

An Introduction to *Perusall*

Gary King¹

Institute for Quantitative Social Science
Harvard University

¹GaryKing.org

What's Perusall ?

- A new type of collaborative e-book reader
 - Based on extensive (patent-pending) data analytics, behavioral science, and educational research
 - Developed at Harvard by Gary King, Brian Lukoff, Eric Mazur, Kelly Miller
- Students
 - Complete readings; learn more
 - Stay engaged and motivated; enjoy the experience
- Instructors:
 - No extra work; save considerable time
 - Improve classroom teaching
- No cost:
 - Perusall itself is free
 - Readings you have rights to: drag and drop to perusall.com
 - Books or articles for purchase: Perusall obtains content & permissions from publishers. Students usually pay less than for print version.

What Happens When You're Assigned Reading?

● Students

- They triage — let some courses die to save others
- How many comply with reading assignments: 20-30%
- How many buy the book? <50%
- How to get them to read?
 - Frequent quizzes, extra homework
 - Make grades depend on them
 - Test very specific points from the readings
 - Spend huge amounts of your time!

● Faculty

- When motivated (preparing for class, or to learn): we do all the reading
- When unmotivated, forced, or assigned readings?
 - We're the same as the students! Want proof?
 - Do you do human subjects research?
 - Did you take the CITI training?
 - Did you do the reading before trying the test?
 - Did you look at the readings after getting some answers wrong?

- **Perusall**: students do >90% of the reading

So what's wrong? Why don't they read?

Intrinsic motivators are backwards:

- Humans value **collective** experiences
(e.g., why the concert costs more than the iTunes download)
- Reading is a **solitary** experience
(e.g., the stalling MOOC revolution: **Collective** \rightsquigarrow **Solitary**)

Extrinsic motivators (grades) are weak, unless:

- Faculty waste time preparing quizzes
- Students waste time taking them
- Everyone wastes class or prep time

The **Perusall** Solution

- Transform the reading experience: **Solitary** \rightsquigarrow **Collective**
- Extrinsic motivation (no instructor effort): Automated grading

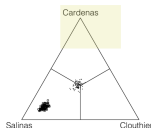
How does it work?

Dominguez and McCann in the first place: the electoral outcome itself. In particular, if every voter thought the PRI was weakening, which candidate would have won the presidency? To answer this question, we coded each voter as thinking that the PRI was weakening and let other characteristics of the voter take on their true values. Then we used the predicted value algorithm to simulate the vote for each person in the sample and used the votes to run a mock election. We repeated this exercise 100 times to generate 100 simulated election outcomes. For comparison, we also coded each voter as thinking the PRI was strengthening and simulated 100 election outcomes conditional on those beliefs.

Figure 3 displays our results. The figure is called a "ternary plot" (see Miller 1977; Katz and King 1999), and coordinates in the figure represent predicted fractions of the vote received by each candidate under a different simulated election outcome. Roughly speaking, the closer a point appears to one of the vertices, the larger the fraction of the vote going to the candidate whose name appears on the vertex. A point near the middle indicates that the simulated election was a dead heat. We also added "win lines" to the figure that divide the ternary diagram into areas that indicate which candidate receives a plurality and thus wins the simulated election (e.g., points that appear in the top third of the triangle are simulated election outcomes where Cárdenas receives a plurality).

In this figure, the o's (all near the bottom left) are simulated outcomes in which everyone thought the PRI was strengthening, while the dots (all near the center) correspond to beliefs that the PRI was weakening. The figure shows that when the country believes the PRI is strengthening, Salinas wins hands down; in fact, he wins every one of the simulated elections. If voters believe the PRI is weakening, however, the 1988 election is a toss-up, with each candidate having an equal chance of victory.

FIGURE 3 Simulated Electoral Outcomes



Coordinates in this ternary diagram are predicted fractions of the vote received by each of the three candidates. Each point is an election outcome drawn randomly from a world in which all voters believe Salinas' PRI party is strengthening (for the "o's" in the bottom left) or weakening (for the "dots" in the middle), with other variables held constant at their means.

question by estimating a [censored Weibull regression \(a form of duration model\)](#) on a dataset in which the dependent variable, Y_i , measures the number of years that leader i remains in office following the onset of war. For fully observed cases (the leader had left office at the time of the study), the model is

$$Y_i \sim \text{Weibull}(\mu, \sigma)$$

$$\mu_i = E(Y_i | X_i) = (e^{X_i \beta})^{-\sigma} \Gamma(1 + \sigma) \quad (6)$$

where σ is an ancillary shape parameter and Γ is the gamma function, an interpolated factorial that works for continuous values of its argument. The model includes four explanatory variables: the leader's pre-war tenure in years, an interaction between pre-war tenure and democ-

number of battle deaths per 10,000 inhabitants, a variable indicating whether the leader won. The authors find that leaders who waged forced to lose their grip on power at home, but leaders with a long pre-war tenure were in office longer than others.

Mesquita and Siverson discuss the merits of their explanatory variables by computing "rate" associated with each variable. Hazard is a traditional method of interpretation in the understanding then requires considerable knowledge. Simulation can help us calculate

- Assign readings & annotations
- Students share questions, answers, perspectives, external knowledge in threads
- Can annotate text, images, or equations
- Classmates motivate
- Perusall grades engagement (better than TAs can)
- Non-adversarial grading; Perusall nudges students not keeping up

What are the advantages of using a Weibull model as opposed to kinds we have discussed in class (like Poisson, Normal, etc)? Feb 26 6:18 pm

As it notes, the Weibull is a time to event model (a duration model), so it's a natural fit for looking at the years someone remains in office following a war. More broadly, it's related to the Exponential distribution, which is typically what you think of when you're trying to model time. But unlike the Exponential, the Weibull has a shape and scale parameter (whereas in the Exponential the shape is always presumed to be 1). Feb 29 11:35 am

What happens when motivated students get stuck?

- Presently: Eyes off the page

- Stop reading, hope you get it in class
- Make believe you understand it and keep reading
- Wait until office hours
- Find another student
- Ask a question via email or a forum

- **Perusall**: Eyes on the page

- Annotate the readings: ask questions in context
- Out-of-class experience centered on the readings
- Get response immediately from other students
- If you've left, expect email with an answer (can respond within email)
- When students figure it out together: Learning is deeper and remembered longer

At the start of class, “Any Questions?”

- Presently:

- Hearing questions and confusions: hugely important in learning
- They have lots of questions, but no one moves
- They’ve just tramped across campus, thinking about their last class, plopped their backpacks and coats down, expecting to be entertained.
- A big missed opportunity

- **Perusall** :

- Just before class, print a **Student Confusion Report** with the top 3–4 topics of confusion or engagement, and the best student annotations
- Walk into class; skip the “any questions” game
- Go through the topics, recognizing students with good questions or comments

Example Student Confusion Report

Perical

Advanced Quantitative Research Methodology, Gov2001

Settings

Gradebook

Router

Confusion report for Making the Most of Statistical Analyses, Entire document

Confusion 1

Making the Most of Statistical Analyses:
Improving Interpretation and Presentation

Gary King
Michael Tomez
Jason Wittenberg

Harvard University
Harvard University
Harvard University

Statistical scientists rarely take full advantage of the information available in their statistical results. As a consequence, they miss opportunities to present quantities that are of greatest academic interest to their research and represent the appropriate degree of certainty about their quantities. In this article, we offer an approach, built on the technique of statistical simulation, to extract the currently overlooked information from any statistical method used to interpret and present data in a modern scientific research setting. Using this technique, we present some quantities which we try to provide herein, but its application should extend to the results of quantitative analyses more informative and transparent. To facilitate our recommendations, we replicate the results of several published works, illustrating in each case how the use of simulation can clarify and improve the presentation of statistical results.

MH

Maybe there are specific examples later in this paper, but I think they would help me understand what he is suggesting. Normally, I would just report an estimate for my beta coefficient with a 95% confidence interval. But this seems to be going beyond that using simulation?

PN

This is because we are not actually drawing from the population; rather, we are drawing multiple times from that one sample that we got from the population. So it's more of a sampling distribution right? The sample we are simulating from could perhaps be thought of as a quasi-population.

NG

Actually, this is a key distinction that has been troubling me on this week's problem set. What exactly are the interpretive differences between theta or theta hat serving as the random variable? Would be useful to review.

Show more...

Confusion 2

IMPROVING STATISTICAL INTERPRETATION AND PRESENTATION

We applied the predicted value algorithm to predict the number of government employees in a state with million people and an 80 percent Democratic lean. We used the statistical software described in the article to generate the log linear model and simulate our results for the other coefficients. β_0 and the other variables are listed in Table 1. The results are presented in Table 2. Next, we set the state explanatory variables at $P_1 = 10,000$ and $P_2 = 10,000$, so we could construct \hat{Y} and compare it to Y . We then drew our value of \hat{Y} from the normal distribution $N(\hat{Y}, \sigma^2)$. In our case, we calculated $\sigma^2 = 1$, so we drew our standard value into the actual number of government employees, a quantity that cannot have noninteger values in natural logarithms. In reporting this process, 10,000 times, we generated 1000 predicted values, which we sorted from lowest to highest. The numbers in the 25th and the 75th positions represented the upper and lower bounds of a 50 percent confidence interval. Thus, we predicted that 50 percent confidence that the state government would employ between 73,000 and 119,000.

ER

What does a logit model look like and what makes it particularly useful in this case of binary data? Why would we know a logit model is better to use here than a normal distribution, for instance? Would we have to have some prior understanding of the nature of the data we are looking at to determine a logit model is optimal?

CC

Intuitively, I have trouble understanding exactly how simulation helps us. Are we using parameters from the data to simulate potential outcomes that give us probabilities? More generally, how does the simulation relate to the actual data?

- One page, easy to digest before class
- See confusions or engagements in context
- Annotations remain live
- Highlights best student annotations

Automated Engagement

- **Classes of any size**
 - “Annotation groups” (of about 20) constructed from your class automatically to optimize engagement
 - Perusall can seed your class with annotations from another predicted to generate engagement
- **If a student skips pp.61-67**
 - Perusall checks for important concepts missed
 - Student gets a private nudge about the point on p.63
- **As class nears, Perusall monitors continuously**
 - Who hasn't done the reading or isn't engaged
 - Who isn't going to get 100% on the annotation assignment
 - Carefully timed, unobtrusive private notes with help and direction
- **Instructor gets: a window on student learning**
 - A dashboard with grading suggestions
 - Info about individual student engagement & performance
 - Alerts for students with serious problems

Administrative

- Works differently across fields:
 - In technical fields, students use Perusall to understand what the text explains, to get past hurdles
 - In the humanities, Perusall engages students with the meaning of the text itself
- Integrates with your university's LMS for single sign-on
- FERPA compliant; university administration friendly; no advertising.
- Student identities not shared with corporate recruiters or anyone outside of class
- Students buy content through Perusall (online or via a code from your bookstore); usually pay less
- Publishers love it (keeps their content central; 100% sell-through, no resales, the ultimate solution to IP piracy)
- Many other features; under active development; suggestions welcome

Sign up at Perusall.com