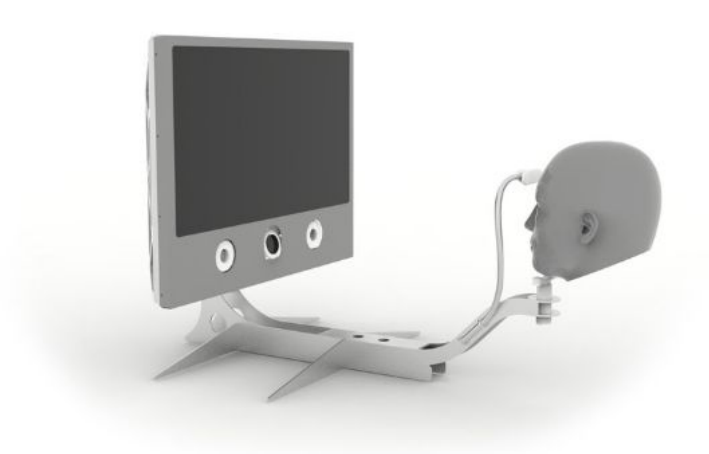
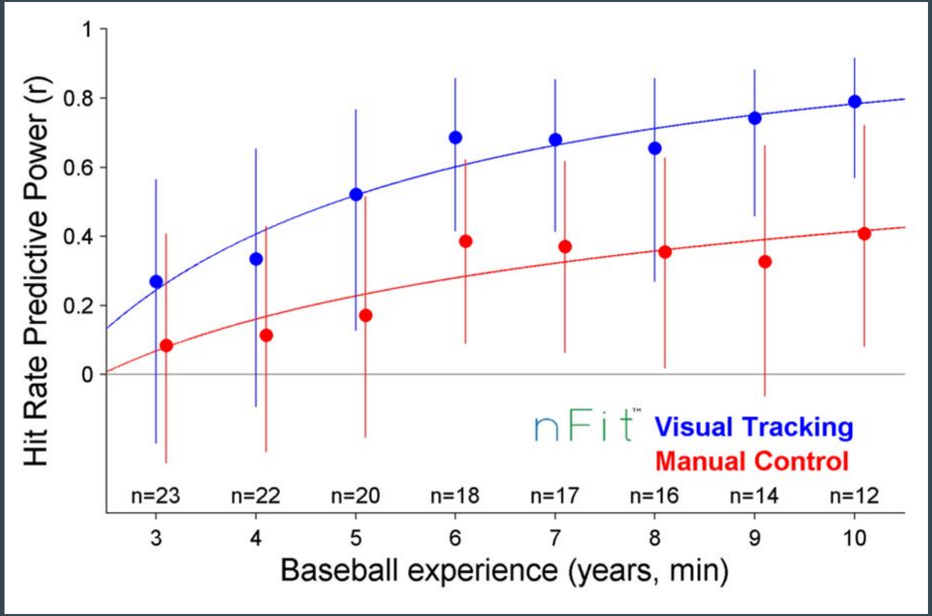
Data 8 Connector: Sports Analytics

Random Coincidence

neuroFit





A Review…

Our main tools so far:

1. Expected value/Standard Deviation
2. Correlation
3. Linear fit/regression
4. Descriptive statistics

Expected Value

We all have our own expectations for what will or will not happen

Expected value is the primary concept for evaluating an observation

- ➔ Did the batter’s performance exceed average performance?
- ➔ Did the shooter underperform from 3?
- ➔ How many possessions do we think a player used given his total FTAs? (.44 × FTA)

A lot of our modeling relies on expected values

We used data to compute expectations

- ➔ Utilize sufficient conditioning or bucketing for reliable effects

Related: we show dispersion around EV with standard deviation

Standard Deviation

Other than the expected value, we need to know the dispersion of data

Standard deviation is a widely used metric for measuring dispersion

Our notable usage was for quantifying errors in relations between metrics and runs scored

Correlation

Correlation is for measuring the strength of association between two variables

A powerful tool for tracking how relevant a metric is with the target

- ➔ BA, OBP, SLG and other metrics related to run scoring
- ➔ FG%, eFG%, TS% and offensive rating
- ➔ Dean Oliver's four factors and net rating

Correlation is not directional, but logic/knowledge can infer direction

- ➔ Batting performance leads to run scoring, hence BA/OBP/SLG → Run scoring
- ➔ More efficient shooting than opposition (first factor) → positive net rating
- ➔ However: batting performance does not perfectly correlate due to imperfect measurement and sequencing
- ➔ Shooting performance does not perfectly correlate due to other factors

Linear Fit/Regression

A linear fit encodes the linear relationship between two variables

Correlation is closely linked to linear fits

- ➔ Small errors and higher slope → stronger correlation

Correlation quantifies the association, linear fit gives the functional relationship

We used linear fits for

- ➔ Estimating relationship between runs/points and wins
- ➔ Measuring performance of batting metrics
- ➔ Explanatory power of Four Factor model

Descriptive Statistics

Overall, descriptive statistics is a lot of what we've been doing

What is "Descriptive Statistics"?

Using statistical methodology to describe observed phenomena

- Often as simple as summaries like mean, median, range, etc

We did a lot of powerful descriptive statistics to reveal the nature of sports

- In case you were thinking descriptive statistics is somehow bad or inferior

Descriptive Statistics

We described...

- ➔ The relationship between runs/points and wins
- ➔ Expected runs and run values for events
- ➔ Relationship between metrics and performance
- ➔ Shooting performance in relation to shot distance and defender distance
- ➔ And so on

New Tools

We need some new tools to go further

1. Regression (multiple input values instead of one)
2. Sampling: bootstrapping, permutations, cross-validation
3. Prediction/projection/forecasting (regression to the mean)
4. Inferential statistics

We'll outline here and go through them over the next few weeks

Multiple Regression

Multiple regression is a powerful tool for more sophisticated modeling of relationships

The relationship between an observation and inputs is given by:

$$\text{Observation} = \alpha + \beta_1 \cdot \text{Input}_1 + \cdots + \beta_k \cdot \text{Input}_k + \text{Error}$$

We did something very similar with run values (note the similarity to LWTS)

But we can go further by using regression modeling

https://www.inferentialthinking.com/chapters/17/6/multiple-regression.html

Multiple Regression

Regression modeling will allow us to handle more complicated situations

Capture the marginal impact of different inputs that are acting simultaneously

- Useful when you can’t isolate a variable to test its impact
- Ex: we can’t control lineups and swap a player on/off to test impact in a controlled environment

Of course, there is no such thing as a free lunch

We will also see some good examples of how the model fit can “fail” and offer some remedies

Sampling Methods

We only see one of an infinite number of possible outcomes to a game/season

Randomization lets us explore more than one outcome

- We randomize our data to “shake” our analytic results

We can then quantify uncertainty or test hypotheses using the new datasets

https://www.inferentialthinking.com/chapters/13/2/bootstrap.html

Sampling Methods

Randomization for quantifying our uncertainty

- When we quote an advanced stat, we quote a single number
- Given all the underlying measurements, how much uncertainty/variation is there in the stat?
- Among all the possible outcomes, what sorts of values for the stat *could* we have seen?

Testing hypotheses

- Hypotheses are often binary: is there a difference between two populations or not?
- Does a phenomenon exist or not?
    - For example: are the leagues different? Do players get “hot”?

Prediction/Projection

We’ve done a lot of *ex-post* analysis using descriptive statistics

If we’re building a team, we need to predict or project performance

Perhaps the most fundamental approach is *regression to the mean*

- ➔ Observations cluster around the expected value of the population (true value)
- ➔ The observation after a higher than usual observation will likely be lower
- ➔ Similar for lower than usual values

Inferential Statistics

We won’t abandon descriptive statistics, but we want to start inferring or generalizing

Our data is just a sample of observations from a *population*

Suppose we poll 1,000 random Americans

➔ Our sample is the 1,000 people
➔ The population is all Americans

We observe one season in the NBA

➔ Our sample is all the games, player performances, etc
➔ The population is the infinite number of games that *could* be played between all the teams and players

Inferential Statistics

A couple types of inference problems:

- ➔ Test a hypothesis about the population
  Ex: there is a “hot hand” phenomenon in NBA shooting performance
- ➔ Provide an interval estimate incorporating uncertainty
  Ex: Aaron Judge had 8.2 WAR in 2017. What’s a plausible uncertainty interval?

Testing Hypotheses: The Hot Hand

The Hot Hand

What is the hot hand?

The basic idea is that a player can get “hot” and have a higher likelihood of hitting shots

A player like Klay is commonly said to be a streaky shooter

- And therefore he gets hot. And apparently when he’s hot he’s better than anyone ever

A good example is the NBA Jam On Fire mode

The Hot Hand

There’s been a lot of work on the Hot Hand

One of the most famous is from Amos Tversky (of The Undoing Project)

Tversky popularized the idea that the Hot Hand is a cognitive error

- ➔ Kahneman and Tversky did extensive work on cognitive bias/error/fallacies
- ➔ Humans (even fairly well trained statistical types) don’t intuitively “understand” randomness
- ➔ Ex: we underestimate how much “clumping” there is in random data

More recent work suggests there is a hot hand

- ➔ Using play-by-play or player tracking datasets, detected recent performance has predictive power on the next at-bat or shot taken

The Hot Hand

What's the conclusion? Some analyses say yes, some say no.

An analysis from Harvard shows player's believe in the hot hand

➔ There is a heat check shot: their shot difficulty goes up if they think they're hot

The analysis also shows (accounting for shot difficulty and defender distance), recent performance in the last 4 shots increases the likelihood of making the next shot

If you make 1 extra shot in your last 4 shots. How much do we expect your shooting percentage to increase on the next shot?

The Hot Hand

1 made shot in the last 4 yields increase in shooting percentage of about 1.2 pct points

The baseball study shows increase of about 30 points in batting average when a player is hot (or about 3 pct points)

In my opinion, these effects may well be correct but they aren’t capturing the popular notion of the hot hand (which is a very large effect)

Testing the Hot Hand

How do we test the hot hand?

What do we measure, ie what is our “test statistic”?

How does one test a hypothesis in general?

- What’s the data?
- What’s the null/alternative hypothesis?
- What’s the statistic? Does the statistic make sense for what we want to measure?
- How should the statistic behave under the null hypothesis?
- How does the observed statistic from data compare to the statistic under the null hypothesis?
- Other considerations: what chance do we have of even rejecting the hypothesis?

Testing the Hot Hand

To test a hypothesis, we first have to

- ➔ Determine our data: we’re going to use shooting performances from Klay
- ➔ Determine our null/alternative hypothesis: there is no hot hand vs there is a hot hand

But there are a lot of other things to address:

- ➔ What statistic should we use to test the hypothesis? Does the test statistic make sense for what we want to measure?
- ➔ How does the test statistic behave under the null hypothesis?
- ➔ How does the observed value from data compare to the test statistic under the null hypothesis?
- ➔ Given the size of our dataset and the potential size of the hot hand effect, what chance do we have of even detecting the hot hand and thus rejecting the null hypothesis?

Testing the Hot Hand

Let’s address in a demo where it’ll make a lot more sense