Implications of Big Data for Statistics Instruction

Mark L. Berenson
Montclair State University
MSMESB Mini-Conference
DSI - Baltimore
November 17, 2013

Teaching Introductory Business Statistics to Undergraduates in an Era of Big Data

"The integration of business, Big Data and statistics is both necessary and long overdue."

Kaiser Fung (*Significance*, August 2013)

Computer Scientists and Statisticians Must Coordinate to Accomplish a Common Goal: Making Reliable Decisions from the Available Data.

- **Computer Scientist's Concern is Data Management**
- **Statistician's Concern is Data Analysis**
- **Computer Scientist's Interest is in Quantity of Data**
- **Statistician's Interest is in Quality of Data**
- **Computer Scientist's Decisions are Based on Frequency of Counts**
- **Statistician's Decisions are Based on Magnitude of Effect**

Kaiser Fung (*Significance*, August 2013)

Bigger *n* Doesn't Necessarily Mean Better Results

- 128,053,180 was the USA population in 1936
- 78,000,000 were Voting Age Eligible (61.0%)
- 27,752,648 voted for Roosevelt (60.8%)
- 16,681,862 voted for Landon (36.5%)
- 10,000,000 received mailed surveys from *Literary Digest*
- 2,300,000 responded to the mailed survey

A Proposal for an Introductory Undergraduate Business Statistics Course in an Era of Big Data

- Course Constraints:
  - Type: One 3-Credit Core-Required Course
  - Prerequisite: Intermediate Algebra
  - Articulation: 19 NJ community colleges
  - Software: Excel
  - Student Body: Preparation assessment

A New Business Statistics Course

- ***Note:*** In the next several slides all statements in **GREEN font** reflect items or topics relevant to a proposed introductory business statistics course in an era of Big Data that are not typically taught at the present time.

Goals for the New Business Statistics Course

Numerical literacy is essential in business and the discipline of statistics enables students to learn to:

- Visualize Data
- Draw Inferences
- Make Predictions
- Manage Processes

regardless of sample size

Course Topics
A. Introduction to Business Statistics

- An historical Introduction to the subject and practice of statistics as it evolves into an era of Big Data
- A list of key terms used in the practice of statistics
- A review of the fundamentals of business numeracy
  - Proportions, Percentages, Percentage Change, Rates, Ratios, Odds, Index Numbers

Course Topics
B. Obtaining Data

- **Data Types:**
  - **Structured vs. Unstructured.**
- **Data Sources:**
  - **Surveys**
  - **Experiments**
  - **Observational Studies**
  - **Primary/Secondary Source Acquisitions**
  - **Transactions**
  - **Log Data**
  - **Email**
  - **Social Media**
  - **Sensors**
  - **Free-form Text**
  - **Geospatial**
  - **Audio**
  - **Image**

Course Topics
B. Obtaining Data

- Data Outcomes:
  - Categorical
  - Numerical
  - Other

Course Topics
C. Categorical Data Visualization and Description

- Summary Table; Bar Chart and Pareto Chart
  -Mode
- Cross-classification Table; Side-by-Side Bar Chart
- Pivot Table
- Dashboards
- Report Cards
- Interactive Graphs

Course Topics:
D. Numerical Data Visualization and Description

- Ordered Array
- Frequency and Percentage Distributions using Sturges' Rule for Bin Groupings
- Histogram and Percentage Polygon
- Boxplot with Five-Number Summary
- Deciles and Percentiles
- Central Tendency: Mean, Median, Trimean, 10 % Trimmed Mean
- Variation: Standard Deviation, IQR, IDR, CV
- Skewness: Stine & Foster K_3
- Kurtosis: Stine & Foster K_4
- Searching for Outliers: Z value, Tukey EDA methods

Course Topics
E. Probability

- Definitions: *A Priori*, Empirical, Subjective
- Marginal, Joint and Conditional Probability
- Bayes Theorem

Course Topics
F. Probability Distributions

- Discrete: Definition and Examples
- Continuous: Definition and Examples
- The Standardized Normal Distribution
- Assessing Normality

Course Topics
G. Sampling Distributions

- Sampling Distribution of the Mean
- Sampling Distribution of the Proportion
- The Central Limit Theorem

Course Topics
H. Inference

- Probability Sampling in Surveys and Randomization in Experiments
  - C. I. E. of the Population Mean
  - C. I. E. of the Population Proportion
  - Concept of Effect Size for Comparing Two Groups (A/B Testing)
  - C. I. E. of the Difference in Two Independent Group Means
  - C. I. E. of the Standardized Mean Difference Effect Size
  - C. I. E. of the Population Point Biserial Correlation Effect Size
  - C. I. E. of the Difference in Two Independent Group Proportions
  - Phi-Coefficient Measure of Association in 2x2 Tables
  - C. I. E. of the Population Odds Ratio Effect Size

Course Topics:
I. Simple Linear Regression Modeling

- Descriptive Analysis and Assessment of Model Appropriateness
- Effect Size for the Slope and for the Coefficient of Correlation
- Confidence Interval Estimate of the Mean Response
- Prediction Interval Estimate of the Individual Response

Course Topics
J. Quality Management

- Introduction to Process Management
- The Use of Control Charts

Summary and Conclusions

- A course in Business Statistics needs to be modified to maintain its relevance in an era of Big Data.
- Business statistics textbooks must adapt its topic coverage to introduce methodology relevant to a Big Data environment – the subject of inference must be re-engineered.
- The time has come for AACSB-accredited undergraduate programs to include a core-required course in Business Analytics as a sequel to a course in Business Statistics.

Implications of Big Data for Statistics Instruction

Mark L. Berenson
Montclair State University

MSMESB Mini-Conference
DSI - Baltimore
November 17, 2013

Teaching Introductory Business Statistics to Undergraduates in an Era of Big Data

"The integration of business, Big Data and statistics is both necessary and long overdue."

Kaiser Fung (*Significance*, August 2013)

Computer Scientists and Statisticians Must Coordinate to Accomplish a Common Goal: Making Reliable Decisions from the Available Data.

- **Computer Scientist's Concern is Data Management**
- **Statistician's Concern is Data Analysis**
- **Computer Scientist's Interest is in Quantity of Data**
- **Statistician's Interest is in Quality of Data**
- **Computer Scientist's Decisions are Based on Frequency of Counts**
- **Statistician's Decisions are Based on Magnitude of Effect**

Kaiser Fung (*Significance*, August 2013)

Bigger *n* Doesn't Necessarily Mean Better Results

- 128,053,180 was the USA population in 1936
- 78,000,000 were Voting Age Eligible (61.0%)
- 27,752,648 voted for Roosevelt (60.8%)
- 16,681,862 voted for Landon (36.5%)
- 10,000,000 received mailed surveys from *Literary Digest*
- 2,300,000 responded to the mailed survey

A Proposal for an Introductory Undergraduate Business Statistics Course in an Era of Big Data

- Course Constraints:
  - Type: One 3-Credit Core-Required Course
  - Prerequisite: Intermediate Algebra
  - Articulation: 19 NJ community colleges
  - Software: Excel
  - Student Body: Preparation assessment

A New Business Statistics Course

- ***Note:*** In the next several slides all statements in **GREEN font** reflect items or topics relevant to a proposed introductory business statistics course in an era of Big Data that are not typically taught at the present time.

Goals for the New Business Statistics Course

Numerical literacy is essential in business and the discipline of statistics enables students to learn to:

- Visualize Data
- Draw Inferences
- Make Predictions
- Manage Processes

regardless of sample size

Course Topics

A. Introduction to Business Statistics

- An historical Introduction to the subject and practice of statistics as it evolves into an era of Big Data
- A list of key terms used in the practice of statistics
- A review of the fundamentals of business numeracy
  - Proportions, Percentages, Percentage Change, Rates, Ratios, Odds, Index Numbers

Course Topics

B. Obtaining Data

- **Data Types:**
  - **Structured vs. Unstructured.**
- **Data Sources:**
  - **Surveys**
  - **Experiments**
  - **Observational Studies**
  - **Primary/Secondary Source Acquisitions**
  - **Transactions**
  - **Log Data**
  - **Email**
  - **Social Media**
  - **Sensors**
  - **Free-form Text**
  - **Geospatial**
  - **Audio**
  - **Image**

Course Topics
B. Obtaining Data

- Data Outcomes:
  - Categorical
  - Numerical
  - Other

Course Topics

C. Categorical Data Visualization and Description

- Summary Table; Bar Chart and Pareto Chart
  - -Mode
- Cross-classification Table; Side-by-Side Bar Chart
- Pivot Table
- Dashboards
- Report Cards
- Interactive Graphs

Course Topics:
D. Numerical Data Visualization and Description

- Ordered Array
- Frequency and Percentage Distributions using Sturges' Rule for Bin Groupings
- Histogram and Percentage Polygon
- Boxplot with Five-Number Summary
- Deciles and Percentiles
- Central Tendency: Mean, Median, Trimean, 10 % Trimmed Mean
- Variation: Standard Deviation, IQR, IDR, CV
- Skewness: Stine & Foster K_3
- Kurtosis: Stine & Foster K_4
- Searching for Outliers: Z value, Tukey EDA methods

Course Topics
E. Probability

- Definitions: *A Priori*, Empirical, Subjective
- Marginal, Joint and Conditional Probability
- Bayes Theorem

Course Topics

F. Probability Distributions

- Discrete: Definition and Examples
- Continuous: Definition and Examples
- The Standardized Normal Distribution
- Assessing Normality

Course Topics

G. Sampling Distributions

- Sampling Distribution of the Mean
- Sampling Distribution of the Proportion
- The Central Limit Theorem

Course Topics

H. Inference

- Probability Sampling in Surveys and Randomization in Experiments
  - C. I. E. of the Population Mean
  - C. I. E. of the Population Proportion
  - Concept of Effect Size for Comparing Two Groups (A/B Testing)
  - C. I. E. of the Difference in Two Independent Group Means
  - C. I. E. of the Standardized Mean Difference Effect Size
  - C. I. E. of the Population Point Biserial Correlation Effect Size
  - C. I. E. of the Difference in Two Independent Group Proportions
  - Phi-Coefficient Measure of Association in 2x2 Tables
  - C. I. E. of the Population Odds Ratio Effect Size

Course Topics:

I. Simple Linear Regression Modeling

- Descriptive Analysis and Assessment of Model Appropriateness
- Effect Size for the Slope and for the Coefficient of Correlation
- Confidence Interval Estimate of the Mean Response
- Prediction Interval Estimate of the Individual Response

Course Topics

J. Quality Management

- Introduction to Process Management
- The Use of Control Charts

Summary and Conclusions

- A course in Business Statistics needs to be modified to maintain its relevance in an era of Big Data.
- Business statistics textbooks must adapt its topic coverage to introduce methodology relevant to a Big Data environment – the subject of inference must be re-engineered.
- The time has come for AACSB-accredited undergraduate programs to include a core-required course in Business Analytics as a sequel to a course in Business Statistics.