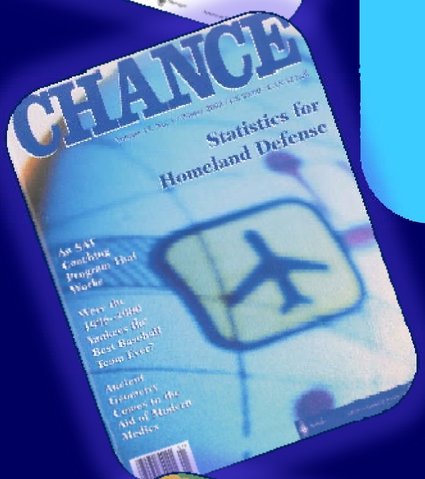
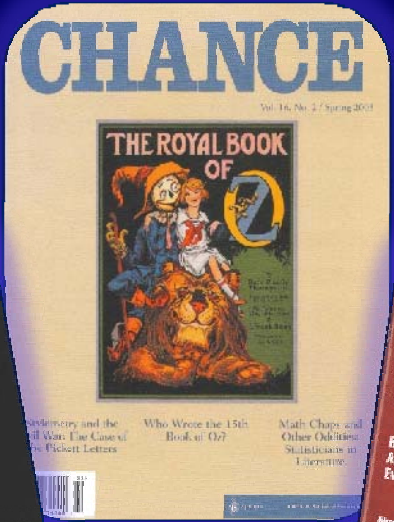
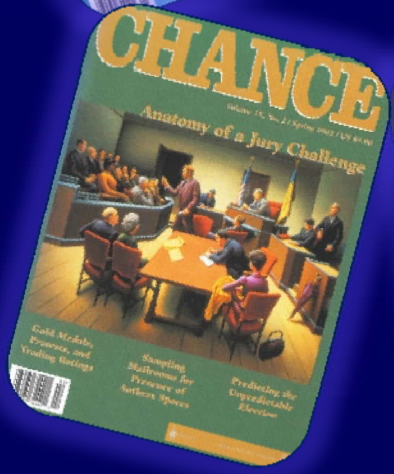
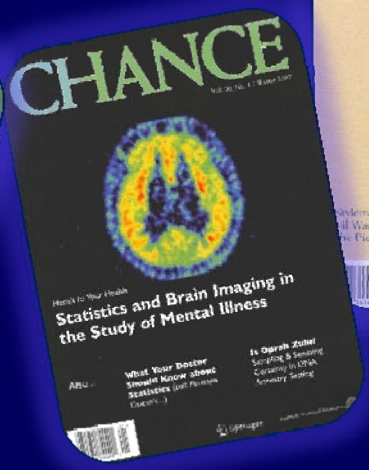
Using *Chance* Magazine to Engage AP and Undergraduate Students in the Study of Statistics: An Update

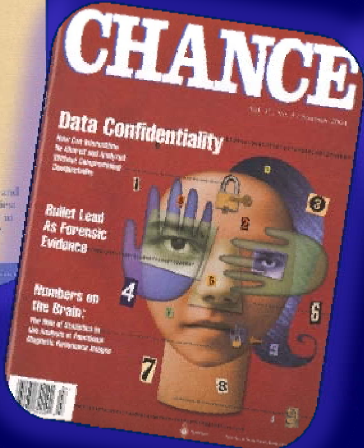
Dalene Stangl
Duke University





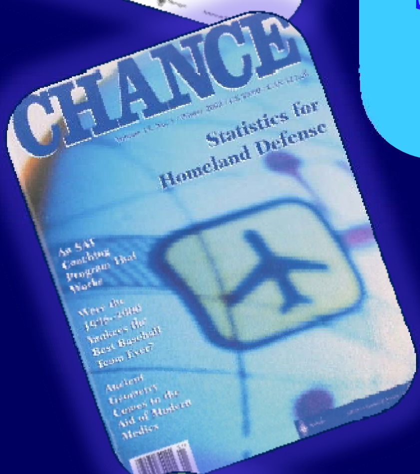


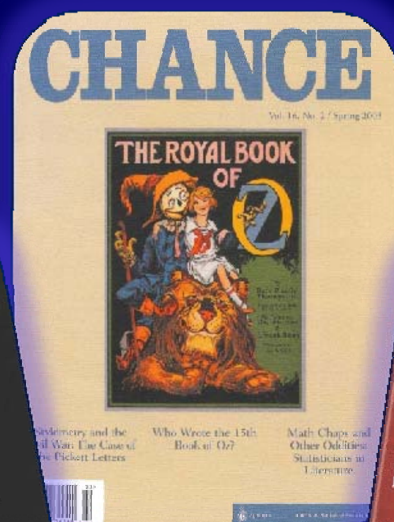


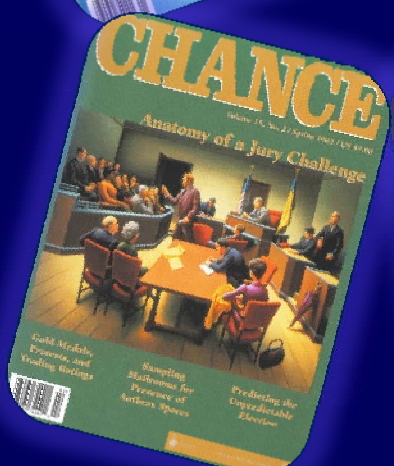


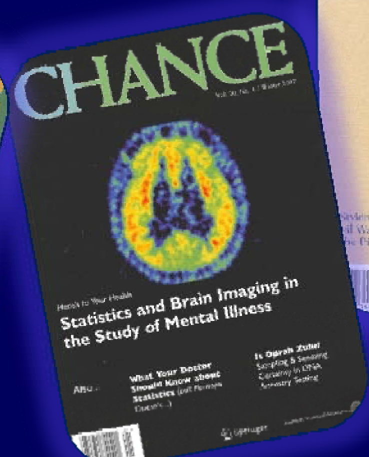
Chance magazine is a joint venture of
ASA
and
Springer Science + Business Media, LLC

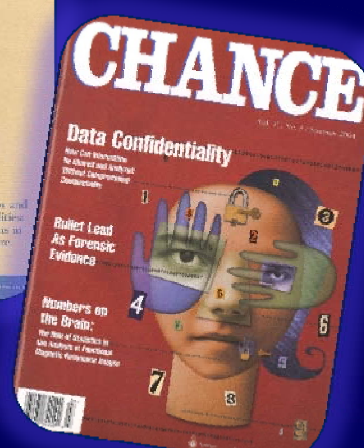
All images within this webinar are from *Chance* unless otherwise noted.

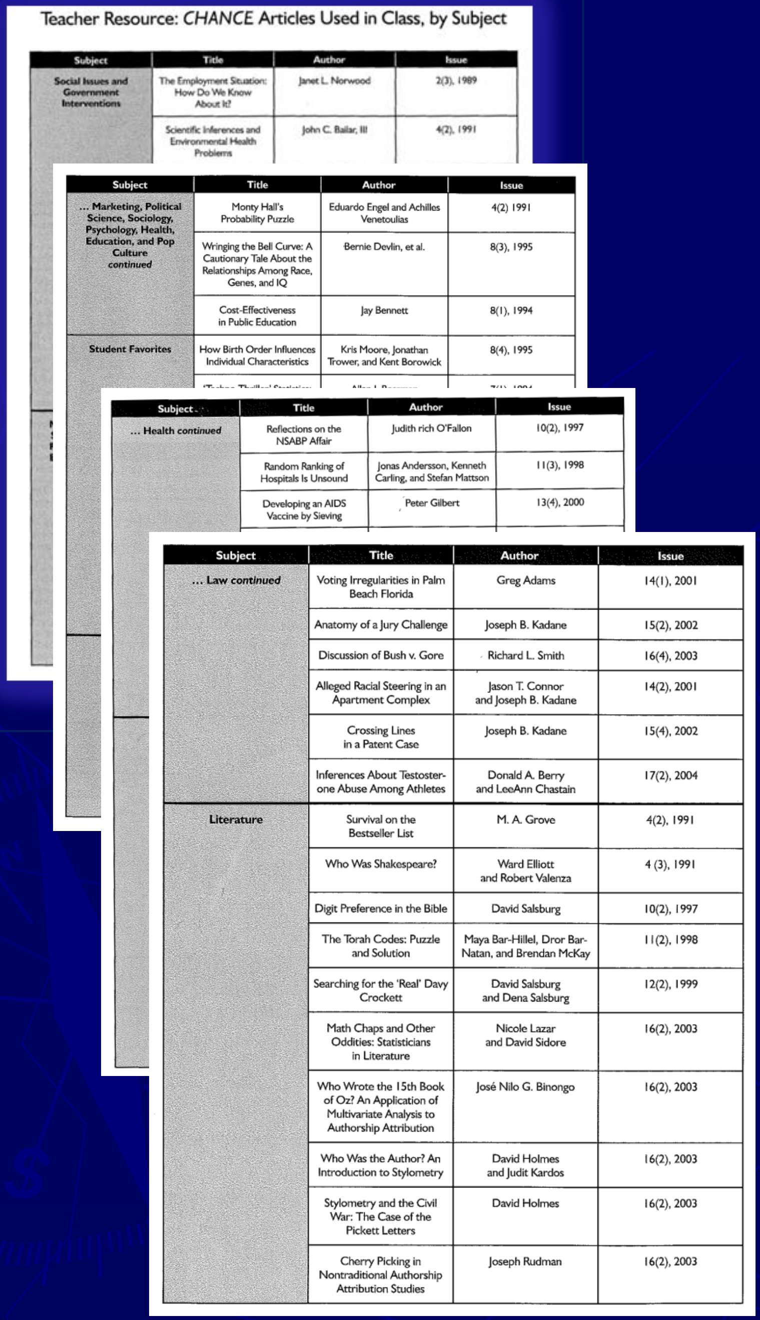












Teacher Resource: *CHANCE* Articles Used in Class, by Subject

Subject	Title	Author	Issue
Social Issues and Government Interventions	The Employment Situation: How Do We Know About It?	Janet L. Norwood	2(3), 1989
	Scientific Inferences and Environmental Health Problems	John C. Bailar, III	4(2), 1991

Subject	Title	Author	Issue
... Marketing, Political Science, Sociology, Psychology, Health, Education, and Pop Culture *continued*	Monty Hall's Probability Puzzle	Eduardo Engel and Achilles Venetoulias	4(2) 1991
	Wringing the Bell Curve: A Cautionary Tale About the Relationships Among Race, Genes, and IQ	Bernie Devlin, et al.	8(3), 1995
	Cost-Effectiveness in Public Education	Jay Bennett	8(1), 1994
Student Favorites	How Birth Order Influences Individual Characteristics	Kris Moore, Jonathan Trower, and Kent Borowick	8(4), 1995

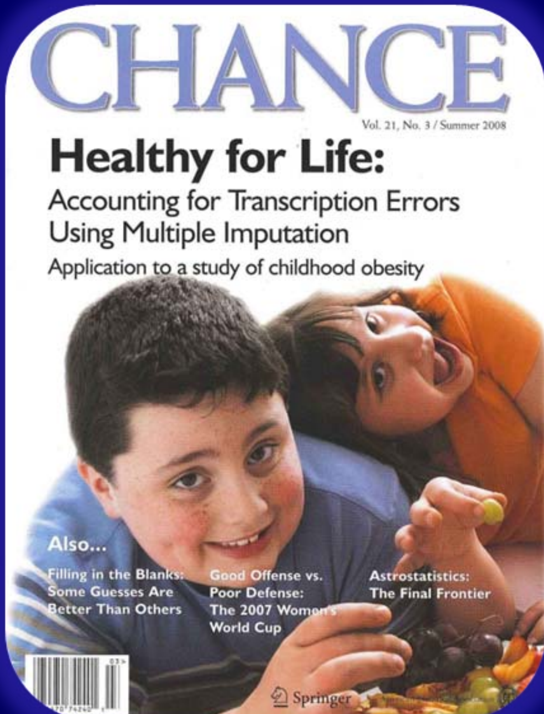
Subject	Title	Author	Issue
... Health *continued*	Reflections on the NSABP Affair	Judith rich O'Fallon	10(2), 1997
	Random Ranking of Hospitals Is Unsound	Jonas Andersson, Kenneth Carling, and Stefan Mattson	11(3), 1998
	Developing an AIDS Vaccine by Sieving	Peter Gilbert	13(4), 2000

Subject	Title	Author	Issue
... Law *continued*	Voting Irregularities in Palm Beach Florida	Greg Adams	14(1), 2001
	Anatomy of a Jury Challenge	Joseph B. Kadane	15(2), 2002
	Discussion of Bush v. Gore	Richard L. Smith	16(4), 2003
	Alleged Racial Steering in an Apartment Complex	Jason T. Connor and Joseph B. Kadane	14(2), 2001
	Crossing Lines in a Patent Case	Joseph B. Kadane	15(4), 2002
	Inferences About Testosterone Abuse Among Athletes	Donald A. Berry and LeeAnn Chastain	17(2), 2004
Literature	Survival on the Bestseller List	M. A. Grove	4(2), 1991
	Who Was Shakespeare?	Ward Elliott and Robert Valenza	4 (3), 1991
	Digit Preference in the Bible	David Salsburg	10(2), 1997
	The Torah Codes: Puzzle and Solution	Maya Bar-Hillel, Dror Bar-Natan, and Brendan McKay	11(2), 1998
	Searching for the 'Real' Davy Crockett	David Salsburg and Dena Salsburg	12(2), 1999
	Math Chaps and Other Oddities: Statisticians in Literature	Nicole Lazar and David Sidore	16(2), 2003
	Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution	José Nilo G. Binongo	16(2), 2003
	Who Was the Author? An Introduction to Stylometry	David Holmes and Judit Kardos	16(2), 2003
	Stylometry and the Civil War: The Case of the Pickett Letters	David Holmes	16(2), 2003
	Cherry Picking in Nontraditional Authorship Attribution Studies	Joseph Rudman	16(2), 2003

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.

Chance
Summer 2008, Vol. 21, No. 3
Table of Contents



Volume 21, Number 3, 2008

CHANCE

A Magazine of the American Statistical Association

Articles

Columns

Departments

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

Cover design: Melissa Muko

7

Tom Krenzke and David Judkins show how working with imputation algorithms can help alleviate the challenge of complex skip and nonresponse patterns in data.

14

Working on the Healthy for Life project of the University of Pennsylvania and Children's Hospital of Philadelphia, statisticians collaborate with clinical scientists and develop new methods to fight childhood obesity and learn to be healthy for life.

24

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 Astro's pitcher Roger Clemens.

Giving the Finger to Dating Services

Grace Lee, Paul Velleman, Howard Wainer

match.com

eHarmony

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 that 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

Giving the Finger to Dating Services

Grace Lee, Paul Velleman, Howard Wainer

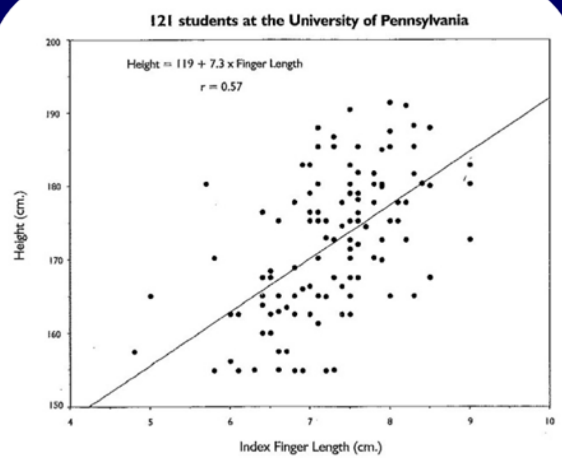


Figure 1. The bivariate relationship between height and the length of index finger among a convenience sample of 121 students at the University of Pennsylvania in 2008

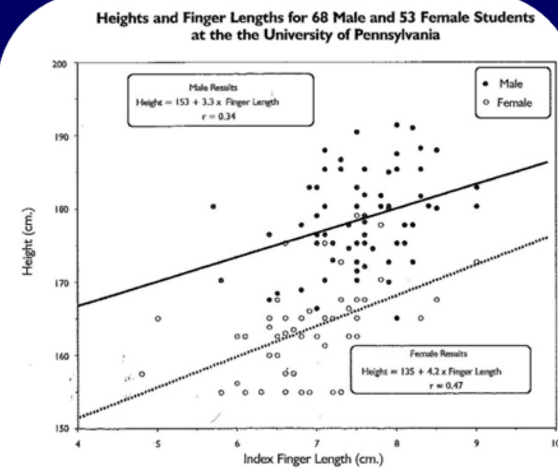


Figure 2. The bivariate relationship between height and the length of index finger among a convenience sample of 68 male students and 53 female students at the University of ennsylvania in 2008

Mean height=172 cm, sd=9.7cm
Sd of prediction error using index finger length is 8.0cm. Hence not much aid in correcting fradulently reported height.

Unconditional sd for men's height is 6.3cm, the conditional using index finger length is 5.9cm. Though better still not likely to give more accurate height predictions.

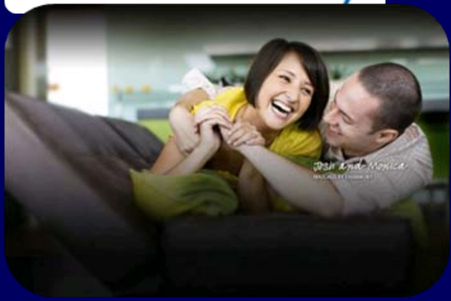
Giving the Finger to Dating Services

Grace Lee, Paul Velleman, Howard Wainer

match.com

Taking height a step further:
What is your height and the desired height for your ideal spouse/partner?

eHarmony



Giving the Finger to Dating Services

Grace Lee, Paul Velleman, Howard Wainer

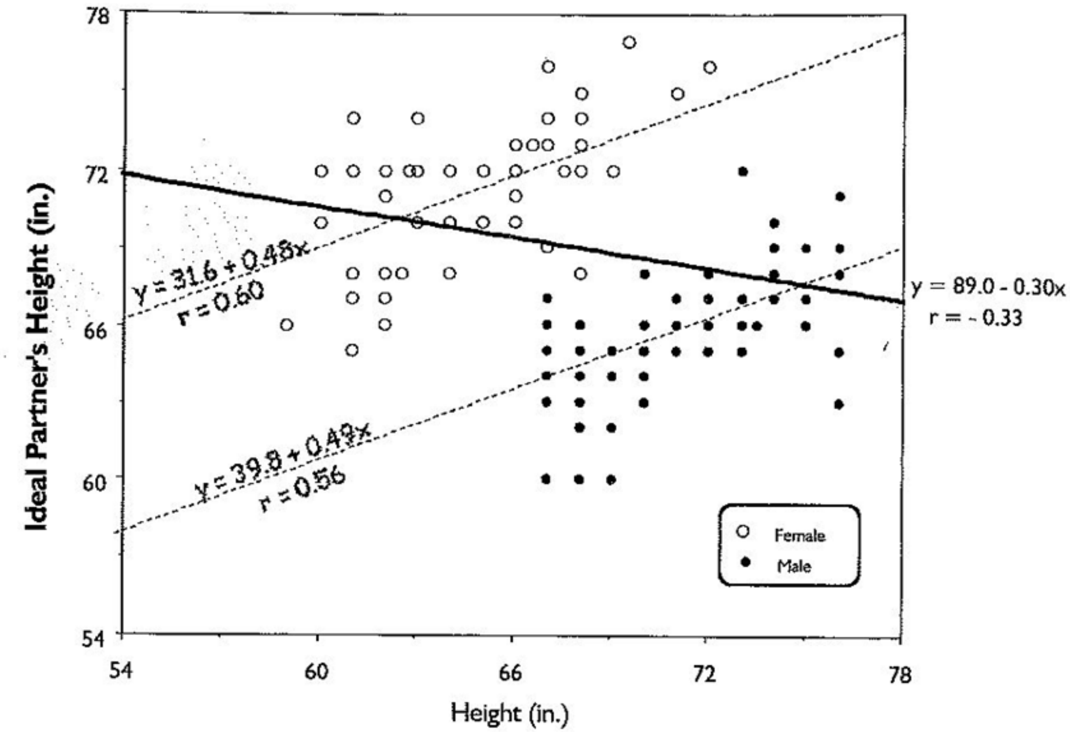


Figure 3. A display of the height and sex of 147 Cornell students, along with their estimates of the ideal height of their future spouse/partner. Shown are the separate regression lines for each sex and the overall regression. The changing sign of the slope indicates the existence of Simpson's Paradox.

A Statistical Look at Roger Clemens' Pitching Career

Eric T. Bradlow, Shane T. Jensen, Justin Wolfers, and Abraham J. Wyner

… 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.

A Statistical Look at
Roger Clemens' Pitching Career

Eric T. Bradlow, Shane T. Jensen, Justin Wolfers, and Abraham J. Wyner

Baseball is America's pastime
interest at an all-time high
business. Furthermore, ma
records (the yearly home run r
record, the 500 home run club,
passed at a pace never before see
use and accusations documented
performance-enhancing substanc
these accomplishments is receivin
than the breaking of the records.
A particularly salient examp
released report by Hendricks Spo
led to widespread national cover
baseball statistics, including ER
allowed per nine innings pitch
rate per nine innings pitched),
Clemens' career to those of other
era (i.e., Randy Johnson, Nolan R
proclaims that Roger Clemens' ca
sures is not atypical. Based on this
the pitching data are not an indic
proof) of Clemens' guilt; in fact,
While we concur with the Hen
analysis of Clemens' career can pr
(and a valuable lens with which to
approach provides a new look at his career
tory using a broader set of measures, as well as a broader comparison set of pitchers. This is important, as there has been a lot of recent research as to what are the most reliable and stable measures of pitching performance. Our attempt is to be inclusive in this regard.

Even more important, one of the pitfalls all analyses of extraordinary events (the immense success of Clemens as a pitcher) have is "right-tail self-selection." If one compares extraordinary players only to other extraordinary players, and selects that set of comparison players based on their behavior on that extraordinary dimension, then one does not obtain a representative (appropriate) comparison set. By focusing on only pitchers who pitched effectively into their mid-40s, the Hendricks report minimized the possibility that Clemens would look atypical.

Here, we use more reasonable criteria for pitchers that are based on their longevity and the number of innings pitched in their career to form the comparison set, rather than performance at any specific point. Thus, the focus of this paper is analysis of Clemens' career using a more sophisticated and

Houston Astros pitcher Roger Clemens throws a pitch against the St. Louis Cardinals during the fifth inning of their Major League game September 24, 2006, in Houston.

n that, what one can say

discussion, we first take
be sure, this unavoidable
r research method, and
s in the order in which
ge fan, the most salient
entage and ERA, which
for each game, there is a
(hence 0.5 is the average
ERA varies between 4.00
le over the last 30 years
ure 1 how extraordinary
f his career.
s is that Clemens quickly
the early 1990s, he lost
s with the Red Sox were
history), so much so that
red to be in the "twilight
h clearly demonstrates,
new heights at the com-
w years showed a second

Now, any well-read student of baseball understands that winning percentage and ERA are fairly noisy measures of quality. Both are readily affected by factors outside a pitcher's ability, such as fielding and the order in which batting events occur. Additionally, winning percentage critically depends on run support. Analysts who specialize in pitching evaluation use measures of component events instead, such as rates of strike outs (K) and walks (BB). We graph the career trajectory of K rate and BB rate for Clemens (Figure 2) and note his career average values

A Statistical Look at Roger Clemens' Pitching Career

Eric T. Bradlow, Shane T. Jensen, Justin Wolfers, and Abraham J. Wyner

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.

A Statistical Look at Roger Clemens' Pitching Career

Eric T. Bradlow, Shane T. Jensen, Justin Wolfers, and Abraham J. Wyner

Baseball is America's pastime
interest at an all-time high
business. Furthermore, ma
records (the yearly home run r
record, the 500 home run club,
passed at a pace never before see
use and accusations documented
performance-enhancing substanc
these accomplishments is receivin
than the breaking of the records.
A particularly salient examp
released report by Hendricks Spo
led to widespread national cover
baseball statistics, including ER
allowed per nine innings pitch
rate per nine innings pitched),
Clemens' career to those of other
era (i.e., Randy Johnson, Nolan R
proclaims that Roger Clemens' ca
sures is not atypical. Based on this
the pitching data are not an indic
proof) of Clemens' guilt; in fact,
While we concur with the Hen
analysis of Clemens' career can pr
(and a valuable lens with which to
approach provides a new look at his career
tory using a broader set of measures, as well as a broader comparison set of pitchers. This is important, as there has been a lot of recent research as to what are the most reliable and stable measures of pitching performance. Our attempt is to be inclusive in this regard.

Even more important, one of the pitfalls all analyses of extraordinary events (the immense success of Clemens as a pitcher) have is "right-tail self-selection." If one compares extraordinary players only to other extraordinary players, and selects that set of comparison players based on their behavior on that extraordinary dimension, then one does not obtain a representative (appropriate) comparison set. By focusing on only pitchers who pitched effectively into their mid-40s, the Hendricks report minimized the possibility that Clemens would look atypical.

Here, we use more reasonable criteria for pitchers that are based on their longevity and the number of innings pitched in their career to form the comparison set, rather than performance at any specific point. Thus, the focus of this paper is analysis of Clemens' career using a more sophisticated and

Houston Astros pitcher Roger Clemens throws a pitch against the St. Louis Cardinals during the fifth inning of their Major League game September 24, 2006, in Houston.

n that, what one can say

discussion, we first take
be sure, this unavoidable
r research method, and
s in the order in which
ge fan, the most salient
entage and ERA, which
for each game, there is a
(hence 0.5 is the average
ERA varies between 4.00
le over the last 30 years
ure 1 how extraordinary
f his career.
s is that Clemens quickly
the early 1990s, he lost
s with the Red Sox were
history), so much so that
red to be in the "twilight
h clearly demonstrates,
new heights at the com-
w years showed a second

Now, any well-read student of baseball understands that winning percentage and ERA are fairly noisy measures of quality. Both are readily affected by factors outside a pitcher's ability, such as fielding and the order in which batting events occur. Additionally, winning percentage critically depends on run support. Analysts who specialize in pitching evaluation use measures of component events instead, such as rates of strike outs (K) and walks (BB). We graph the career trajectory of K rate and BB rate for Clemens (Figure 2) and note his career average values

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

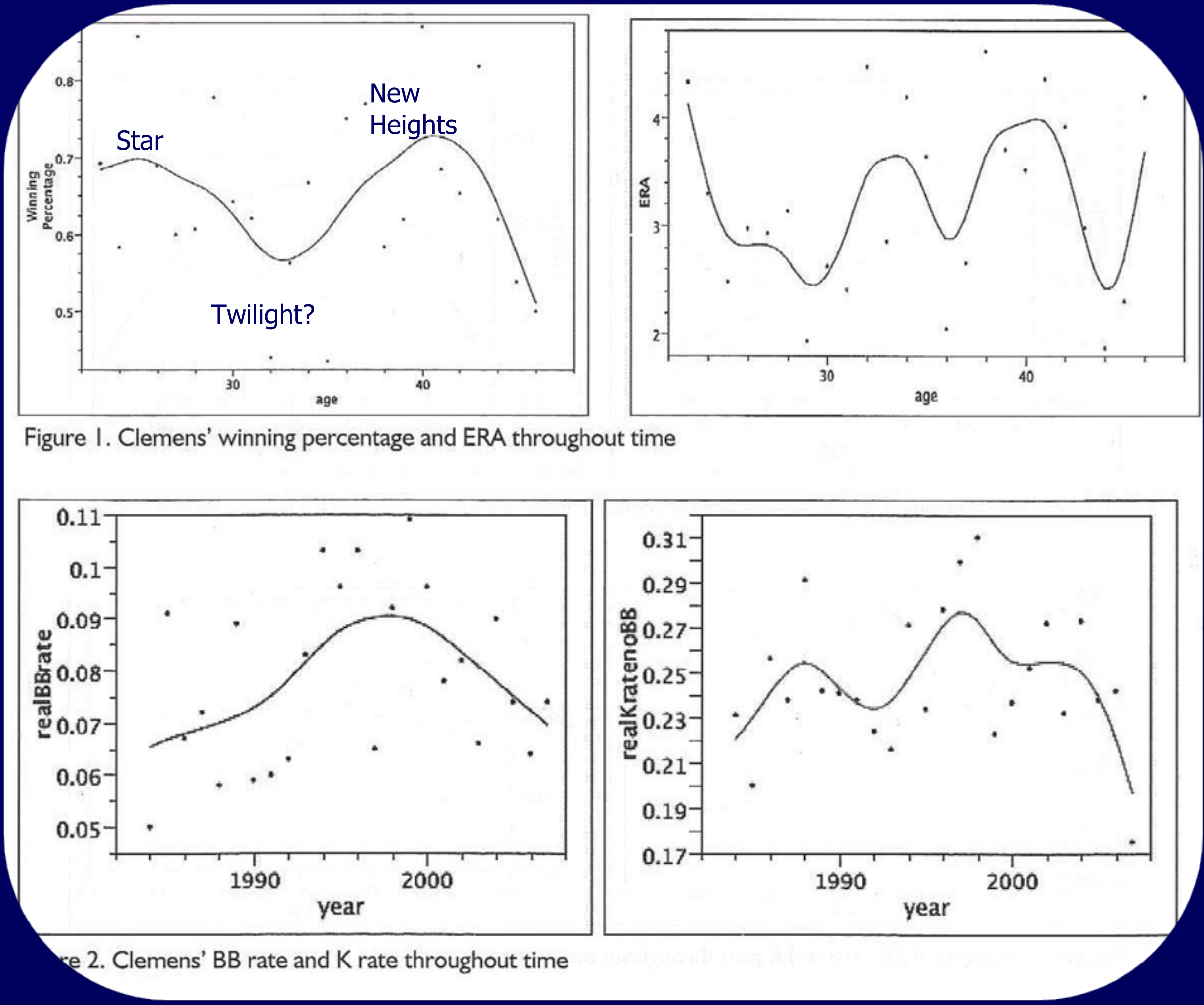


Figure 1. Clemens' winning percentage and ERA throughout time

re 2. Clemens' BB rate and K rate throughout time

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?

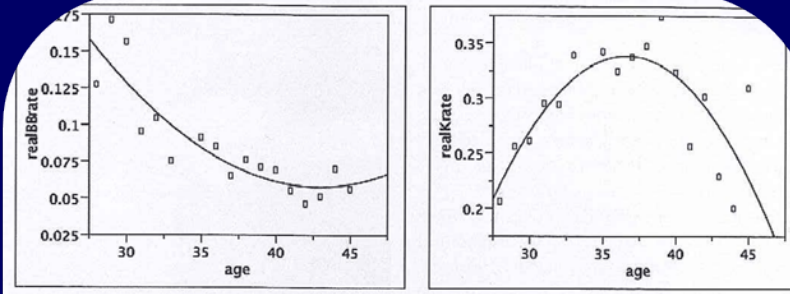


Figure 3. Randy Johnson BB rate and K rate throughout time

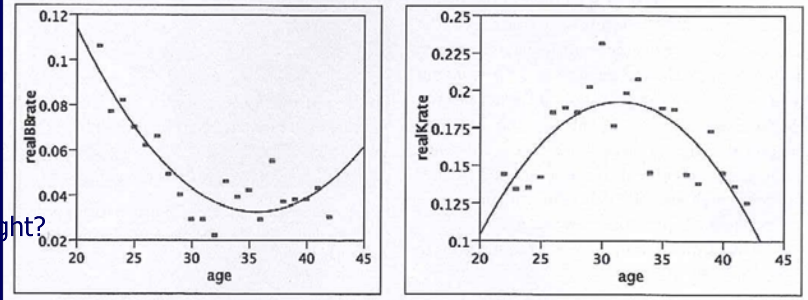


Figure 4. Greg Maddux BB rate and K rate throughout time

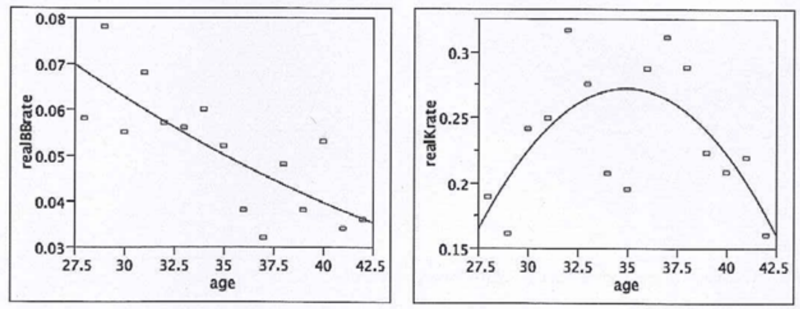
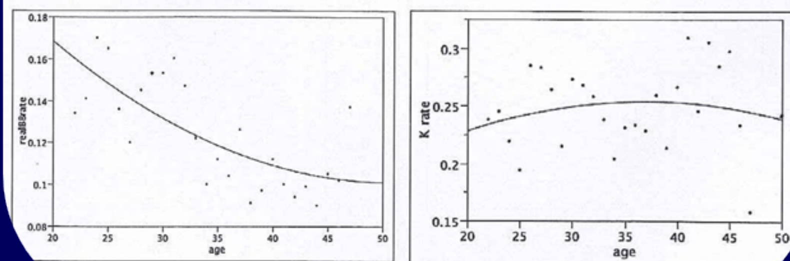


Figure 5. Curt Schilling BB rate and K rate throughout time



6. Nolan Ryan BB rate and K rate throughout time

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

$S_{ijt}=\beta_{0ij}+\beta_{1ij}Age_{it}+\beta_{2ij}Age^{2}_{ij}+\varepsilon_{ijt}$

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

Interest focuses on β_{2ij}

$\beta_{2ij}=0$ linear

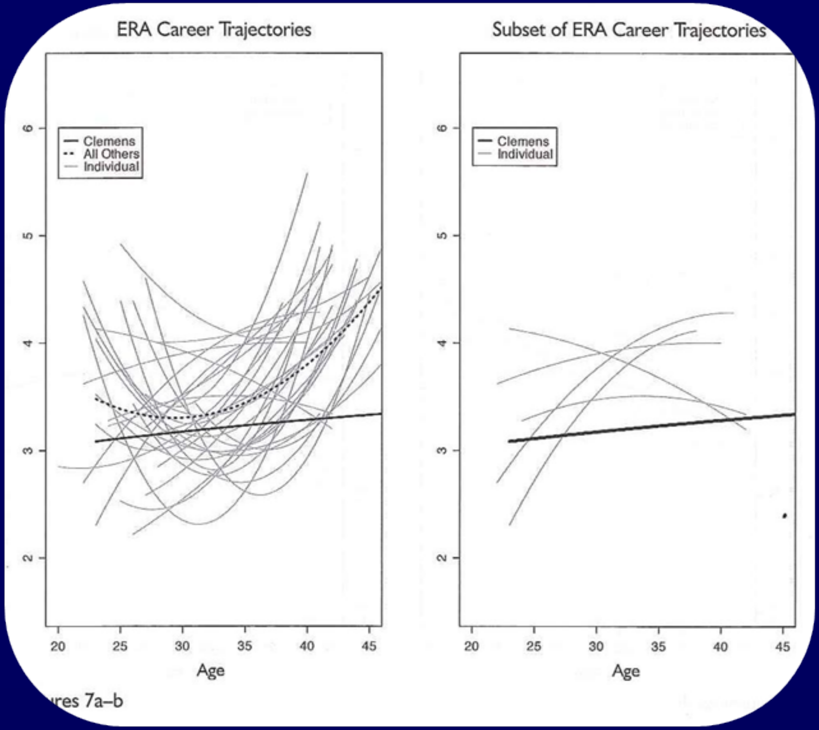
$\beta_{2ij}<0$ hump

$B_{2ij}>0$ 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	$B_{2ij}<0$

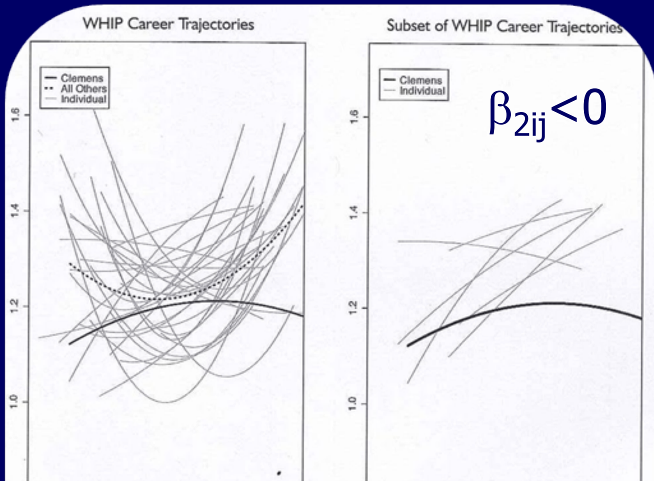
Data from Hendricks report, using ERA



All 32 players

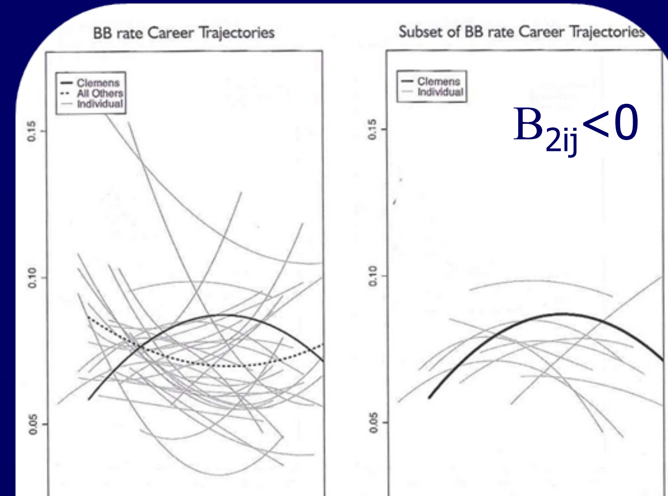
6 atypical players, $\beta_{2ij}<0$

WHIP



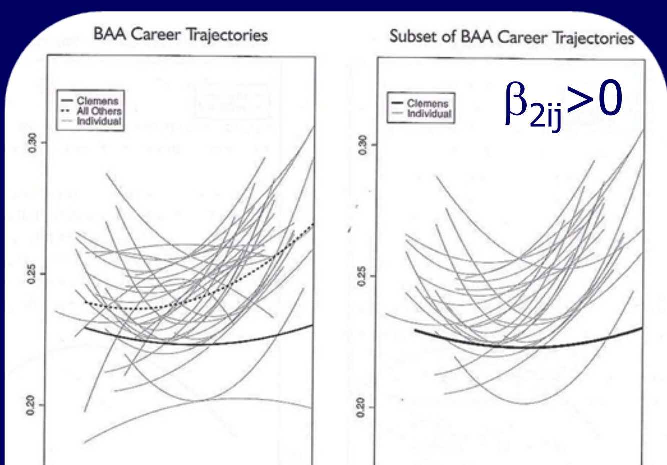
Clemens is only pitcher to get worse mid-career and then better at end

BB rate



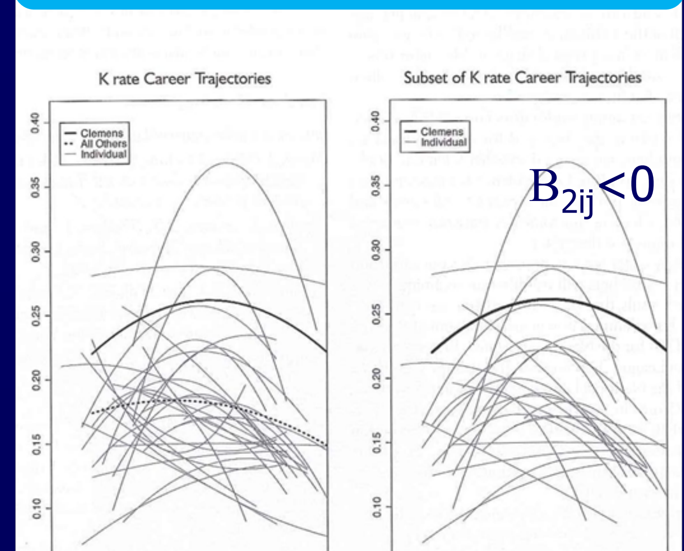
Steepness of Clemens' curve is noticeable in later years.

BAA



Clemens' trajectory has similar shape albeit flatter.

K rate



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."

Misreporting, Missing Data, and Multiple Imputation: Improving Accuracy of Cancer Registry Databases

Yulei He, Recai Yucel, and Alan M. Zaslavsky

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.

Mark Glickman,
Column Editor

Misreporting, Missing Data, and Multiple Imputation: Improving Accuracy of Cancer Registry Databases

Yulei He, Recai Yucel, and Alan M. Zaslavsky

Cancer registries collect information on type of cancer; histological characteristics; stage at 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.

Past literature has documented the inaccuracy of registry records on adjuvant, or supplemental, chemotherapy and radiotherapy. The Quality of Cancer Care (QOCC) project used data from the California Cancer Registry—the largest geographically contiguous, population-based cancer registry n the world—to study the patterns of receiving and reporting juvant therapies for stage II/III colorectal cancer patients. The study surveyed the treating physicians for a subsample of the patients in the registry to obtain more accurate reports of whether they received adjuvant therapies. This study confirmed the inaccuracy of the registry data in favor of under-reporting. Table 1 (line 2 vs. line 1), which is based on this study, implies substantial under-reporting of 20% and 13% in chemotherapy and radiotherapy rates, respectively.

Given that the registry is a valuable data source in health services research, how can we improve quality of inferences using the comprehensive but inaccurate registry database? Consider, for example, that our goal is to obtain accurate estimates of treatment rates from the misreported records in the registry. A simple approach is to use only the validation sample (i.e., the physician survey data collected in the QOCC project). However, due to logistic reasons, the survey sample (< 2000 patients) was much smaller than the registry sample (> 12,000 patients) used in the study; hence, analyzing the validation sample alone would greatly reduce precision, especially for complex estimands such as regression estimates.

Another approach, the errors-in-variables method, would analyze the registry data while adjusting for reporting error. This approach typically involves modeling the relationship between the correct values and misreported ones, represented here by the validation sample and corresponding registry data.

Table 1— Adjuvant Therapy Rates % (SE)

Sample	Chemotherapy		Radiotherapy	
Survey	73.3	(1.16)	25.4	(1.14)
Registry (in the survey region)	57.9	(0.79)	22.2	(0.67)
Registry (statewide)	51.4	(0.45)	19.6	(0.35)
Imputed Registry (statewide)	61.2	(0.77)	23.1	(0.61)

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 1— Adjuvant Therapy Rates % (SE)

Sample	Chemotherapy		Radiotherapy	
Survey	73.3	(1.16)	25.4	(1.14)
Registry (in the survey region)	57.9	(0.79)	22.2	(0.67)
Registry (statewide)	51.4	(0.45)	19.6	(0.35)
Imputed Registry (statewide)	61.2	(0.77)	23.1	(0.61)

Misreporting, Missing Data, and Multiple Imputation: Improving Accuracy of Cancer Registry Databases

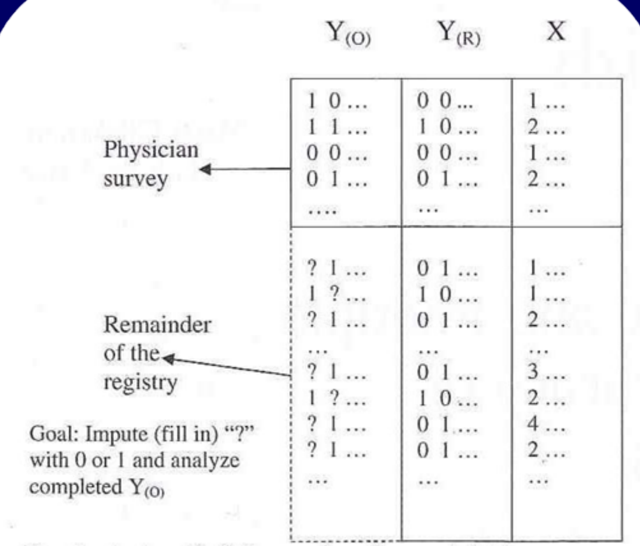
Yulei He, Recai Yucel, and Alan M. Zaslavsky

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



$Y_{(O)}$= true treatment (radiotherapy, chemotherapy, etc.) status
$Y_{(R)}$= reported treatment status
X = covariates

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. $Y_{(R)}$ is the matrix of reported treatment status for various treatments. $Y_{(O)}$ is the true treatment. $Y_{(O)}$ is observed in the survey. An observed value of 1 is assumed to be true, but a value of 0 might be incorrect.

Misreporting, Missing Data, and Multiple Imputation: Improving Accuracy of Cancer Registry Databases

Yulei He, Recai Yucel, and Alan M. Zaslavsky

- 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.

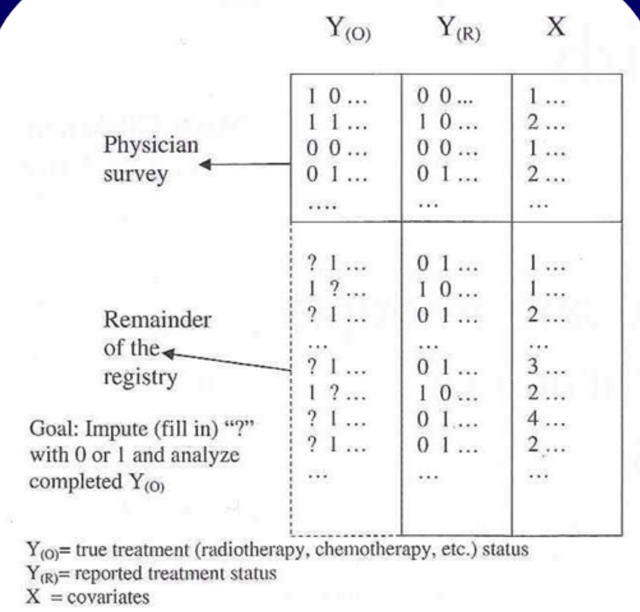


Figure 1. An illustration of using imputation to correct for under-reporting. X is a matrix of covariate variable values with one row for each person in the registry. $Y_{(R)}$ is the matrix of reported treatment status for various treatments. $Y_{(O)}$ is the true treatment. $Y_{(O)}$ is observed in 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
*www.stat.duke.edu/~er**vance**/2002-2003/UNCgame/*

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

Assumes under reporting in the registry

$Y_{(O)}$ truth
$Y_{(R)}$ reported in registry
$X_{(i)}$ covariate

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

Surveyed Students ⟹

$Y_{(O)}$	$Y_{(R)}$	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$
Yes	Yes	Male	G	1400
No	No	Male	N	?
Yes	Yes	Female	N	1380
Yes	No	Male	G	1410
Yes	No	Female	G	?
…				

Remaining Registry ⟹

$Y_{(O)}$	$Y_{(R)}$	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$
Yes	Yes	?	?	1460
?	No	Female	?	1400
Yes	Yes	Male	N	1360
?	No	Female	G	1550
?	No	Male	G	1490
…				

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

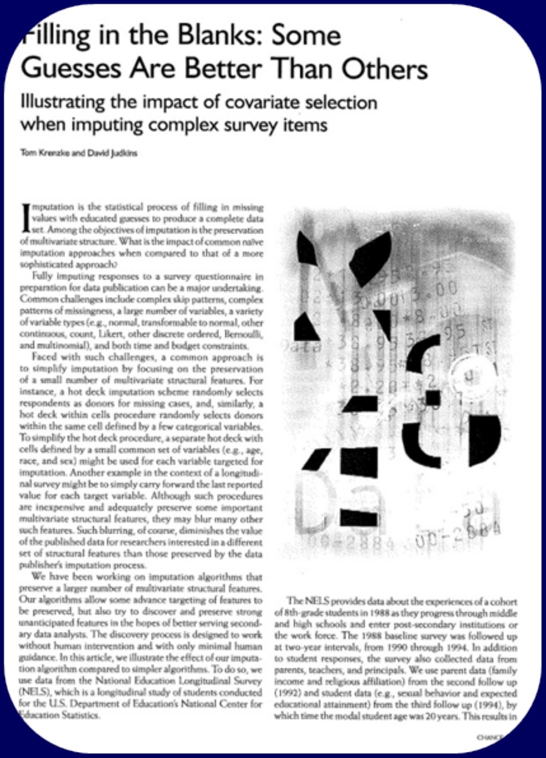
Are particular subgroups more likely to be under reported?

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



Filling in the Blanks: Some Guesses Are Better Than Others

Illustrating the impact of covariate selection when imputing complex survey items

Tom Krenzke and David Judkins

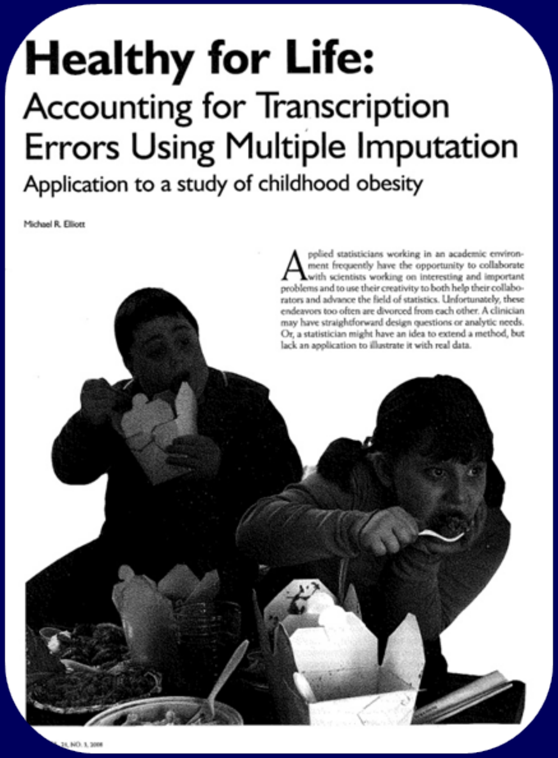
Imputation is the statistical process of filling in missing values with educated guesses to produce a complete data set. Among the objectives of imputation is the preservation of multivariate structure. What is the impact of common naïve imputation approaches when compared to that of a more sophisticated approach?

Fully imputing responses to a survey questionnaire in preparation for data publication can be a major undertaking. Common challenges include complex skip patterns, complex patterns of missingness, a large number of variables, a variety of variable types (e.g., normal, transformable to normal, other continuous, count, Likert, other discrete ordered, Bernoulli, and multinomial), and both time and budget constraints.

Faced with such challenges, a common approach is to simplify imputation by focusing on the preservation of a small number of multivariate structural features. For instance, a hot deck imputation scheme randomly selects respondents as donors for missing cases, and, similarly, a hot deck within cells procedure randomly selects donors within the same cell defined by a few categorical variables. To simplify the hot deck procedure, a separate hot deck with cells defined by a small common set of variables (e.g., age, race, and sex) might be used for each variable targeted for imputation. Another example in the context of a longitudinal survey might be to simply carry forward the last reported value for each target variable. Although such procedures are inexpensive and adequately preserve some important multivariate structural features, they may blur many other such features. Such blurring, of course, diminishes the value of the published data for researchers interested in a different set of structural features than those preserved by the data publisher's imputation process.

We have been working on imputation algorithms that preserve a larger number of multivariate structural features. Our algorithms allow some advance targeting of features to be preserved, but also try to discover and preserve strong unanticipated features in the hopes of better serving secondary data analysts. The discovery process is designed to work without human intervention and with only minimal human guidance. In this article, we illustrate the effect of our imputation algorithm compared to simpler algorithms. To do so, we use data from the National Education Longitudinal Survey (NELS), which is a longitudinal study of students conducted for the U.S. Department of Education's National Center for Education Statistics.

The NELS provides data about the experiences of a cohort of 8th-grade students in 1988 as they progress through middle and high schools and enter post-secondary institutions or the work force. The 1988 baseline survey was followed up at two-year intervals, from 1990 through 1994. In addition to student responses, the survey also collected data from parents, teachers, and principals. We use parent data (family income and religious affiliation) from the second follow up (1992) and student data (e.g., sexual behavior and expected educational attainment) from the third follow up (1994), by which time the modal student age was 20 years. This results in

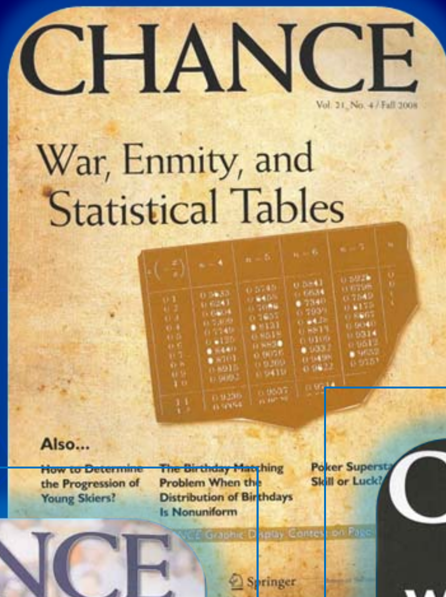


Healthy for Life: Accounting for Transcription Errors Using Multiple Imputation

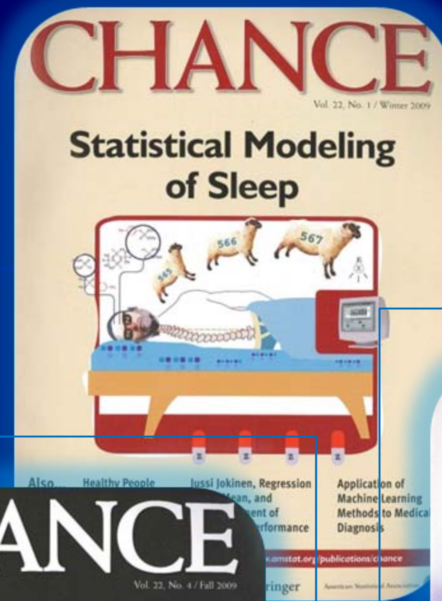
Application to a study of childhood obesity

Michael R. Elliott

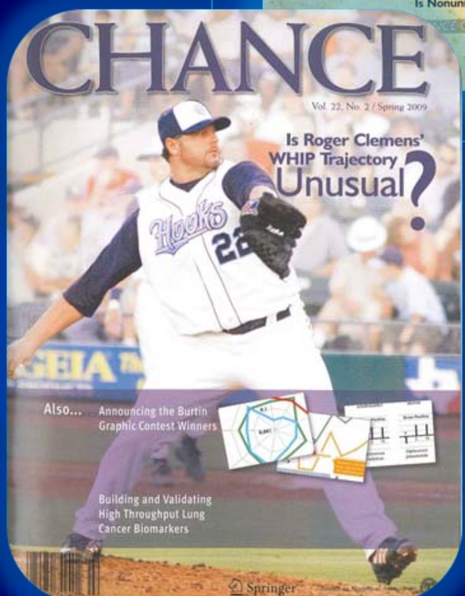
Applied statisticians working in an academic environment frequently have the opportunity to collaborate with scientists working on interesting and important problems and to use their creativity to both help their collaborators and advance the field of statistics. Unfortunately, these endeavors too often are divorced from each other. A clinician may have straightforward design questions or analytic needs. Or, a statistician might have an idea to extend a method, but lack an application to illustrate it with real data.



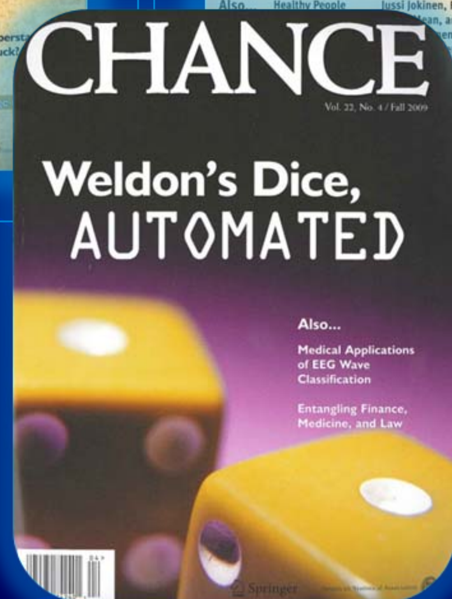
CHANCE
Vol. 21, No. 4 / Fall 2008
War, Enmity, and Statistical Tables
Also...
How to Determine the Progression of Young Skiers?
The Birthday Matching Problem When the Distribution of Birthdays Is Nonuniform
Springer



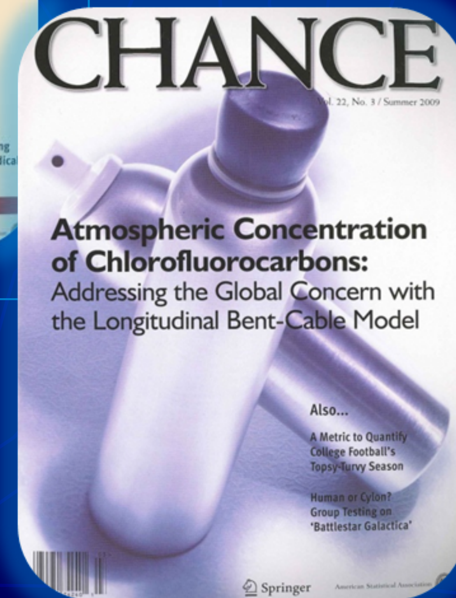
CHANCE
Vol. 22, No. 1 / Winter 2009
Statistical Modeling of Sleep
Application of Machine Learning Methods to Medical Diagnosis



CHANCE
Vol. 22, No. 2 / Spring 2009
Is Roger Clemens' WHIP Trajectory Unusual?
Also...
Announcing the Burtin Graphic Contest Winners
Building and Validating High Throughput Lung Cancer Biomarkers
Springer



CHANCE
Vol. 22, No. 4 / Fall 2009
Weldon's Dice, AUTOMATED
Also...
Medical Applications of EEG Wave Classification
Entangling Finance, Medicine, and Law



CHANCE
Vol. 22, No. 3 / Summer 2009
Atmospheric Concentration of Chlorofluorocarbons: Addressing the Global Concern with the Longitudinal Bent-Cable Model
Also...
A Metric to Quantify College Football's Topsy-Turvy Season
Human or Cylon? Group Testing on 'Battlestar Galactica'
Springer

Probability, Statistics, Evolution, and Intelligent Design

Peter Olofsson

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.

Probability, Statistics, Evolution, and Intelligent Design

Peter Olofsson

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.

Probability, Statistics, and Evolution

The theory of evolution states in part that traits of organisms are passed on to successive generations through genetic material and that modifications in genetic material cause changes in appearance, ability, function, and survival of organisms. Genetic changes that are advantageous to successful reproduction over time dominate and new species evolve. Charles Darwin (1809–1892) is famously credited with originating and popularizing the idea of speciation through gradual change after observing animals on the Galapagos Islands.

Today, the theory of evolution is the scientific consensus concerning the development of species, but is nevertheless routinely challenged by its detractors. The National Academy of Sciences and Institute of Medicine (NAS/IM) recently issued a revised and updated document, titled "Science, Evolution, and Creationism," that describes the theory of evolution and investigates the relation between science and religion. Although the latter topic is of interest in its own right, in fairness to ID proponents, it should be pointed out that many of them do not employ religious arguments against evolution and this article does not deal with issues of faith and religion.

How do probability and statistics enter the scene? In statistics, hypotheses are evaluated with data collected in a way that introduces as little bias as possible and with as much precision as possible. A hypothesis suggests what we would expect to observe or measure, if the hypothesis were true. If such predictions do not agree with the observed data, the hypothesis is rejected and more plausible hypotheses are suggested and evaluated. There are many statistical techniques and methods that may be used, and they are all firmly rooted in the theory of probability, the "mathematics of chance."

An ID Hypothesis Testing Challenge to Evolution

In his book *The Design Inference*, William Dembski introduces the "explanatory filter" as a device to rule out chance explanations and infer design of observed phenomena. The filter also appears in his book *No Free Lunch*, where the description differs slightly. In essence, the filter is a variation on statistical hypothesis testing with the main difference being that it aims at ruling out chance altogether, rather than just a specified null hypothesis. Once all chance explanations have been ruled out, 'design' is inferred. Thus, in this context, design is merely viewed as the complement of chance.

To illustrate the filter, Dembski uses the example of Nicholas Caputo, a New Jersey Democrat who was in charge of putting together the ballots in his county. Names were to be listed in random order, and, supposedly, there is an advantage in having the top line of the ballot. As Caputo managed to place a Democrat on the top line in 40 out of 41 elections, he was suspected of cheating. In Dembski's terminology, cheating now plays the role of design, which is inferred by ruling out chance.

Let us first look at how a statistician might approach the Caputo case. The way in which Caputo was supposed to draw names gives rise to a null hypothesis $H_0 : p = 1/2$ and an alternative hypothesis $H_A : p > ½$, where p is the probability of drawing a Democrat. A standard binomial test of $p = 1/2$ based on the observed relative frequency $\hat{p} = 40/41 \approx 0.98$ gives a solid rejection of H_0 in favor of H_A with a p-value of less than 1 in 50 billion, assuming independent drawings. A statistician could also consider the possibility of different values of p in different drawings, or dependence between listings for different races.

What then would a 'design theorist' do differently? To apply Dembski's filter and infer design, we need to rule out all chance explanations, that is, we need to rule out both H_0 and H_A. There is no way to do so with certainty, and, to continue, we need to use methods other than probability calculations. Dembski's solution is to take Caputo's word that he did not use a flawed randomization device and conclude that the only relevant chance hypothesis is H_0. It might sound questionable to trust a man who is charged with cheating, but as it hardly makes a difference to the case whether Caputo cheated by "intelligent design" or by "intelligent chance," let us not quibble, but generously accept that the explanatory filter reaches the same conclusion as the test: Caputo cheated. The shortcomings of the filter are nevertheless obvious, even in such a simple example.

In *No Free Lunch*, Dembski attempts to apply the filter to a real biological problem: the evolution of the bacterial flagellum, the little whip-like motility device some bacteria

Probability, Statistics, Evolution, and Intelligent Design

Peter Olofsson

An ID Hypothesis Testing Challenge to Evolution

William Dembski

Once all chance explanations have been ruled out 'design' is inferred

$(\text{Chance})^c = \text{Design}$

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.

Probability, Statistics, Evolution, and Intelligent Design

Peter Olofsson

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.

Probability, Statistics, and Evolution

The theory of evolution states in part that traits of organisms are passed on to successive generations through genetic material and that modifications in genetic material cause changes in appearance, ability, function, and survival of organisms. Genetic changes that are advantageous to successful reproduction over time dominate and new species evolve. Charles Darwin (1809–1892) is famously credited with originating and popularizing the idea of speciation through gradual change after observing animals on the Galapagos Islands.

Today, the theory of evolution is the scientific consensus concerning the development of species, but is nevertheless routinely challenged by its detractors. The National Academy of Sciences and Institute of Medicine (NAS/IM) recently issued a revised and updated document, titled "Science, Evolution, and Creationism," that describes the theory of evolution and investigates the relation between science and religion. Although the latter topic is of interest in its own right, in fairness to ID proponents, it should be pointed out that many of them do not employ religious arguments against evolution and this article does not deal with issues of faith and religion.

How do probability and statistics enter the scene? In statistics, hypotheses are evaluated with data collected in a way that introduces as little bias as possible and with as much precision as possible. A hypothesis suggests what we would expect to observe or measure, if the hypothesis were true. If such predictions do not agree with the observed data, the hypothesis is rejected and more plausible hypotheses are suggested and evaluated. There are many statistical techniques and methods that may be used, and they are all firmly rooted in the theory of probability, the "mathematics of chance."

An ID Hypothesis Testing Challenge to Evolution

In his book *The Design Inference*, William Dembski introduces the "explanatory filter" as a device to rule out chance explanations and infer design of observed phenomena. The filter also appears in his book *No Free Lunch*, where the description differs slightly. In essence, the filter is a variation on statistical hypothesis testing with the main difference being that it aims at ruling out chance altogether, rather than just a specified null hypothesis. Once all chance explanations have been ruled out, 'design' is inferred. Thus, in this context, design is merely viewed as the complement of chance.

To illustrate the filter, Dembski uses the example of Nicholas Caputo, a New Jersey Democrat who was in charge of putting together the ballots in his county. Names were to be listed in random order, and, supposedly, there is an advantage in having the top line of the ballot. As Caputo managed to place a Democrat on the top line in 40 out of 41 elections, he was suspected of cheating. In Dembski's terminology, cheating now plays the role of design, which is inferred by ruling out chance.

Let us first look at how a statistician might approach the Caputo case. The way in which Caputo was supposed to draw names gives rise to a null hypothesis $H_0 : p = 1/2$ and an alternative hypothesis $H_A : p > 1/2$, where p is the probability of drawing a Democrat. A standard binomial test of $p = 1/2$ based on the observed relative frequency $\hat{p} = 40/41 \approx 0.98$ gives a solid rejection of H_0 in favor of H_A with a p-value of less than 1 in 50 billion, assuming independent drawings. A statistician could also consider the possibility of different values of p in different drawings, or dependence between listings for different races.

What then would a 'design theorist' do differently? To apply Dembski's filter and infer design, we need to rule out all chance explanations; that is, we need to rule out both H_0 and H_A. There is no way to do so with certainty, and, to continue, we need to use methods other than probability calculations. Dembski's solution is to take Caputo's word that he did not use a flawed randomization device and conclude that the only relevant chance hypothesis is H_0. It might sound questionable to trust a man who is charged with cheating, but as it hardly makes a difference to the case whether Caputo cheated by "intelligent design" or by "intelligent chance," let us not quibble, but generously accept that the explanatory filter reaches the same conclusion as the test: Caputo cheated. The shortcomings of the filter are nevertheless obvious, even in such a simple example.

In *No Free Lunch*, Dembski attempts to apply the filter to a real biological problem: the evolution of the bacterial flagellum, the little whip-like motility device some bacteria

Probability, Statistics, Evolution, and Intelligent Design

Peter Olofsson

An ID Hypothesis Testing Challenge to Evolution

William Dembski

Once all chance explanations have been ruled out 'design' is inferred

$(\text{Chance})^c = \text{Design}$

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.

Probability, Statistics, Evolution, and Intelligent Design

Peter Olofsson

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.

Probability, Statistics, and Evolution

The theory of evolution states in part that traits of organisms are passed on to successive generations through genetic material and that modifications in genetic material cause changes in appearance, ability, function, and survival of organisms. Genetic changes that are advantageous to successful reproduction over time dominate and new species evolve. Charles Darwin (1809–1892) is famously credited with originating and popularizing the idea of speciation through gradual change after observing animals on the Galapagos Islands.

Today, the theory of evolution is the scientific consensus concerning the development of species, but is nevertheless routinely challenged by its detractors. The National Academy of Sciences and Institute of Medicine (NAS/IM) recently issued a revised and updated document, titled "Science, Evolution, and Creationism," that describes the theory of evolution and investigates the relation between science and religion. Although the latter topic is of interest in its own right, in fairness to ID proponents, it should be pointed out that many of them do not employ religious arguments against evolution and this article does not deal with issues of faith and religion.

How do probability and statistics enter the scene? In statistics, hypotheses are evaluated with data collected in a way that introduces as little bias as possible and with as much precision as possible. A hypothesis suggests what we would expect to observe or measure, if the hypothesis were true. If such predictions do not agree with the observed data, the hypothesis is rejected and more plausible hypotheses are suggested and evaluated. There are many statistical techniques and methods that may be used, and they are all firmly rooted in the theory of probability, the "mathematics of chance."

An ID Hypothesis Testing Challenge to Evolution

In his book *The Design Inference*, William Dembski introduces the "explanatory filter" as a device to rule out chance explanations and infer design of observed phenomena. The filter also appears in his book *No Free Lunch*, where the description differs slightly. In essence, the filter is a variation on statistical hypothesis testing with the main difference being that it aims at ruling out chance altogether, rather than just a specified null hypothesis. Once all chance explanations have been ruled out, 'design' is inferred. Thus, in this context, design is merely viewed as the complement of chance.

To illustrate the filter, Dembski uses the example of Nicholas Caputo, a New Jersey Democrat who was in charge of putting together the ballots in his county. Names were to be listed in random order, and, supposedly, there is an advantage in having the top line of the ballot. As Caputo managed to place a Democrat on the top line in 40 out of 41 elections, he was suspected of cheating. In Dembski's terminology, cheating now plays the role of design, which is inferred by ruling out chance.

Let us first look at how a statistician might approach the Caputo case. The way in which Caputo was supposed to draw names gives rise to a null hypothesis $H_0 : p = 1/2$ and an alternative hypothesis $H_A : p > 1/2$, where p is the probability of drawing a Democrat. A standard binomial test of $p = 1/2$ based on the observed relative frequency $\hat{p} = 40/41 \approx 0.98$ gives a solid rejection of H_0 in favor of H_A with a p-value of less than 1 in 50 billion, assuming independent drawings. A statistician could also consider the possibility of different values of p in different drawings, or dependence between listings for different races.

What then would a 'design theorist' do differently? To apply Dembski's filter and infer design, we need to rule out all chance explanations; that is, we need to rule out both H_0 and H_A. There is no way to do so with certainty, and, to continue, we need to use methods other than probability calculations. Dembski's solution is to take Caputo's word that he did not use a flawed randomization device and conclude that the only relevant chance hypothesis is H_0. It might sound questionable to trust a man who is charged with cheating, but as it hardly makes a difference to the case whether Caputo cheated by "intelligent design" or by "intelligent chance," let us not quibble, but generously accept that the explanatory filter reaches the same conclusion as the test: Caputo cheated. The shortcomings of the filter are nevertheless obvious, even in such a simple example.

In *No Free Lunch*, Dembski attempts to apply the filter to a real biological problem: the evolution of the bacterial flagellum, the little whip-like motility device some bacteria