 1)$$

which is true because there are more possibilities for τ when $W_t = 1$.

My ultimate conclusions on the Hot Hand…



4

Measuring Streakiness

So what makes a good measure for “streakiness” or “clumpiness” in sports?

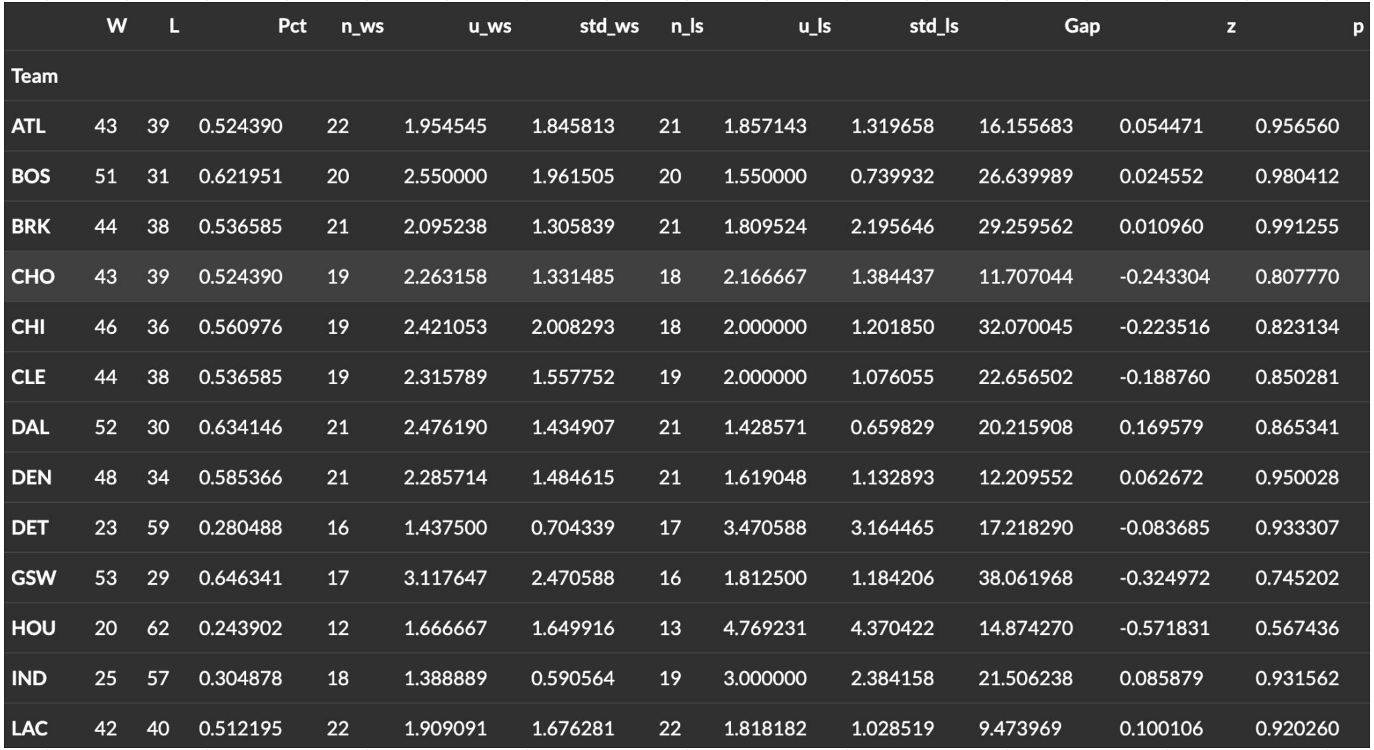
Wald-Wolfowitz Runs Test

This test counts the number of *runs* or *streaks*, i.e. subsequences made up of equal values, e.g.

L W W W W W L W L W

with the following properties

- **Null hypothesis:** The elements are independently drawn from the same distribution, i.e. the sequence is binomial
- **Mean:** $\mu_r = \frac{2n_W n_L}{n_W+n_L} + 1$
- **Variance:** $\sigma_r^2 = \frac{2n_W n_L(2n_W n_L-(n_W+n_L))}{(n_W+n_L)^2(n_W+n_L-1)} = \frac{(\mu-1)(\mu-2)}{n_W+n_L-1}$
- **Standard score:** $z = \frac{r-\mu_r}{\sigma_r}$
- **p-value:** $p = \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{-|z|} e^{-x^2/2} \mathrm{d}x$



	W	L	Pct	n_ws	u_ws	std_ws	n_ls	u_ls	std_ls	Gap	z	p
Team												
ATL	43	39	0.524390	22	1.954545	1.845813	21	1.857143	1.319658	16.155683	0.054471	0.956560
BOS	51	31	0.621951	20	2.550000	1.961505	20	1.550000	0.739932	26.639989	0.024552	0.980412
BRK	44	38	0.536585	21	2.095238	1.305839	21	1.809524	2.195646	29.259562	0.010960	0.991255
CHO	43	39	0.524390	19	2.263158	1.331485	18	2.166667	1.384437	11.707044	-0.243304	0.807770
CHI	46	36	0.560976	19	2.421053	2.008293	18	2.000000	1.201850	32.070045	-0.223516	0.823134
CLE	44	38	0.536585	19	2.315789	1.557752	19	2.000000	1.076055	22.656502	-0.188760	0.850281
DAL	52	30	0.634146	21	2.476190	1.434907	21	1.428571	0.659829	20.215908	0.169579	0.865341
DEN	48	34	0.585366	21	2.285714	1.484615	21	1.619048	1.132893	12.209552	0.062672	0.950028
DET	23	59	0.280488	16	1.437500	0.704339	17	3.470588	3.164465	17.218290	-0.083685	0.933307
GSW	53	29	0.646341	17	3.117647	2.470588	16	1.812500	1.184206	38.061968	-0.324972	0.745202
HOU	20	62	0.243902	12	1.666667	1.649916	13	4.769231	4.370422	14.874270	-0.571831	0.567436
IND	25	57	0.304878	18	1.388889	0.590564	19	3.000000	2.384158	21.506238	0.085879	0.931562
LAC	42	40	0.512195	22	1.909091	1.676281	22	1.818182	1.028519	9.473969	0.100106	0.920260

Empirical calculations of WW Runs Test statistics for NBA teams in 2021-22

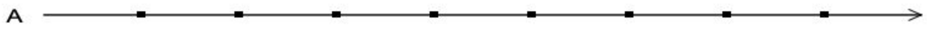
Potential Deviations from Binomial

1. **Autocorrelation:** the dependence of outcomes on the previous outcome
   a. E.g. a team who wins more frequently after recent success than after a losing spell
2. **Non-stationarity:** the probability of success changes across the trials
   a. E.g. a team who makes a trade or has injuries

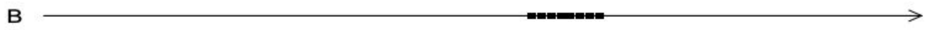
The **Runs Test** does not address **non-stationarity** well given its exclusive focus on runs, resulting in **low power** to reject the null hypothesis.

Desired Properties of Measures

1. **Minimum:** The measure should be minimized when events (wins/losses) are equally spaced

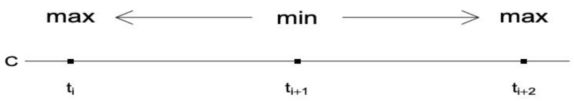


2. **Maximum:** The measure should be maximized if all events are clumped together

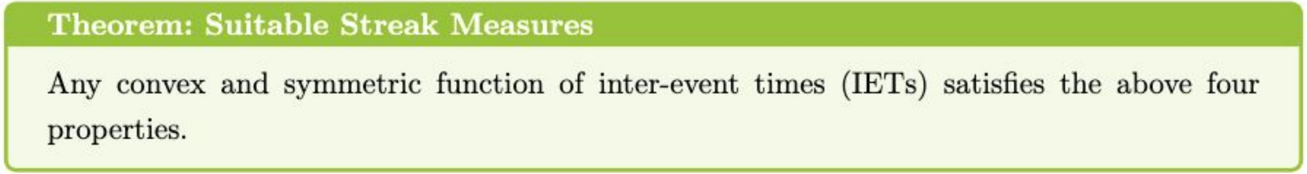


3. **Continuity:** A small shift in events should only change the measure a small amount

4. **Convergence:** As events move closer, the measure should decrease



Potential Measure Candidates



Theorem: Suitable Streak Measures

Any convex and symmetric function of inter-event times (IETs) satisfies the above four properties.

1. **Second moment:** Often utilized to describe and distinguish probability distributions

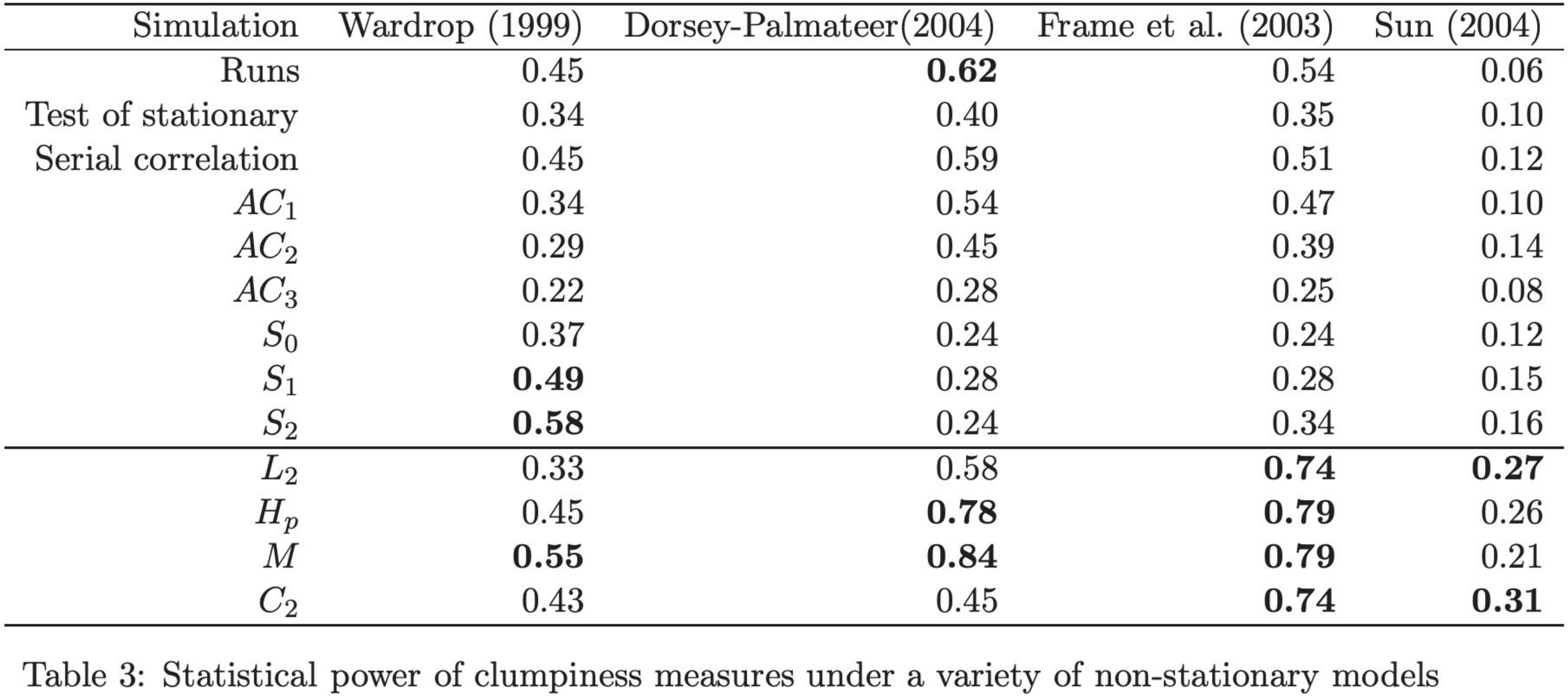
$$L_2 = \sum_{i=1}^{n+1} x_i^2$$

2. **Entropy:** A measure of uncertainty and disorder used in information theory

$$H_p = \sum_{i=1}^{n+1} x_i \log x_i$$

3. **Log utility:** Used in economics, it “normalizes” the ranges being considered

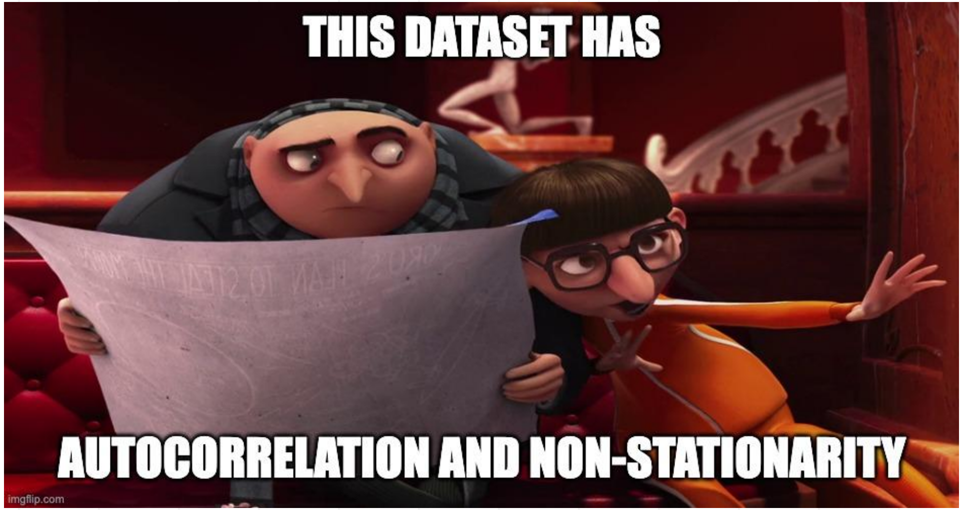
$$M = -\sum_{i=1}^{n+1} \log(x_i)$$

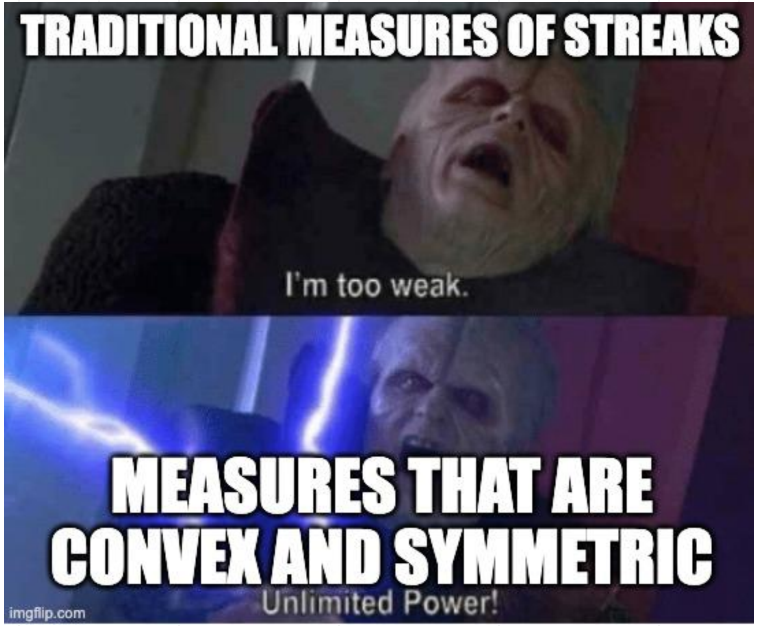


Simulation	Wardrop (1999)	Dorsey-Palmateer(2004)	Frame et al. (2003)	Sun (2004)
Runs	0.45	**0.62**	0.54	0.06
Test of stationary	0.34	0.40	0.35	0.10
Serial correlation	0.45	0.59	0.51	0.12
AC_1	0.34	0.54	0.47	0.10
AC_2	0.29	0.45	0.39	0.14
AC_3	0.22	0.28	0.25	0.08
S_0	0.37	0.24	0.24	0.12
S_1	**0.49**	0.28	0.28	0.15
S_2	**0.58**	0.24	0.34	0.16
L_2	0.33	0.58	**0.74**	**0.27**
H_p	0.45	**0.78**	**0.79**	0.26
M	**0.55**	**0.84**	**0.79**	0.21
C_2	0.43	0.45	**0.74**	**0.31**

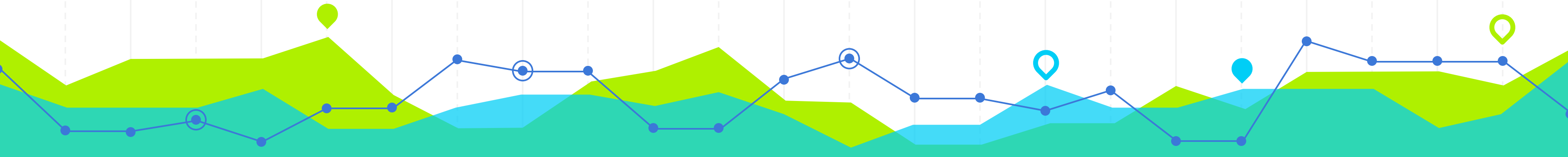
Table 3: Statistical power of clumpiness measures under a variety of non-stationary models

Chart from Zhang, Bradlow, and Small (2013) showing underpowered streak tests




TRADITIONAL MEASURES OF STREAKS
I'm too weak.
MEASURES THAT ARE CONVEX AND SYMMETRIC
Unlimited Power!
imgflip.com

More conclusions…



5 Simulations and Results

How can we utilize simulations to apply these measures to real data?

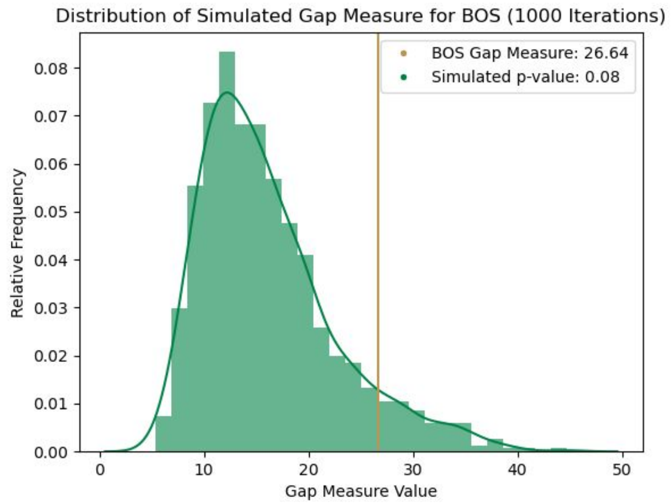
Permutation and Monte Carlo Tests

A **permutation test** is an exact statistical hypothesis test that establishes a distribution of the test statistic by calculating **all possible values** under **rearrangements** of the **observed data**. It is

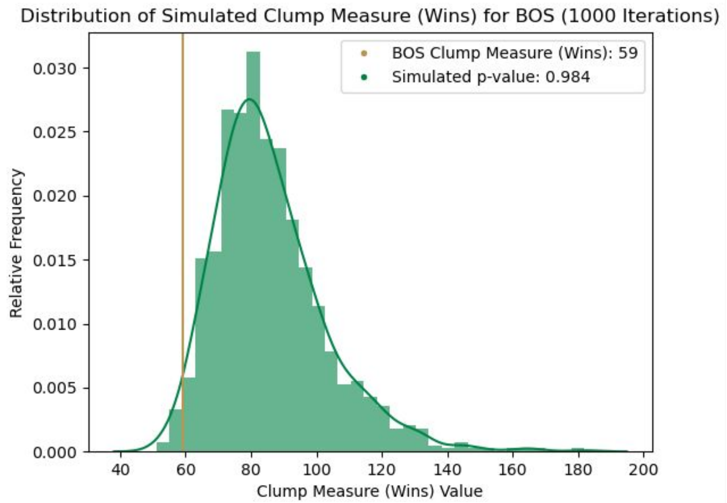
- Non-parametric (unlike traditional t-test, F-test, z-test)
- Computationally intensive and challenging to get confidence intervals and other distribution information

Monte Carlo sampling is an **asymptotically equivalent** permutation test that takes a **random sample** of possible permutations. It can be used for a **confidence interval** for the p-value of the Binomial distribution, e.g. for 10000 random permutations and an estimated p-value of 0.05, the CI would be

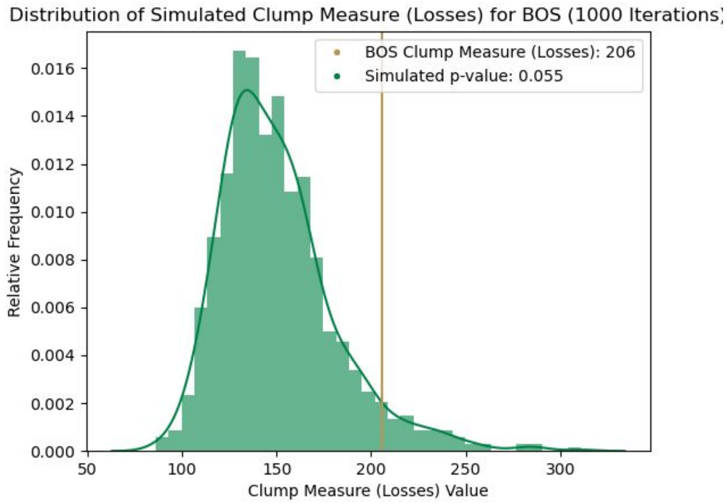
$$\left[\hat{p} - z\sqrt{\frac{0.05(1-0.05)}{10000}}, \hat{p} + z\sqrt{\frac{0.05(1-0.05)}{10000}}\right] = [0.045, 0.055]$$



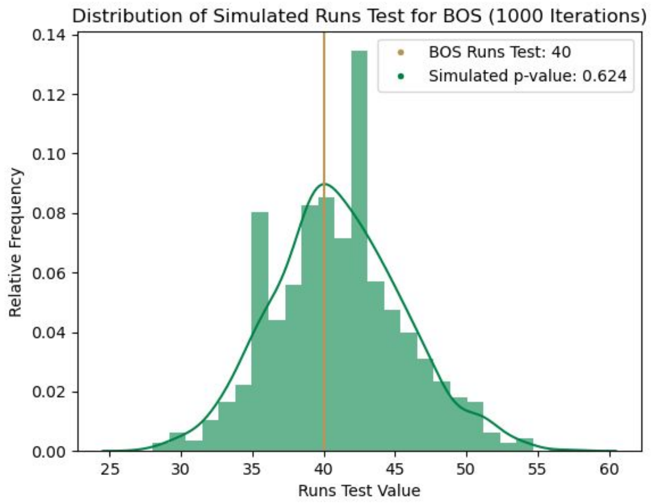
Distribution of Simulated Gap Measure for BOS (1000 Iterations)



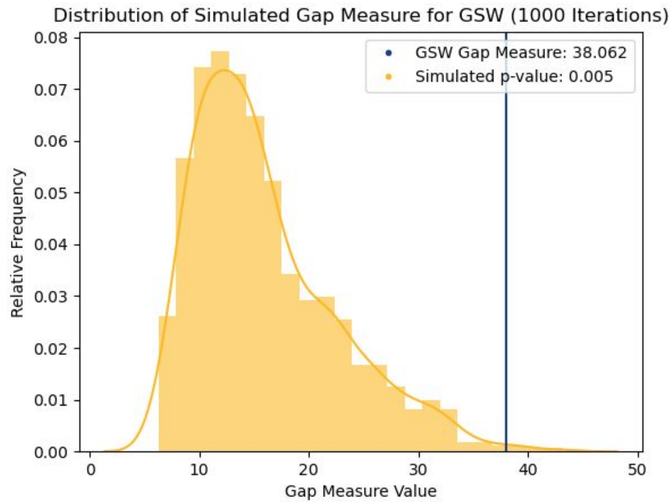
Distribution of Simulated Clump Measure (Wins) for BOS (1000 Iterations)



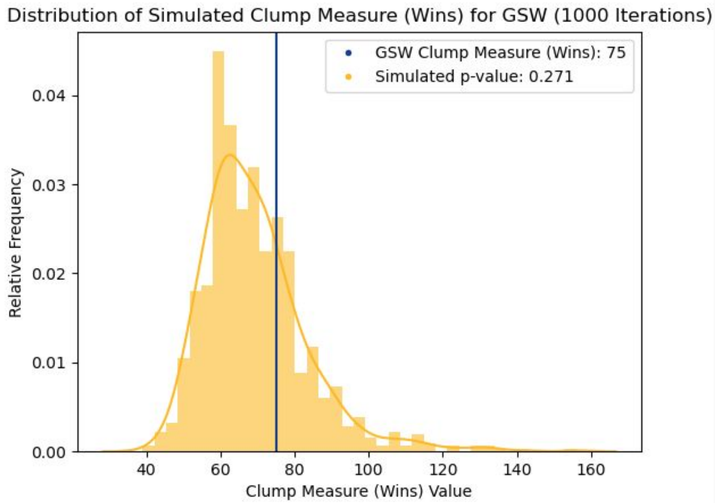
Distribution of Simulated Clump Measure (Losses) for BOS (1000 Iterations)



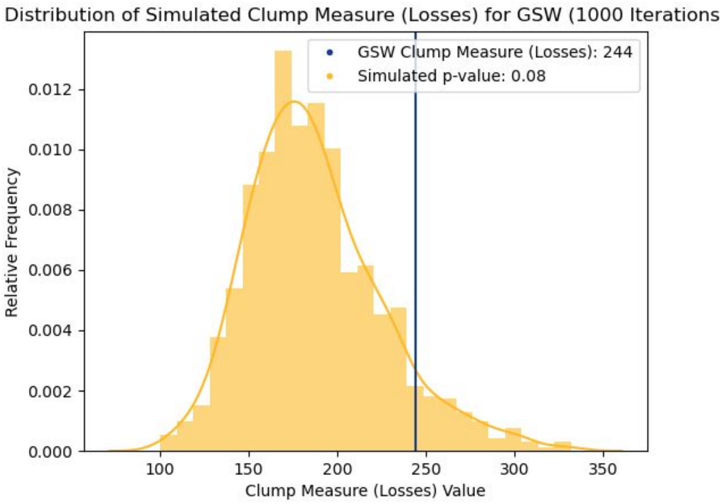
Distribution of Simulated Runs Test for BOS (1000 Iterations)



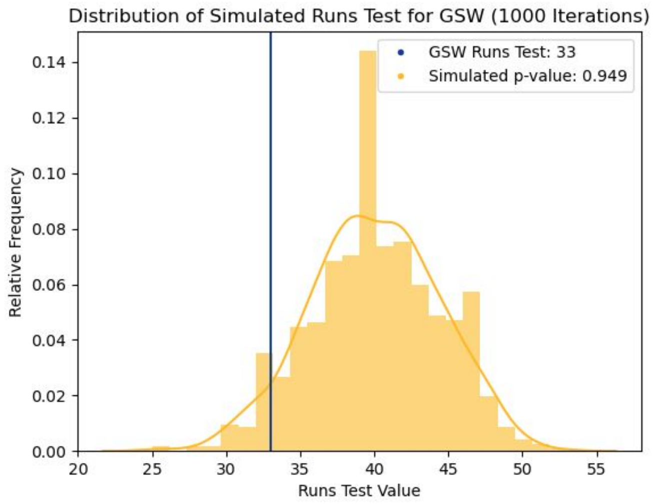
Distribution of Simulated Gap Measure for GSW (1000 Iterations)



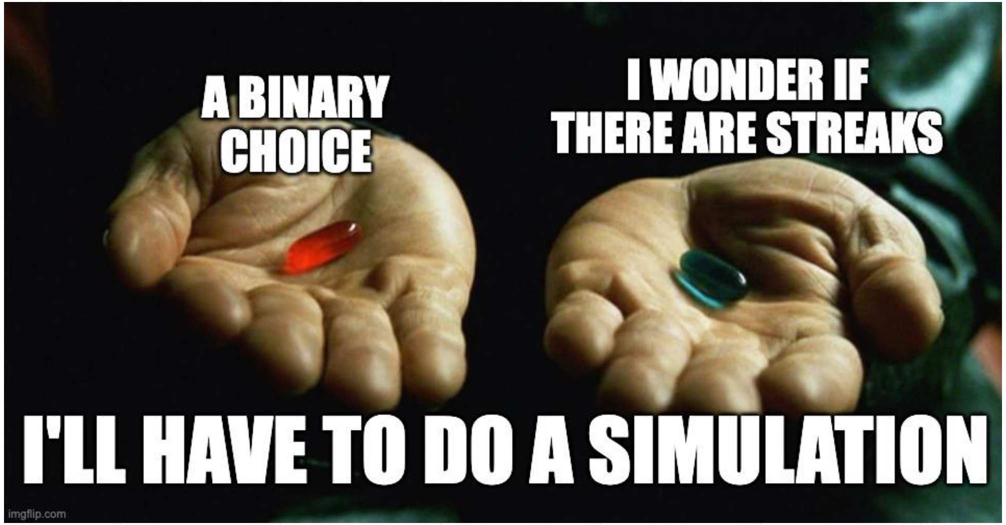
Distribution of Simulated Clump Measure (Wins) for GSW (1000 Iterations)



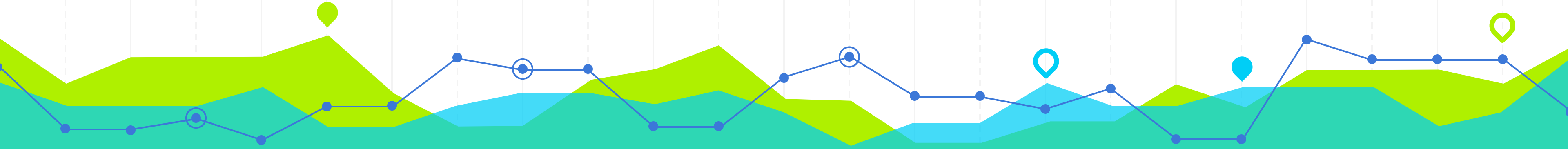
Distribution of Simulated Clump Measure (Losses) for GSW (1000 Iterations



Distribution of Simulated Runs Test for GSW (1000 Iterations)



Some more conclusions…



6 Future Directions and Applications to Modeling

Where do we go from here?

Future Directions and Applications

- **Gap Measure:** A measure of how far your local win percentages are from your expectation up to that point

$$\text{gap} = \sqrt{\textstyle\sum_{i=1}^{n} \left[\left(\sum_{j=1}^{i} w_j\right) - i * p\right]^2}$$

- **Refining the Elo models:** Determine the best way to implement these measures to increase accuracy of Elo models
  - Set *k* to be equal to one of the measures or some function of them
  - Categorize which *k* to use based on streakiness
- **Test with more real data:** determine where these new models excel and have deficiencies

Bibliography

[1] NBA Logo. url: https://www.nba.com.

[2] Cal Ripken Quote. AZ Quotes. url: https://www.azquotes.com/quote/712065.

[3] Bethlehem Shoals. The 25 Most Unbreakable Streaks in Sports. GQ. url: https://www.gq.com/gallery/greatest-sports-streaks.

[4] Raghav Mittal. What is an ELO Rating? Medium. url: https://medium.com/purple-theory/what-is-elo-rating-c4eb7a9061e0.

[5] Nate Silver and Reuben Fischer-Baum. How We Calculate NBA Elo Ratings. FiveThirtyEight. url: https://fivethirtyeight.com/features/how- we- calculate- nba- elo-ratings/.

[6] How Our NFL Predictions Work. FiveThirtyEight. url: https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/.

[7] ESPN Analytics. ESPN’s Basketball Power Index. ESPN. url: https://www.espn.com/nba/story/_/page/Basketball-Power-Index/espn-nba-basketball-power-index.

[8] Mike Lynch. SRS Calculation Details. Sports Reference. url: https : / / www.sports-reference.com/blog/2015/03/srs-calculation-details/.

[9] Sharon Katz. How ESPN’s NFL Football Power Index was developed. ESPN. url: https://www.espn.com/nfl/story/_/id/13539941/how- espn- nfl- football- power-index-was-developed-implemented.

[10] Sascha Wilkens. “Sports prediction and betting models in the Machine Learning Age: The case of tennis”. In: Journal of Sports Analytics 7.2 (2021), pp. 99–117. doi: 10.3233/jsa-200463

[11] Joshua B. Miller and Adam Sanjurjo. “Surprised by the Gambler’s and Hot Hand fallacy? A Truth in the Law of Small Numbers”. In: Econometrica 86.6 (Dec. 2018), pp. 2019–2047. doi: 10.2139/ssrn.2627354.

[12] Thomas Gilovich, Robert Vallone, and Amos Tversky. “The hot hand in basketball: On the misperception of random sequences”. In: Cognitive Psychology 17.3 (1985), pp. 295–314. doi: 10.1016/0010-0285(85)90010-6.

[13] Yao Zhang, Eric T. Bradlow, and Dylan S. Small. “New Measures of Clumpiness for Incidence Data”. In: Journal of Applied Statistics 40.11 (2013), pp. 2533–2548. doi: 10.1080/02664763.2013.818627.

[14] Patrick R. McMullen. “Standardization of winning streaks in sports”. In: Applied Mathematics 08.03 (2017), pp. 344–357. doi: 10.4236/am.2017.83029.

[15] Jim Albert. Streaky and Consistent Teams? Sept. 2017. url: https://baseballwithr.wordpress.com/2017/09/18/streaky-and-consistent-teams/.

Bernie
I am once again asking
for you to laugh at my memes and
clap for the end of my presentation...
imgflip.com

Last one I promise…

THANKS!

Any questions?

You can also email:

harry.jain@yale.edu

bit.ly/3Fd3wlU