Harvard School of Public Health
PH525.1x: Statistics and R for the Life Sciences

Instructor
Rafa Irizarry, PhD
Professor of Biostatistics
Department of Biostatistics
Harvard School of Public Health

Teaching Fellow
Michael Love, PhD
Postdoctoral Fellow
Department of Biostatistics
Harvard School of Public Health

Suggested pre-requisites

- **PH207x:** Health in Numbers: Quantitative Methods in Clinical and Public Health Research. This is another HarvardX course, which we recommend, but it is not a strict pre-requisite.
- **Basic programming skills.** We will assume that learners are familiar with very basic programming concepts (variables, functions).
- **Familiarity with the R language**. The course will use R in order to demonstrate data analyses. In the first week, we will have a refresher on the commands in R which you will need to use in the following weeks, but this is not a comprehensive R course, and we will not go in depth on R syntax. Please see below for online R resources.

Related Resources

Online R resources:

- R reference card (PDF) by Tom Short (more can be found under Short Documents and Reference Cards here)
- Quick-R: Quick online reference for data input, basic statistics, and plots
- Thomas Girke's R & Bioconductor manuals
- R programming class on Coursera, taught by Roger Peng, Jeff Leek, and Brian Caffo
- The free "try R" class from Code School is also a good place to start: http://tryr.codeschool.com/
- swirl: learn R interactively from within the R console

R Books:

- Software for Data Analysis: Programming with R (Statistics and Computing) by John M. Chambers (Springer)
- S Programming (Statistics and Computing) Brian D. Ripley and William N. Venables (Springer)

- Programming with Data: A Guide to the S Language by John M. Chambers (Springer)

Course Description

We will learn the basics of statistical inference in order to understand and compute p-values and confidence intervals. We will provide examples by programming in R in a way that will help make the connection between concepts and implementation. Problem sets requiring R programming will be used to test understanding and ability to implement basic data analyses. We will use visualization techniques to explore new data sets and determine the most appropriate approach. We will describe robust statistical techniques as alternatives when data do not fit assumptions required by the standard approaches. We will also introduce the basics of using R scripts to conduct reproducible research.

More details on the content for each week is provided below.

Graded Assignments

You will be assessed by homeworks within each subsection (each week is divided into several subsections). You will be given 5 attempts per question to enter a numeric value, which will be compared to the correct value. Upon entering the correct vaule, or after the 5th attempt, the code we used to produce the correct answer will be displayed. Each subsection counts as one homework, which counts equally toward the final score. So if there are 10 subsections in the course and you get 6/7 points within a single subsection, this adds 1/10 * 6/7 to your final score. You must score overall at or above 70% in order to pass the course and earn either the Honor Code or Verified Certificate of Achievement.

Enrollment Options

You may choose to audit this course or earn either a Verified Certificate of Achievement or the edX Honor Code Certificate of Achievement. The deadline for signing up for Verified is 4/27/2015. You can change your enrollment option on the course site, or get a refund (if applicable) before April 27, 2015 by contacting billing@edx.org and providing the email address used to register for the course.

- Audit: By auditing this course, you will have access to videos, labs, assessments and the discussion board and can participate as much, or as little, as you like. There is no penalty for registering for PH525.1x and not completing the assessments.
- Honor Code: An Honor Code Certificate of Achievement certifies that you have successfully completed a course, but does not verify your identity. Honor Code Certificates are currently free.
- Verified: A Verified Certificate of Achievement shows that you have successfully completed an edX course and verifies your identity through your photo and ID.

Questions

All questions should be made on the discussion board. You are encouraged to discuss homework problems on the discussion board, although try to avoid providing the exact code needed to produce the correct answer. We might delete posts which give the exact code, ready for copy-

paste into R, which produces the correct answer. Given the high enrollment in the class, emails sent directly to the course teaching staff will not be answered.

Course Schedule By Week

Course content will be discussed on a weekly basis with the following schedule:

Week 1 : Introduction : January 19, 2015

- Using Rstudio
- R programming skills
- Getting organized

Week 2 : Probability Distributions : January 26, 2015

- Introduction to random variables
- Introduction to the null distribution
- Probability distributions
- The normal distribution

Week 3 : Inference : February 2, 2015

- t-tests
- The Central Limit Theorem
- Association tests
- Monte Carlo methods
- Permutation tests
- Power

Week 4 : Exploratory Data Analysis and Robust Summaries : February 9, 2015

- Exploratory data analysis
  - histogram
  - QQ-plot
  - boxplot
  - scatterplot
  - log transformation
- Robust summaries
  - Median, MAD and Spearman correlation
  - Mann-Whitney-Wilcoxon test

All assignments must be completed before May 23, 2015 at 5:00 UTC.

Research Disclaimer

This course is hosting a study on participation and learning. There will be treatment groups and a control group. We will also send a short, optional survey to get your feedback. Your information would be completely confidential. Your being included in the experiment would benefit future students through insights into encouraging learning. If you would rather not be included in this study, you'll have the chance to opt-out through the course update emails. Otherwise, we greatly appreciate your willingness to improve online education by being included.

Please read the edX **Privacy Policy** for more information regarding the processing, transmission and use of data collected through the edX platform.