Biostatistics: Paving the Way in the Big Data Era

DuBois Bowman, PhD
Department of Biostatistics
Professor and Chairman

July 21, 2016
Data
Data Are ...

- Ubiquitous
- Increasing in complexity
- Rapidly expanding in scale
  - Quickly and inexpensively generated by emerging technologies
  - Increased storage capacity
  - Increased number of experimental units
  - Increased temporal scales
  - Increased spatial scales
Big Data
Big Data

Google Trends

Explore

Compare

Search terms

big data

+ Add term

Interest over time

Big Data

A Revolution

• A total of 2.5 quintillion bytes of data were generated every day in 2012 alone

• Decoding...
  o 2.5 trillion megabytes
  o If these units (megabytes) were time...say seconds...

  It would take 4125 years to elapse
Big Data

Electronic medical/health records

- What type of care would improve health outcomes?
- How to minimize risk for readmission?
- Identify high risk patients
- from data.

Bowman, D. 2016 Pipelines into Biostatistics
Big Data

Personal smart devices

- Physical activity
- Medication use / compliance
- Monitor symptoms of medical conditions
Big Data

Genetic/Genomic Analyses

• Genetic basis of human diseases and traits
Big Data

**Neuroimaging**

- Localize brain regions altered in neurological or psychiatric diseases.
- Identify precursors of disease prior to clinical symptoms.
- Determine which patients are likely to respond to particular treatments.
"Smart" devices are creating a flood of data. Now hospitals need to figure out how to manage it all.

**Hospitals: Big data use is a 'significant challenge'**

August 13, 2013 | By Dan Bowman

Second, companies are struggling to find the right talent capable of both working with new technologies and of interpreting the data to find meaningful business insights.
Need More than Just Data

The 3 big problems in big data (hint: They all involve people)

2. Data Scientists

A typical enterprise generally has 10x more IT employees than analysts or data scientists. The process of analysis starts with a line of business request. IT collects data from various databases and transfers it to data scientists. Large teams of data scientists are deployed who spend months (or sometimes years) querying the data. Hiring data scientists (with advanced background in statistics, computer science, and some functional expertise) to accelerate the process is difficult because people with these skills are extremely scarce. The demand and interest in data scientists is skyrocketing, as Google Trends can attest, while we are producing fewer of them. What we need is a new class of technologies that amplify the impact data scientists and allow more people to become data scientists.
Need: Biostatistics / Statistics
Biostatistics

- Recent focus on data science, analytics, etc.
- Statistical thinking is fundamental to all of these areas
  - Rests upon well-established principles of estimation, inference, and prediction
  - Head start:
    - Long history...dating back to 1749
    - American Statistical Association was founded in 1839 (Over 175 years!)
    - Many seminal results established in early 1900’s
What is Biostatistics?
What is Biostatistics?

Definition of Statistics

1: a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data
What is Biostatistics?

Branch of mathematics dealing with gathering, analyzing, and making inferences from data. Originally associated with government data (e.g., census data), the subject now has applications in all the sciences. Statistical tools not only summarize past data through such indicators as the mean (see **mean**, **median**, and **mode**) and the standard deviation but can predict future events using **frequency distribution** functions. Statistics provides ways to design efficient experiments that eliminate time-consuming trial and error.
What is Biostatistics?

Definition of Biostatistics

statistics applied to the analysis of biological data

Bowman, D. 2016 Pipelines into Biostatistics
Biostatistics

Helps to gain insights into health and medicine.

- **Quantify uncertainty**: how likely or unlikely something is to occur (or via variability).

- **Detect differences**, e.g. between patients treated with a new cholesterol lowering medication and those receiving standard therapy.
Biostatistics

• **Estimate unknown quantities:**
  - **WHO** is at high risk for developing Alzheimer’s disease?
  - **WHAT** is the safest and most effective dose for a targeted cancer therapy?
  - **WHERE** (in which neighborhoods) are there disproportionate levels of environmental pollutants causing respiratory diseases and other illnesses?
  - **WHEN** is the optimal time to intervene with a vaccination to protect against widespread outbreak of a potentially fatal infectious disease?
  - **WHY** are the rates for autism increasing?

• **Predict**, e.g. health outcomes (progression of diseases, relapse, cure).
Big Data

Magnetic Resonance Imaging (MRI)
Big Data

Magnetic Resonance Imaging (MRI)

\[ \int E \cdot ds = -\frac{d}{dt} \int B \cdot dA \]
\[ \int B \cdot dA = 0 \]
Brain Imaging
- Depression
- Schizophrenia
- Bipolar Disorder
- Social Anxiety
- Cocaine dependence
- Parkinson’s Disease
- Alzheimer’s Disease
- Stroke Recovery
- Autism
- Moral Reasoning

Cardiac Imaging
- Recovery following a heart attack

Cancer Imaging
- Breast cancer
  - Early diagnosis
  - Monitoring treatment
- Prostate cancer
The Neural Correlates of Social Anxiety Disorder and Response to Pharmacotherapy

**Figure 3** Greater ($p < 0.005$, uncorrected) provoked social anxiety-related rCBF before compared to after treatment of individuals with gSAD with nefazodone. Common effects are illustrated in parasagittal views for the social anxiety imagery (top) and confrontational mental arithmetic task provocation (bottom) conditions. Abbreviations refer to the right brainstem/midbrain (BrS), left middle (F46), and right medial (F9) frontal gyri, left precentral gyrus (PrG6), and left temporal pole (T38).
Predicting future brain activity in patients with Alzheimer’s disease
Age and task-related differences in functional brain networks using awFC

Bowman et al., 2012, NeuroImage
Biostatistics: Neuroimaging

Multimodal neuroimaging data

AAL Subregions
Subregions Seeds
Subregions Spheres

ALFF
VBM

Functional Connectivity

Structural Connectivity

Bowman, D.

2016 Pipelines into Biostatistics
Markers of PD:

- Model 2
  - FC: Amygdala R x Angular R
  - FC: Calcarine L x Thalamus L
  - FC: Frontal Sup Orb L x Temporal Sup R

- Model 3
  - FC: Cuneus R x Precuneus R
  - FC: Frontal Sup Medial L x Cingulum Ant L
  - FC: Frontal Sup Orb L x Temporal Pole Sup L

Bowman et al., 2016, Front in Neurosc

Bowman, D.
The Future is Bright
Biostatistics

Best Graduate Degrees for Jobs in 2016:

• **#1: Master’s in Biostatistics**
  - Median Salary: $106K
  - Projected Job Growth by 2024: 23%
  - Percent Highly Satisfied: 85%
  - Low Stress: 57%

• 2. Master’s, Statistics

• 3. Ph.D., Computer Science

• 4. Ph.D., Economics

• 5. Master’s, Applied Mathematics

Fortune magazine, 2016


**Job Outlook**

Employment of statisticians is projected to grow 34 percent from 2014 to 2024, much faster than the average for all occupations. Growth is expected to result from more widespread use of statistical analysis to make informed business, healthcare, and policy decisions. In addition, the large increase in available data from the Internet will open up new areas for analysis.

A substantial amount of data is generated from Internet searching and the use of social media, smartphones, and other mobile devices. Businesses, particularly those in the retail, finance, and insurance industries, will increasingly need statisticians to organize, analyze, and sort through the data for commercial reasons. Analyses will help companies improve their business processes, design and develop new products, and advertise products to potential customers.

In addition, statisticians will be needed in the pharmaceutical industry. The aging of the U.S. population will encourage pharmaceutical companies to develop new treatments and medical technologies. Biostatisticians will be needed to conduct the research and clinical trials necessary for companies to obtain approval for their products from the Food and Drug Administration.

The occupation will also see growth in research and development in the physical, engineering, and life sciences, fields in which statisticians’ skills in designing tests and assessing results are highly useful.

---

**Statisticians**

Percent change in employment, projected 2014-24

- **Statisticians**: 34%
- **Mathematical science occupations**: 28%
- **Total, all occupations**: 7%

*Note: All Occupations includes all occupations in the U.S. Economy.*

*Source: U.S. Bureau of Labor Statistics, Employment Projections program*
My Journey
My Educational Background

Undergraduate
- Morehouse College, B.S. Mathematics
  - COR (MARC) Program

Graduate
- University of Michigan, M.S. Biostatistics
- UNC at Chapel Hill, Ph.D. Biostatistics
My Faculty Career

Department of Biostatistics

Emory University
- Assistant Professor
- Associate Professor
- Full Professor

Columbia University
- Full Professor and Chairman
Networking

Professional and Personal Networks

Parents

Mentors

Teachers

Scholars

Friends

Classmates

Bowman, D.
Diversity
### Underrepresentation in Academia:

#### Faculty

<table>
<thead>
<tr>
<th>Year, sex, and academic rank</th>
<th>Total</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>Pacific Islander</th>
<th>American Indian/Alaska Native</th>
<th>Two or more races</th>
<th>Race/ethnicity unknown</th>
<th>Non-resident alien</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>791,391</td>
<td>575,491</td>
<td>43,188</td>
<td>33,217</td>
<td>71,038</td>
<td>1,208</td>
<td>3,538</td>
<td>5,291</td>
<td>20,013</td>
<td>38,407</td>
</tr>
<tr>
<td>[Pct of Faculty]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professors</td>
<td>181,530</td>
<td>148,577</td>
<td>6,665</td>
<td>5,604</td>
<td>15,247</td>
<td>170</td>
<td>573</td>
<td>852</td>
<td>2,323</td>
<td>1,519</td>
</tr>
<tr>
<td>[Pct of Profs]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Associate Professors</td>
<td>155,095</td>
<td>116,817</td>
<td>8,812</td>
<td>6,381</td>
<td>15,626</td>
<td>183</td>
<td>591</td>
<td>987</td>
<td>2,859</td>
<td>2,839</td>
</tr>
<tr>
<td>[Pct of Assoc. Profs]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assistant Professors</td>
<td>166,045</td>
<td>112,262</td>
<td>10,542</td>
<td>7,130</td>
<td>18,070</td>
<td>332</td>
<td>683</td>
<td>1,254</td>
<td>5,695</td>
<td>10,077</td>
</tr>
<tr>
<td>[Pct of Assist Profs]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructors</td>
<td>99,304</td>
<td>73,859</td>
<td>7,448</td>
<td>6,340</td>
<td>4,950</td>
<td>286</td>
<td>879</td>
<td>781</td>
<td>3,180</td>
<td>1,581</td>
</tr>
<tr>
<td>Lecturers</td>
<td>36,728</td>
<td>27,453</td>
<td>1,728</td>
<td>2,015</td>
<td>2,403</td>
<td>33</td>
<td>117</td>
<td>295</td>
<td>1,151</td>
<td>1,533</td>
</tr>
<tr>
<td>Other faculty</td>
<td>152,689</td>
<td>96,523</td>
<td>7,993</td>
<td>5,747</td>
<td>14,742</td>
<td>204</td>
<td>695</td>
<td>1,122</td>
<td>4,805</td>
<td>20,858</td>
</tr>
</tbody>
</table>
Underrepresentation in Academia:

**Faculty (2013)**
- 791,391 full-time faculty at degree-granting institutions
  - 43,188 (5.5%) were Black [Includes faculty at HBCUs]
  - 33,217 (4.2%) were Hispanic
- Full professor: 6,665 (3.7%) Black / 5,604 (3.1%) Hispanic (of 181,530)
- Associate professor: 8,812 (5.7%) Black / 6,381 (4.1%) Hispanic (of 155,095)
- Assistant professor: 10,542 (6.3%) Black / 7,130 (4.3%) Hispanic (of 166,045)

**The Pool (2012)**
- 224,000 African Americans held doctorates.
- 217,000 AAs held a professional degree
- 1,364,000 AAs with a master’s degree (as the highest degree).
- 3.5 million AAs with a bachelor’s degree (but no graduate degree)
- **10.1%** of all degrees granted at 4 year institutions

[Faculty, Racial Gap, Research & Studies on July 12, 2012]
### Underrepresentation in Academia:

<table>
<thead>
<tr>
<th>Institution</th>
<th>All Applicants</th>
<th>Total Applicants Accepted</th>
<th>Overall Student Acceptance Rate</th>
<th>Black Applicants</th>
<th>Blacks Accepted</th>
<th>Black Acceptance Rate</th>
<th>Black Enrolees</th>
<th>Black Student Yield*</th>
<th>Black % of First-Year Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbia University</td>
<td>31,851</td>
<td>2,362</td>
<td>7.4</td>
<td>3,881</td>
<td>**</td>
<td>**</td>
<td>201</td>
<td>**</td>
<td>14.2</td>
</tr>
<tr>
<td>Duke University</td>
<td>31,566</td>
<td>4,037</td>
<td>12.9</td>
<td>3,215</td>
<td>**</td>
<td>**</td>
<td>198</td>
<td>**</td>
<td>11.5</td>
</tr>
<tr>
<td>Univ. of Pennsylvania</td>
<td>31,218</td>
<td>3,935</td>
<td>12.6</td>
<td>2,766</td>
<td>**</td>
<td>**</td>
<td>272</td>
<td>**</td>
<td>11.1</td>
</tr>
<tr>
<td>Univ. of N.C Chapel Hill</td>
<td>29,947</td>
<td>7,847</td>
<td>26.2</td>
<td>3,254</td>
<td>799</td>
<td>24.6</td>
<td>382</td>
<td>47.8</td>
<td>9.8</td>
</tr>
<tr>
<td>Vanderbilt University</td>
<td>28,348</td>
<td>4,034</td>
<td>14.2</td>
<td>2,507</td>
<td>411</td>
<td>16.4</td>
<td>155</td>
<td>37.7</td>
<td>9.6</td>
</tr>
<tr>
<td>Brown University</td>
<td>28,742</td>
<td>2,759</td>
<td>9.6</td>
<td>2,348</td>
<td>331</td>
<td>14.1</td>
<td>146</td>
<td>44.1</td>
<td>9.5</td>
</tr>
<tr>
<td>Harvard University</td>
<td>34,302</td>
<td>2,032</td>
<td>5.9</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>150</td>
<td>**</td>
<td>9.4</td>
</tr>
<tr>
<td>Yale University</td>
<td>28,977</td>
<td>2,043</td>
<td>7.1</td>
<td>2,035</td>
<td>**</td>
<td>**</td>
<td>122</td>
<td>**</td>
<td>9.0</td>
</tr>
<tr>
<td>Emory University</td>
<td>17,500</td>
<td>4,599</td>
<td>26.3</td>
<td>2,337</td>
<td>460</td>
<td>19.7</td>
<td>120</td>
<td>26.1</td>
<td>8.6</td>
</tr>
<tr>
<td>Mass. Inst. of Technology</td>
<td>18,109</td>
<td>1,620</td>
<td>8.9</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>91</td>
<td>8.0</td>
</tr>
<tr>
<td>Stanford University</td>
<td>36,632</td>
<td>2,423</td>
<td>6.6</td>
<td>**</td>
<td>233</td>
<td>**</td>
<td>142</td>
<td>60.9</td>
<td>8.0</td>
</tr>
<tr>
<td>Cornell University</td>
<td>37,808</td>
<td>6,259</td>
<td>16.6</td>
<td>3,201</td>
<td>597</td>
<td>18.7</td>
<td>257</td>
<td>43.0</td>
<td>7.9</td>
</tr>
<tr>
<td>Dartmouth College</td>
<td>23,110</td>
<td>2,262</td>
<td>9.8</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>87</td>
<td>7.9</td>
</tr>
<tr>
<td>Northwestern University</td>
<td>32,069</td>
<td>4,324</td>
<td>15.4</td>
<td>2,043</td>
<td>465</td>
<td>22.8</td>
<td>161</td>
<td>34.6</td>
<td>7.9</td>
</tr>
<tr>
<td>Johns Hopkins University</td>
<td>20,502</td>
<td>3,626</td>
<td>17.7</td>
<td>1,711</td>
<td>381</td>
<td>22.3</td>
<td>102</td>
<td>26.8</td>
<td>7.7</td>
</tr>
<tr>
<td>Princeton University</td>
<td>26,664</td>
<td>2,084</td>
<td>7.9</td>
<td>1,546</td>
<td>**</td>
<td>**</td>
<td>102</td>
<td>**</td>
<td>7.5</td>
</tr>
<tr>
<td>University of Chicago</td>
<td>25,307</td>
<td>3,363</td>
<td>13.3</td>
<td>2,309</td>
<td>**</td>
<td>**</td>
<td>113</td>
<td>**</td>
<td>7.4</td>
</tr>
<tr>
<td>Wake Forest University</td>
<td>11,408</td>
<td>3,875</td>
<td>34.0</td>
<td>1,612</td>
<td>299</td>
<td>18.5</td>
<td>87</td>
<td>29.1</td>
<td>7.0</td>
</tr>
<tr>
<td>Rice University</td>
<td>15,133</td>
<td>2,528</td>
<td>16.7</td>
<td>948</td>
<td>200</td>
<td>21.1</td>
<td>64</td>
<td>32.0</td>
<td>6.8</td>
</tr>
<tr>
<td>Carnegie Mellon Univ.</td>
<td>17,313</td>
<td>4,748</td>
<td>27.4</td>
<td>812</td>
<td>326</td>
<td>40.1</td>
<td>96</td>
<td>29.4</td>
<td>6.7</td>
</tr>
<tr>
<td>Washington University</td>
<td>27,265</td>
<td>4,876</td>
<td>17.9</td>
<td>2,149</td>
<td>364</td>
<td>16.9</td>
<td>104</td>
<td>28.6</td>
<td>6.3</td>
</tr>
<tr>
<td>University of Virginia</td>
<td>28,251</td>
<td>8,031</td>
<td>28.4</td>
<td>1,708</td>
<td>508</td>
<td>29.7</td>
<td>188</td>
<td>37.0</td>
<td>5.5</td>
</tr>
<tr>
<td>University of Notre Dame</td>
<td>16,957</td>
<td>3,947</td>
<td>23.3</td>
<td>830</td>
<td>267</td>
<td>32.2</td>
<td>110</td>
<td>41.2</td>
<td>5.4</td>
</tr>
<tr>
<td>Univ. of Southern Calif.</td>
<td>54,367</td>
<td>11,957</td>
<td>22.0</td>
<td>4,470</td>
<td>696</td>
<td>15.6</td>
<td>248</td>
<td>35.6</td>
<td>5.3</td>
</tr>
<tr>
<td>Tufts University</td>
<td>15,378</td>
<td>3,504</td>
<td>21.4</td>
<td>948</td>
<td>233</td>
<td>24.6</td>
<td>67</td>
<td>28.8</td>
<td>5.1</td>
</tr>
<tr>
<td>University of Michigan</td>
<td>42,544</td>
<td>15,551</td>
<td>36.6</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>323</td>
<td>5.2</td>
</tr>
<tr>
<td>Georgetown University</td>
<td>20,116</td>
<td>3,413</td>
<td>17.0</td>
<td>1,931</td>
<td>335</td>
<td>17.3</td>
<td>126</td>
<td>37.6</td>
<td>4.9</td>
</tr>
<tr>
<td>Univ. of Calif. Los Angeles</td>
<td>52,231</td>
<td>9,263</td>
<td>17.7</td>
<td>3,036</td>
<td>348</td>
<td>11.5</td>
<td>181</td>
<td>52.0</td>
<td>4.4</td>
</tr>
<tr>
<td>Univ. of Calif. Berkeley</td>
<td>41,159</td>
<td>9,348</td>
<td>22.7</td>
<td>2,394</td>
<td>324</td>
<td>13.5</td>
<td>152</td>
<td>46.9</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Universities are listed by highest percentage of black freshmen.

**Declined to provide information to the JBHE research department.

Data for UCLA and Berkeley are for California residents only.

*Yield = percentage of accepted students who enroll.

Source: Survey conducted by JBHE RESEARCH DEPARTMENT.
Opportunities

• Clear demand for biostatisticians

• Challenging road, but very rewarding

• Opportunity to make a difference!!
Opportunities

• Programmatic efforts
  o Recruitment
  o Funding
  o Fostering environment for success

• Students: in your hands!
Success in graduate school: Staying the Course

- Strategy
- Navigate the maze of graduate school
  - Course selection
  - Selecting an advisor
  - Dissertation topics
- Work smarter AND harder
- “He who fails to plan, plans to fail.”
Success in graduate school: 
Staying the Course

• Today’s preparation determines tomorrow’s success.
  o Qualifying exams
  o Dissertation research
  o Coursework
  o Meetings, ...

• “Luck is what happens when preparation meets opportunity.”
Staying the Course

• Your ability: **Confidence**

• Your work
  - Personal standard of excellence
Success in graduate school: Staying the Course

- Disciplined
  - Time management
- Help you stay afloat during overwhelming periods
Success in graduate school: Staying the Course

- Take ownership of your graduate education
- Take initiative
- Utilize the resources around you (students, faculty, labs, courses, jobs, potential mentors)
Success in graduate school: Staying the Course

• Determine what things motivate you. Passion will follow from these.
  o Learning
  o Discovery
  o Challenge
  o Purpose

• “Passion is the genesis of genius” [Anthony Robbins]
Success in graduate school: Staying the Course

• *WHEN* you fall down, get up!!

• Your ultimate career trajectory depends heavily on how you respond to challenging times.

• “There are no secrets to success. It is the result of preparation, hard work, and learning from failure.” [Colin Powell]
Biostatistics

• Provides a pathway to a bright future

• Increasing demand for biostatisticians, particularly in the big data era
  o Strong job market for master’s and doctoral level statisticians
Biostatistics

The field needs you!
Thank you!

DuBois Bowman, Ph.D.
Columbia University
Department of Biostatistics
Professor and Chairman