Lecture 1

Population and Sample
Lecture Summary

• We have a **population** to conduct our study.

• Often, we *can’t* gather information from every member of the population. Therefore, we **sample**!

• From the sample, we investigate various **features of the population**, called **parameters**

• We do this by creating **statistics** based on the sample
Population

• **Population**: A collection of objects for study

• Example 1:
  – **Goal**: Study the efficacy of a new malaria vaccine
  – **Population**: Individuals prone to malarial infection
  – Why not just have all individuals as the population?

• Example 2:
  – **Goal**: Study the pattern of spam mail in Gmail
  – **Population**: All the possible spam mail that are (and will be in Google’s servers)
  – **Note**: objects in the population may not exist!
See any Patterns?

Weekend

<table>
<thead>
<tr>
<th>Email Provider</th>
<th>Email Subject</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>RussianBrides</td>
<td>Flirt Live with Sexy Girls on RussianBrides</td>
<td>May 28</td>
</tr>
<tr>
<td>RussianBrides</td>
<td>Designing Group Projects: Strategies for Effective Student Collaboration - 06/14 Webinar</td>
<td>May 28</td>
</tr>
<tr>
<td>RussianBrides</td>
<td>Live with Sexy Girls on RussianBrides</td>
<td>May 28</td>
</tr>
<tr>
<td>RussianBrides</td>
<td>4 foods that KILL fat and 7 food chemicals that CAUSE it</td>
<td>May 28</td>
</tr>
<tr>
<td>FT Loan Form</td>
<td>(no subject) - We offer 2% interest rate loan, contact us for more details</td>
<td>May 28</td>
</tr>
<tr>
<td>Rate Marketplace</td>
<td>Too busy to REFI - even at 2.50%? - Thank you for subscribing.</td>
<td>May 28</td>
</tr>
<tr>
<td>Re-test Refinance</td>
<td>Seeking low rates to refinance, purchase a home or equity?</td>
<td>May 28</td>
</tr>
<tr>
<td>DirectBuy Free Pass Cen.</td>
<td>Get the Kitchen you deserve - Thank you for subscribing.</td>
<td>May 28</td>
</tr>
<tr>
<td>Belly Fat Blast</td>
<td>4 foods that KILL fat and 7 food chemicals that CAUSE it</td>
<td>May 26</td>
</tr>
<tr>
<td>Belly Fat Blast</td>
<td>4 foods that KILL fat and 7 food chemicals that CAUSE it</td>
<td>May 26</td>
</tr>
<tr>
<td>UCnet.com Specials</td>
<td>UCnet.com Specials (Week of May 28, 2012) - UCnet.com University City Specials</td>
<td>May 26</td>
</tr>
<tr>
<td>RussianBrides</td>
<td>Flirt Live with Sexy Girls on RussianBrides</td>
<td>May 25</td>
</tr>
<tr>
<td>Waterproof Spacesbags</td>
<td>Triple offer - get 3 sets for the price of 1 - Thank you for subscribing.</td>
<td>May 24</td>
</tr>
<tr>
<td>RussianBrides</td>
<td>Designing Group Projects: Strategies for Effective Student Collaboration - 06/14 Webinar</td>
<td>May 23</td>
</tr>
<tr>
<td>Colorful Hummingbird Vine</td>
<td>TVs hummingbird trumpet vine - Thank you for subscribing.</td>
<td>May 23</td>
</tr>
<tr>
<td>RussianBrides</td>
<td>RussianBrides Only $19.95 - Thank you for subscribing.</td>
<td>May 23</td>
</tr>
<tr>
<td>Rate Marketplace</td>
<td>Too busy to REFI - even at 2.50%? - Thank you for subscribing.</td>
<td>May 23</td>
</tr>
<tr>
<td>Instant Check Mate</td>
<td>The #1 Source of Background Checks: Who has searched for you?</td>
<td>May 23</td>
</tr>
<tr>
<td>Norton Anti-Virus Protoc.</td>
<td>Warning - You may not be protected by Norton. Update Now.</td>
<td>May 23</td>
</tr>
<tr>
<td>DirectBuy Free Pass Cen.</td>
<td>Limited time offer - Free DirectBuy Pass</td>
<td>May 23</td>
</tr>
<tr>
<td>Online Dating</td>
<td>Dating News: 1 in 5 Relationships Start Online - Meet Singles Today!</td>
<td>May 22</td>
</tr>
<tr>
<td>RussianBrides</td>
<td>Flirt Live with Sexy Girls on RussianBrides</td>
<td>May 22</td>
</tr>
<tr>
<td>Auto Price Finder</td>
<td>Blowout Saving on all Chevys! - Thank you for subscribing.</td>
<td>May 22</td>
</tr>
</tbody>
</table>

“It’s Friday, Friday, Gotta get down on Friday…”
Sample

• Often, we can’t take measurements for every single object in the population
  – Expensive, morally unjustified, etc.
  – May not even exist yet!

• Sample: A manageable subset of the population that is representative of the population
  – Size of subset denoted as $n$
  – Measurements from sample denoted as $X_1, \ldots, X_n$
Parameters

- **Parameters**: numerical features/descriptions/characteristics of the population, usually unknown
  - From example 1 (malaria vaccine efficacy):
    - Distribution of body temperature for all individuals after vaccination
    - Average difference in parasite levels for all individuals before and after vaccination
  - From example 2 (Gmail spam pattern):
    - Average word count in spam
    - Frequency of spam for each day of the week
**Statistic**

- **Statistic**: a function of the sample that is used to estimate/infer about the unknown parameters!
  - Examples: Sample mean, sample variance, empirical distribution/frequency, etc.

- Generally a statistic is denoted as $T(X_1, \ldots, X_n)$ or $T$ where $T$ is a function of the sample
Population/Parameter and Sample/Statistic

Features of the population (parameters)

Mean: $\mu = 4.6364$

Distribution:

<table>
<thead>
<tr>
<th>Red</th>
<th>DBlue</th>
<th>LBlue</th>
<th>Green</th>
<th>Purple</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Estimates of the features (statistics)

Mean: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = 4.619048$

Empirical Distribution/Frequency

<table>
<thead>
<tr>
<th>Red</th>
<th>DBlue</th>
<th>LBlue</th>
<th>Green</th>
<th>Purple</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>
Population/Sample with Malaria

Parameter

Distribution of body temperature for all individuals after vaccination

- \( F(x) \): cdf of \( X \)

Average difference in parasite levels for all individuals before vaccination

- \( \mu = E(X) \)
- \( X \sim F_\mu \), independent and identically distributed

Statistic

Empirical distribution of body temperature for vaccinated individuals in the sample

- \( T(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \)

Sample average difference in parasite levels before vaccination

- \( T(X_1, \ldots, X_n) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \)
How old am I?

1) What is the population
2) What is my sample
3) What parameters am I interested in
4) What statistics should I use to estimate the parameters?
Summary

• **Population**: a collection of units
  – **Parameters**: numerical description of the collection
    • E.g. Mean, variance, cumulative distribution function, etc.

• **Sample**: a manageable and representative collection of units
  – We derive **statistics** that **estimate** the parameters
    • E.g. Sample mean, sample variance, empirical distribution function, etc.
Extra Slides
Representative Sampling Strategies

• **Simple Random Sampling (SRS)**: randomly sample \( n \) objects from the population
  – Any \( n \)-subset of the population is equally likely
  – If objects are randomly sampled with replacement or if the population size is infinite, it is i.i.d. (independent and identically distributed...more on this later)

• **Stratified Sampling**: divide the population into \( K \) homogenous groups and perform SRS on each group
  – **Example 1**: Efficacy of malaria vaccine
  – Divide the population into children and adults.