Using *Chance* Magazine to Engage AP and Undergraduate Students in the Study of Statistics: An Update

Dalene Stangl
Duke University
Chance magazine is a joint venture of ASA and Springer Science + Business Media, LLC
All images within this webinar are from Chance unless otherwise noted.
During the past 20 years, undergraduate education has shifted from student as passive recipient of information to student as active participant in the classroom. I wrote an article for *Chance* magazine’s 20th anniversary issue titled, “Using *Chance* to Engage Undergraduates in the Study of Statistics.” The article gave examples of activities inspired by *Chance* magazine articles from the last 20 years.

This webinar will take articles from a recent issue of *Chance* and demonstrate the ease with which any issue can be used to develop class activities that are fun for high school students and undergraduates whether the course is a basic quantitative literacy course, an AP statistics course, an introductory course for non-statistics majors, or a core or elective course for the statistics major.
Healthy for Life: Accounting for Transcription Errors Using Multiple Imputation
Application to a study of childhood obesity

Articles
7 Filling in the Blanks: Some Guesses Are Better Than Others
Illustrating the Impact of Erroneous Selection When Implying
Complex Survey Items
Tom Kenneke and David Juddens
14 Healthy for Life: Accounting for Transcription Errors Using Multiple Imputation
Application to a Study of Childhood Obesity
Michael R. Elliot
24 A Statistical Look at Roger Clemens' Pitching Career
Eric T. Bradlow, Shane J. Jensen, Justin Wolfers, and
Abraham J. Wyner
31 Astrostatistics: The Final Frontier
Peter Freeman, Joseph Richards, Chad Schaefer, and Ann Lee
36 Stochastic Stamps: A Philosophical Introduction to Chance
Simo Pantanen and George P. H. Styan
42 Probability, Statistics, Evolution, and Intelligent Design
Peter Oliphant

Columns
46 A Statistician Reads the Sports Pages, Phil Everton, Column Editor
Good Defense vs. Poor Defense: The 2007 Women's World Cup
55 Here's to Your Health, Mark Chickens, Column Editor
Misreporting, Missing Data, and Multiple Imputation: Improving
Accuracy of Cancer Registry Databases
Yulei He, Recal Yurel, and Alan M. Zaslavsky
59 Visual Revelations, Howard Wainer, Column Editor
Giving the Finger to Dating Services
Grace Lee, Paul Welleran, and Howard Wainer
62 Goodness of Wit Test, Jonathan Berkowitz, Column Editor
Goodness of Wit Test #1

Departments
3 About the Authors
5 Editor's Letter
6 Letter to the Editor

Indexed in Academic Abstracts, Academic Search, Current Index to Statistics, and MasterFILE

Cover design: Melissa Miko

Using a comprehensive database, statisticians Eric Bradlow, Shane Jensen, Justin Wolfers, and Abraham Wyner take a close look at the extremely successful career of Houston Astros' pitcher Roger Clemens.
Giving the Finger to Dating Services
Grace Lee, Paul Velleman, Howard Wainer
A computer dating service asked:

What is the length of your index finger?

Why ask this?
Proxy for height because people exaggerate their height
Distract attention from other, more meaningful questions
Other plausible explanations?

How do you test these plausible explanations?
Gather students’ reported height followed by actual height and index finger length
Look at scatterplots, correlation, and regression
Is prediction from the regression more accurate than reported height?
What if we subgroup by gender?
Discuss change in means, sd, correlation, regression, rms-error resulting from conditioning on index finger length and gender
Mean height = 172 cm, sd = 9.7 cm
Sd of prediction error using index finger length is 8.0 cm. Hence not much aid in correcting fraudulently reported height.

Unconditional sd for men’s height is 6.3 cm, the conditional using index finger length is 5.9 cm. Though better still not likely to give more accurate height predictions.
Giving the Finger to Dating Services
Grace Lee, Paul Velleman, Howard Wainer

Taking height a step further:
What is your height and the desired height for your ideal spouse/partner?
Figure 3. A display of the height and sex of 147 Cornell students, along with their estimates of the ideal height of their future spouse/partner. Shown are the separate regression lines for each sex and the overall regression. The changing sign of the slope indicates the existence of Simpson’s Paradox.
... a recently released report by Hendricks Sports Management ... Using well-established baseball statistics including ERA (number of earned runs allowed per nine innings pitched) and K-rate (strikeout rate per nine innings pitched), the report compares Roger Clemens’ career to those of other great power pitchers of his era (i.e. Randy Johnson, Nolan Ryan, and Curt Schilling) and proclaims that Roger Clemens’ career trajectory on these measures is not atypical. Based on this finding, the report suggests the pitching data are not an indictment (nor do they provide proof) of Clemens’ guilt [use of performance enhancing substances] in fact they suggest the opposite.
The authors concur with the Hendricks report that a statistical analysis of Clemens’ career can provide prima facie ‘evidence’, but their approach provides a new look at his career pitching trajectory using a broader set of measures and a broader set of comparison pitchers.

By focusing on only pitchers who pitched effectively into their mid-40s, the Hendricks report minimized the possibility that Clemens would look atypical.

This paper provides a more sophisticated and comprehensive analysis of Clemens’ career.
ERA: number of earned runs allowed per nine innings pitched (an average ERA varies btw 4.00 and 5.00)

K-rate: strikeouts

BB: walks
To put Clemens’ trajectory into an appropriate context requires a comparison group.

Star level contemporaries:
Randy Johnson
Greg Maddux
Curt Schilling
Nolan Ryan

Their trajectories fit nicely with quadratic curves. This is in stark contrast to Clemens’ trajectory. The “second act” for Clemens is unusual compared to these greats.

How unusual is it for a durable pitcher to have suffered a mid-career decline only to recover in his mid- and late 30s?
More complete analysis

Lahman Database, Version 5.5
www.baseball1.com

All Major League Baseball pitchers
whose careers were contained in years
1969-2007

Included pitchers who played ≥15 full
seasons as a starter (10+ games/yr)
and > 3000 innings

N=31+Clemens

Pitching stats used
WHIP=Walks+hits per inning pitched
BAA=Batting avg for hitters facing given pitcher
ERA=Earned run average per nine innings pitched
BB =Walk Rate
K = Batter strike-out rate per plate appearance

For each stat, fit a quadratic to each of the 32
pitchers data at year t
For each statistic, fit a quadratic to each of the 32 pitchers data at year t

\[ S_{ijt} = \beta_{0ij} + \beta_{1ij} \text{Age}_{it} + \beta_{2ij} \text{Age}^2_{ij} + \epsilon_{ijt} \]

A quadratic curve may not be best for every pitcher, but the goal is only to detect those patterns that stick out as highly unusual with respect to a quadratic reference.

Interest focuses on \( \beta_{2ij} \)

\( \beta_{2ij} = 0 \) linear
\( \beta_{2ij} < 0 \) hump
\( \beta_{2ij} > 0 \) U-shaped

For pitcher hitting a mid-career prime

WHIP \( B_{2ij} > 0 \)
BAA \( B_{2ij} > 0 \)
ERA \( B_{2ij} > 0 \)
BB \( B_{2ij} > 0 \)
K \( B_{2ij} < 0 \)

Data from Hendricks report, using ERA

All 32 players
6 atypical players, \( \beta_{2ij} < 0 \)
Clemens is only pitcher to get worse mid-career and then better at end.

Steepness of Clemens’ curve is noticeable in later years.

Clemens’ trajectory has similar shape albeit flatter.

Clemens has overall higher K rate, but his trajectory is a similar shape.
Conclusion: “Through the use of simple exploratory curve fitting applied to a number of pitching statistics, and for a well-defined set of long-career pitchers, we assessed whether Clemens’ pitching trajectories were atypical. Our evidence is suggestive that while most long-term pitchers have peaked mid-career and decline thereafter, Clemens (for some key statistics: WHIP, BB, and ERA) worsened mid-career and improved thereafter.”

“We emphasize that our analysis is entirely exploratory---we don not believe there exists a reasonable probability model to test relevant hypotheses by calculating significance levels. The data does not exonerate (nor does it indict) Clemens, as an exploratory statistical analysis of this type never proves innocence or guilt. After analyzing this data set, there are at least as many questions remaining as before.”
Cancer registries collect information on type of cancer, histological characteristics, stage of diagnosis, patient demographics, initial course of treatment including surgery, radiotherapy, and chemotherapy, and patient survival. Such information can be valuable for studying the patterns of cancer epidemiology, diagnosis, treatment, and outcome. However, misreporting of registry information is unavoidable; therefore, studies based solely on registry data would lead to invalid results.
Misreporting, Missing Data, and Multiple Imputation: Improving Accuracy of Cancer Registry Databases
Yulei He, Recai Yucel, and Alan M. Zaslavsky

California Cancer Registry

Largest, geographically contiguous population-based cancer registry in the world

Data on patterns of receiving and reporting adjuvant therapies for cancer patients

Table compares physician survey (subsample of patients in the registry) to registry. Registry data has considerable under-reporting.

How can we use valuable registry data, but improve inferences?

<table>
<thead>
<tr>
<th>Sample</th>
<th>Chemotherapy</th>
<th>Radiotherapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>73.3 (1.16)</td>
<td>25.4 (1.14)</td>
</tr>
<tr>
<td>Registry (in the survey region)</td>
<td>57.9 (0.79)</td>
<td>22.2 (0.67)</td>
</tr>
<tr>
<td>Registry (statewide)</td>
<td>51.4 (0.45)</td>
<td>19.6 (0.35)</td>
</tr>
<tr>
<td>Imputed Registry (statewide)</td>
<td>61.2 (0.77)</td>
<td>23.1 (0.61)</td>
</tr>
</tbody>
</table>
How can we use valuable registry data, but improve inferences?

1. Use only validation survey? 2K vs 12K, loses precision

2. Error-in-variables method – analyze registry data while adjusting for reporting error – model the relationship between correct values and misreported ones – requires statistical expertise to implement

3. Multiple imputation – fill in missing values several times to create multiple complete data sets – combine results from separate sets into a single inference using simple rules
Multiple imputation – fill in missing values several times to create multiple complete data sets – combine results from separate sets into a single inference using simple rules.

The imputation model characterizes the measurement error process and makes the adjustment. The imputation may incorporate additional information to further improve the analyses.

Figure 1. An illustration of using imputation to correct for under-reporting. X is a matrix of covariate variable values with one row for each person in the registry. $Y_{o(i)}$ is the matrix of reported treatment status for various treatments. $Y_{o(i)}$ is the true treatment. $Y_{o(i)}$ is observed in the survey. An observed value of 1 is assumed to be true, but a value of 0 might be incorrect.
Student Example:
What percentage of Duke students are basketball fans?

Picture from album of Eric Vance
Student Example:
What percentage of Duke students are basketball fans?

Assumes under reporting in the registry

\[ Y_{(O)} \text{ truth} \]
\[ Y_{(R)} \text{ reported in registry} \]
\[ X_{(i)} \text{ covariate} \]

Yes in registry is assumed true, while No may be incorrect (under reporting)

<table>
<thead>
<tr>
<th>( Y_{(O)} )</th>
<th>( Y_{(R)} )</th>
<th>( X_{(1)} )</th>
<th>( X_{(2)} )</th>
<th>( X_{(3)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>G</td>
<td>1400</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Male</td>
<td>N</td>
<td>?</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Female</td>
<td>N</td>
<td>1380</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Male</td>
<td>G</td>
<td>1410</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Female</td>
<td>G</td>
<td>?</td>
</tr>
</tbody>
</table>

Surveyed Students

<table>
<thead>
<tr>
<th>( Y_{(O)} )</th>
<th>( Y_{(R)} )</th>
<th>( X_{(1)} )</th>
<th>( X_{(2)} )</th>
<th>( X_{(3)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>?</td>
<td>?</td>
<td>1460</td>
</tr>
<tr>
<td>?</td>
<td>No</td>
<td>Female</td>
<td>?</td>
<td>1400</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>N</td>
<td>1360</td>
</tr>
<tr>
<td>?</td>
<td>No</td>
<td>Female</td>
<td>G</td>
<td>1550</td>
</tr>
<tr>
<td>?</td>
<td>No</td>
<td>Male</td>
<td>G</td>
<td>1490</td>
</tr>
</tbody>
</table>

Remaining Registry
Student Example:
What percentage of Duke students are basketball fans?

*Simple Imputation Scheme*

*Set initial values*

- \( P(Y_{(O)} = \text{Yes}) \)
- \( P(Y_{(R)} = \text{No} \mid Y_{(O)} = \text{Yes}) \)

*Compute*

- \( P(Y_{(O)} = \text{Yes} \mid Y_{(R)} = \text{No}) \) (Bayes Theorem)

*Impute*

- \( Y_{(O)} \) from \( P(Y_{(O)} = \text{Yes} \mid Y_{(R)} = \text{No}) \) (Bernoulli draws)

*Iterate between imputing missing values of \( Y_{(O)} \) and estimating \( P(Y_{(O)} = \text{Yes}) \) and \( P(Y_{(O)} = \text{Yes} \mid Y_{(R)} = \text{No}) \) using imputed values of \( Y_{(O)} \).

After many iterations probabilities converges to target distribution and a final draw of \( Y_{(O)} \) produces a complete set of imputations. Repeat multiple times.

Work through one iteration, then go to JMP and generate draws from Beta and Binomial distributions.
Student Example:
What percentage of Duke students are basketball fans?

More Complex Imputation Scheme – Adding Covariates

Are particular subgroups more likely to be under reported?

- Gender?
- Greek Affiliation?
- SAT score?
- Major?

Impute
\[ Y_{(O)} \] from \( P(Y_{(O)} = \text{Yes} \mid Y_{(R)} = \text{No}, X, \text{model parameters linking} \ X \ \text{and} \ Y) \)
Two other articles in this issue discuss imputation

- **Filling in the Blanks: Some Guesses Are Better Than Others**

- **Healthy for Life: Accounting for Transcription Errors Using Multiple Imputation**
In the last decades, arguments against Darwinian evolution have become increasingly sophisticated, replacing Creationism by Intelligent Design (ID) and the book of Genesis by biochemistry and mathematics. As arguments claiming to be based in probability and statistics are being used to justify the anti-evolution stance, it may be of interest to readers of CHANCE to investigate methods and claims of ID theorists.
An ID Hypothesis Testing Challenge to Evolution

William Dembski

Once all chance explanations have been ruled out ‘design’ is inferred (Chance) = Design

Ex.1 Nicholas Caputo was a NJ democrat in charge of election ballots. Names were to be listed in random order. In 41 elections a Democrat was listed 1st in 40. Dembski’s argument here would be if we can rule out chance, Caputo cheated. If p = .5, the probability of 40 out of 41 is on the order of magnitude 1 in 50 billion, so infer Caputo cheated.
An ID Hypothesis Testing Challenge to Evolution

William Dembski

Once all chance explanations have been ruled out ‘design’ is inferred (Chance)\(c=\)Design

Ex.2 The evolution of bacterial flagellum. The probability that a random configuration will produce the number and types of proteins needed to form different parts of the flagellum is so extremely improbable, that design must be inferred.