Introduction to Biostatistics PhD programs

Takumi Saegusa

University of Washington
Department of Biostatistics

April 17 2013
Outline

- What is Biostatistics?
- Applying to Graduate Schools
- Survival Guide
WHAT IS BIOSTATISTICS?
Statistics = Theory + Methods + Applications

- Balanced Development: Math + Science + YOU
- Methodological research
Sequential Clinical trial

- FDA requires clinical trials for a new drug approval
- Sample size calculation
- Mean vs Median (Doesn’t matter? Really?)
Sequential Clinical trial

- Should we stop early if a new drug is “clearly” beneficial or harmful?
Network estimation

- Central Dogma: Gene $\rightarrow$ RNA $\rightarrow$ Protein
- Abundance of thousands of mRNAs is measured by microarray at once
Network estimation

- Estimate covariance matrix in high dimension
- Regularization (penalized MLE)
Air Pollution Study
Air Pollution Study

- Data are scattered over a region
- Data are scattered over time points (different measurements)
- Data are measured with errors
HIV clinical trials

- censoring: no information after time $T = \text{information that she live more than } T$
- death: censoring or outcome?
HIV clinical trials

- Breastfeeding was confirmed as a risk factor of the mother-to-child transmission (MTCT) in a randomized clinical trial around 2000.
- WHO does not recommend mixed feeding (breastfeeding + formula feeding).
- $T$: time to infection.
- $0 = t_0 < t_1 < t_2 < \ldots < t_p$: observation times.
- Infection-free probability at time $t_i$ (no censoring):

$$S(t_i) = P(\text{infection free more than time } t_i)$$

$$= P(T > t_i)$$

$$= P(T \geq t_{i+1})$$

$$= P(T \geq t_{i+1}| T \geq t_i)P(T \geq t_i| T \geq t_{i-1}) \ldots$$

$$\times P(T \geq t_1| T \geq t_0)P(T = t_0)$$

$$\approx \frac{\# \text{ uninfected at } t_{i+1}}{\# \text{ uninfected at } t_i} \times \ldots \frac{\# \text{ uninfected at } t_1}{\# \text{ uninfected at } t_0}$$
hypothetical example: breastfeeding

<table>
<thead>
<tr>
<th>time</th>
<th>0</th>
<th>(t_1)</th>
<th>(t_2)</th>
<th>(t_3)</th>
<th>(\ldots)</th>
</tr>
</thead>
<tbody>
<tr>
<td># uninfected</td>
<td>100</td>
<td>90</td>
<td>70</td>
<td>50</td>
<td>(\ldots)</td>
</tr>
<tr>
<td># infected</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>(\ldots)</td>
</tr>
<tr>
<td># missing before (t_i)</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(P(T \geq t_i</td>
<td>T \geq t_{i-1}))</td>
<td>NA</td>
<td>90/100</td>
<td>70/(100-10)</td>
<td>50/(90-10-10)</td>
</tr>
<tr>
<td>(S(t))</td>
<td>1</td>
<td>(1 \times .9)</td>
<td>(.9 \times 7/9 = .7)</td>
<td>(.7 \times 5/7 = .5)</td>
<td>(\ldots)</td>
</tr>
</tbody>
</table>
**hypothetical example: formula feeding**

<table>
<thead>
<tr>
<th>time</th>
<th>0</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$\cdots$</th>
</tr>
</thead>
<tbody>
<tr>
<td># uninfected</td>
<td>100</td>
<td>80</td>
<td>60</td>
<td>40</td>
<td>$\cdots$</td>
</tr>
<tr>
<td># infected</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\cdots$</td>
</tr>
<tr>
<td># missing before $t_i$</td>
<td>0</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$P(T \geq t_i \mid T \geq t_{i-1})$</td>
<td>NA</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$S(t)$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>
hypothesetical example: mixed feeding

<table>
<thead>
<tr>
<th>time</th>
<th>0</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td># uninfected</td>
<td>100</td>
<td>80</td>
<td>60</td>
<td>40</td>
<td>...</td>
</tr>
<tr>
<td># infected</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>...</td>
</tr>
<tr>
<td># missing before $t_i$</td>
<td>0</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>...</td>
</tr>
<tr>
<td>$P(T \geq t_i</td>
<td>T \geq t_{i-1})$</td>
<td>NA</td>
<td>80/100</td>
<td>60/(100-20)</td>
<td>40/(80-20)</td>
</tr>
<tr>
<td>$S(t)$</td>
<td>1.0</td>
<td>.8</td>
<td>.6</td>
<td>.4</td>
<td>...</td>
</tr>
</tbody>
</table>
Career as a Statistician

- Website of the American Statistical Association (http://www.amstat.org/careers)
- Government survey
- Industry (drug development, quality control, market research)
- Academia (statistical research, scientific research)
- etc.
## Table 1—2012–2013 Salaries of Academic Statisticians Based on Type of Institution

<table>
<thead>
<tr>
<th>Institution Type</th>
<th>Title</th>
<th>Years in Rank</th>
<th>Count</th>
<th>1st Quartile</th>
<th>Median</th>
<th>3rd Quartile</th>
<th>90th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>College</td>
<td>Assistant Professor</td>
<td>0+</td>
<td>10</td>
<td>$58,500</td>
<td>$63,300</td>
<td>$64,800</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Associate Professor</td>
<td>0+</td>
<td>18</td>
<td>$60,600</td>
<td>$68,400</td>
<td>$80,000</td>
<td>NA</td>
</tr>
<tr>
<td>Professor</td>
<td>0+</td>
<td>27</td>
<td></td>
<td>$76,500</td>
<td>$82,900</td>
<td>$106,000</td>
<td>$138,800</td>
</tr>
<tr>
<td>Research University</td>
<td>Assistant Professor</td>
<td>0–1</td>
<td>63</td>
<td>$75,000</td>
<td>$82,000</td>
<td>$85,000</td>
<td>$89,600</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>25</td>
<td></td>
<td>$72,300</td>
<td>$75,500</td>
<td>$84,000</td>
<td>$96,000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>23</td>
<td></td>
<td>$76,800</td>
<td>$81,500</td>
<td>$90,800</td>
<td>$94,300</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>23</td>
<td></td>
<td>$76,100</td>
<td>$80,500</td>
<td>$86,800</td>
<td>$93,400</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>30</td>
<td></td>
<td>$72,700</td>
<td>$79,400</td>
<td>$83,500</td>
<td>$86,000</td>
</tr>
<tr>
<td></td>
<td>6+</td>
<td>19</td>
<td></td>
<td>$67,700</td>
<td>$79,100</td>
<td>$82,500</td>
<td>NA</td>
</tr>
<tr>
<td>Associate Professor</td>
<td>0–1</td>
<td>45</td>
<td></td>
<td>$77,100</td>
<td>$86,300</td>
<td>$92,000</td>
<td>$102,200</td>
</tr>
</tbody>
</table>
### Table 1—Results from the Fall 2012 Salary Survey of Biostatistics and Other Biomedical Statistics Departments and Units

<table>
<thead>
<tr>
<th>Rank/Years in Rank</th>
<th>Percentile</th>
<th>Fall 2006 (Sample Size)</th>
<th>Fall 2007 (Sample Size)</th>
<th>Fall 2008 (Sample Size)</th>
<th>Fall 2009 (Sample Size)</th>
<th>Fall 2010 (Sample Size)</th>
<th>Fall 2011 (Sample Size)</th>
<th>Fall 2012 (Sample Size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant 1-3</td>
<td>25th</td>
<td>$82,400</td>
<td>$86,000</td>
<td>$89,200</td>
<td>$89,100</td>
<td>$89,500</td>
<td>$91,900</td>
<td>$92,700</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>85,000</td>
<td>88,452</td>
<td>93,600</td>
<td>93,500</td>
<td>93,600</td>
<td>97,400</td>
<td>96,900</td>
</tr>
<tr>
<td></td>
<td>75th</td>
<td>90,000</td>
<td>92,869</td>
<td>98,300</td>
<td>99,100</td>
<td>102,600</td>
<td>104,300</td>
<td>106,300</td>
</tr>
<tr>
<td></td>
<td>90th</td>
<td>(112)</td>
<td>(106)</td>
<td>(69)</td>
<td>(82)</td>
<td>(78)</td>
<td>(83)</td>
<td>(73)</td>
</tr>
<tr>
<td>4 or more</td>
<td>25th</td>
<td>$84,476</td>
<td>$87,400</td>
<td>$90,500</td>
<td>$89,400</td>
<td>$91,000</td>
<td>$93,000</td>
<td>$96,800</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>88,471</td>
<td>92,000</td>
<td>95,500</td>
<td>95,500</td>
<td>97,300</td>
<td>97,200</td>
<td>99,500</td>
</tr>
<tr>
<td></td>
<td>75th</td>
<td>94,819</td>
<td>98,220</td>
<td>106,200</td>
<td>100,700</td>
<td>103,000</td>
<td>104,600</td>
<td>106,300</td>
</tr>
<tr>
<td></td>
<td>90th</td>
<td>(48)</td>
<td>(65)</td>
<td>(62)</td>
<td>(87)</td>
<td>(101)</td>
<td>(97)</td>
<td>(80)</td>
</tr>
<tr>
<td>Associate 0-2</td>
<td>25th</td>
<td>$89,937</td>
<td>$102,525</td>
<td>$102,500</td>
<td>$102,400</td>
<td>$103,000</td>
<td>$109,600</td>
<td>$105,800</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>100,441</td>
<td>110,493</td>
<td>110,800</td>
<td>114,000</td>
<td>110,100</td>
<td>118,000</td>
<td>115,500</td>
</tr>
<tr>
<td></td>
<td>75th</td>
<td>113,000</td>
<td>118,900</td>
<td>127,000</td>
<td>129,200</td>
<td>122,400</td>
<td>130,600</td>
<td>125,000</td>
</tr>
<tr>
<td></td>
<td>90th</td>
<td>(46)</td>
<td>(50)</td>
<td>(36)</td>
<td>(36)</td>
<td>(46)</td>
<td>(56)</td>
<td>(54)</td>
</tr>
</tbody>
</table>
Salary (Industry)

<table>
<thead>
<tr>
<th>Employer</th>
<th>n</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Respondents</td>
<td>1,615</td>
<td>87.0</td>
<td>110.0</td>
<td>138.0</td>
<td>176.0</td>
<td>240.0</td>
</tr>
<tr>
<td>Federal Government</td>
<td>320</td>
<td>89.0</td>
<td>105.5</td>
<td>127.0</td>
<td>149.0</td>
<td>162.5</td>
</tr>
<tr>
<td>State or Local Government</td>
<td>37</td>
<td>55.0</td>
<td>72.0</td>
<td>87.0</td>
<td>107.0</td>
<td>185.0</td>
</tr>
<tr>
<td>For-Profit Business or Industry</td>
<td>964</td>
<td>97.0</td>
<td>120.0</td>
<td>150.0</td>
<td>200.0</td>
<td>265.0</td>
</tr>
<tr>
<td>Nonprofit Organization</td>
<td>172</td>
<td>70.0</td>
<td>90.0</td>
<td>120.0</td>
<td>156.0</td>
<td>190.0</td>
</tr>
<tr>
<td>Self Employed or Private Consultant</td>
<td>92</td>
<td>80.0</td>
<td>102.0</td>
<td>150.0</td>
<td>201.5</td>
<td>350.0</td>
</tr>
<tr>
<td>Other</td>
<td>24</td>
<td>80.0</td>
<td>97.5</td>
<td>122.5</td>
<td>160.0</td>
<td>171.0</td>
</tr>
<tr>
<td>South Atlantic (DE, DC, GA, FL, MD, NC, SC, VA, WV)</td>
<td>538</td>
<td>91.0</td>
<td>111.0</td>
<td>136.0</td>
<td>160.0</td>
<td>200.0</td>
</tr>
<tr>
<td>Middle Atlantic (NJ, NY, PA)</td>
<td>285</td>
<td>100.0</td>
<td>125.0</td>
<td>163.0</td>
<td>220.0</td>
<td>300.0</td>
</tr>
<tr>
<td>East North Central (IL, IN, MI, OH, WI)</td>
<td>174</td>
<td>80.0</td>
<td>102.0</td>
<td>132.5</td>
<td>180.0</td>
<td>234.0</td>
</tr>
<tr>
<td>Pacific (AK, CA, HI, OR, WA)</td>
<td>232</td>
<td>87.0</td>
<td>105.0</td>
<td>140.0</td>
<td>185.0</td>
<td>260.0</td>
</tr>
<tr>
<td>New England (CT, MA, ME, NH, RI, VT)</td>
<td>119</td>
<td>87.0</td>
<td>117.0</td>
<td>156.0</td>
<td>210.0</td>
<td>300.0</td>
</tr>
<tr>
<td>West North Central (IA, KS, MN, MO, ND, NE, SD)</td>
<td>95</td>
<td>80.0</td>
<td>96.0</td>
<td>120.0</td>
<td>145.0</td>
<td>176.0</td>
</tr>
<tr>
<td>Other (states not listed above)</td>
<td>165</td>
<td>75.0</td>
<td>93.0</td>
<td>121.0</td>
<td>155.0</td>
<td>200.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geographic Region</th>
<th>n</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>South Atlantic (DE, DC, GA, FL, MD, NC, SC, VA, WV)</td>
<td>538</td>
<td>91.0</td>
<td>111.0</td>
<td>136.0</td>
<td>160.0</td>
<td>200.0</td>
</tr>
<tr>
<td>Middle Atlantic (NJ, NY, PA)</td>
<td>285</td>
<td>100.0</td>
<td>125.0</td>
<td>163.0</td>
<td>220.0</td>
<td>300.0</td>
</tr>
<tr>
<td>East North Central (IL, IN, MI, OH, WI)</td>
<td>174</td>
<td>80.0</td>
<td>102.0</td>
<td>132.5</td>
<td>180.0</td>
<td>234.0</td>
</tr>
<tr>
<td>Pacific (AK, CA, HI, OR, WA)</td>
<td>232</td>
<td>87.0</td>
<td>105.0</td>
<td>140.0</td>
<td>185.0</td>
<td>260.0</td>
</tr>
<tr>
<td>New England (CT, MA, ME, NH, RI, VT)</td>
<td>119</td>
<td>87.0</td>
<td>117.0</td>
<td>156.0</td>
<td>210.0</td>
<td>300.0</td>
</tr>
<tr>
<td>West North Central (IA, KS, MN, MO, ND, NE, SD)</td>
<td>95</td>
<td>80.0</td>
<td>96.0</td>
<td>120.0</td>
<td>145.0</td>
<td>176.0</td>
</tr>
<tr>
<td>Other (states not listed above)</td>
<td>165</td>
<td>75.0</td>
<td>93.0</td>
<td>121.0</td>
<td>155.0</td>
<td>200.0</td>
</tr>
</tbody>
</table>
APPLYING TO GRAD SCHOOLS
Before Applying...

Ask yourself
- Do you have passion?
- Do you have patience?
Which Schools to Apply?

Get rankings (usnews, www.phds.org, etc.) (≠ difficulty of getting in)
- Is funding available to all students (including master’s students)?
- Proportions of PhD students successfully completing the program
- Average years of completion.
- Careers after graduate schools
- Possible to change from a master’s program to a PhD program?
- Male vs. Female, American vs Intl’ Students, Class size, etc.
- culture, weather, etc.

Visit campus if you can (ask graduate programs)
What does an admission committee want to know about you?
What does an admission committee want to know about you?

- You know what you are going to do. (\( \Rightarrow \text{Strong Interests} \))
What does an admission committee want to know about you?

- You know what you are going to do. (Strong Interests)
  - Classes (stat)
  - Research Experience in Statistics and Related Fields
  - Consulting Experience
  - Statement of Purpose
What does an admission committee want to know about you?

- You know what you are going to do. (\(=\) **Strong Interests**)
  - Classes (stat)
  - Research Experience in Statistics and Related Fields
  - Consulting Experience
  - Statement of Purpose
- You can do it. (\(=\) **High Potentials**)

Takumi Saegusa (UW Biostat)
Biostat PhD
April 17 2013
What does an admission committee want to know about you?

- **You know what you are going to do.** (\(\equiv\) **Strong Interests**)
  - Classes (stat)
  - Research Experience in Statistics and Related Fields
  - Consulting Experience
  - Statement of Purpose

- **You can do it.** (\(\equiv\) **High Potentials**)
  - GPA
  - Letters of Recommendation
  - Publications
  - Research Experience in Other Fields
  - Classes in Math and Science
  - GRE, etc.
What does an admission committee want to know about you?

- **You know what you are going to do.** (= **Strong Interests**)
  - Classes (stat)
  - Research Experience in Statistics and Related Fields
  - Consulting Experience
  - Statement of Purpose

- **You can do it.** (= **High Potentials**)
  - GPA
  - Letters of Recommendation
  - Publications
  - Research Experience in Other Fields
  - Classes in Math and Science
  - GRE, etc.
Classes

Of course, all applicants take probability and statistics classes. Analysis (Advanced Calculus) and Linear Algebra are required for admission. Taking graduate level Real Analysis and Probability are plus (you need to learn it by yourself anyway). Take computer science class to show your programming skills (in C, C++, etc., not in commercial statistical softwares). Take science classes where you want to apply statistics and show how serious you are about applications.
GET GOOD GRADES.
GET GOOD GRADES.

- Of course, all applicants take probability and statistics classes.
- Analysis (Advanced Calculus) and Linear Algebra are required for admission.
- Taking graduate level Real Analysis and Probability are plus (you need to learn it by yourself anyway).
- Take computer science class to show your programming skills (in C, C++, etc., not in commercial statistical softwares).
- Take science classes where you want to apply statistics and show how serious you are about applications.
Research Experience

Research Experience provides

- the evidence of your strong interests
- letters from supervisors verify your competency

Various ways to obtain research experience:

- Participate in the REU (strongly recommended).
  - Google “Summer Institute Biostatistics”, “Research Experience for Undergraduate Statistics” etc.
  - Find one for students where your expense will be covered.

- Seek an opportunity on campus where statistical analysis would be of help (computational biology lab, clinical and epidemiological research, research in sociology, economics, etc.)

- Set up consulting services on campus.

- Independent study with faculty.
Can you explain how your interest in (bio)statistics has been shaped in a UNIQUE (=not superficial) and DETAILED way?

Can you show how serious you are about (bio)statistics or the field to which you want to apply statistical methods with EVIDENCE?

Can you briefly write about your potential to do research with EVIDENCE?

Can you list some research directions you want to pursue in a COHERENT way to the above?

Can you convince an admission committee that you have STRONG INTERESTS and HIGH POTENTIAL?
Biostatistics vs. Statistics

Biostatistics
- Admit more students
- More funding available through research assistantship
- Educational emphases on methods and applications (scientific relevance more emphasized)
- More job opportunities in biology-related fields
- Background in Biology, Medicine, Math, CS, etc.

Statistics
- Selective in admission
- Limited funding through teaching assistantship (teaching experience)
- Educational emphases on theory (not always) and methods (mathematical rigor more emphasized)
- Job opportunities in finance, engineering etc.
- Background in Math and CS
References

- Ask other alumni (Jimmy Doi, strongly recommended, several ppt files found at http://statweb.calpoly.edu/jdoi/web/research/index.htm).
- Go online to find CV’s of the first or second year grad students and find out their strengths you can learn to get.
- Find books on how to apply to graduate schools
  Note: Most tips might be appropriate for other majors (biology, CS, etc.)
Find a role model among students.
Study

- Slow understanding does not hurt, but lack of patience to get complete understanding does.
- Math vs. Engineering
  - Math (less emphasized in general): Rigorously prove theorems with minimum conditions.
  - Engineering: Explain what conditions are not likely to hold in practice and what happens to your statistical procedures if conditions fails.
  - Engineering: Explain your statistical procedure to non-statisticians in a concise way and convince them of its usefulness.
- Forming a study group is often very helpful (take an initiative to form one).
- Have fun!
Advisor

Choice of a thesis advisor is the most important than grades, your school, etc.

- How successful their students are (Mathematical Genealogy Project or CV’s)
- Publication records
- Rumor from senior students
- Get to know your potential advisors (through classes, departmental events, etc.)
- Do readings with several potential advisors.
- Personality match matters more because your interests change to be more sophisticated...
Research

- Very different from studying
- Discontinuous progress = long time of getting nothing
- Hot topics: severe competitions vs. job opportunities
- Many projects vs. one deep research question
- Different focus: mathematical skills, computational skills, deep understanding of applications, etc.
- Take advantage of RA to do something different from your thesis
GOOD LUCK!