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

RussianBrides	**Flirt Live with Sexy Girls on RussianBrides** - Thank you for subscribing. You can alter your subscription options bel	May 28
audio@bestprofessortrain.	**Designing Group Projects: Strategies for Effective Student Collaboration - 6/14 Webinar** - Hyunseung Kang, For t	May 28
RussianBrides	**Flirt Live with Sexy Girls on RussianBrides** - Thank you for subscribing. You can alter your subscription options bel	May 28
Belly Fat Blast	**4 foods that KILL fat and 7 food chemicals that CAUSE it** - Thank you for subscribing. You can alter your subscripti	May 28
FT Loan Form	**(no subject)** - We offer 2% interest rate loan contact us more details	May 27
Rate Marketplace	**Too busy to REFI - even at 2.50%?** - Thank you for subscribing, . You can alter your subscription options below ...	May 26
Ra-tesite Refin-ance	**Seeking lo-an ra-tes to refi-nance, pur-chase a ho-me or equity?** - Thank you for subscribing, . You can alter your s	May 26
DirectBuy Fr.ee Pass Cen.	**Get the Kitchen you deserve** - Thank you for subscribing, . You can alter your subscription options below ...	May 26
Belly Fat Blast	**4 foods that KILL fat and 7 food chemicals that CAUSE it** - Thank you for subscribing. You can alter your subscrip	May 26
UCnet.com Specials	**UCnet.com Specials (Week of May 28, 2012)** - UCnet.com: University City Specials: Trouble reading this message? T	May 26
Belly Fat Blast	... your subscripti	May 25
RussianBrides	... iption options be	May 24
Waterproof Spacebags	**Triple of-fer - get 3 sets for the price of 1** - Thank you for subscribing. You can alter your subscription options below	May 24
Belly Fat Blast	**4 foods that KILL fat and 7 food chemicals that CAUSE it** - Thank you for subscribing. You can alter your subscripti	May 24
Colorful Hummingbird Vine	**TVs hummingbird trumpet vine** - Thank you for subscribing. You can alter your subscription options below ...	May 23
Premium Cigars	**Premium Cigars Only $19.95** - Thank you for subscribing. You can alter your subscription options below ...	May 23
Rate Marketplace	**Too busy to REFI - even at 2.50%?** - Thank you for subscribing. You can alter your subscription options below ...	May 23
Instant Check Mate	**The #1 Source of Background Checks: Who has searched for you?** - Thank you for subscribing. You can alter your	May 23
Matrix Direct - Convenie.	**Secure $250K coverage for ju-st 12.82$ per month!** - Thank you for subscribing. You can alter your subscription opti	May 23
Norton Anti-Virus Protec.	**Warning- You may not be protected by Norton. Update Now.** - Thank you for subscribing. You can alter your subsc	May 23
DirectBuy Fr.ee Pass Cen.	**Limited time offer - Free DirectBuy Pass** - Thank you for subscribing. You can alter your subscription options below .	May 23
Online Dating	**Dating News: 1 in 5 Relationships Start Online - Meet Sin-gles To-day!** - Thank you for subscribing, . You can alter	May 22
RussianBrides	**Flirt Live with Sexy Girls on RussianBrides** - Thank you for subscribing. You can alter your subscription options bel	May 22
Auto Price Finder	**Blowout Saving on all Chevys!** - Thank you for subscribing, . You can alter your subscription options below ...	May 22

"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 $X_1, \dots, X_n$

Parameters

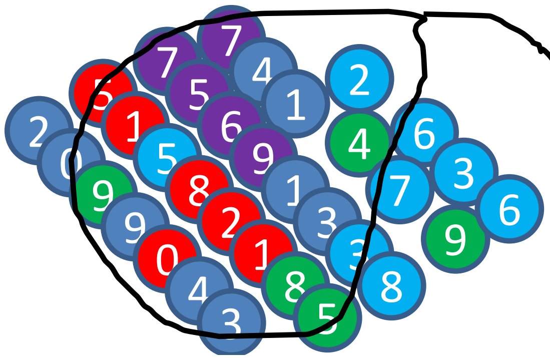
- **Parameters**: numerical features/descriptions/characteristics of the population, usually unknown
  - From example 1 (malaria vaccine efficacy):
    - Distribution of body temperature for all individuals after vaccination
    - Average difference in parasite levels for all individuals before and after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

Statistic

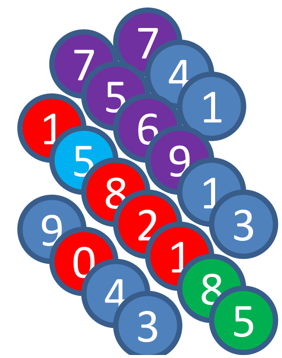
- **Statistic**: a function of the sample that is used to estimate/infer about the unknown **parameters**!
  - Examples: Sample mean, sample variance, empirical distribution/frequency, etc.

- Generally a statistic is denoted as $T(X_1, \dots, X_n)$ or T where T is a function of the sample

Population/Parameter and Sample/Statistic



Features of the population
(parameters)



Estimates of the features
(statistics)

Mean: $\mu = 4.6364$
Distribution:

Red	DBlue	LBlue	Green	Purple
6	9	8	5	5

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

Red	DBlue	LBlue	Green	Purple
5	7	1	2	5

Population/Sample with Malaria

Parameter

Distribution of body temperature for all individuals after vaccination

- $F(x)$: cdf of X

Statistic

Empirical distribution of body temperature for vaccinated individuals in the sample

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

Average difference in parasite levels for all individuals before vaccination

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

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
  - <u>Example 1</u>: Efficacy of malaria vaccine
  - Divide the population into children and adults.