Science and Statistical Detection

Themes in Data Science and Predictive Analytics

Kenneth W. Regan[1]
University at Buffalo (SUNY)

WNY STEM Forum, September 2019

[1] With grateful acknowledgment to co-authors and UB's Center for Computational Research (CCR)

Predictive Analytics

A Predictive Analytic Model:

- Addresses events or decisions with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Should also assign *confidence intervals* for p_j and those quantities.

Examples of areas that use predictive models:

- Insurance
- Weather forecasting
- Investment managing
- Equity markets
- Betting—in particular, setting initial odds in horse racing etc.

In my model, the m_j are possible moves in a given chess position.

"Probable Structure" versus "Sure Structure"

Who has the highest need for precision and accuracy?

- An investment manager or CEO/CFO *should* play "60% shots."
- Weather forecaster saying "60% chance of rain"
  - would like to be accurate for a given day;
  - *needs* to be accurate over periods of time (or over geographic areas).
- Insurance company needs to gauge risk accurately to price policies competitively.
- Bookies need to set betting lines accurately. Else, **arbitrage**.
- But bookie can give long shots higher chances since betting *against* long shots gives minimal *leverage*.
- **My chess model needs total assurance with accusations that invove long odds.**
- Extreme corner of Data Science.

Prediction Factors and Skill

Two particular modeling situations:

- **Aptitude Model**: Projections depend primarily on estimates of the agents' fitness or skill:
  - E.g. *handicap* in golf or horse racing, PECOTA in baseball, QB rating in football...
  - **Elo Rating** in chess: Just One Number.
  - Difference $R_1 - R_2$ used to forecast match between players 1 and 2.
  - **FiveThirtyEight** uses Elo ratings for all sports forecasts.
- **Decision Model**: Project human choices, e.g. elections, toothpaste brands, ways to go downtown.
- Combination is an **Aptitude Decision Model**. E.g. for **Exams**.
- Aptitude measure can be **GPA**, **SAT scores**, **IQ**...
- Chess ratings based on results of games, and forecast results.
- Examinations judged by answer keys [and part-credit rubrics].
- Chess skill can be measured both ways. ("Intrinsic Performance Ratings")

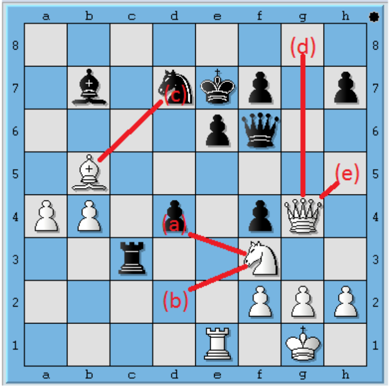
Chess and Tests: Prediction ≈ Grading

The ____ of drug-resistant strains of bacteria and viruses has ____ researchers' hopes that permanent victories against many diseases have been achieved.

- (a) vigor . . corroborated
- (b) feebleness . . dashed
- (c) proliferation . . blighted
- (d) destruction . . disputed
- (e) disappearance . . frustrated

(source: itunes.apple.com)

=



How the Chess Model Operates

(A) For each position π and legal move m_j—and for some setting of parameters correspondng to a player's rating etc.:

- Generate the projected probability p_j.
- Literally paint $1000 \cdot p_j$ faces of a 1,000-sided die with the move m_j. (Numerical precision to 3-4 places like FIDE used to do with ratings.) Then:
- *The die is cast.*

(B) This presumes **independence** between positions. Strictly speaking this doesn't hold—e.g., Carlsen-Anand double blunder involved Anand's fixation on pushing his a-pawn. But it is a *sparse dependence* that can be accounted as a reduction in the effective number N of game turns.

(C) The rest—including all statistical inference—is just analysis of (loaded) dice, known since the 1700s ("multinomial Bernoulli trials").

Confidence Intervals and Z-Scores

- The cheating-test quantities MM, ASD, etc., are all *averages* of (presumed-)independent events, hence by the Central Limit Theorem they conform as N grows to *normal distribution* (also called *Gaussian* or *standard* distribution or just "The Bell Curve").
- The theorem does not need the distributions $[\vec{p}_j]$ of moves in individual positions to be Gaussian or theoretically known at all—you just have to be good at projecting them.
- **Validation** checks the conformance empirically.
- Hence can use *z-scores*, not just general "p-values."

$$z = \frac{actual - projected}{\sigma}.$$

- The σ can be $\sigma_{projected}$ or σ_{actual} (from ambient "clean" data).

Using Z-Scores

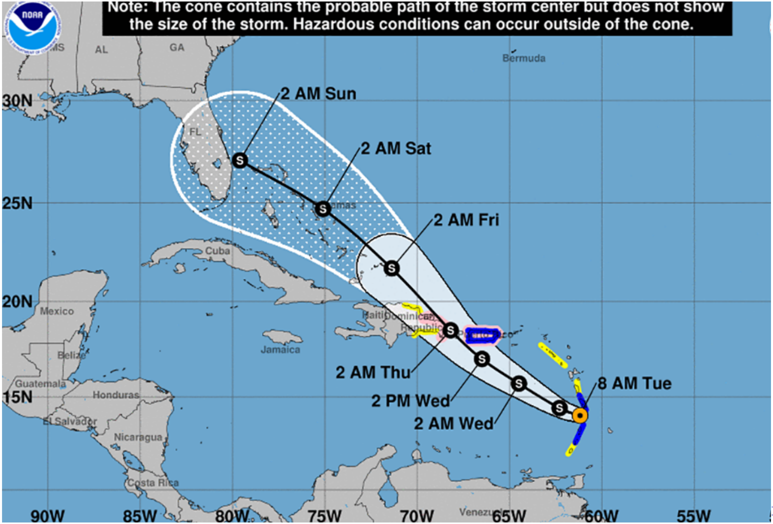
- A z-value expresses the deviation as a multiple of σ.
- Adjustment to z == adjustment to σ, which is $\propto \sqrt{N}$.
- The z-value gives "Face-Value odds" against the *null hypothesis* of the deviation occurring by natural chance.
- **z = 2.00: 1-in-44** odds, 2.275% natural frequency.
- **z = 3.00: 1-in-741** odds, 0.135% natural frequency.
- **z = 4.00: 1-in-31,754** odds, 3.167/100,000 natural frequency.
- **z = 5.00: 1-in-3,486,914** odds, 2.87/10,000,000 natural freq.
- Rough but helpful analogy to a Richter scale.
- But face-value odds need to be tempered against selection bias. Which can often be estimated, but mostly depends on *other evidence*.

Margin of Error As 95% Confidence

- Social convention enshrined in various policies. Usually works well.
- Almost coincides with $z \leq 2$ ("Two Sigma") when two-sided.
- Flip coin 100 times, $\sigma = \frac{\sqrt{100}}{2} = 5$, so interval is 40–60.
- Poll 1,600 in tossup election, $\sigma = 20$, so interval is $\pm 40 = \pm 2.5\%$.
- Poll only 900 people, $\sigma = 15$, $2\sigma = 30 = \pm 3.3\%$.
- Polls in-between all say they have a "3% margin of error."
- One-directional confidence "should be" $p < 0.025$ but academic publishing standards shade that down to $p < 0.05$, which is discounting 5% chance of being outside the 90% interval in the other direction.
- Civic law convention: results above two-sigma are admissible as evidence of irregularity.
- Social convention: OK to neglect events outside interval.
- **When do we need far more than 95% confidence?**

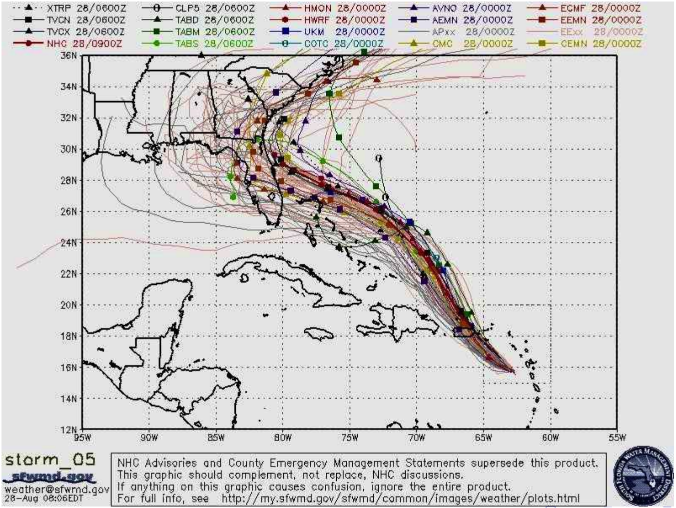
Example: Hurricanes 95% Uncertainty Cone

Hurricane Dorian, Friday Aug. 30, 2019.

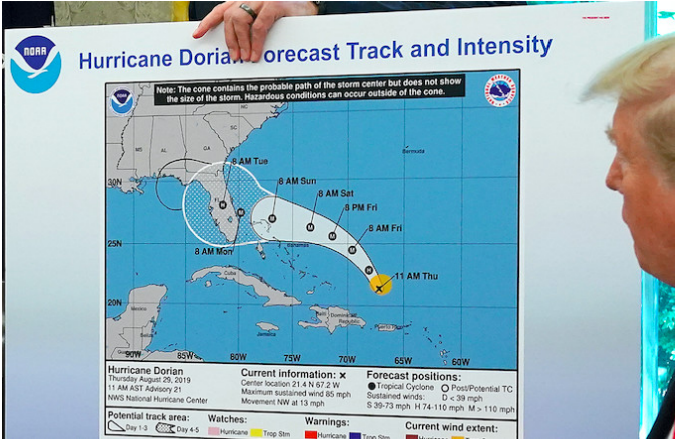


How Cone Was Obtained

Thousands of simulations of weather prediction model:



Cone Does Not Show Momentum



Other Topics For Discussion

- How to Expect the Unexpected.
- Is the “Hot Hand” Systematic?
- Confirmation Bias
- Selection Bias
- The Reproducibility Problem
- The P-Hacking Problem
- How to Judge Outliers
- Cross-Checking a Model
- Bias and Social Fairness
- Predictive Scoring

Decision Theory Models

- Usual situation is "one choice π, many choosers P":
  - brands of toothpaste
  - election candidates
  - bus or metro or cab (etc.) to downtown. (Daniel McFadden, 1960s BART study)
- Chess has "one chooser P, many choices π."
- Few other such situations have large data like mine (says colleague).
- But for each (P, π) choice, both situations modeled same way.
- **Log-Linear Model** (a.k.a. "multinomial logit"):

$$\begin{aligned} \log(p_j) &= \mathit{linear}(\text{utility } u_j), \quad \text{so} \\ p_j &= \mathit{softmax}(\vec{u}_j) = \mathit{normalize}(e^{\alpha+\beta\vec{u}_j}). \end{aligned}$$

- Won the 2000 Economics Nobel for McFadden—but **fails** a basic "sanity check" in chess.

LogLog-Linear Model

$$\log\log(\frac{1}{p_j}) = linear(u_j), \quad \text{so}$$
$$p_j = p_1^{\left(e^{\alpha+\beta\vec{u}_j}\right)}.$$

- The *normalize* step goes into determining p_1 first.
- Double-decker exponentiation—which invites dynamical chaos.
- Has just 1 mention in a 960-page textbook used for Machine Learning at UB.
- But works well in chess—deployed model is hard to improve on.
- I used to say it's like the Marshall with 11...c6 rather than 11...Nf6.
- Now that analogy understates the importance of sanity checks.

How well does your model perform on neighboring tests that it isn't specifically trained for? Like a cross-examination or stress test.

Cross-Validation Within the Model

- **MM** and **ASD** are expressly trained to be *unbiased estimators*. Means solving 2 equations in the two parameter unknowns s, c.
- The **EV** test is not directly fitted but is consistently biased *against false positives*—so safe to use. Does not give away too much.
- *Frequency of predicting the second move* (**M2**): actual typically 17–19% regardless of rating. Log-linear often projects under 12% (!!), deployed model usually close.
- **M3**, **M4**, **M5**... **Pass (quite close conformance)**.
- Predicting errors that are *slight*; *moderate*; *big*; *blunders*: **Pass** (here must expect normal variation, cannot "improve on God").
- *Predicting inferior moves as most likely*: **Pass**—by happy accident!
- Other cross-checks... [show demo of program output].

Phenomena With Force of Natural Law

- Linear$^{(*)}$ relation to rating. (Asterisk under 1600 Elo).
- Error rate linear in position value—but corrected *logarithmically*.
- Preference about 58% for the first-listed of equal value moves:
  - Deployed model uses a patch.
  - New model handles naturally.
- Swing effects at lower depths among non-optimal moves:
  - Deployed model: no.
  - New model: yes.
- Prediction accuracy on favoring inferior moves:
  - Deployed model: n.a..
  - New model: yes—not by hyp. that thinking depths vary by rating, but by giving everyone the same peak depths as the engines...
- New model captures every major "datum of experience" I know.

If your model contravenes or ignores or ignores these laws, this work argues that it is *wrong*. If it doesn't cross-validate (so well), then it can still be right for trained tasks but not more generally.