LS88: Sports Analytics

New Toolbox

The Toolbox So Far

Current Tools

- ➔ Probability
- ➔ Expected Value
- ➔ Standard Deviation
- ➔ Correlation
- ➔ Regression/Linear Model/Line-of-Best-Fit
- ➔ Visualizations like histograms and heatmaps

Probability

The likelihood of an event

- ➔ Coin flip: .5 (50%) probability of coming heads or tails
- ➔ Weighted coin flip: general probability *p* of coming heads or tails
- ➔ HR / PA: How likely a HR is to be hit in a PA

Can weight the value of an event by it's likelihood:

$$\text{RE of a Steal Attempt} = p_{\text{Steal}} RE_{\text{After Steal}} + (1 - p_{\text{Steal}}) RE_{\text{Caught Stealing}}$$

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

Measure 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
- ➔ Performance does not perfectly correlate due to imperfect measurement and sequencing
- ➔ Shooting performance does not perfectly correlate due to other factors

Linear Fit/Regression

Encodes the linear relationship between two variables

Correlation is closely linked to linear fits

- ➔ Small errors and higher slope → stronger correlation

Correlation gives 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

Descriptive statistics is 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 location, distance and defender proximity
- ➔ 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

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 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

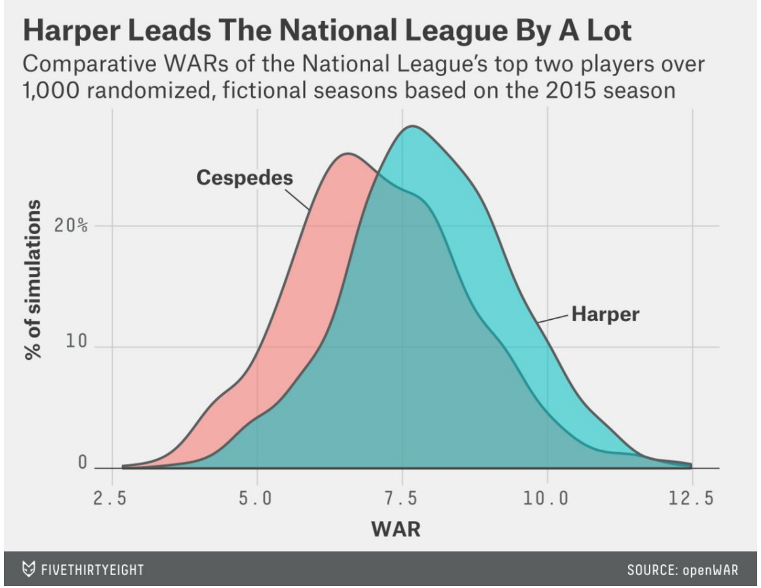
We observe one season in the NBA

- ➔ 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?


Harper Leads The National League By A Lot
Comparative WARs of the National League’s top two players over 1,000 randomized, fictional seasons based on the 2015 season
Cespedes
Harper
% of simulations
WAR
FIVETHIRTYEIGHT
SOURCE: openWAR