

SUMMER STATISTICS BOOTCAMP

Summer 2020

Instructors:	Professors J. Sun, S. Bruce, P. Bagchi	Time:	W 7:00PM - 9:00PM
Emails:	jsun21@gmu.edu , sbruce7@gmu.edu , pbagchi@gmu.edu	Place:	Online (Zoom)
Dates:	July 15th - August 12th	Office Hours:	TBD and by appt.

Textbook: *R for Everyone: Advanced Analytics and Graphics* by Jared P. Lander, 2nd Edition, Addison Wesley Data and Analytics Series, 2017. [Book Website](#).

Software: R will be used to perform data analysis. This training will be taught using RStudio accessible via free download. For details on accessing and using R and RStudio, see the first two chapters of the text.

Communication: E-mail will be the primary mode of communication and will be used for all training communication (e.g. send Zoom meeting links, provide announcements and updates, etc.).

Training Material: All training materials (e.g. slides, handouts, recordings, etc.) will be posted to a shared folder. Access will be provided only to individuals participating in this training, and participants are prohibited from sharing training materials with others who are not participating in the current training session. Ownership and dissemination of training materials created by instructors shall be in accordance with George Mason University (GMU) [Policy 4002](#) and the Data Use Agreement between Inova and GMU.

Training Description: This training will provide participants with a hands-on introduction to basic study design considerations and basic statistical analyses using R and RStudio. This training is designed to improve participants' abilities to 1) recognize challenges associated with study design and statistical analysis, 2) work with statisticians more effectively to address these challenges, and 3) produce basic preliminary data visualizations and analyses using R and RStudio. During training sessions, participants will analyze sample data sets using sample R scripts provided. Outside of the weekly training sessions, each participant will analyze a given (or his/her own) data set using the techniques covered in the training by adapting the sample scripts provided. Office hours will be available to assist each participant in carrying out the analysis of his or her own data set using R and RStudio.

Topics Covered:

- Challenges and considerations in study design, sampling, and sources of bias in scientific studies.
- Prepare and read data into RStudio in order to conduct statistical analyses using R.
- Produce descriptive statistics in R for variables of different types (e.g. means for continuous variables, frequency tables for categorical variables, etc.).
- Produce basic data visualizations using R for exploratory data analysis (e.g. box plots, histograms, bar charts, line plots, etc.).
- Conduct hypothesis tests for various basic analyses (e.g. one sample vs. two sample vs. paired, parametric vs. nonparametric).
- Fit and interpret basic generalized linear regression and survival analysis models (e.g. linear regression, logistic regression, Poisson regression, Cox proportional hazards).

Tentative Training Schedule:

Week	Instructor	Topic	Reading
1 (7/15)	J. Sun	Introduction to study design, sampling, and bias	
2 (7/22)	S. Bruce	Introduction to R and RStudio <ul style="list-style-type: none"> • Syntax • Functions • Reading in data • Descriptive statistics and data visualizations 	Ch 1-7
3 (7/29)	P. Bagchi	Inferential statistics in R <ul style="list-style-type: none"> • Hypothesis testing • Confidence intervals • Sample size determinations 	Ch 18
4 (8/5)	P. Bagchi	Linear models in R <ul style="list-style-type: none"> • Simple/multiple linear regression • Logistic regression • Poisson regression • Variable selection 	Ch 19, 20.1-20.3, 21
5 (8/12)	S. Bruce	Survival analysis in R <ul style="list-style-type: none"> • Kaplan-Meier curves • Cox proportional hazards regression 	Ch 20.4