MATH 4720 / MSSC 5720

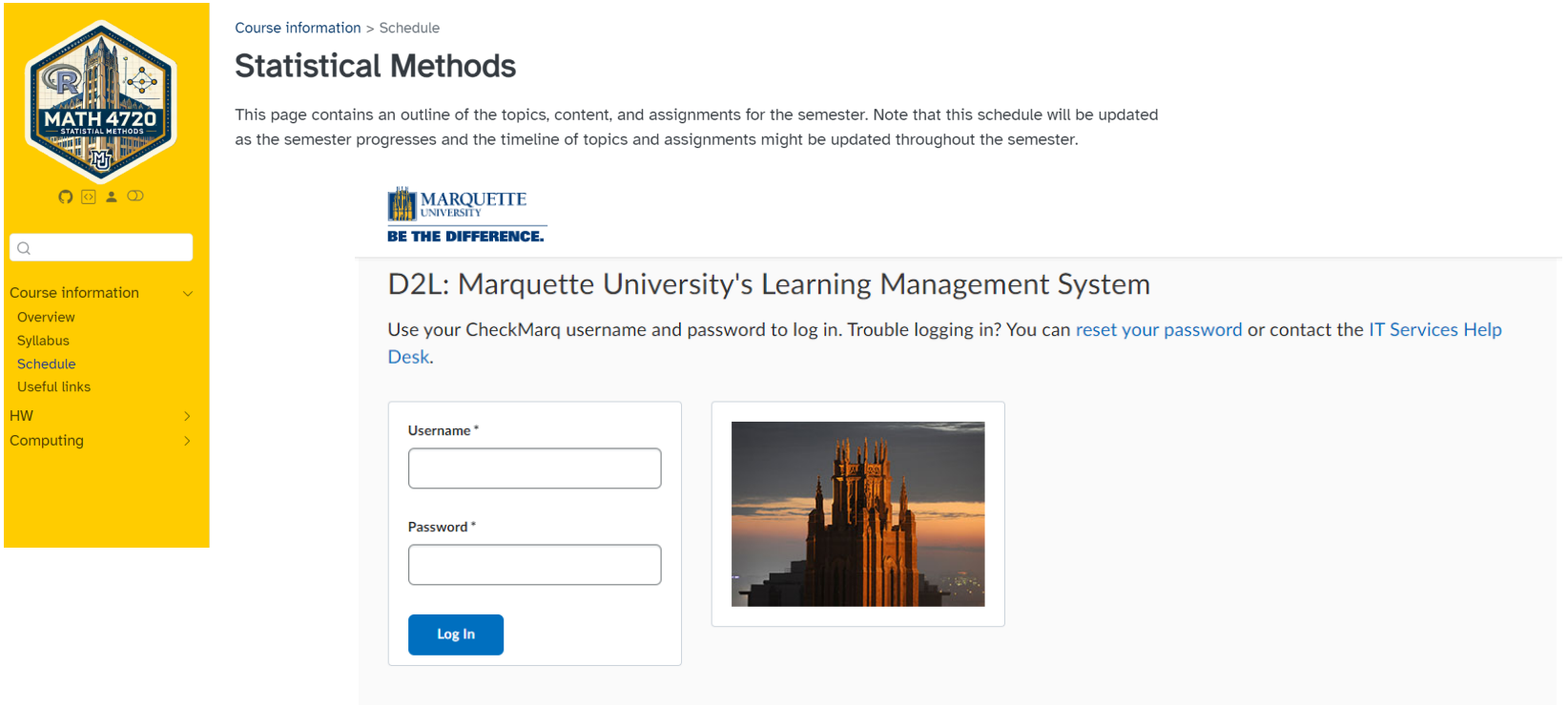
Instructor: Mehdi Maadooliat

Lecture 1

Department of Mathematical and Statistical Sciences

SYLLABUS – WEBSITE - D2L

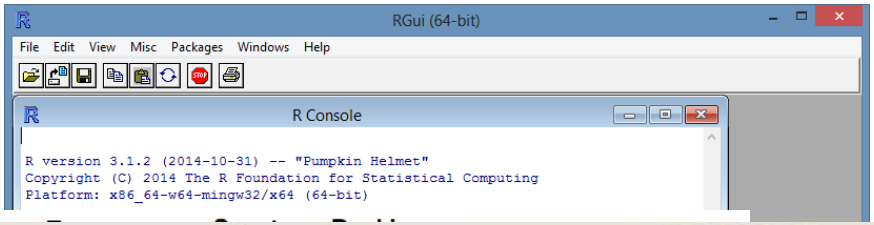
- **Syllabus**
- **Course Materials in http://tinyurl.com/Stat-Meth**
- **Homework and Discussions: http://d2l.mu.edu**



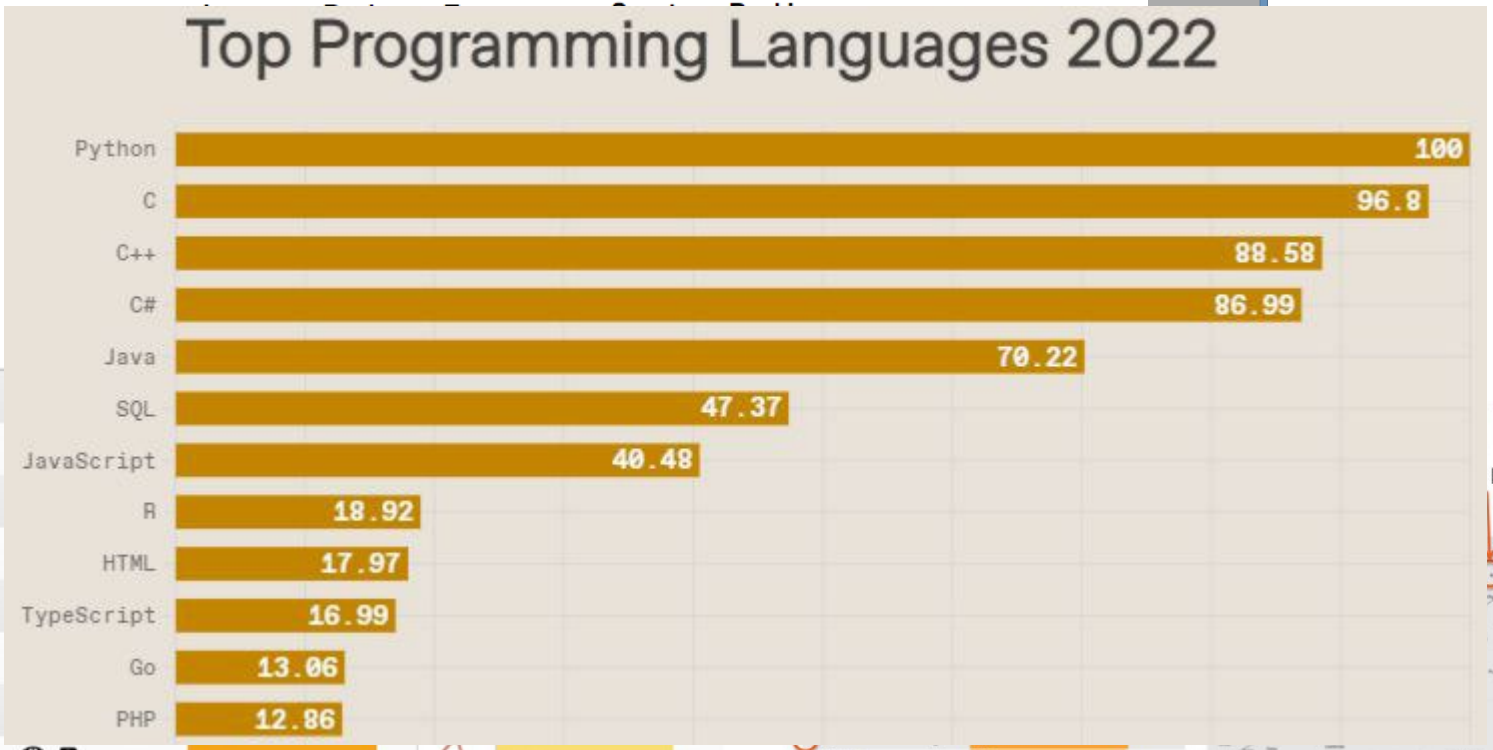
R IS THE STATISTICAL SOFTWARE THAT WE MOSTLY USE IN THIS COURSE

MARQUETTE UNIVERSITY
Be The Difference.

- **It is free!!**



[IEEE Spectrum, July 2016]



Language Rank				
1. Java				
2. C				
3. C++				
4. Python				
5. C#				
6. R				
7. PHP				
8. JavaScript		83.0	78.9	
9. Ruby		76.2	74.3	
10. Matlab		72.4	72.8	

- IEEE Spectrum, July 2014
- RedMonk Programming Language Rankings, 2013

[IEEE Spectrum, July 2015]

HISTORY OF R (FROM WIKIPEDIA)

MARQUETTE UNIVERSITY
Be The Difference.

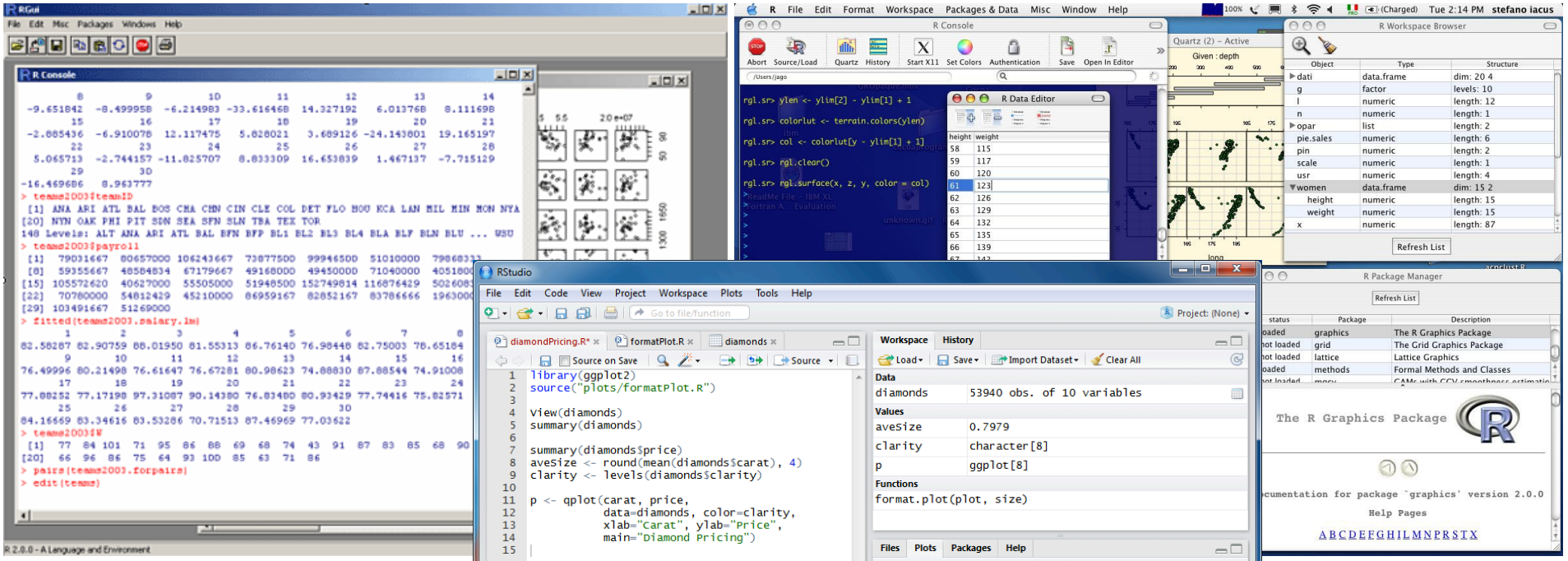
- R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.
- R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme. S was created by John Chambers while at Bell Labs. There are some important differences, but much of the code written for S runs unaltered.
- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the *R Development Core Team*, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of S.
- R is a GNU project. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. R uses a command line interface; there are also several graphical front-ends for it.

R : THE STATISTICAL SOFTWARE

MARQUETTE UNIVERSITY
Be The Difference.

- **Download and Install:** [Comprehensive R Archive Network]

Windows **Mac**

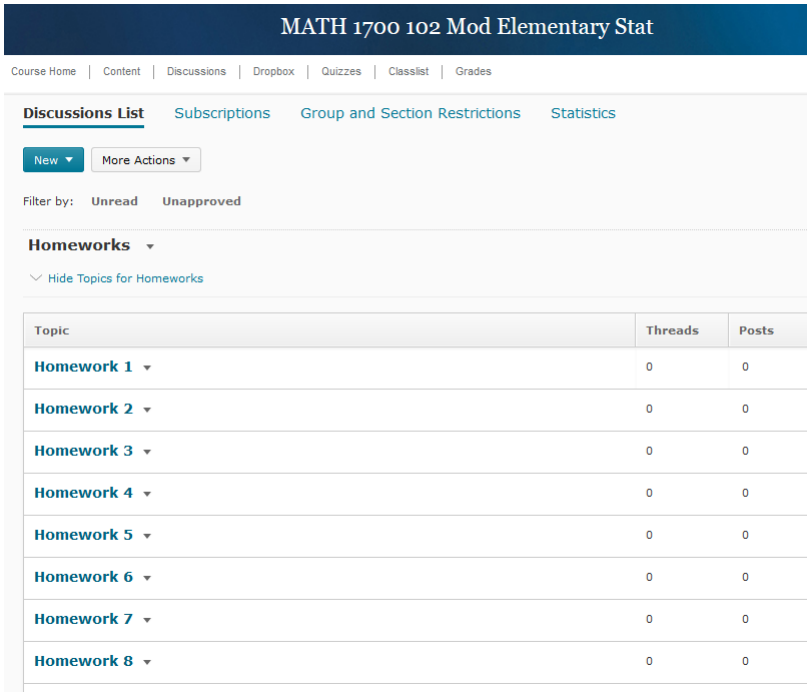


- **RStudio:**
  - [Download]

DTSC 4997 CONT…

MARQUETTE UNIVERSITY
Be The Difference.

- **Any General questions about Homework, and Projects:**
  - **SHOULD be posted in D2L Discussion Board.**
  - **I will NOT answer general emails about Homework and/or Projects.**



- **Homework and Projects:**
  - **Should be submitted as a PDF file (Otherwise you will get ZERO):**
    - **How to Combine Images into a PDF file [FREE & EASY + No Software] (Youtube)**
    - **Microsoft Word to PDF in 10 Seconds (Youtube)**
    - **How to: convert Images to PDF in Macbook/iMac (Youtube)**
      - http://apple.stackexchange.com/questions/11163/how-do-i-combine-two-or-more-images-to-get-a-single-pdf-file

MARQUETTE UNIVERSITY
Be The Difference.

TOPIC 1: CHAPTER 1-2

- **Statistics and Data Description**
- **Populations and Samples**
- **Types of Studies**
- **Confounding Variable**

MARQUETTE UNIVERSITY
Be The Difference.

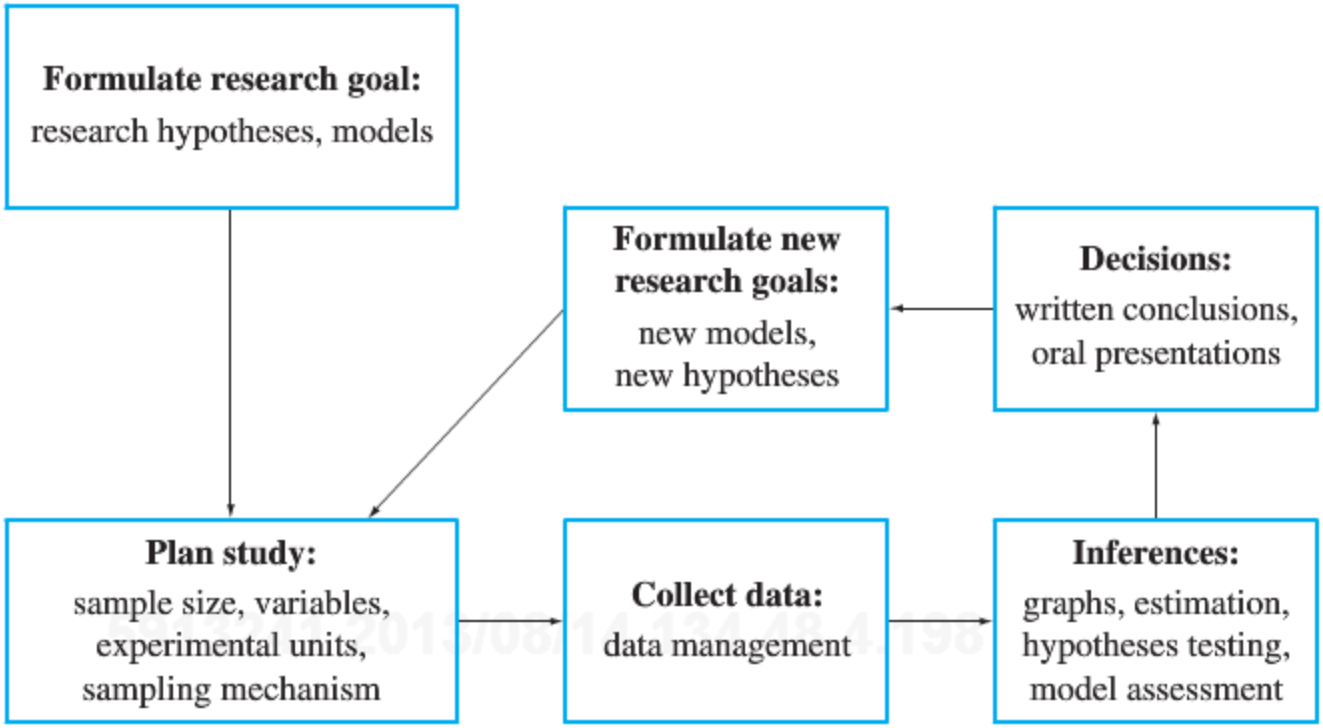
WHAT IS STATISTICS?

- What do you think of when you hear the word "statistics"?

- Statistics: The science of collecting, classifying, and interpreting data.

- Anticipated learning outcomes:
  - appreciate and apply basic statistical methods in their scientific field
  - appreciate and apply basic statistical methods in an everyday life setting

HOW TO LEARN FROM DATA?



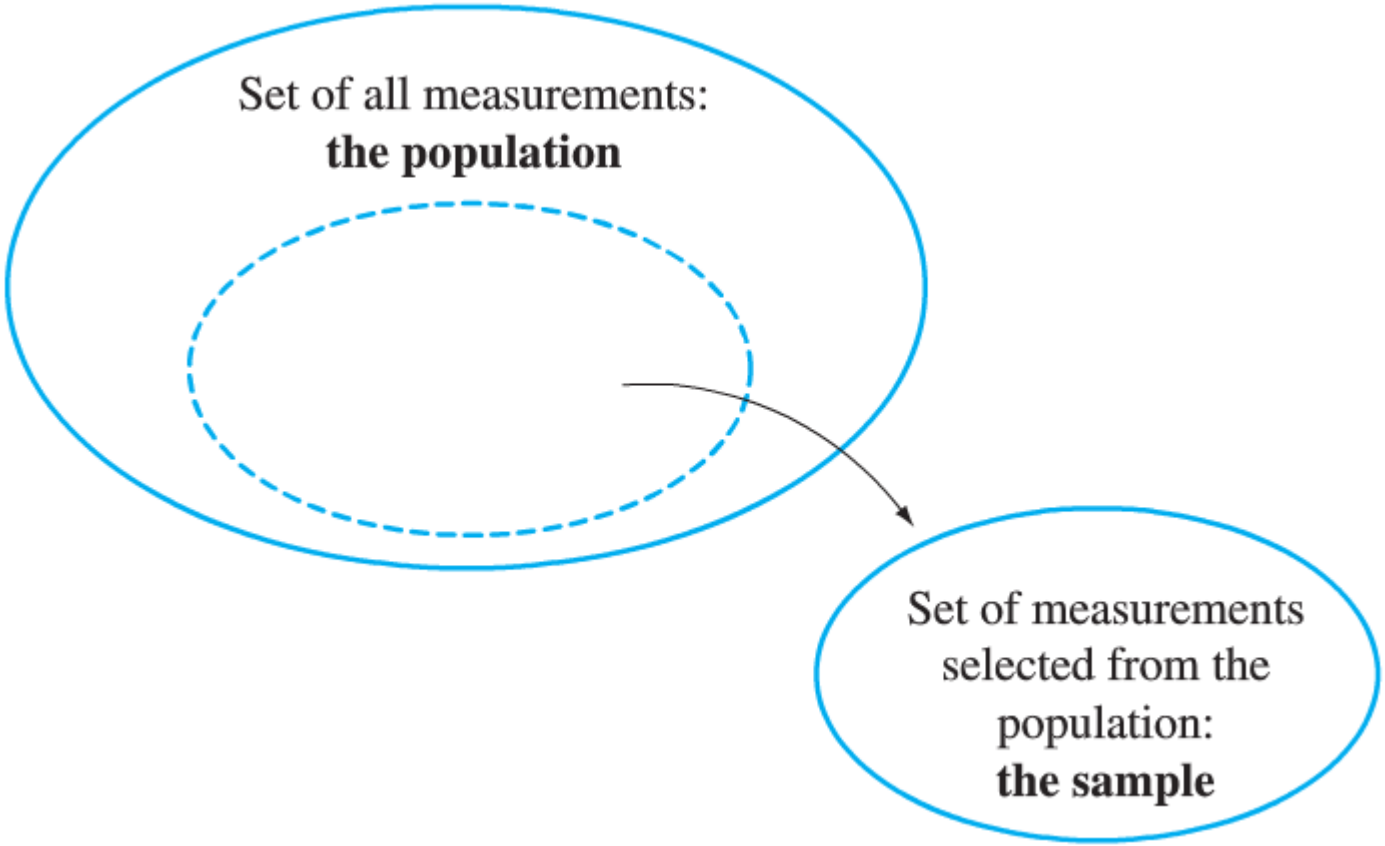
MARQUETTE UNIVERSITY
Be The Difference.

WHAT SHOULD YOU EXPECT?

The Four-Step Process	Chapters
1 Defining the Problem	1 Statistics and the Scientific Method
2 Collecting the Data	2 Using Surveys and Experimental Studies to Gather Data
3 Summarizing the Data	3 Data Description
	4 Probability and Probability Distributions
4 Analyzing the Data, Interpreting the Analyses, and Communicating the Results	5 Inferences about Population Central Values
	6 Inferences Comparing Two Population Central Values
	7 Inferences about Population Variances
	8 Inferences about More Than Two Population Central Values
	9 Multiple Comparisons
	10 Categorical Data
	11 Linear Regression and Correlation
	12 Multiple Regression and the General Linear Model

POPULATION VS. SAMPLE

- **Population**: The entire group of interest
- **Sample**: A part of the population selected to draw conclusions about the entire population



COLLECTING DATA

- **Observational study**: Observe a group and measure quantities of interest. This is passive data collection in that one does not attempt to influence the group. The purpose of the study is to describe the group.

- **Experiment**: Deliberately impose treatments on groups in order to observe responses. The purpose is to study whether the treatments cause a change in the responses.

EXPERIMENT TERMS

- **Experimental Group**: A collection of experimental units subjected to a real treatment.

- **Control Group**: A collection of experimental units subjected to the same conditions as those in an experimental group except that no treatment is imposed.

- This design helps control for potential **confounding** effects.

WHAT IS CONFOUNDING?

- The Variables that are not in control of the researcher.
- A variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.
- A perceived relationship between an dependent(response) variable and a independent(explanatory) variable that has been misestimated due to the failure to account for a confounding factor is termed a spurious relationship
  - Socioeconomic status and Life expectancy
  - Berkeley gender bias case (http://en.wikipedia.org/wiki/Simpson's_Paradox)

LURKING VARIABLE AND SIMPSON'S PARADOX

MARQUETTE UNIVERSITY
Be The Difference.

- **Lurking variable:** A variable that is not included in a study but has an effect on the variables of the study and makes it appear that those variables are related.
- **Simpson's Paradox:** An association or comparison that holds for all of several groups can reverse direction when a lurking variable is present.
- **Example: Kidney stone treatment(**Br Med J (Clin Res Ed) 292 (6524): 879–882**)**

	Treatment A	Treatment B
Small Stones	*Group 1* 93% (81/87)	*Group 2* 87% (234/270)
Large Stones	*Group 3* 73% (192/263)	*Group 4* 69% (55/80)
Both	78% (273/350)	83% (289/350)

- http://en.wikipedia.org/wiki/Simpson's_Paradox

ASSOCIATION BETWEEN CELL PHONE USE AND THE OCCURRENCE OF CANCER?

- Recent USA Today article.
- Three studies:
  - A German study (Stang et al., 2001) compared 118 patients with a rare form of eye cancer to 475 healthy patients who did not have the eye cancer. The patients cell phone use was measured using a questionnaire. The eye cancer patients used cell phones more often, on the average.
  - A British study (Hepworth et al., 2006) compared 966 patients with brain cancer to 1716 patients who did not have brain cancer. The patients cell phone use was measured using a questionnaire. The two groups' use of cell phones was similar.
  - An Australian study (Repacholi, 1997) conducted an experiment with 200 transgenic mice, specially bred to be susceptible to cancers of the immune system. One hundred mice were exposed for two-half hour periods a day to the same kind of microwaves with roughly the same power as that transmitted from a cell phone. The other 100 were not exposed. After 18 months, the brain tumor rate for the mice exposed to radiation was twice as high as the brain tumor rate for the unexposed mice.

Source: Agresti, A & Franklin C (2009) Statistics: The Art and Science of Learning from Data (2nd edn) Pearson Prentice Hall

REFERENCES FOR CELL PHONE & CANCER STUDIES

- Hepworth, SJ et al., (2006) Mobile phone use and risk of glioma in adults: case control study. British Medical Journal, 332, 883-887.
- Repacholi, HM (1997) Radio frequency field exposure and cancer. Environ. Health Prospect, 105, 1565-1568.
- Stang A et al., (2001) The possible role of radio frequency radiation in the development of uveal melanoma. Epidemiology, 12(1), 7-12.

QUESTIONS TO CONSIDER ABOUT THESE THREE STUDIES

- **How do the three studies differ?**
  - In studies 1 and 2 no treatments are assigned. Patients are merely questioned. Thus, studies 1 and 2 are observational studies.
  - Study 3 uses experiments on mice with the hope of generalizing to humans.
- **Why do the results of different medical studies sometimes disagree?**
  - Differing types of studies, data collection, sample frames.
  - Sampling variability.
- **Could the third study have used human subjects instead?**
  - No, because it would be unethical to knowingly expose humans to possibly harmful waves.

INFERENTIAL STATISTICS

- **Example (1988, the Steering Committee of the Physicians' Health Study Research Group)**
  Question: Can aspirin reduce the risk of heart attack in humans?

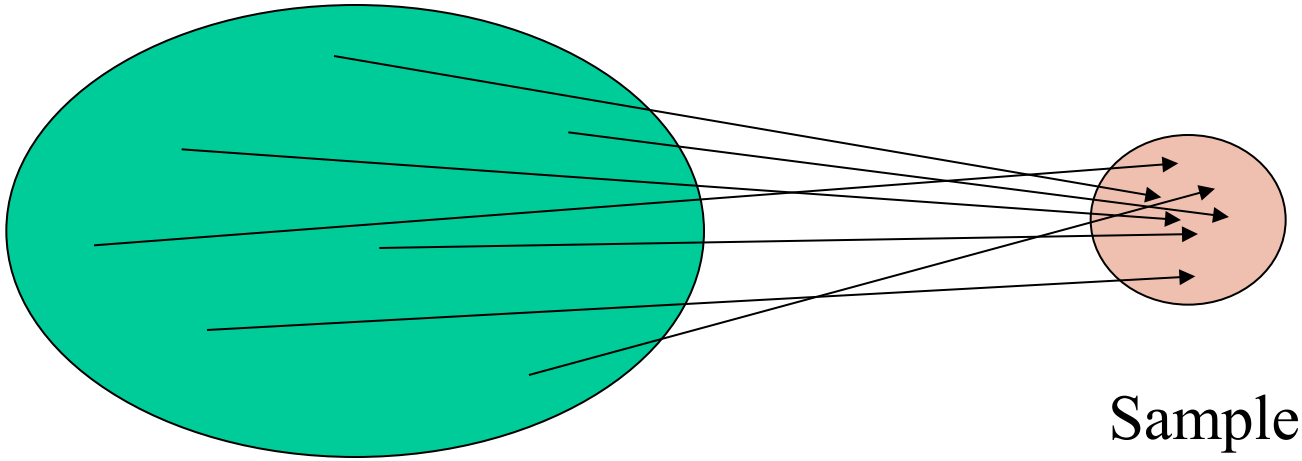
  - **Sample**: Sample of 22,071 male physicians between the ages of 40 and 84, randomly assigned to one of two groups. One group took an ordinary aspirin tablet every other day (headache or not). The other group took a placebo every other day. This group is the control group.

  - **Summary statistic**: The rate of heart attacks in the group taking aspirin was only 55% of the rate of heart attacks in the placebo group.

  - **Inference to population**: Taking aspirin causes lower rate of heart attacks in humans.

SAMPLING A SINGLE POPULATION

- **Sampling Techniques**
  - **Simple Random Sample (SRS): every member of the population has an equal chance of being selected.**



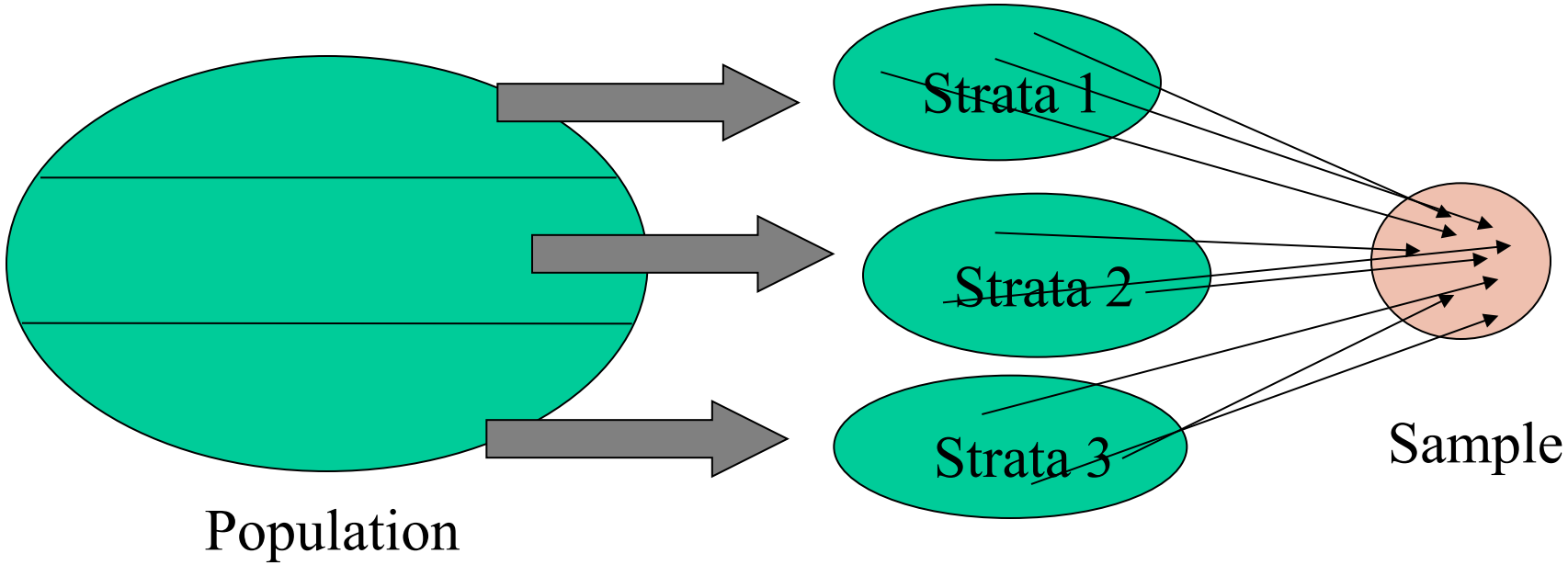
- **Simple Random Sample**

MARQUETTE UNIVERSITY
Be The Difference.

IS IT THAT EASY?

SAMPLING A SINGLE POPULATION

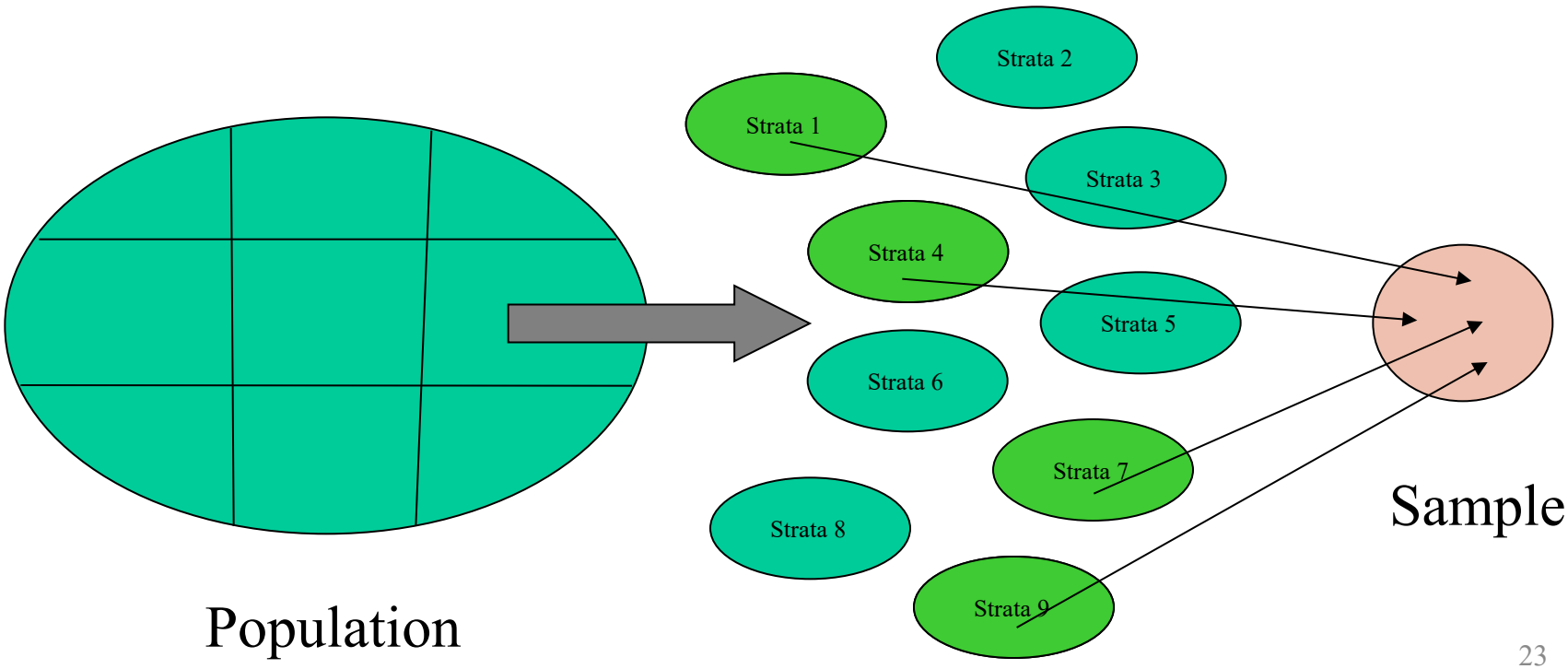
- **Sampling Techniques**
  - **Stratified Random Sample: Divide the sample into several strata. Then take a SRS from each stratum.**



- **Advantage: Each stratum is guaranteed to be randomly sampled**
- **Example: Obtain a list of all SSN for individuals in the U.S. who are over 65. Divide up the SSNs into region of the country (time zones). Then randomly sample 30 from each time zone.**

SAMPLING A SINGLE POPULATION

- **Sampling Techniques**
  - **Cluster Sample: Divide the sample into several strata or clusters. Then take a SRS of clusters.**



SAMPLING A SINGLE POPULATION

- **Sampling Techniques**
  - **Cluster Sample**
    - **Advantage: May be the only feasible method, given resoures.**
    - **Example: Obtain a list of all SSNs for individuals in the U.S. who are over 65. Sort the SSNs by the last 4 digits making each set of 100 a cluster. Use a random number table to pick the clusters. You may get the 4100's, 5600's and 8200's for example.**

INFERENCE OVERVIEW

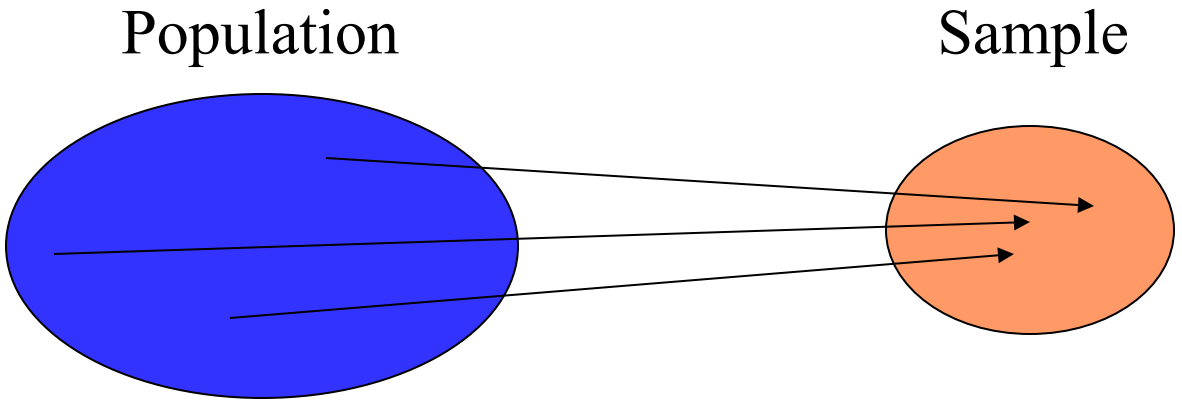
- **Describing a Population**
  - It is common practice to use Greek letters when talking about a population.
  - We call the mean of a population μ.
  - We call the standard deviation of a population σ and the variance σ^2.
  - When we are talking about percentages, we call the population proportion π. (or pi).
  - It is important to know that for a given population there is only **one** true mean and **one** true standard deviation and variance or **one** true proportion.
  - There is a special name for these values: **parameters**.

INFERENCE OVERVIEW

- **Describing a Sample**
  - We call the mean of a sample $\overline{x}$.
  - We call the standard deviation of a sample s and the variance s^2.
  - When we are talking about percentages, we call the sample proportion $\hat{\pi}$.
  - There are many different possible samples that could be taken from a given population. For each sample there may be a **different** mean, standard deviation, variance, or proportion.
  - There is a special name for these values: **statistics**.

INFERENCE OVERVIEW

- **We use sample statistics to make inference about population parameters**



Mean:	μ	$\bar{x}$
Standard Deviation:	σ	s
Proportion:	π	$\hat{\pi}$