Introduction of opportunity and challenge in Biostatistics and Bioinformatics to Math major students

George C. Tseng
Department of Biostatistics
Department of Human Genetics
University of Pittsburgh
Outline

- Possible applications of probability and statistics
- Biostatistics
  - Academic research
  - Industry
- Bioinformatics
- Transitions
  - Application => Ph.D. student => Research/Job
- Some final words
  - Curriculum and preparation
  - Studying abroad??
<table>
<thead>
<tr>
<th>Year(s)</th>
<th>Degree</th>
<th>University</th>
</tr>
</thead>
<tbody>
<tr>
<td>93-97</td>
<td>B.S. Math.</td>
<td>National Taiwan Univ.</td>
</tr>
<tr>
<td>97-99</td>
<td>M.S. Stat.</td>
<td>National Taiwan Univ.</td>
</tr>
<tr>
<td>99-00</td>
<td>Stat.</td>
<td>UCLA</td>
</tr>
<tr>
<td>00-03</td>
<td>Ph.D. Biostat.</td>
<td>Harvard University</td>
</tr>
<tr>
<td>03~</td>
<td>Biostat.</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>Brain</td>
<td>Income</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Mathematician</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Applied Mathematician</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistician</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biostatistician</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Epidemiologist</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Physician</td>
<td>High</td>
</tr>
</tbody>
</table>
I. Applications of statistics

- Agricultural science
- Social science: education, psychology,…
- Financial mathematics
- Actuarial science
- Biomedical science
  - Biostatistics, medical imaging, Biomath, Biophysics...
  - Bioinformatics, Computational Biology

……
II. Biostatistics

- Statistical research usually motivated by applications of public health, medicine or genetics.
- Research results should at least have one area of application.
II. Biostatistics

Research Areas: (from the dept. website)

- Dept. of Biostatistics at Harvard
  - AIDS research
  - Cancer research
  - Computational biology & Bioinformatics
  - Environmental statistics
  - Genetic epidemiology
  - Neurostatistics
  - Psychiatric biostatistics
II. Biostatistics

Research Areas: (from the dept. website)

- Dept. of Biostatistics at Univ. Pittsburgh
  - Cancer treatment trials
  - Health outcomes/health services research
  - Environmental & occupational epidemiology
  - Radiological imaging system
  - Psychiatric research
  - Computational biology & Bioinformatics
  - Statistical methodology
II. Biostatistics

A simple example of survival analysis:

A new drug and an old drug are applied to cancer patients. Survival time of each patients are recorded after treatment. The study was terminated at 60 months.

<table>
<thead>
<tr>
<th>ID</th>
<th>group</th>
<th>relapse</th>
<th>survival</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>60</td>
</tr>
</tbody>
</table>

New drug (1): 196 patients
Old drug (0): 35 patients
Relapse (1): died
Relapse (0): survived over

Q: How do we rigorously and confidently decide the new drug is better than the old drug?
II. Biostatistics

A simple example of survival analysis:

Kaplan-Meier curve

- n=196
- p<0.0001
- n=35

survival %

survival (month)
II. Biostatistics

A simple example of survival analysis:

- Compare the difference of two survival curves.
- Modelling censoring and survival model.
  - Early drop out patients
  - Patients participate in interim of study
- Experimental design
  - Case-control matched study
  - Early termination
II. Biostatistics

Employment of alumni (Dept. of Biostatistics, Univ. of Pittsburgh)

<table>
<thead>
<tr>
<th>Type of Employment</th>
<th>M.S./M.P.H.</th>
<th>Ph.D/Sc.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic Institutions</td>
<td>52</td>
<td>48</td>
</tr>
<tr>
<td>Government Agencies</td>
<td>28</td>
<td>11</td>
</tr>
<tr>
<td>Other Health Research Groups*</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>Private Industry</td>
<td>31</td>
<td>23</td>
</tr>
<tr>
<td>Other (includes students continuing for doctoral degree)</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Unknown</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>Deceased</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>181</td>
<td>96</td>
</tr>
</tbody>
</table>
II. Biostatistics: working in university

- Tenure track
  - Research (publication and academic activity)
    - Methodology research
    - Collaborative research
  - Teaching
  - Grant proposals
  - Service (committees, advising students…)

- Research track
  - Research
    - Collaborative
    - Methodology
  - Grant proposals
II. Biostatistics: working in government

Centers for Disease Control
National Institutes of Health
U.S. Census Bureau
National Center for Health Statistics
Food and Drug Administration
II. Biostatistics: working in pharmaceutical company

**Merck:** one of the largest drug companies in the US

- Global, research-driven pharmaceutical company
  - ~62,000 employees worldwide in 26 countries
  - In 2004, $22.9 billion in sales, $5.8 billion in net income, $3 billion invested in research
- Broad range of products
- Ranked in “100 Best to Work For” and “America’s Most Admired” and “Global Most Admired”

*Info. from Merck & Co., Inc.*
II. Biostatistics: working in pharmaceutical company

Areas of Application

- Clinical Trials
- Epidemiology
- Data Management
- Manufacturing/Quality Control
- Research Administration
- Market Research
- Genomics
- Management
- Discovery
- Pharmacology/Toxicology
- Regulatory Affairs

Statistician

Info. from Merck & Co., Inc.
II. Biostatistics: working in pharmaceutical company

New Drug Development

**Creation**
- Drug discovery
- Chemical synthesis
- Laboratory testing
- Animal testing
- Formulation of ingredients

(2 - 4 Years)

**Role of Statistician**
- Analyze high throughput screening results
- Design screening strategies and select analogs
- Analyze dose-response studies
- Employ bioassay techniques
- Evaluate carcinogenic potential
- Evaluate reproductive and genetic toxicology

Info. from Merck & Co., Inc.
## II. Biostatistics: working in pharmaceutical company

### New Drug Development

<table>
<thead>
<tr>
<th>Human Testing</th>
<th>Role of Statistician</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase I - Safety</td>
<td>Propose statistical methodology</td>
</tr>
<tr>
<td>Phase II a - Proof of Concept</td>
<td>Approve study protocols</td>
</tr>
<tr>
<td>Phase II b - Dose-Ranging</td>
<td>Interact with Project Team</td>
</tr>
<tr>
<td>Phase III - Safety and Efficacy</td>
<td>Analyze and interpret early studies</td>
</tr>
</tbody>
</table>

(3 - 7 Years)

Info. from Merck & Co., Inc.
II. Biostatistics:
working in pharmaceutical company

New Drug Development

- FDA submission and review
- New drug available to patients and physicians

(1 - 3 years to prepare, 1 year to review)

New Drug Application
Role of Statistician

- Summarize across studies
- Prepare statistical technical section
- Present methodology and results to FDA

Info. from Merck & Co., Inc.
II. Biostatistics: working in pharmaceutical company

New Drug Development

- Ongoing
- Additional uses
- Additional side effects
- Modification of dosage or form

Further Evaluation

Role of Statistician

- Design and analyze post-marketing studies
- Submit papers for publication

Info. from Merck & Co., Inc.
II. Biostatistics: working in pharmaceutical company

Discovery and Development of a Successful Drug

Source: Based on PhRMA analysis, updated for data per Tufts Center for the Study of Drug Development (CSDD) database.

Info. from Merck & Co., Inc.
Combinatorial Gene Regulation

- A microarray experiment showed that when gene X is knocked out, 20 other genes are not expressed

- How can one gene have such drastic effects?

From http://www.bioalgorithms.info/
III. Bioinformatics
A simple example of motif finding

Regulatory Proteins

- Gene X encodes regulatory protein, a.k.a. a transcription factor (TF)

- The 20 unexpressed genes rely on gene X’s TF to induce transcription

- A single TF may regulate multiple genes

From http://www.bioalgorithms.info/
III. Bioinformatics
A simple example of motif finding

Transcription Factors and Motifs

From http://www.bioalgorithms.info/
III. Bioinformatics
A simple example of motif finding

Motifs and Transcriptional Start Sites

From http://www.bioalgorithms.info/
III. Bioinformatics

A simple example of motif finding

Motif Logos: An Example

From http://www.bioalgorithms.info/
Random Sample

III. Bioinformatics
A simple example of motif finding

Random Sample

at gaccgggatatctgataccgtatttggcctaggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccgaccctattttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatactgggcataaggtactagtgccttttgggaacactatagtgctctcccgatttttgaatatgtaggatcattcgccagggtccgagctgagaatttggtagctgttctgtctgcccaatctacggactttaatggcccacttagtccacttataggacgtccttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatggcccacttagtccacttataggacgtccttttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatggcccacttagtccacttataggtcaatcatgttcttgtgaatggatttttaactgagggcatagaccgcttggcgcacccaaattcagtgtgggcgagcgcaacggttttggcccttgttagaggcccccgtactgatggaaactttcaattatgagagagctaatctatcgcgtgcgtgttcatacttgagttggtttcgaaaatgctctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgtattggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatttcaacgtatgccgaaccgaaagggaagctgggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttctgggtactgatagcatgct cctt t t ggcggt aat gt gccgggaggct ggt t acgt agggagccct aacggact t aat ggcccact t agt ccact t at aggt caat cat g t c t t g t g a a g g a g g g c t g t t a a c t g a g g g c a t a g a c c g c t t g g c g a c c c a a a t t c a g t g t g g g g a g c g c a a

cggttttggcccttgttagaggcccccgtactgatggaaactttcaattatgagagagctaatctatcgcgtgcgtgttcatacttgagttggtttcgaaaatgctctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgtattggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatttcaacgtatgccgaaccgaaagggaagctgggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttctgggtactgatagcatgct cctt t t ggcggt aat gt gccgggaggct ggt t acgt agggagccct aacggact t aat ggcccact t agt ccact t at aggt caat cat g t c t t g t g a a g g a g g g c t g t t a a c t g a g g g c a t a g a c c g c t t g g c g a c c c a a a t t c a g t g t g g g g a g c g c a a
Implanting Motif

III. Bioinformatics

A simple example of motif finding
III. Bioinformatics
A simple example of motif finding

Where is the Implanted Motif?

```plaintext
Where is the Implanted Motif?

at gaccgggat act gat AAAAAAAAGGGGGGggcgt acacat tagat aaacgt at gaagt acgt t agact cggcgcgcgcg
acccct attttt gagcagat ttagt gacct ggaaaaaatttt tagt aacaaaactttt ccgaat aAAAAAAAGGGGGGGGa
tgagt at ccct gggat gact t AAAAAAAAGGGGGGGt gct ct cccgat tttt gaat at gt aggat cat t gcgccaggtt ccga
gct gagaat t ggt gAAAAAAAGGGGGGGt cccagcaat cgccgaaccaacgcggaccccaaggaag accgat aaaggaga
tccctttt gcggt aat gt gccgggaggct ggt t acgt agggaagcct aacggactt aat AAAAAAAAGGGGGGGctttatag
gt caat cat gt tct t gt gaat ggtt t AAAAAAAAGGGGGGGgaccgct t ggcgcacccaatt cagt gt gggcgagccaa
cggtttt gcgcctttgtagggcaccctt AAAAAAAAGGGGGGGtcatatgagagagct aat ct at cgccgt gcgt gt t cat
aact t gagt t AAAAAAAAGGGGGGGct ggggacacat aacaagggagtct tct t at cagt t aat gct gt at gacact at gt a
tttgcccat t ggct aaaaagccccactt gacaaat ggaagat agaat cttgc aat AAAAAAAAGGGGGGGgaccgaaaggggaag
cgt ggt gagcaacgcagattt ctt acgt gcattagct cgcct t cccgggagct aat agcacgaagctt AAAAAAAAGGGGGGGGa
```
Implanting Motif AAAAAAGGGGGGGG with Four Mutations

III. Bioinformatics
A simple example of motif finding
Where is the Motif???

III. Bioinformatics
A simple example of motif finding

at gaccgaggat act gat agaagaaaggt t ggggcgt acacat t agat aaacgt at gaagt acgt t agact cggcgcg cgcg
cacct at tt t t t gagcagat t t g t gacct ggaaaaaaat tt gagt acaaaaaact tt t ccaat acaat aaaacggcggga
t gagt at ccc t gggat gact t aaaa at aat ggaat ggt gct ct cccgat t t t t gaat at gt aggat cat t cgccaggtt ccca
gct gagaat t ggt gcaaaaaaagggat tt gt ccacgcaat cgcgaaccaacgcggac ccgg ccaagggcggat aaaggaga
t ccc t tt t ggcgt aat gt gcccggaggtt ggt t acgt agggagccct aagggact t aat at aat aaaggaagggct t at ag
gt caat cat gt t ct t gt gaat ggt t tt aacaat aagggct gggaccgct t ggcgcacccaaat t cagt gt gggcggcggaa
cgg t tt ggcccct tt gt t agaggccccctt at aaac aagggagggccaat t at gagagagct aat ct at cgccgt gcgt t t cat
aact t gagt t aaaaat agggagccct ggggacacat acaagagaggt ct ct cct t at cagt t aat gct gt at gacact at gt a
t t ggccccat t ggct aaaaagcccaact t gacaaat ggaagat agaat cct t gcat act aaaaagagcggaccggaaagggag
ct ggt gagcagacagagat t ct t acgt gcatt agct cgct t cgggggat ct aat agcagagct t act aaaaagagcggga
Why Finding (15,4) Motif is Difficult?

III. Bioinformatics
A simple example of motif finding
Questions:

- How to develop a good probabilistic model for the motifs?
- Is the computation affordable to search the whole genome? (Human genome is around 3 billion base pair long.)
- How to evaluate the statistical significance of the motifs you find?
IV. Transitions

Preparation; Military service

Undergraduate

Ph.D. study

Post-doctoral position

Assistant Professor

Associate Professor

Full Professor

Tenure evaluation

School application

Job application

4-5 yrs

2-3 yrs

6 yrs
IV. Transitions: Application

- GRE, TOEFL, GPA
- Recommendation letter
- Study plan

- Prepare and ask around early: take GRE and TOEFL; identify professors for recommendation letters and advises
- Academia Sinica (a good place to stay for short term transition and preparation)
IV. Transitions: Ph.D. study

- Settle down and enjoy
- Improve English; think open and American
- Professor, classmates, office-mates, colleagues are good assets for your future

Financial situation:
- Stipend (US$1600-$300 tax) from TA or RA
- Rent US$400~500. Living cost $300~500.
IV. Transitions: Research/Job

- Going to academic is usually more busy than going to industry but with more freedom.
- No boss v.s. with a boss
- Irregular/flexible working hour v.s. regular working hour

?? $$$ ??
## IV. Transitions: Research/Job

### University (9 months)

<table>
<thead>
<tr>
<th>Institution Type</th>
<th>Title</th>
<th>Years in Rank</th>
<th>Count</th>
<th>Median</th>
<th>3rd Quartile</th>
<th>90th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research University</td>
<td>Assistant Professor</td>
<td>0 to 1</td>
<td>73</td>
<td>$ 66,000</td>
<td>$ 69,345</td>
<td>$ 77,500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>35</td>
<td>$ 65,000</td>
<td>$ 67,000</td>
<td>$ 69,918</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>35</td>
<td>$ 64,000</td>
<td>$ 67,700</td>
<td>$ 72,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 to 5</td>
<td>47</td>
<td>$ 66,000</td>
<td>$ 69,900</td>
<td>$ 75,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 or more</td>
<td>12</td>
<td>$ 63,900</td>
<td>$ 67,750</td>
<td>$ 69,910</td>
</tr>
<tr>
<td>Associate Professor</td>
<td>0 to 1</td>
<td>22</td>
<td>$ 71,900</td>
<td></td>
<td>$ 82,275</td>
<td>$ 95,134</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 to 3</td>
<td>43</td>
<td>$ 80,000</td>
<td>$ 85,100</td>
<td>$ 89,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 to 5</td>
<td>31</td>
<td>$ 70,000</td>
<td>$ 87,700</td>
<td>$ 91,900</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 to 8</td>
<td>30</td>
<td>$ 68,538</td>
<td>$ 78,021</td>
<td>$ 82,441</td>
</tr>
</tbody>
</table>

From Amstat News
### IV. Transitions: Research/Job Industry

<table>
<thead>
<tr>
<th>Years of Experience</th>
<th>Highest Degree</th>
<th>N</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>55.0</td>
<td>60.0</td>
<td>70.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>72.0</td>
<td>83.0</td>
<td>85.0</td>
</tr>
<tr>
<td>0–1.9</td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>64.0</td>
<td>70.0</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28</td>
<td>85.0</td>
<td>88.0</td>
<td>98.0</td>
</tr>
<tr>
<td>2–3.9</td>
<td></td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>67.0</td>
<td>75.0</td>
<td>85.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>62</td>
<td>85.0</td>
<td>94.0</td>
<td>104.0</td>
</tr>
<tr>
<td>4–7.9</td>
<td></td>
<td>71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>70.0</td>
<td>83.0</td>
<td>94.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45</td>
<td>88.0</td>
<td>98.5</td>
<td>121.0</td>
</tr>
<tr>
<td>8–11.9</td>
<td></td>
<td>36</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*From Amstat News*
## IV. Transitions: Research/Job

### Government

<table>
<thead>
<tr>
<th>Years of Experience</th>
<th>Highest Degree</th>
<th>N</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With No Managerial Responsibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–1.9</td>
<td>MS</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PhD</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2–3.9</td>
<td>BS</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MS</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PhD</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4–7.9</td>
<td>MS</td>
<td>8</td>
<td>78.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PhD</td>
<td>18</td>
<td>85.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8–11.9</td>
<td>MS</td>
<td>5</td>
<td>85.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PhD</td>
<td>11</td>
<td>98.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Amstat News
V. Some final words: course preparation

**Life Sciences**
- Cell Biology/Molecular Biology
- Biochemistry
- Genetics

**Computer Science**
- Intermediate/Advanced Programming (JAVA, C++)
- Fundamental Data Structures and Algorithms
- Algorithms

**Physical Sciences**
- Statistical Thermodynamics or Physical Chemistry

**Mathematics and Statistics**
- Vector Calculus
- Linear Algebra
- Probability & Statistics

**Computational Biology**
- Computational Biology; Bioinformatics
V. Some final words: Taiwan or abroad

- Try to go abroad if possible

- There are very good graduate programs in Taiwan. If you choose to stay, try to apply for a one-year exchange program abroad.
V. Some final words: Preparation

- Course preparation
- Improve English (take GRE and TOEFL early)
- Talk to some researchers in NTU and Sinica
- Get good recommendation letters and write a good essay
- Go to talks (NTU Math, NTU biostatistics, Sinica)
- Apply as many (good) schools as you can.
- Money should not be an issue if you get stipend support.
V. Some final words: after you get there

- Continue to improve English
- Find a good advisor (reputation in research, personality)
- Be collegial and collaborative; change our viewpoint and re-interpret what you see without bias.
Thanks for your attention!

Merry Christmas!!