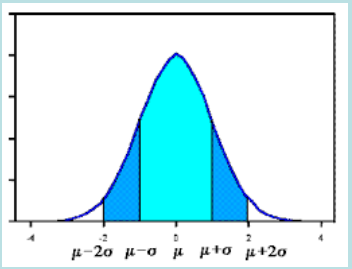
Fall 2020

Advanced Biometry

BIOL 6090



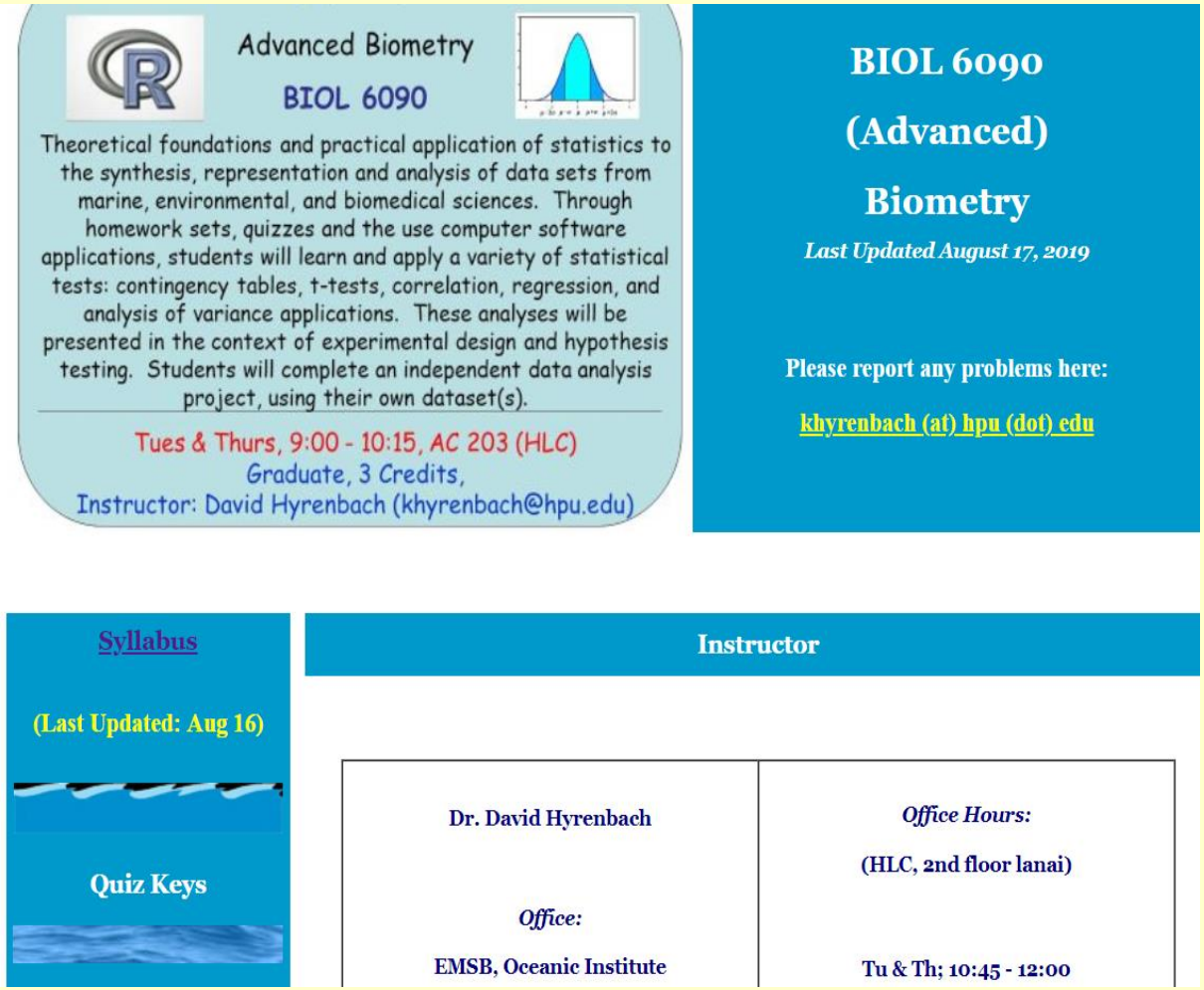
Theoretical foundations and practical application of statistics to the synthesis, representation and analysis of data sets from marine, environmental, and biomedical sciences. Through homework sets, quizzes and the use computer software applications, students will learn and apply a variety of statistical tests: contingency tables, t-tests, correlation, regression, and analysis of variance applications. These analyses will be presented in the context of experimental design and hypothesis testing. Students will complete an independent data analysis project, using their own dataset(s).

Mon & Weds, 9:00 - 10:15, AC 203 (HLC)

Graduate, 3 Credits,

Instructor: David Hyrenbach (khyrenbach@hpu.edu)

Course Web-Site



Lectures (pdfs)

Readings
- Assigned
- Extra

Homework Keys

Quiz Keys

www.pelagicos.net/classes_advancedbiometry_fa20.htm

Course Objectives

The main focus of this course is to provide students with the background of knowledge and the tools necessary to select, perform and interpret statistical analyses of biological data.

Another major focus of this course is to review the philosophical underpinnings of the scientific method and how hypothesis testing is implemented.

Achievement of the learning outcomes will be assessed using assignments, quizzes, an individual project and in-class participation.

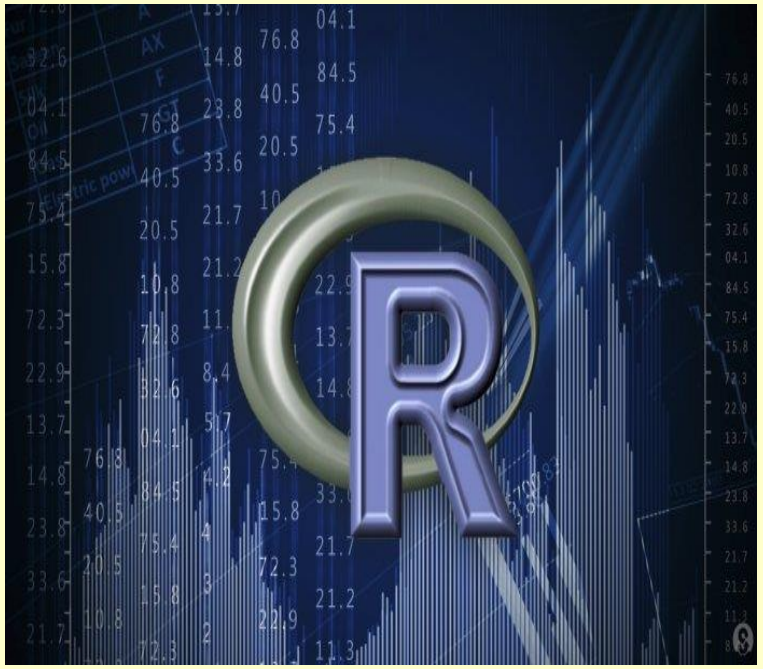
Pragmatic Course Objectives

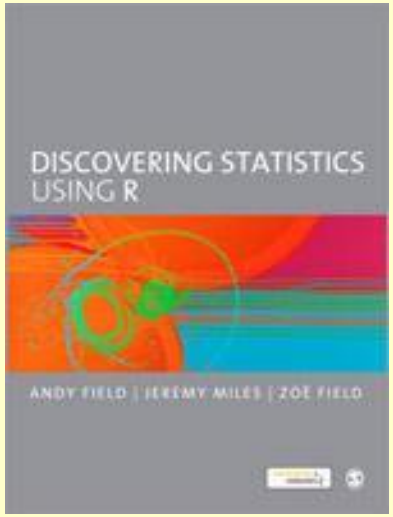
A key goal of this course is to provide students with the tools necessary to perform statistical tests and data analyses…

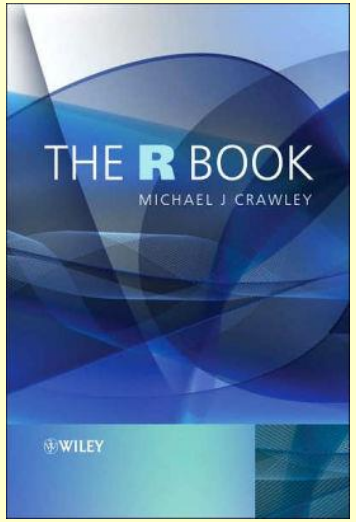
during this course

and

beyond this course







Pragmatic Course Objectives

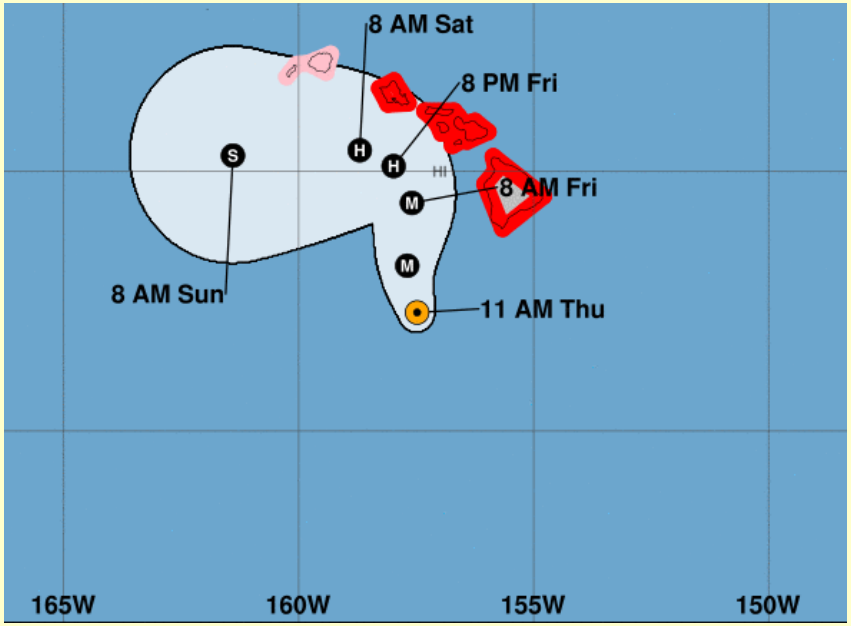
In addition to establishing a foundation of statistical knowledge and methods, this class is built upon 3 key underlying ideas:

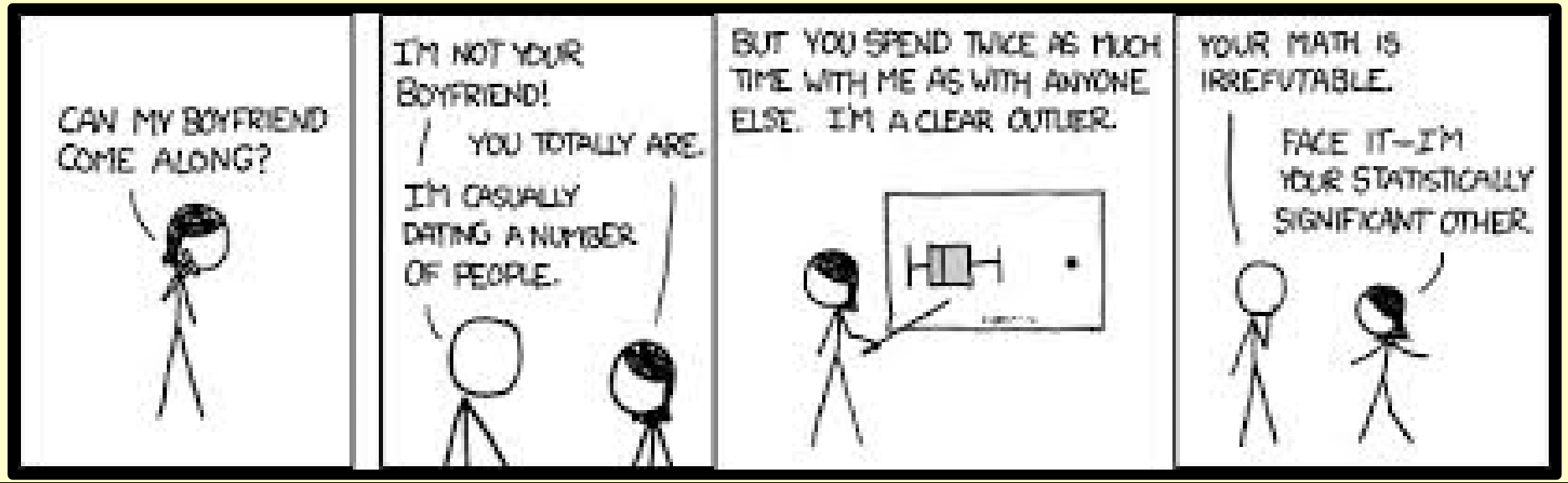
(1) Statistics is an applied field with a wide range of practical applications.

(2) Data are messy, and statistical tools are imperfect.

(3) When you understand the strengths and weaknesses of these tools, you can use them to learn about the real world.

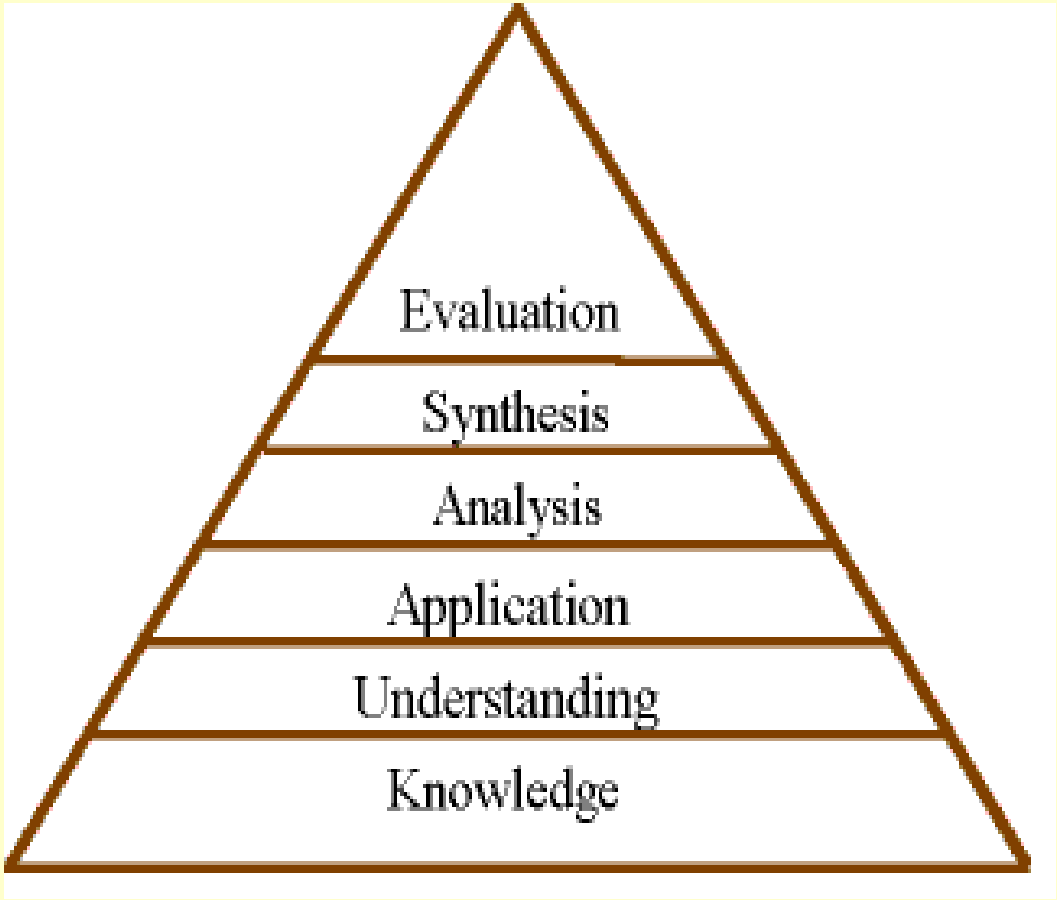
Using Statistics in Every-Day Life





Course Structure

Bloom's Taxonomy: (Bloom, 1956)
describes six levels of cognitive domains



Evaluation:
appraise, argue, evaluate

Synthesis:
arrange, develop, formulate

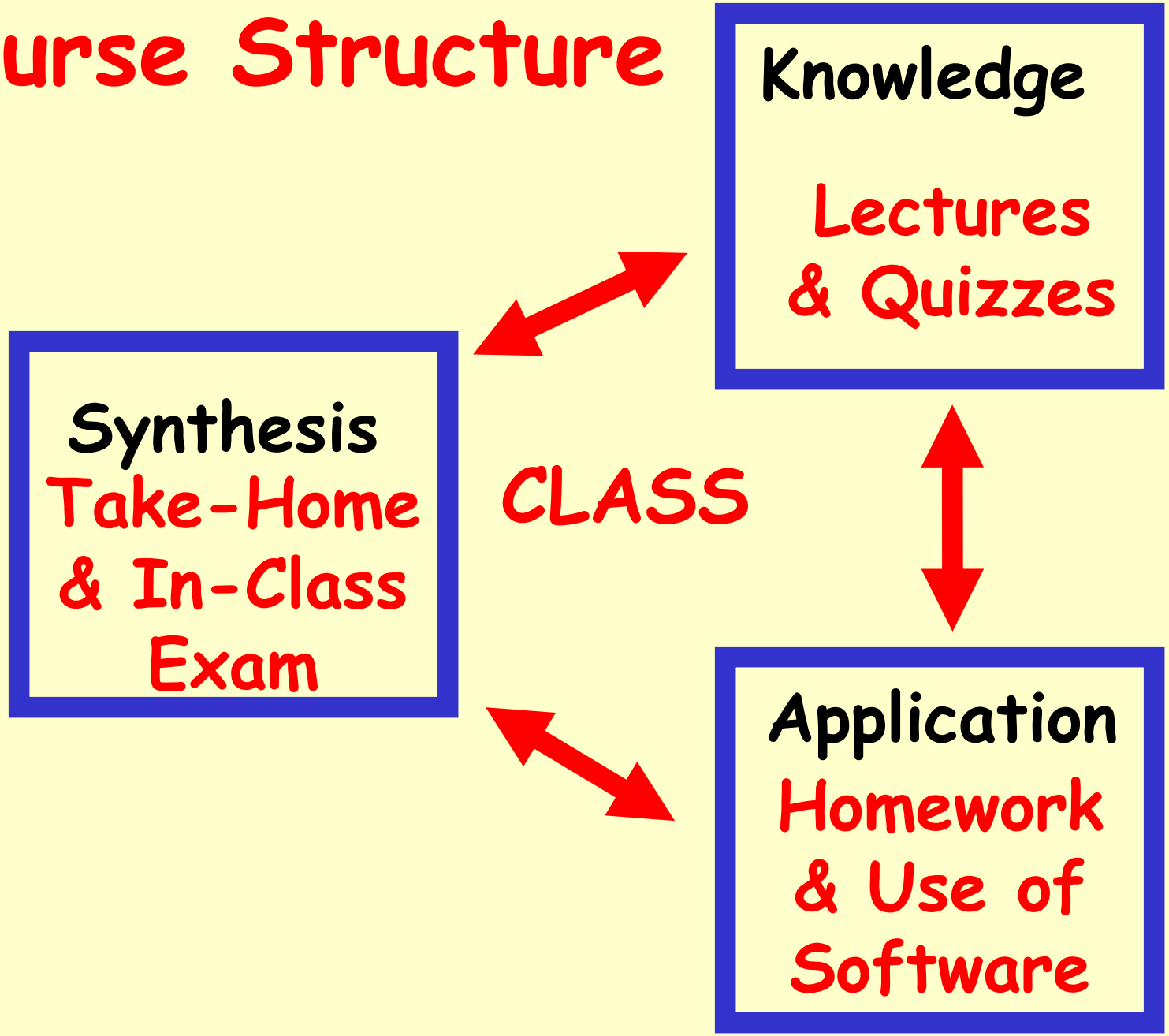
Analysis:
analyze, compare, contrast

Application:
apply, employ, practice

Understanding:
describe, discuss, explain

Knowledge:
define, label, list

Course Structure



Meeting Times / Places

- Meetings: Mon & Weds 9:00 – 10:15

  OLC 103, OI

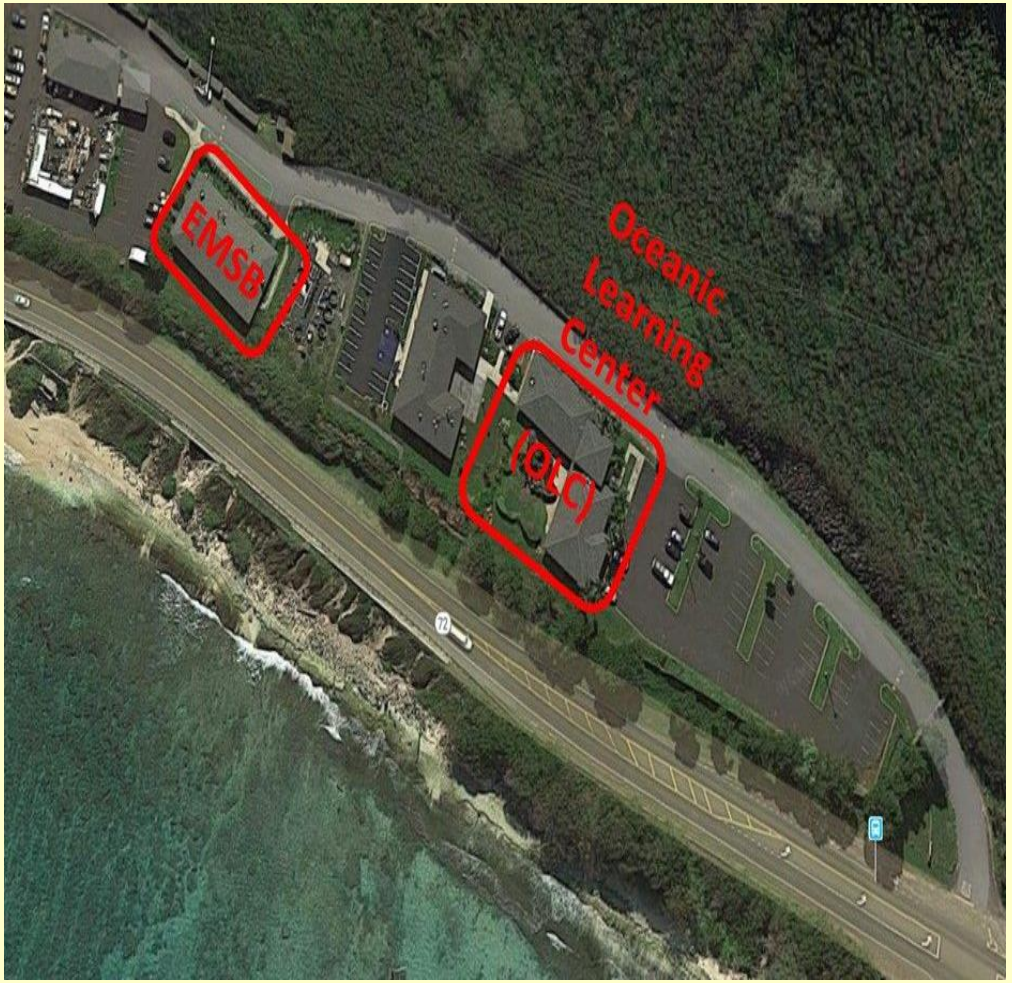
- Office Hours:

  @ EMSB, OI

M & W, 10:30 - 11:45

M & W, 14:00 – 15:00

- Or by appointment…



Required Readings

- Texts:

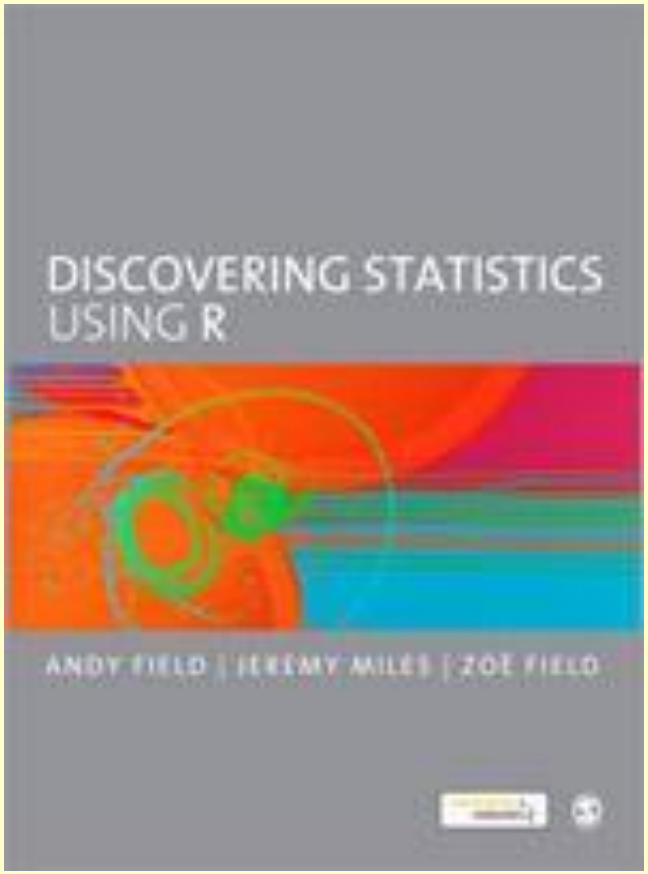
Discovering Statistics Using R, 1st Ed, 2012,

Andy Field

Sage Publications

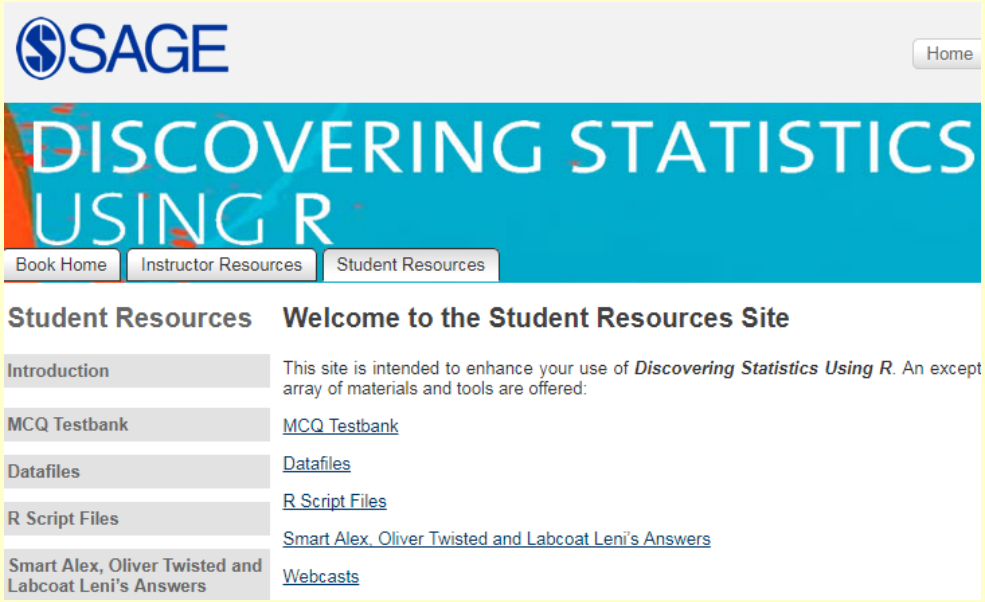
Thousand Oaks

(ISBN 978-1446200469)



Other Resources

Text Web-Site: https://studysites.sagepub.com/dsur/



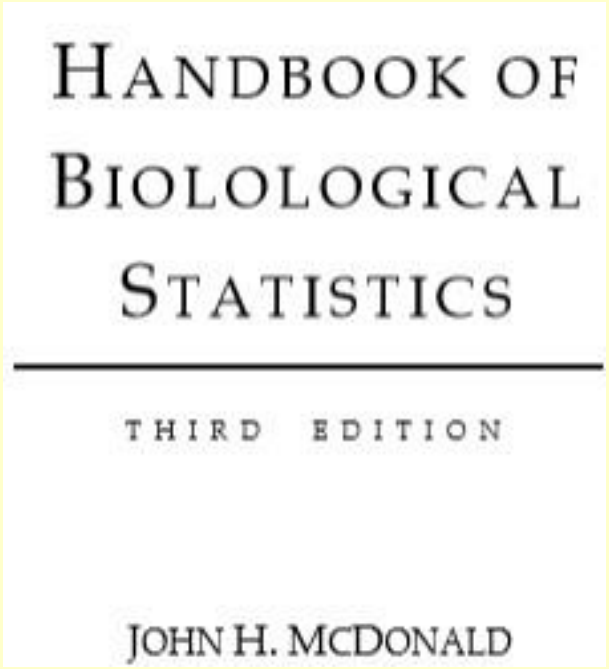
Resources include:

- Interactive MCQs
- Flashcard glossary
- R scripts / Datasets
- Smart Alex's answers

Scientific Articles: 1 – 2 each week

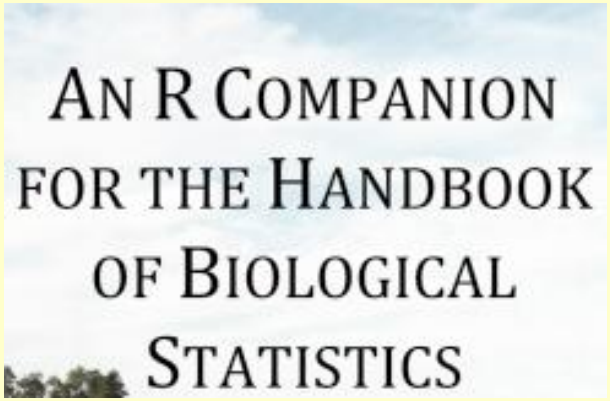
Recommended Readings

- Texts:



2014 (3rd edition)
by McDonald, J.H.

www.biostathandbook.com/



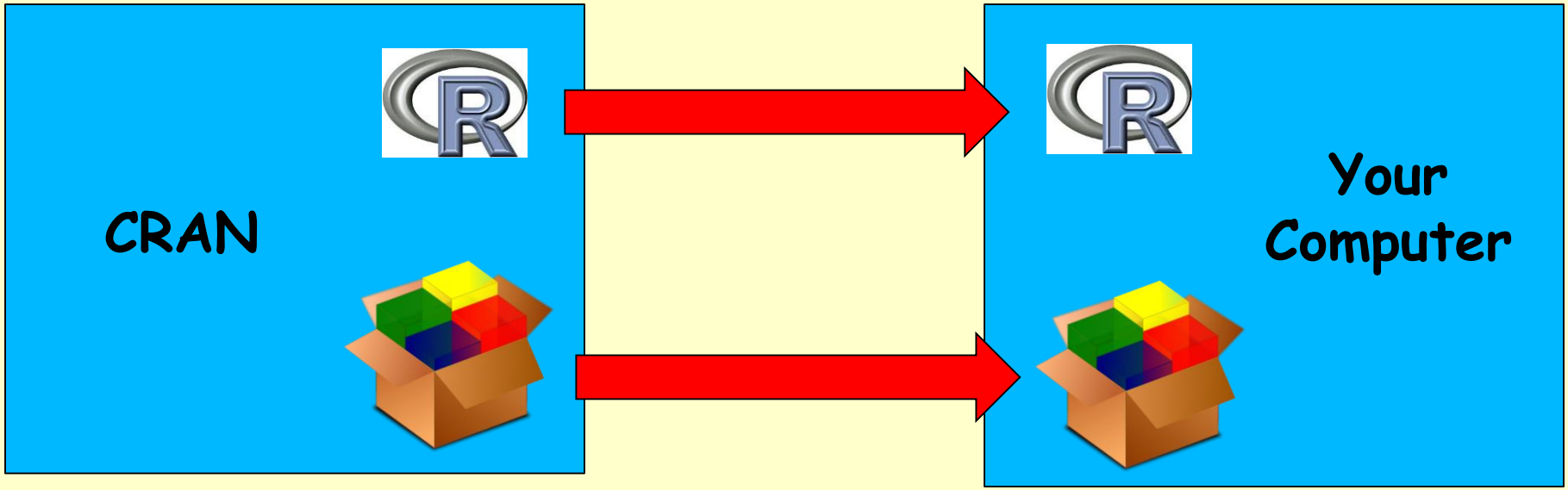
2015 by Salvatore S. Mangiafico.
Rutgers Cooperative Extension,
New Brunswick, NJ.

www.rcompanion.org/rcompanion/

Why R ?

Powerful, Versatile, Dynamic, Free !!!

The R Architecture



R Software

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Developer Pages
R Blog

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

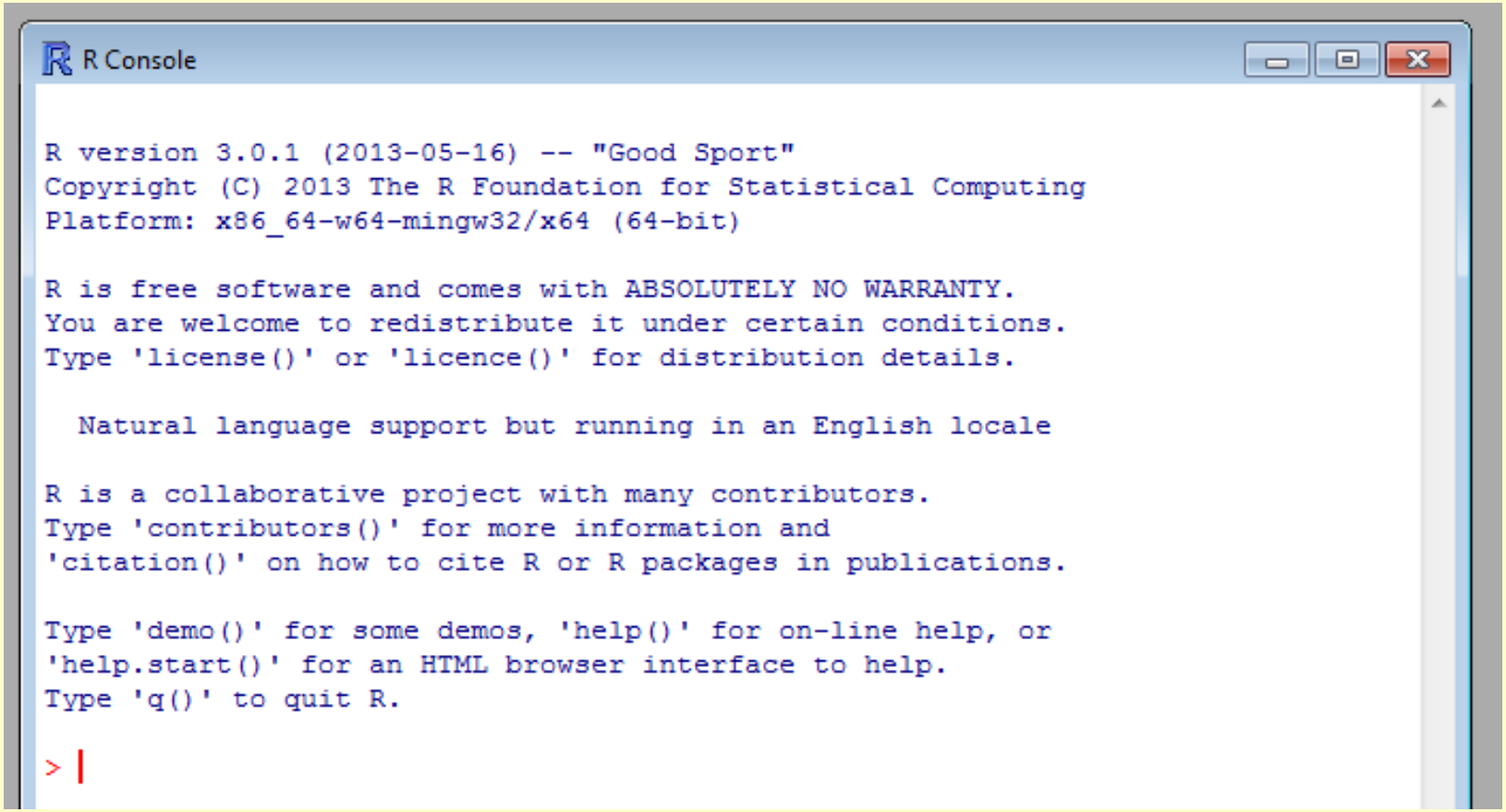
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 4.0.2 (Taking Off Again)** has been released on 2020-06-22.
- useR! 2020 in Saint Louis has been cancelled. The European hub planned in Munich will not be an in-person conference. Both organizing committees are working on the best course of action.
- **R version 3.6.3 (Holding the Windsock)** has been released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a supporting member

http://www.r-project.org/

R Software



R console allows use of interactive Markdown script commands and creation / loading / running of scripts.

Minimal use of menu-driven GUIs

Other Resources

R Studio Freeware:

Windows and Mac OS X versions of R come with simple programming editors, but I strongly recommend using the RStudio interactive development environment (IDE) for command-line use of R.

RStudio supports R Markdown documents and incorporates a powerful editor and easy-to-use file management tools.

To download, visit the RStudio web site at:
https://www.rstudio.com/products/rstudio/
for details, including extensive documentation.

Other Resources

R Commander Package:

R Commander is a basic graphical user interface (GUI) for R, which provides a series of menus that allow users to run many statistics and create graphics without typing code.

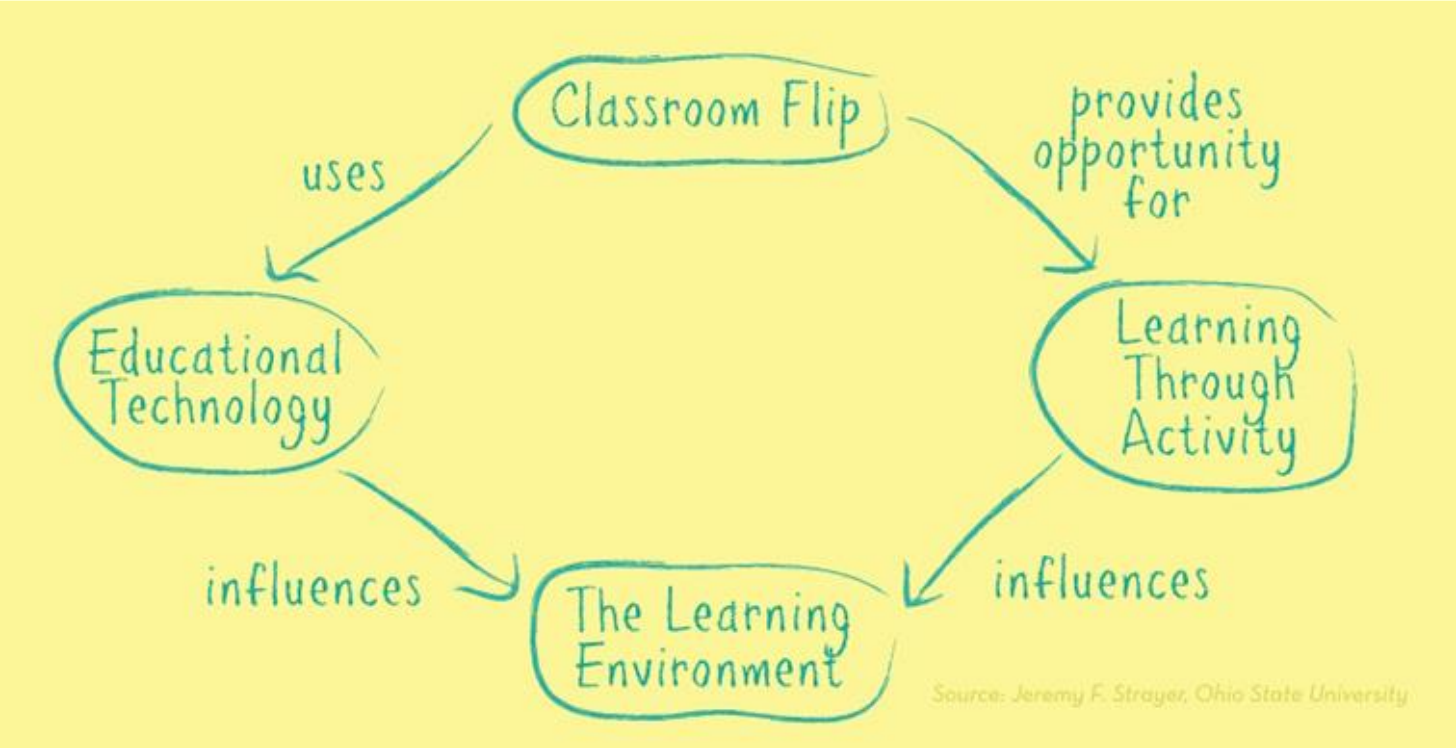
While more advanced features of R are not accessible through R Commander, you can use it for most statistics.

We will use R Commander to get the hang of the R language and to learn how to perform many analyses.

Flipped Teaching

Form of blended learning in which students review lectures online and work on problem sets in class.

This approach allows the instructors to spend more time interacting with students instead of lecturing.



Examinations

- 5 Quizzes
- NO MIDTERMS
- NO FINAL

Grades

- **Final grades will be determined as follows:**

Quizzes - 5	50%
R Homework Sets - 4:	20%
Independent Analysis Project:	25%
Participation: (come prepared to class, seek help)	5%
Total	**100%**

Individual Project

- Students will develop and undertake an independent data analysis class project.
- Analyses may involve datasets from your graduate advisor, internship, or the internet.
- Students will turn in:

A proposal detailing proposed analyses (5 points)
A data screening report, summarizing the data fields, evaluating the data for normality, and performing any needed transformations (5 points)
A 15-minute presentation, outlining the project's goals, approach, and outcomes (5 points)
A write-up, summarizing methods and results (10 points)

Participation

Students will be assessed on their effort and commitment to learning. This includes coming to class prepared, and – if necessary - seeking help in class or during office hours.

Students will review lecture presentations, readings and keys before coming to class to become familiar with the material.

The instructor will organize group activities and will randomly select volunteers to explain concepts and homework problems to the class.

Students will be evaluated qualitatively (+ / -).

Class Policies

Attendance - There will be no make-up quizzes / 5-minute papers except in the case of documented medical / family necessity.

Coming to class late - Tardiness disturbs others. If you must come late or leave early, make as small a disturbance as possible by sitting close to the door.

Cell phones are ***not*** allowed in class; turn them off (make them silent) before entering the room.

Laptops are allowed to take notes / view lecture pdfs. This is a privilege which will be revoked if laptops are used for non-class activities (e.g., email / facebook).

Academic Integrity

It is academically dishonest to try to pass off someone else's intellectual work as your own, or to help someone else to do so. Thus, **there are no circumstances under which including someone else's writing or results in your papers or assignments is permissible**.

Plagiarism will result in a zero on the assignment, and issuance of an academic dishonesty report to the University's Office of Academic Affairs. Serious cases of academic dishonesty will lead to an "F" in the course and may lead to expulsion from the University. Students are expected to comply with HPU's Academic Integrity

(http://www.hpu.edu/StudentServices/AcademicIntegrity/index.html)

Big Data

A Revolution That is Transforming How We Live, Work, and Think

Big Data

www.ted.com/talks/
mona_chalabi_3_ways_to_spot_a_bad_statistic?

What are Statistics ?

My Definition: The study of the collection, organization and interpretation of data.

They are involved in all stages of scientific research, from the planning of data collection (design of surveys and experiments) to the reporting and interpretation of results.

Why use Statistics ?

To plan efficient sampling of patterns.

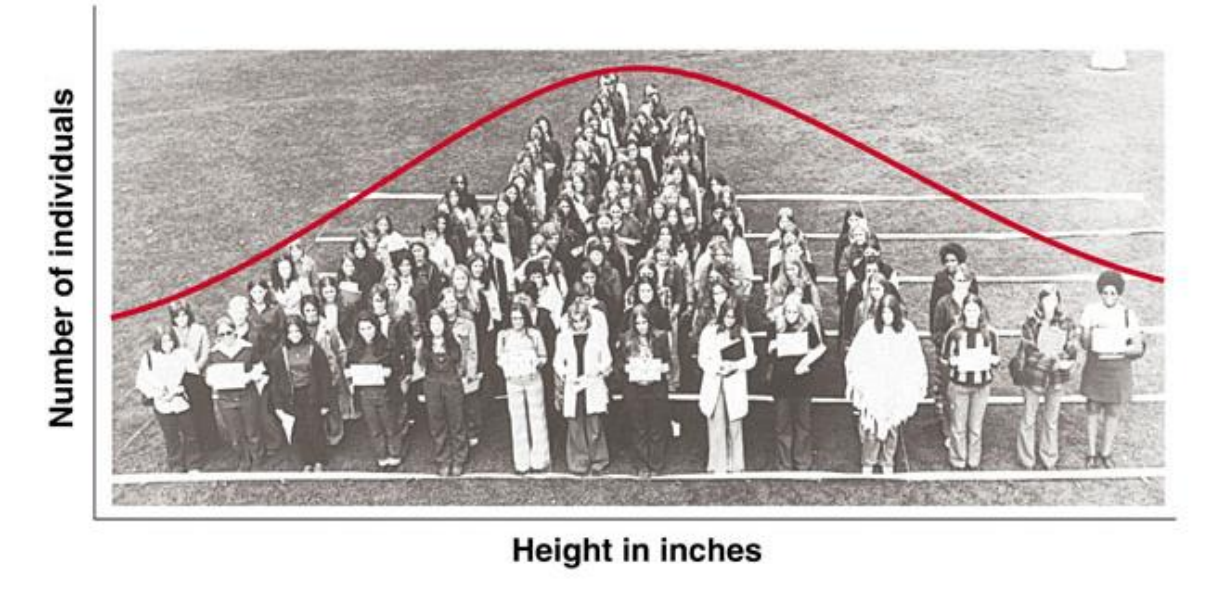
To rigorously quantify and compare observations.

To understand patterns by developing and evaluating models.

"Data don't make any sense,
we will have to resort to statistics."

Describing Patterns

Facilitate description of a single population and the comparison of multiple populations



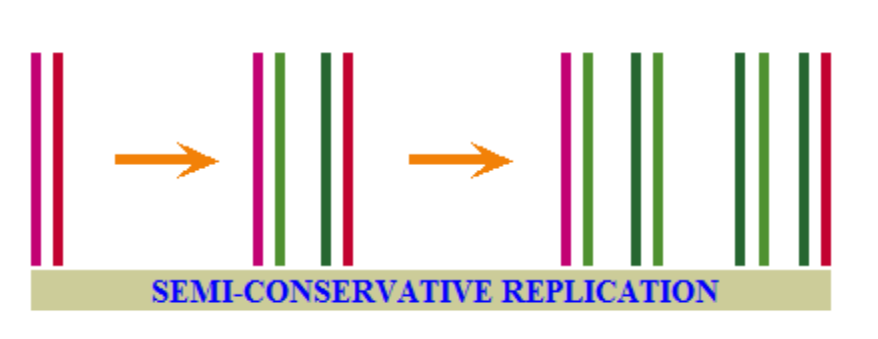
Describe "shape" of frequency distributions (using descriptive statistics)

Compare populations (using statistical tests)

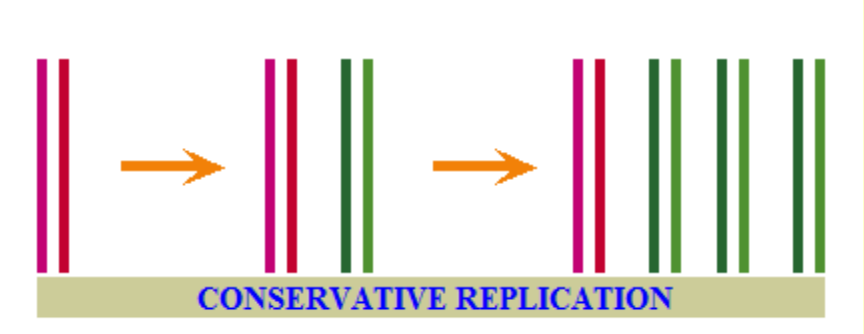
Comparing Predictions & Observations

- Simple systems: Two Models of DNA Replication

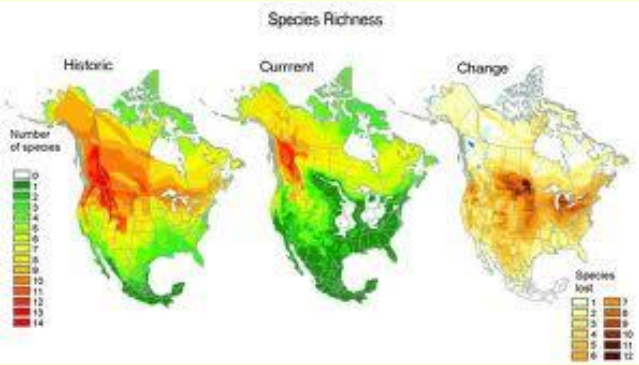
SEMI-CONSERVATIVE



CONSERVATIVE



- Complex systems: Species Distributions



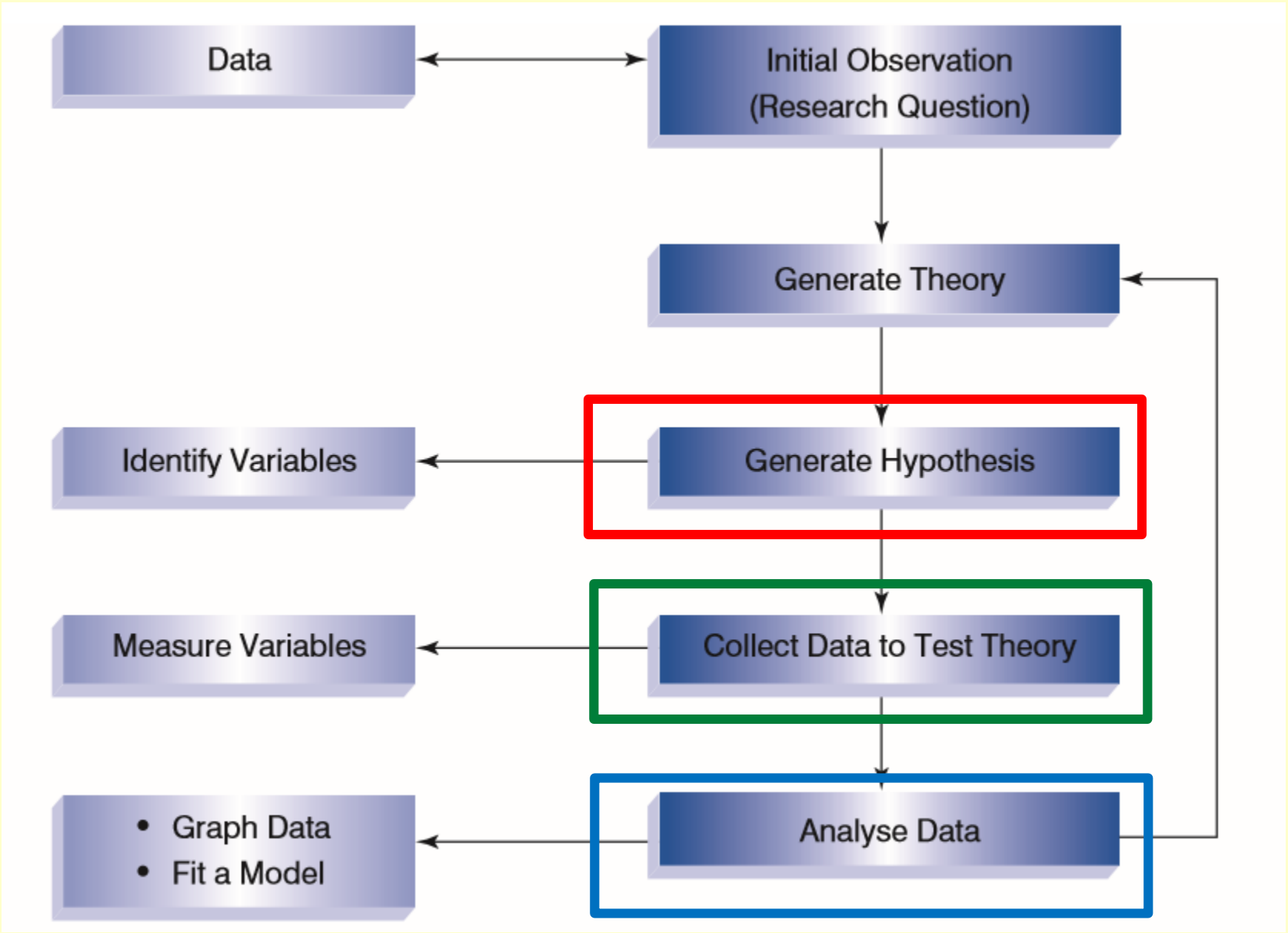
Using Statistics Correctly

Critical evaluation of the scientific literature

Elegant and rigorous study design and analysis

"I can prove it or disprove it! What do you want me to do?"

The Scientific Method



Step-by-step Data Analysis

1. Specify the biological question you are asking.

2. Phrase the question in form of a biological null / alternate hypotheses (the scientific hypothesis).

3. Phrase the question in form of a statistical null / alternate hypotheses.

4. Determine which variables are relevant to the statistical hypotheses (and the biological question).

5. Determine what kind of variables they are: numerical / categorical, discrete / continuous.

6. Design an experiment (or sampling approach) that controls or randomizes the confounding variables.

Step-by-step Data Analysis

7. Choose the best statistical test to use, based on: the hypothesis being tested, the number and kinds of variables, and the expected fit to the assumptions.

8. If possible, do a power analysis to determine the sample size needed for good statistical power.

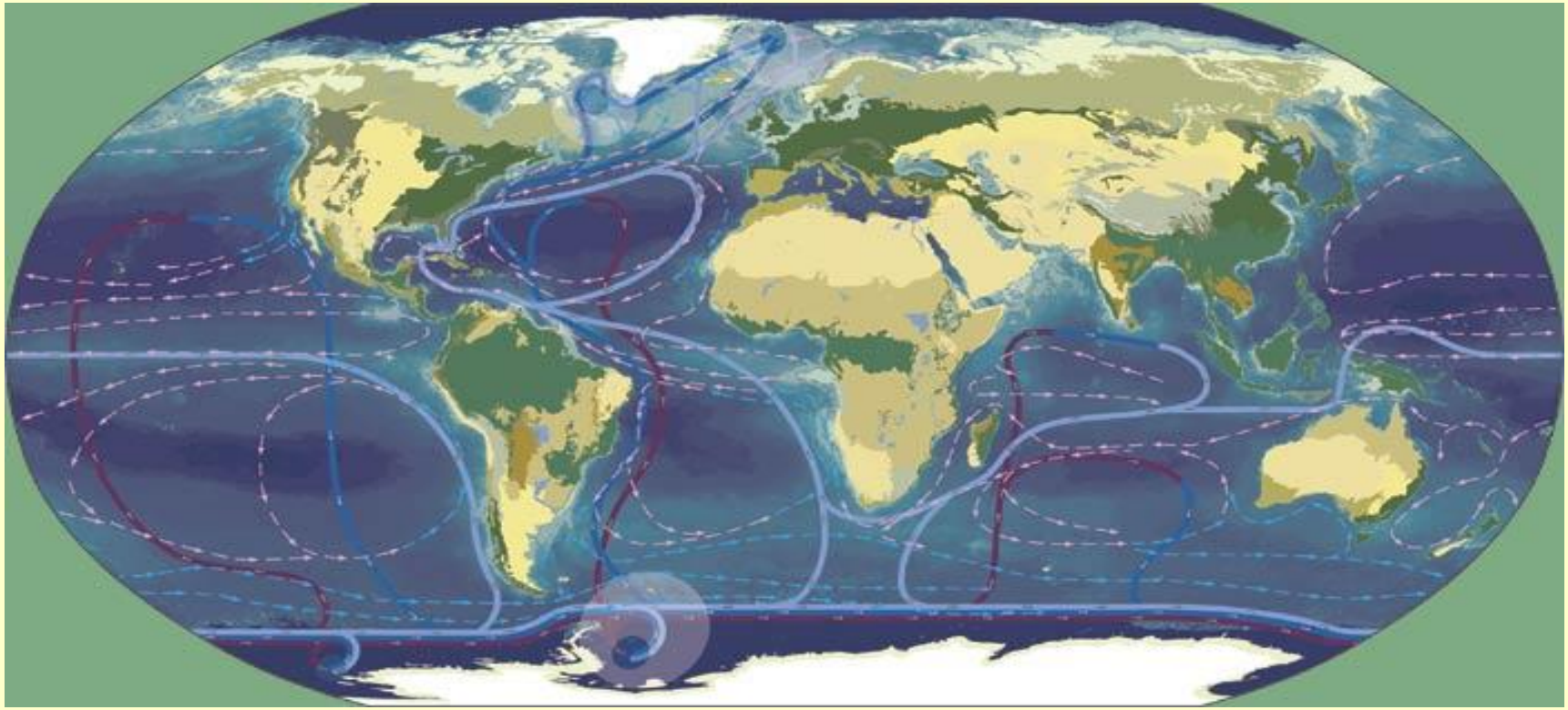
9. Do experiment (Take samples / Make measurements).

10. Examine data to see if they meet the assumptions of the selected statistical test (parametric statistics). If they do not, choose a non-parametric version of test.

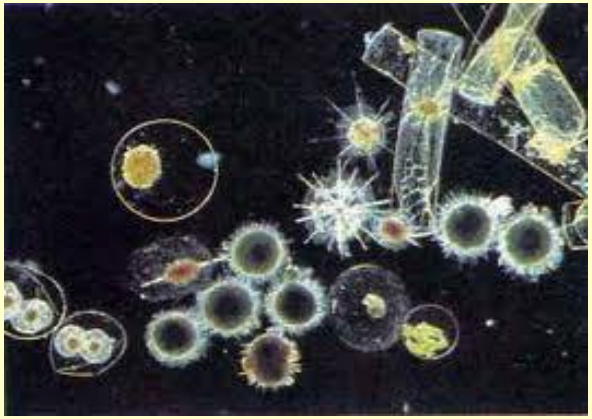
11. Apply the statistical test, and interpret the results.

12. Communicate the outcome, with a graph(s) and table(s).

Step 1. Define Your Question



Is ocean primary production changing due to global warming?



Define the Population of Interest

Who do we want to learn about ?

Biological Population:

All organisms of the same species (interbreed) that live in the same geographical area (given time period).

Statistical Population:

A defined set of entities concerning which statistical inferences are to be drawn, often based on a sample taken from the biological population.

Potential set of all measurements or observations in a sample, including not only the cases actually observed but also those that are potentially observable.

Step 2. Scientific Hypotheses

NULL:

No change in ocean productivity over time (No Trend)

ALTERNATIVE:

Decrease in ocean productivity due to deepening thermocline and increased stratification.

Increase in ocean productivity due to strengthening atmospheric pressure gradients and enhanced upwelling.

Step 3. Statistical Hypotheses

NULL:
The slope of ocean productivity as a function of time (from 1900 to 2100) is equal to 0.

ALTERNATIVE:
The slope of ocean productivity as a function of time (from 1900 to 2100) is not equal to 0:

> 0 (increase)

< 0 (decrease)

Step 4. Identify Variables

Define Inference Space

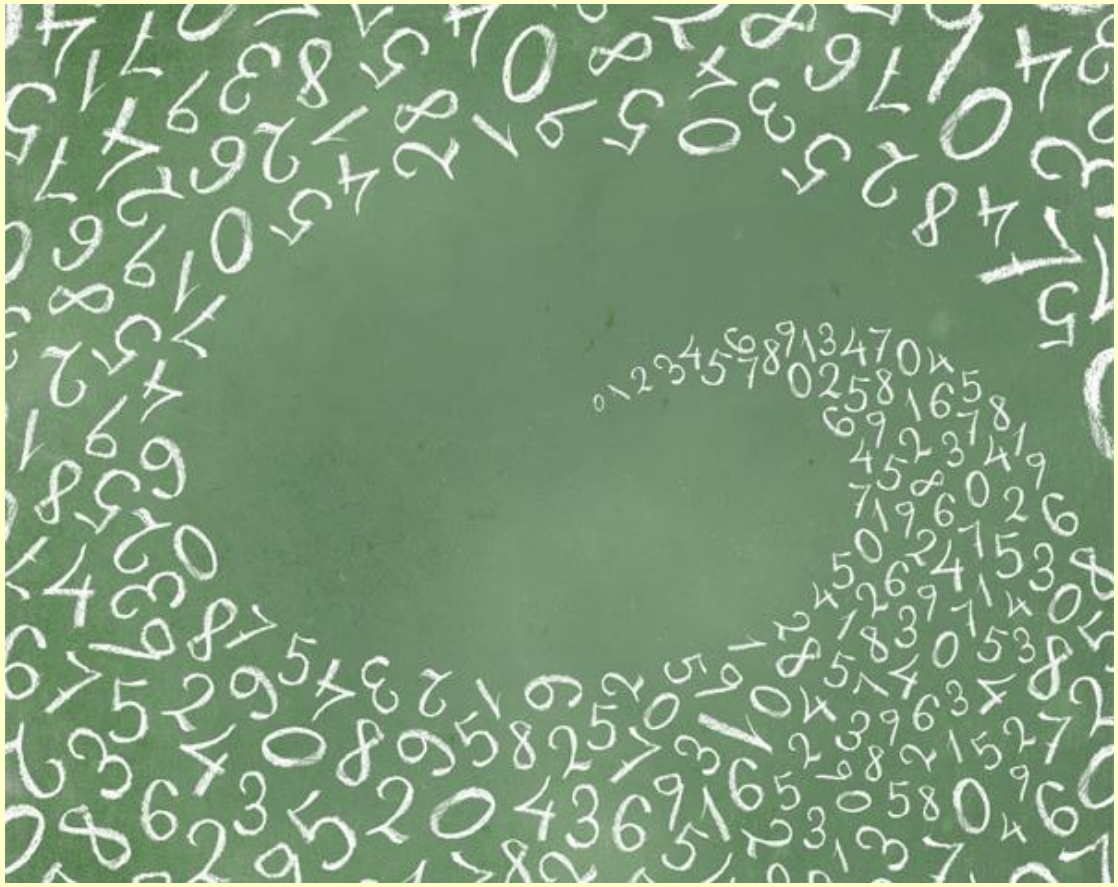
What are we measuring ?
Where ? How often ?

What time period are we considering ?

Design study with standardized samples

Units: Cells per meter 3
mg Chl per meter 2
g C per meter 2 per year

Variables



Definition: Anything that can be measured and (potentially) can differ across entities or over time

Variables and Hypothesis Testing

Testing hypotheses requires making predictions and taking measurements of different variables

Often, the goal is to measure the response of one variable to an experimental change in other variables

Independent	Dependent
Variable denotes the cause (the driver of the pattern)	Variable denotes the effect (responds to the driver)
Termed: predictor variable	Termed: outcome variable

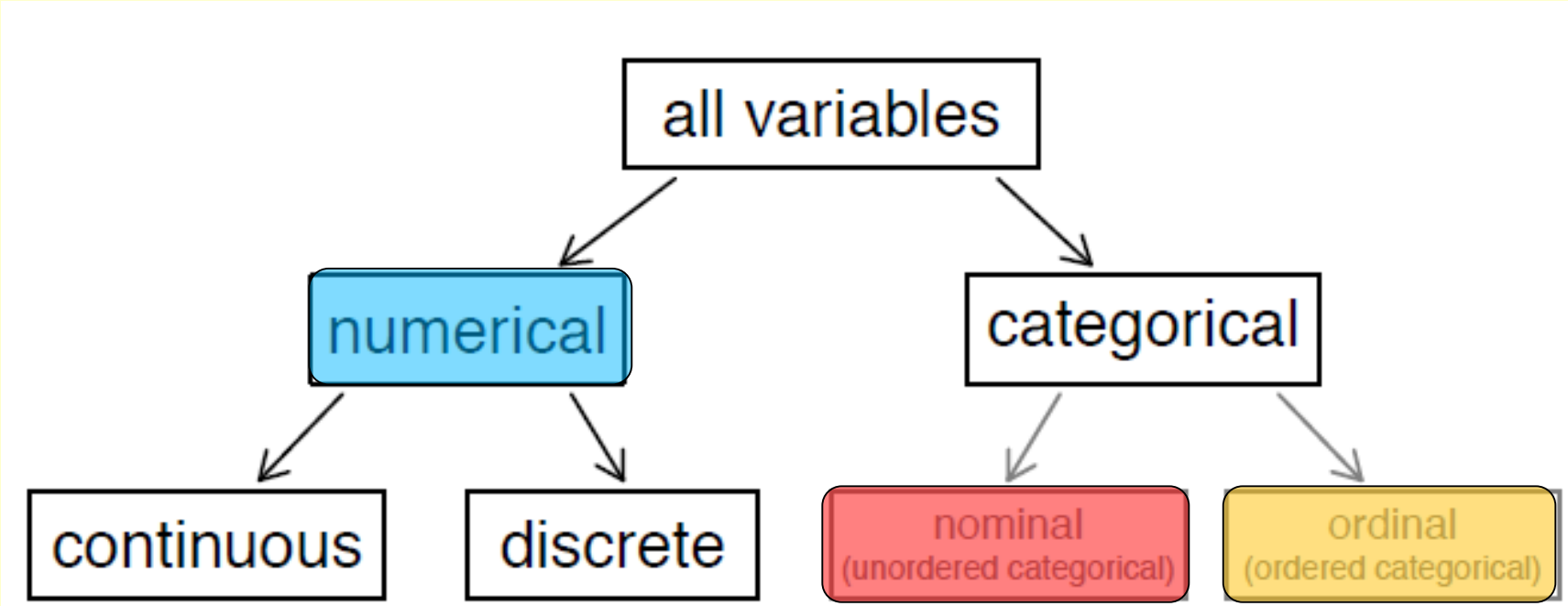
Examples:

Dependent: Primary Productivity Independent: Year (Time)
Dependent: Primary Productivity Independent: Habitat

Types of Variables

McDonald classifies variables into three types: **measurement**, **nominal**, and **ranked (ordinal)**.

You will see other names for these variable types and other ways of classify variables in other statistics references, so try not to get confused.



Numerical (Measurement) Variables

Properties:

Variable takes on numerical values (e.g., ants, arm lengths)

Can be ordered and ranked

Subclasses:

- Discrete: Few possible values (integers)
- Continuous: Measurements take any value within range

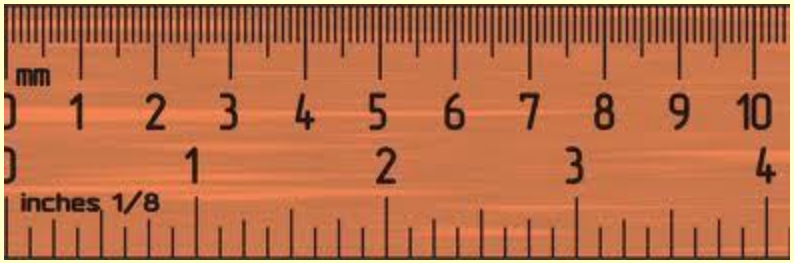
NOTE: Math principles underlying many statistical tests fo measurement variables assume they are continuous. Luckily, these statistical tests work well on discrete measurement variables, so you usually do not need to worry.

However, some tests only work with discrete variables.

Continuous Variables

Interval Variables:

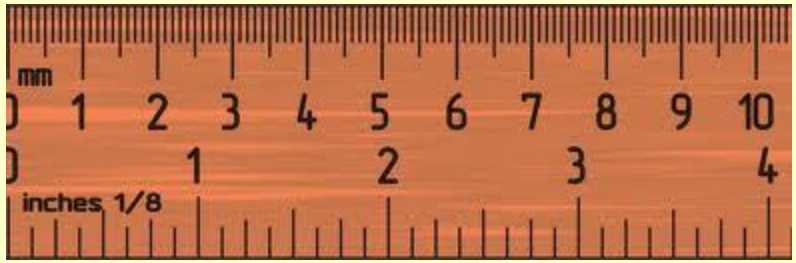
Differences in one unit of measurement equal along entire measurement scale



e.g., Temperature (deg. C)

Ratio Variables:

Differences in one unit of measurement equal along entire measurement scale



There is a real zero value

e.g., Salinity, Money

Types of Variables

Categorical Variables:

Properties:

Variable takes on different categories (e.g., eye color)

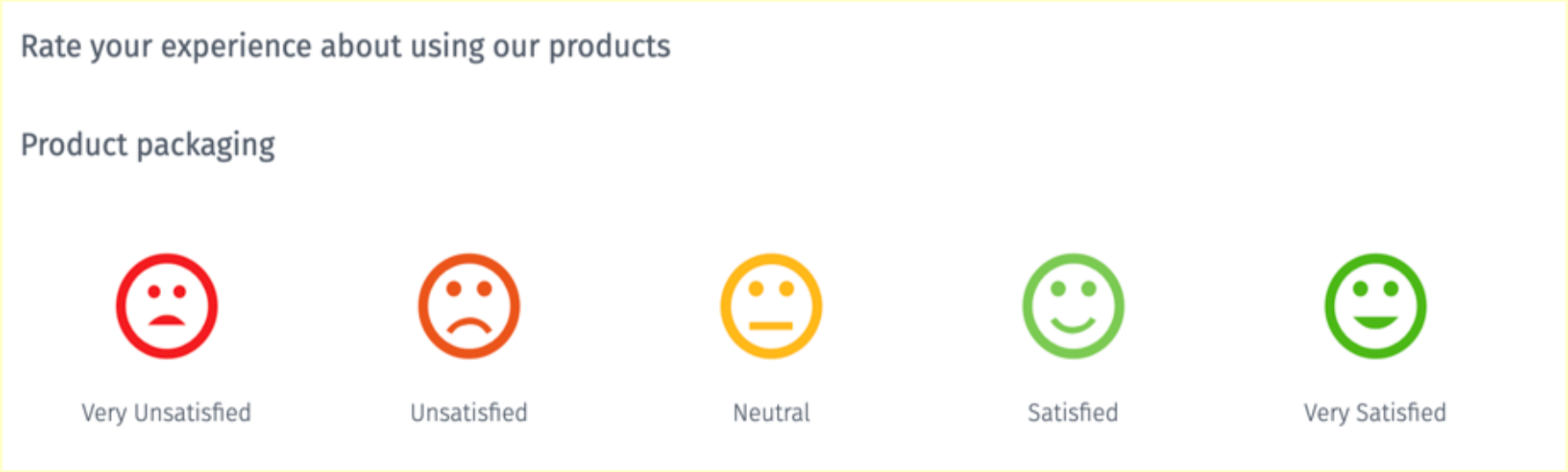
Only ordinal variables can be ordered and ranked

Subclasses:

- Ordinal: Ordered (first, second, third…)
- Nominal: Unordered

Likert Items

A statement that the respondents evaluate by giving a quantitative value on any subjective or objective dimension, with the level of agreement / disagreement being the dimension most commonly used.

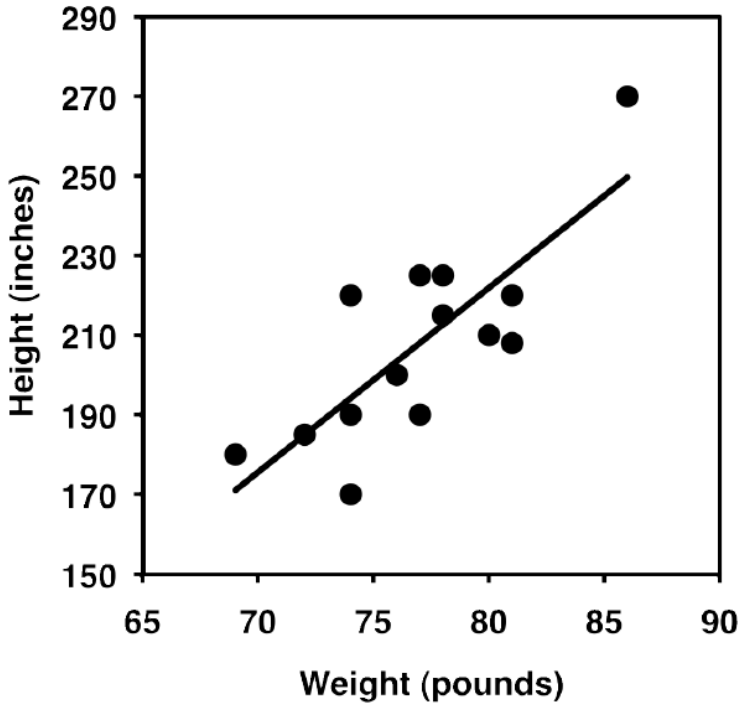


1	2	3	4	5
5	4	3	2	1
25	20	3	2	0

Categorizing Continuous Data

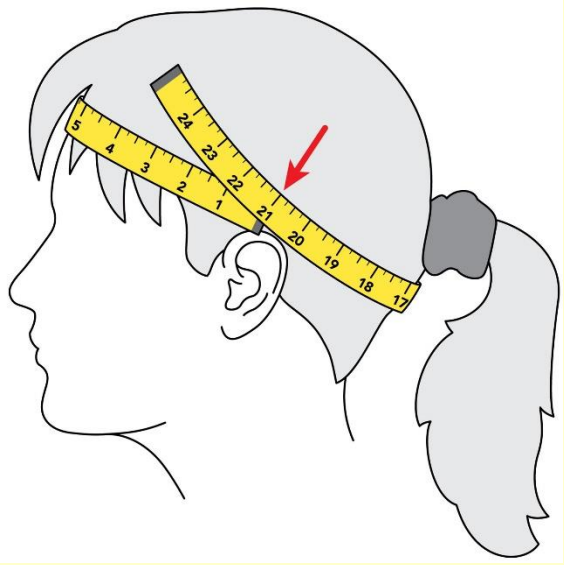
Possible to convert a measurement variable into a nominal variable, dividing the measurements up into two or more classes based on the range of the variable.

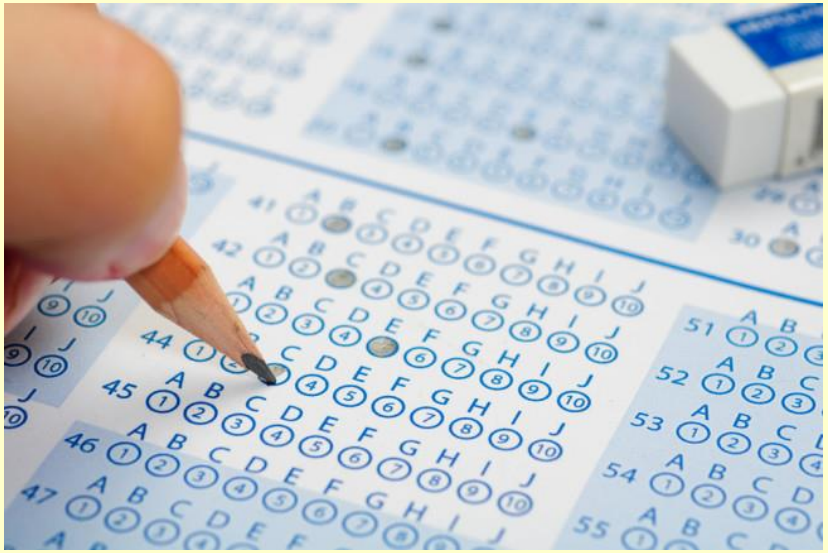
Height (inches)	Weight (pounds)
69	180
72	185
74	170
74	190
74	220
76	200
77	190
77	225
78	215
78	225
80	210
81	208
81	220
86	270



Error Terminology

Validity: Refers to whether an instrument measures what it was designed to measure. The extent to which the scores from a **measure** represent the variable they are intended to.





Error Terminology

Reliability is the ability of the measure to produce the same results under the same conditions.

Precision

Accuracy

Error Terminology

Measurement Error: Discrepancy between the value we use to represent what we are measuring and the actual value of what we are measuring (i.e., the real value).

Systematic Error

Unsystematic Error

Three Statistical Frameworks

Monte Carlo Approach:

Makes no assumptions about underlying data distributions. Uses randomizations of the observed data for inference. Calculates probability that the pattern found in the sample occurred "by chance" (due to sampling).

Bayesian Approach:

Makes assumptions about underlying data distributions. Uses the data to estimate parameters, augmented with additional "previous" knowledge. Assigns probabilities to these parameter estimates.

Fisherian Analysis: (A.K.A. Parametric Statistics)

Statistical Framework for this Class

Fisherian Analysis:

Makes assumptions about the underlying data distributions.

Uses only the data from one experiment / study to estimate parameters.

Calculates probability that the pattern found in the sample occurred "by chance" (sampling).

Sir Ronald Aylmer Fisher (1890 – 1962)

Analysis of variance
Fisher's exact test
F-distribution

Example

Tasked with comparing nest density of ground-foraging ant species in two adjacent habitats: agricultural field / forest.

Delineate Two Study Areas

Sample Randomly (lat / long)

Standardize Survey Effort (1 m x 1 m)

Reminder – Important Statistical Terms

Biological Population:

All ant colonies in the agricultural field and in the adjacent forest we are studying.

Statistical Population:

The ant colonies that we actually sampled using 10 randomly-laid quadrats: 4 in the agricultural field and 6 in the adjacent forest.

The potential set of all measurements or observations, including not only the cases actually observed (in our sample) but also those that are potentially observable.

Data Collection

Field#1: 6 nests / m 2

Field#2: 10 nests / m 2

Field#3: 12 nests / m 2

Field#4: 12 nests / m 2

Forest#1: 6 nests / m 2

Forest#2: 9 nests / m 2

Forest#3: 4 nests / m 2

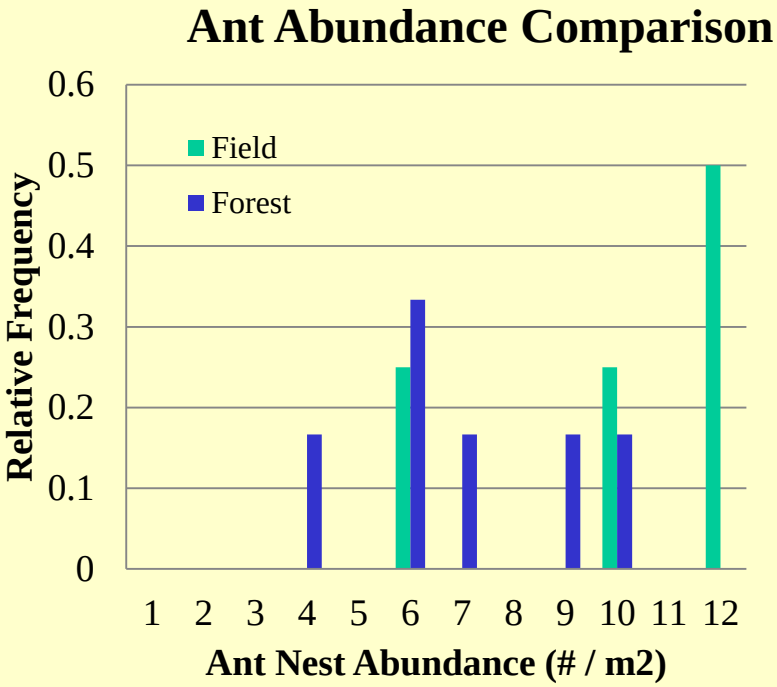
Forest#4: 6 nests / m 2

Forest#5: 7 nests / m 2

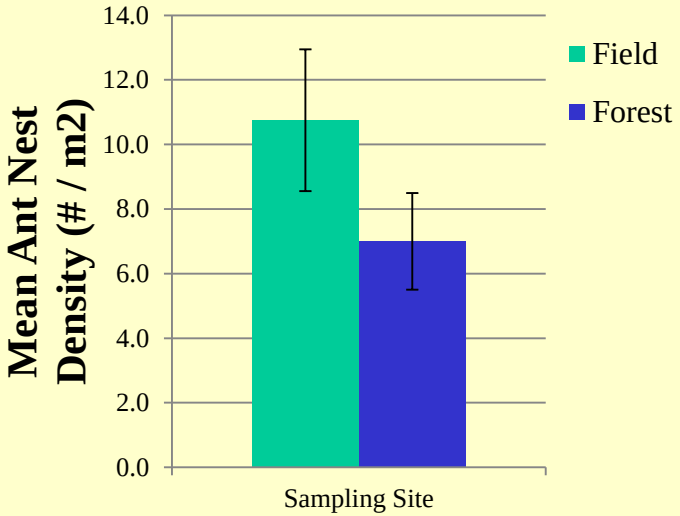
Forest#6: 10 nests / m 2

Data Summarization

What is the simplest way to summarize these observations ?



RANGE	6 - 12	4 - 10
MEAN	10.2	7.0
ST. DEV.	2.8	2.2

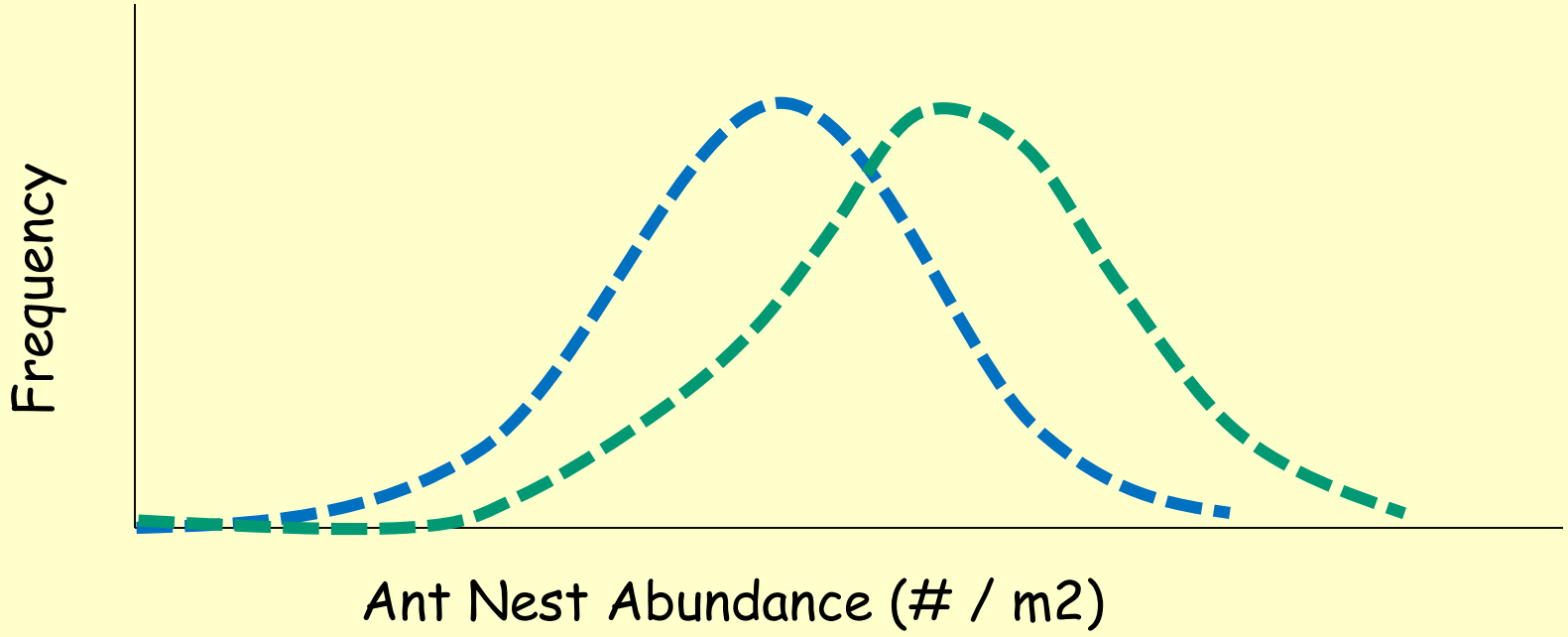


Frequency distribution of ground foraging ant nests at two adjacent sampling sites: an agricultural field and a forest.

Using Parametric Statistics

Based on the assumption that these data were sampled from a specific distribution – a Normal Distribution.

Assume that the parameters describing the frequency distribution of the sample (mean = $\overline{x}$; s.d. = s) are unbiased estimators of population parameters (mean = μ ; s.d. = σ).

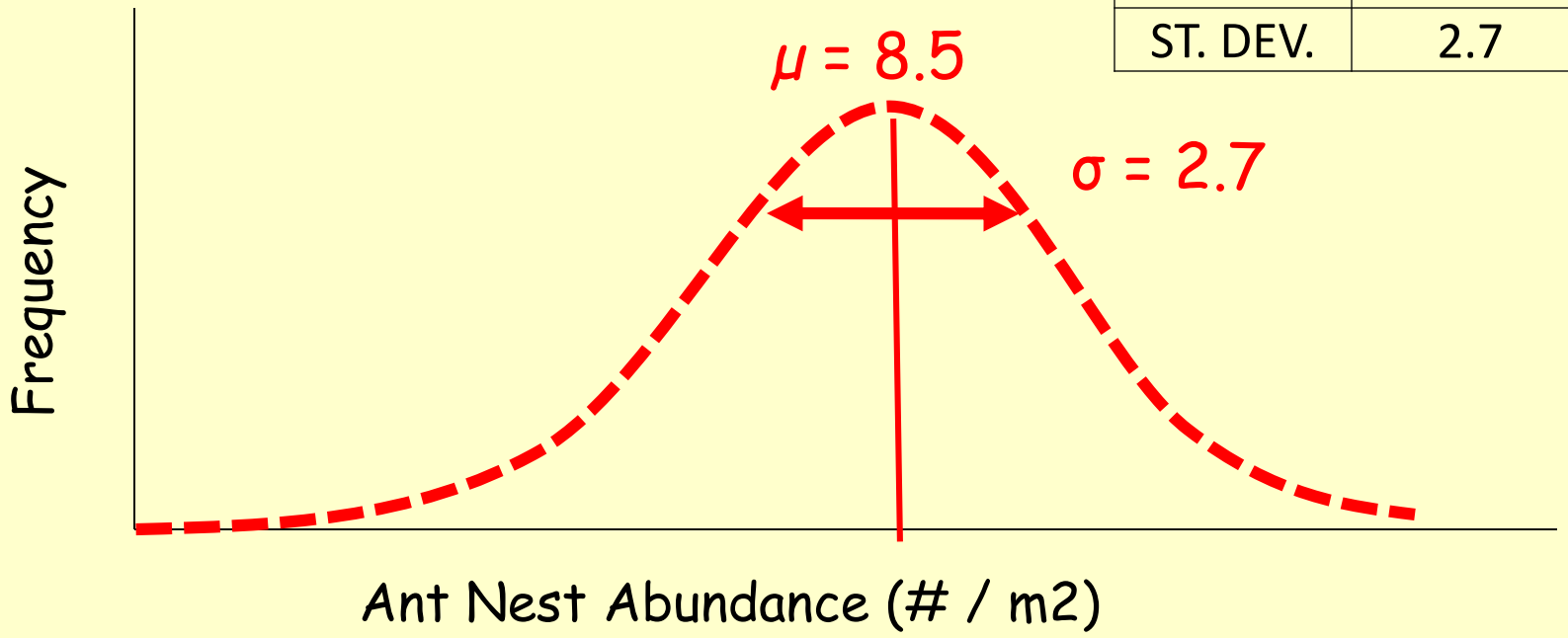


Using Parametric Statistics

Specifying the Null Hypothesis

Null Hypothesis: There is no difference… No Pattern.
(In other words: both populations are really the same)

RANGE	4 - 12
MEAN	8.5
ST. DEV.	2.7

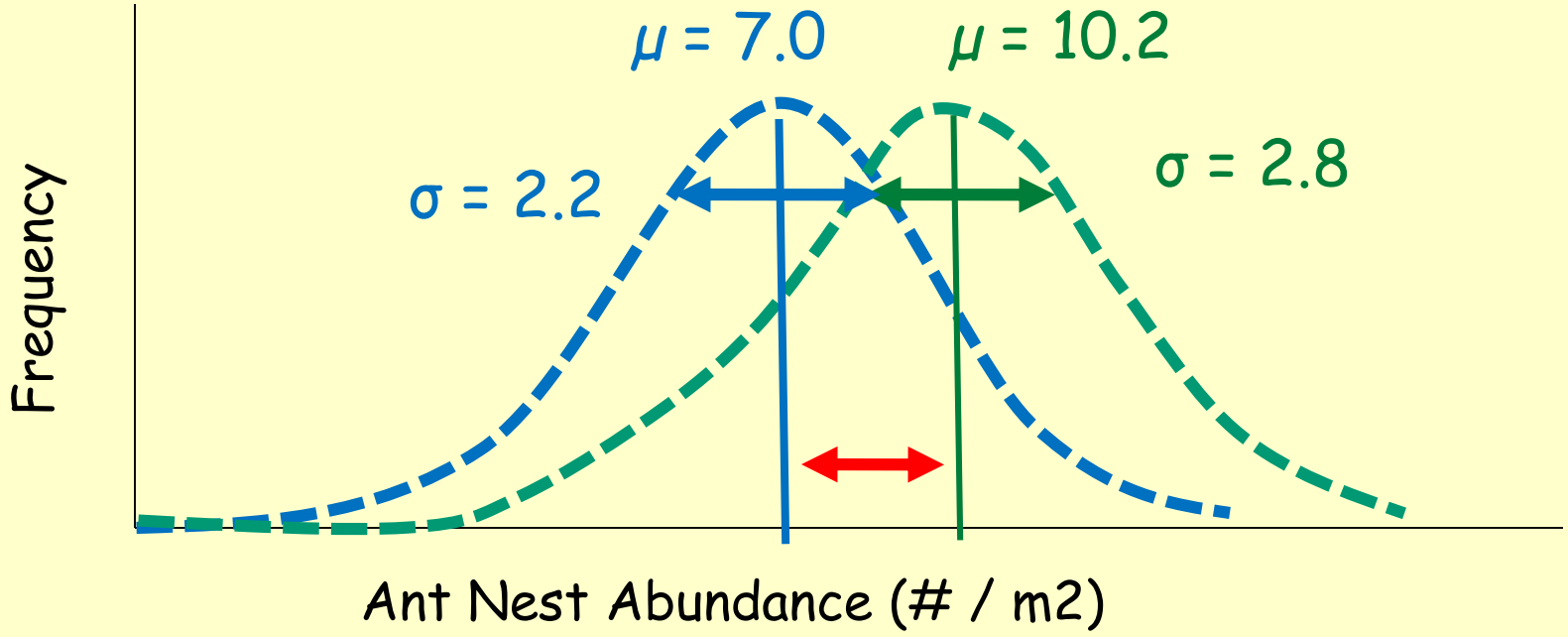


Using Parametric Statistics

Once we have made the assumption of normality, we can proceed with the statistical comparison of the two sites:

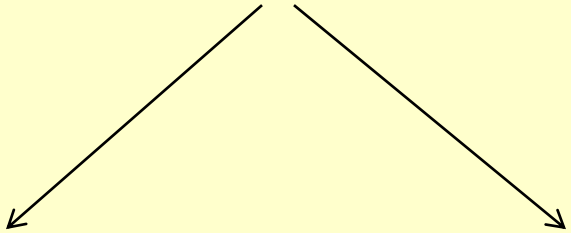
Specifying the test statistic

- Compare means (central tendency)
- Relate to their S.D.s (variability)



Overview

Fisherian Approach used in this class



Parametric statistics
(make assumptions about probability distributions)

Nonparametric statistics
(do not make assumptions, compare ranked data)

Ant Nest Example: How can we improve this study?

- Sampling Effort: More samples to describe distributions

- Inference Space: Clearly consider inference space:

To whom do our results apply?
(those sites, all sites in O'ahu, …)

Your Tasks for This Week

Read textbook chapters and article pdfs

Download software (R and R Studio)

Install and run software (R and R Studio)

PC Installation Instructions

Download R from http://cran.us.r-project.org/ (click on "Download R for Windows" > "base" > "Download R 2.x.x for Windows")

Install R.

Leave all default settings in the installation options.

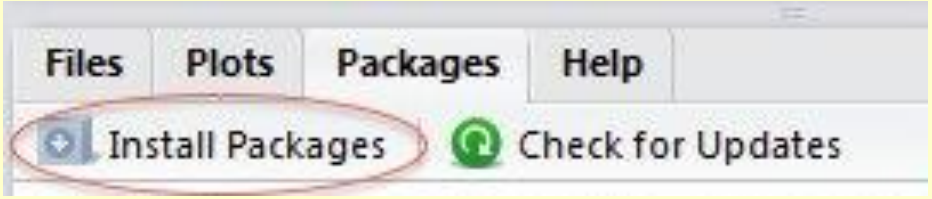
Download RStudio from: http://rstudio.org/download/desktop and install it.

Leave all default settings in the installation options.

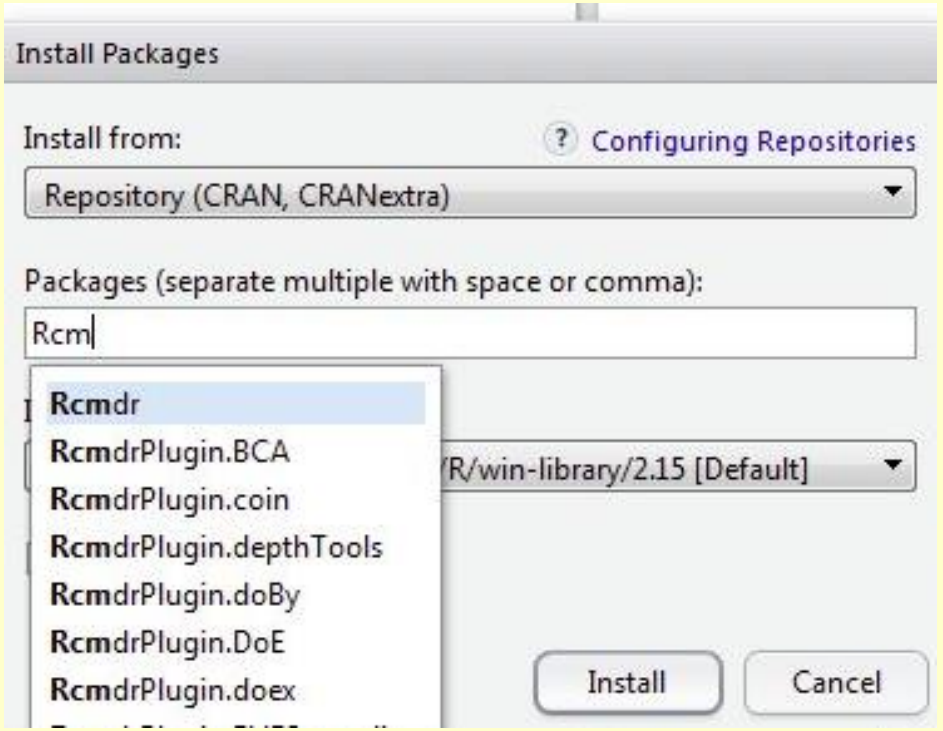
Open RStudio.

PC Installation Instructions

Go to "Packages" tab and click on "Install Packages".



The first time you do this you will be prompted to choose a CRAN mirror. Choose the location closest to you ("USA CA 1" or "USA CA 2", which are housed at UC Berkeley and UCLA, respectively).



R will download all necessary files from the server you select.

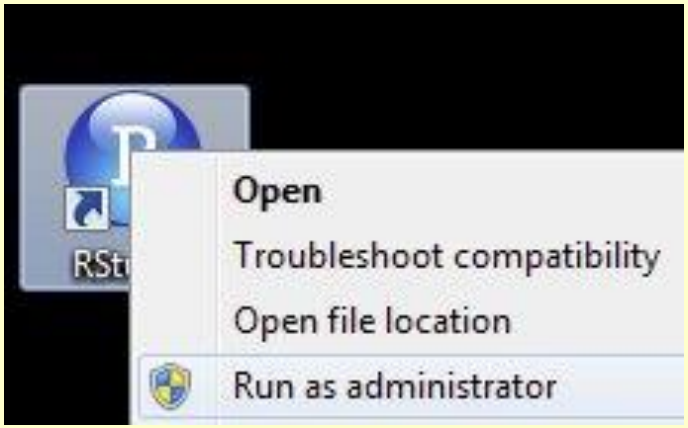
PC Installation Instructions

Typing "Rcmdr" and the package will appear in the list.

Ensure that "Install dependencies" is checked – to install all the needed supporting packages - , and click "Install".

Wait while all the parts of the R Commander package are installed.

NOTE: If you get permission errors while installing packages, close R Studio and reopen it with administrator privileges.



Once you've installed R Commander, you won't have to go through all those steps again !

Mac Installation Instructions

Download R from http://cran.us.r-project.org/
(click on "Download R for MacOS X" > "R-2.x.x.pkg (select latest version)") and Install R.

Download RStudio from:
http://rstudio.org/download/desktop.

Install RStudio by dragging the application icon to your Applications folder.

Download Tcl/Tk from:
http://cran.r-project.org/bin/macosx/tools/
(click on tcltk-8.x.x-x11.dmg; OS X needs this to run R Commander.)

Install Tcl/Tk.

Mac Installation Instructions

Go to your Applications folder and find a folder named Utilities.

Verify you have a program named "X11".

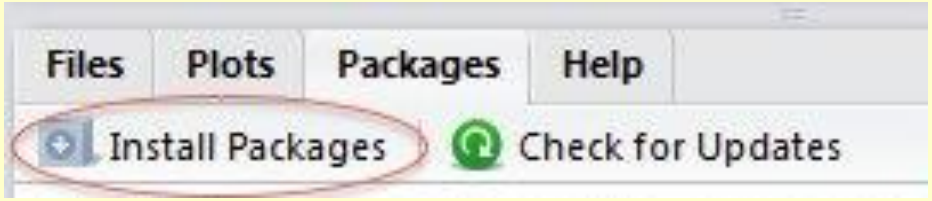
If you do not, go to:
http://xquartz.macosforge.org/

and download and install the latest version of XQuartz.

Mac Installation Instructions

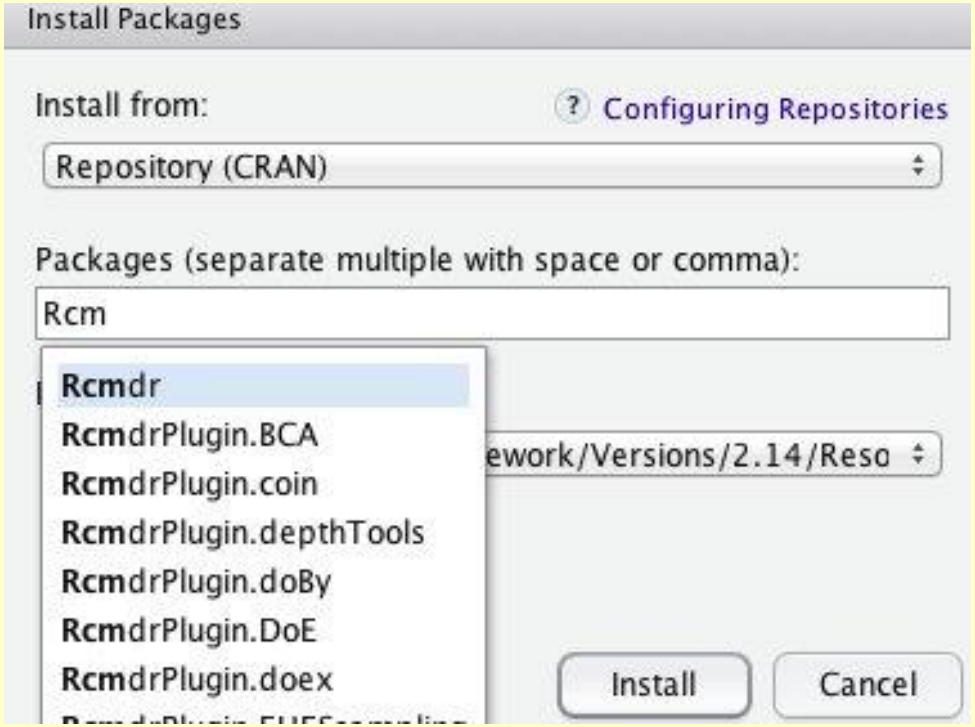
Open RStudio.

Go to "Packages" tab and click on "Install Packages".



The first time you do this you will be prompted to choose a CRAN mirror. R will download all files from the server you select.

Choose closest location ("USA CA 1" or "USA CA 2", housed at UC Berkeley and UCLA).



Mac Installation Instructions

Start typing "Rcmdr" until you see it appear in a list. Select the first option (or finish typing Rcmdr), ensure that "Install dependencies" is checked, and click "Install".

Wait while all the parts of the R Commander package are installed.

Open R Commander in Windows and OS X

NOTE: Once you've installed R Commander, you won't have to go through all those steps again !