Lecture Slides

- Lecture slides will be posted on the course website the day before class.
- The slides posted will NOT be complete (I want you to think during class, so won’t give you the answers to questions posed).
- You are encouraged to take notes on the slides.

Textbook

- **Statistics: Unlocking the Power of Data**
  by Lock, Lock, Lock Morgan, Lock, and Lock

  - **Purchasing options:**
    - Bookstore (new, used)
    - wiley.com (e-book)
    - Amazon.com (new, used, kindle, rent)
    - Wiley Plus (wiley.com): interactive online text (linked videos, practice problems, odd solutions)

Clickers

- Two options:
  - 1) Purchase an i-clicker remote (any version OK)
  - 2) Get the i-clickerGO app for your smartphone


  - The point of clicker questions is to motivate you to think actively about new material as it is being presented

  - Credit simply for clicking in

Class Year

- **What is your class year?**
  - (a) First-year
  - (b) Sophomore
  - (c) Junior
  - (d) Senior
Major
Your primary major (or potential future major) best falls under the category...
(a) Natural Sciences
(b) Arts and Humanities
(c) Social Sciences
(d) Math/Statistics/CS
(e) Other

Support
• My Office Hours: (in Old Chemistry 216)
  ○ Mon 3 – 4 pm, Wed 3 – 4 pm, Fri 1 – 3 pm
• TA Office Hours: tbd
• Statistics Education Center: (in Old Chem 211A)
  ○ 4 – 9 pm Sunday – Thursday in Old Chem 211A
• Email: Email your TA or kari@stat.duke.edu

Grade Breakdown
Labs 26 points (5%)
Homework 50 points (10%)
Clickers 24 points (5%)
Projects 100 points (20%)
Midterms 150 points (30%)
Final Exam 150 points (30%)
TOTAL 500 points
(Up to 10 extra credit points may be earned.)

Labs
• Labs are on Thursdays in Old Chem 101, starting tomorrow
• Statistical software, practice analyzing data
• Labs will be group based
• Labs will use all free software:
  ○ StatKey: lock5stat.com/statkey
  ○ Other free software: tbd

Homework
• Weekly homework due, usually on Mondays
• Point of homework:
  ○ to LEARN!
  ○ to make sure you are keeping up with the material
  ○ to prepare you for projects and exams
• Graded problems and practice problems
• Grading
  ○ Graded on a 6 point scale
  ○ Lowest homework grade dropped
  ○ Penalties for late homework

Projects
• Project 1
  ○ Individual
  ○ EDA, confidence intervals, hypothesis tests
  ○ written report up to 5 pages in length
• Project 2
  ○ with your lab group
  ○ Regression
  ○ written report up to 10 pages in length
Exams

- Midterm Exams: 2/19 and 4/2 in class
- Final: 4/28 2 – 5pm
- Exams are mandatory and must be taken at the given time. Make-up exams will not be given.
- In extreme circumstances (severe illness), midterms may be excused *only in advance*. In this case the grade will be imputed by regression.

Keys to Success

- Come to class ready to think and be engaged
- Come to lab ready to think and be engaged
- Do the homework, trying it by yourself first
- Do lots of practice problems
- Read the textbook
- Stay on top of the material

Why Statistics?

- Statistics is all about DATA
  - Collecting DATA
  - Describing DATA – summarizing, visualizing
  - Analyzing DATA
- Data are *everywhere*! Regardless of your field, interests, lifestyle, etc., you will almost definitely have to make decisions based on data, or evaluate decisions someone else has made based on data

Introduction to Data

SECTION 1.1

- Data
- Cases and variables
- Categorical and quantitative variables
- Using data to answer a question

Data

- Data are a set of measurements taken on a set of individual units
- Usually data is stored and presented in a *dataset*, comprised of variables measured on cases

Cases and Variables

We obtain information about *cases* or *units*.

*Variable* is any characteristic that is recorded for each case.

- Generally each case makes up a row in a dataset, and each variable makes up a column
### Countries of the World

<table>
<thead>
<tr>
<th>Country</th>
<th>Land Area</th>
<th>Population</th>
<th>Rural</th>
<th>Health</th>
<th>Internet</th>
<th>Birth Rate</th>
<th>Life Expectancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>652230</td>
<td>29021099</td>
<td>76</td>
<td>3.7</td>
<td>1.7</td>
<td>46.5</td>
<td>45.9</td>
</tr>
<tr>
<td>Albania</td>
<td>27400</td>
<td>3143291</td>
<td>53.3</td>
<td>8.2</td>
<td>23.9</td>
<td>14.6</td>
<td>76.6</td>
</tr>
<tr>
<td>Algeria</td>
<td>3281740</td>
<td>34973426</td>
<td>34.8</td>
<td>10.6</td>
<td>10.2</td>
<td>20.8</td>
<td>72.4</td>
</tr>
<tr>
<td>American Samoa</td>
<td>200</td>
<td>66107</td>
<td>7.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andorra</td>
<td>470</td>
<td>33830</td>
<td>11.1</td>
<td>21.3</td>
<td>70.5</td>
<td>10.4</td>
<td></td>
</tr>
<tr>
<td>Angola</td>
<td>1246700</td>
<td>18200668</td>
<td>43.3</td>
<td>6.8</td>
<td>3.1</td>
<td>42.9</td>
<td>47</td>
</tr>
<tr>
<td>Antigua and Barbuda</td>
<td>440</td>
<td>86634</td>
<td>69.5</td>
<td>11</td>
<td>75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina</td>
<td>2736690</td>
<td>39982980</td>
<td>8</td>
<td>13.7</td>
<td>28.1</td>
<td>17.3</td>
<td>75.3</td>
</tr>
</tbody>
</table>

### Diet Coke and Calcium

<table>
<thead>
<tr>
<th>Drink</th>
<th>Calcium Excreted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet cola</td>
<td>50</td>
</tr>
<tr>
<td>Diet cola</td>
<td>62</td>
</tr>
<tr>
<td>Diet cola</td>
<td>48</td>
</tr>
<tr>
<td>Diet cola</td>
<td>55</td>
</tr>
<tr>
<td>Diet cola</td>
<td>58</td>
</tr>
<tr>
<td>Diet cola</td>
<td>61</td>
</tr>
<tr>
<td>Diet cola</td>
<td>58</td>
</tr>
<tr>
<td>Water</td>
<td>46</td>
</tr>
<tr>
<td>Water</td>
<td>54</td>
</tr>
<tr>
<td>Water</td>
<td>45</td>
</tr>
<tr>
<td>Water</td>
<td>53</td>
</tr>
<tr>
<td>Water</td>
<td>46</td>
</tr>
<tr>
<td>Water</td>
<td>53</td>
</tr>
<tr>
<td>Water</td>
<td>48</td>
</tr>
</tbody>
</table>

### Intro Statistics Survey Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Gender</th>
<th>SAT</th>
<th>GPA</th>
<th>Sliding Height</th>
<th>Height</th>
<th>Weight</th>
<th>Exercise</th>
<th>TV</th>
<th>Puls</th>
<th>Award</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freshman</td>
<td>F</td>
<td>M</td>
<td>Freshman</td>
<td>F</td>
<td>M</td>
<td>Freshman</td>
<td>F</td>
<td>M</td>
<td>Freshman</td>
</tr>
<tr>
<td>2014</td>
<td>Female</td>
<td>110</td>
<td>25</td>
<td>72</td>
<td>20</td>
<td>10</td>
<td>14</td>
<td>30</td>
<td>150</td>
<td>3</td>
</tr>
<tr>
<td>2015</td>
<td>Male</td>
<td>110</td>
<td>25</td>
<td>72</td>
<td>20</td>
<td>10</td>
<td>14</td>
<td>30</td>
<td>150</td>
<td>3</td>
</tr>
</tbody>
</table>

### Data

- **US News and World Report National University Rankings**
- **Stock Market**
- **Duke Basketball**
- **Unemployment Rate**
- **Hybrid Cars**
- **Public Opinion**
- **Antidepressants and Alzheimer's**

### Data Applicable to You

- Think of a potential dataset (it doesn’t have to actually exist) that you would be interested in analyzing
  - What are the cases?
  - What are the variables?
  - What interesting questions could it help you answer?
**Kidney Cancer**

If the values in the kidney cancer dataset are rates of kidney cancer deaths, then what are the cases?

(a) The people living in the US
(b) The counties of the US

A person either has kidney cancer or doesn’t... a rate must apply to a group of people, such as a county.

---

**Kidney Cancer**

If the values in the kidney cancer dataset are yes/no, then what are the cases?

(a) The people living in the US
(b) The counties of the US

A person either has kidney cancer or doesn’t. Yes/no doesn’t make sense for a county.

---

**Categorical versus Quantitative**

- Variables are classified as either categorical or quantitative:
  - A categorical variable divides the cases into groups
  - A quantitative variable measures a numerical quantity for each case

---

**Categorical**

- Gender

**Quantitative**

- SAT

---

**Categorical**

- Year

**Quantitative**

- GPA, Math, Science, Reading, Writing, Exercise, TV, Pub, Award

---

**Kidney Cancer**

If the cases in the kidney cancer dataset are counties, then the measured variable is...

(a) Categorical
(b) Quantitative

Rates are numbers (quantitative).
Kidney Cancer

If the cases in the kidney cancer dataset are people, then the measured variable is...

(a) Categorical
(b) Quantitative

Either having kidney cancer or not is categorical.

Variables

For each of the following situations:

- What are the variables?
- Is each variable categorical or quantitative?

1. Can eating a yogurt a day cause you to lose weight?
2. Do males find females more attractive if they wear red?
3. Does louder music cause people to drink more beer?
4. Are lions more likely to attack after a full moon?

*(the answer to all of these questions is yes!)*

Summary

- Data are everywhere, and pertain to a wide variety of topics
- A dataset is usually comprised of variables measured on cases
- Variables are either categorical or quantitative
- Data can be used to provide information about essentially anything we are interested in and want to collect data on!

To Do

- Read Section 1.1
- If you haven't already...
  - Get the textbook
  - Get a clicker or app for your smartphone and register it at [http://www.iclicker.com/support/registeryourclicker/](http://www.iclicker.com/support/registeryourclicker/) (for Student ID use your NetID)

Why Statistics?

[http://www.youtube.com/watch?v=nTBZuQR7dRc&feature=youtu.be](http://www.youtube.com/watch?v=nTBZuQR7dRc&feature=youtu.be)