STAT 210
Applied Statistics and Data Analysis

Joaquín Ortega
joaquin.ortegasanchez@kaust.edu.sa
Room 4297, Building 1

King Abdullah University of Science and Technology
Fall Term 2022

Outline

Instructors

Joaquin Ortega
joaquin.ortegasanchez@kaust.edu.sa

Teaching Assistants
TBA

Outline

This course provides fundamentals of probability and statistics for data analysis in research. Topics include

- exploratory data analysis,
- sampling distributions,
- estimation and confidence intervals,
- hypothesis tests,
- elementary simulation and bootstrapping,
- distribution-free techniques,
- linear regression,
- analysis of variance,
- two-way tables, and
- data analysis using statistical software.

Formal modeling of data and formal inference will be covered.

Goals and Objectives:

By the end of this course, the students are expected to have mastered the following:

- Statistical visualization methods
- Framework for statistical modeling of continuous and discrete-valued types of data
- Formal inferential procedures
- Various data analytic techniques

Required Knowledge

- Undergraduate mathematics including calculus and basic matrix algebra

Reference Texts

- **Michael J. Crawley** (2015). *Statistics, An Introduction Using R, Second Edition.*, John Wiley & Sons.
- **S. Weisberg (2013)**. *Applied Linear Regression.* Wiley, 4th edition

Additional Bibliography:

- **Michael J. Crawley** (2013). *The R Book, Second Edition.* John Wiley & Sons.
- **Peter Daalgard (2008)** *Introductory Statistics with R, Second Edition.* Springer.
- **John Maindonald, W. John Braun** (2010). *Data Analysis and Graphics Using R: an Example-Based Approach, Third Edition*, Cambridge University Press.
- **B. Shahbaba** (2012). *Biostatistics with R: An Introduction to Statistics Through Biological Data.* Springer

Survey

Please take a few minutes to answer a few survey questions that will give us information about you and your background so that we can plan the course to meet your needs.

You can find the survey using this link:

▸ Link

You can also use the link in the welcome message in Blackboard.

It is completely painless and should take less than 5 minutes.

Evaluation

- 30% Homework
- 20% First exam
- 20% Second exam
- 30% Course Project

Homework

- Homework will be posted on the BB page and should be submitted through the BB page.
- Late assignment submissions will not be accepted unless prior arrangements have been made (except in university established cases of illness or emergency).
- All homework assignments must be neatly typed (LaTeX or R Markdown are recommended). You should upload a pdf and the code, if it is not included in the pdf.
- All projects and homework assignments are required.

Homework

	Posted		Due	
HW1	Sun	Sep-04	Sun	Sep-11
HW2	Sun	Sep-11	Sun	Sep-18
HW3	Sun	Sep-18	Sun	Sep-25
HW4	Sun	Oct-02	Sun	Oct-09
HW5	Sun	Oct-09	Tue	Oct-18
HW6	Sun	Oct-23	Sun	Oct-30
HW7	Sun	Oct-30	Sun	Nov-06
HW8	Sun	Nov-06	Sun	Nov-13
HW9	Sun	Nov-13	Sun	Nov-20

You are allowed to collaborate on all homework problems according to the following rules:

- You must first attempt to solve each problem on your own.
- If you get stuck, you can then talk to any student currently enrolled in the class about the issue, as well as the instructor or a TA.
- However, solutions and R code should not be exchanged (i.e., you still must work through the details of the problem after you have gotten help, write the final answers alone, and understand them thoroughly).

You cannot share code!

Exams

- Exams will be done outside class hours.
- They will be open-book but not open-internet. You are not allowed to search for answers on the internet.
- Exams are strictly individual. You cannot communicate with anyone other than the instructor or the TAs during the exam.
- They will be based on R.

- Exam 1: Saturday, October 22, 9:00-12:00 am.
- Exam 2: Saturday, November 26, 9:00-12:00 am.

Project

- All students must do a project on a subject related to the course.
- Projects will be done in groups of 4/5 students.
- You can team up with students from other sections.
- There are a project guide and rubric on the Blackboard page.
- Try to find a problem that is challenging and important for you.

Important Dates

- Sep. 30: email with project topic and group.
- Oct. 20: proposal (5% of final grade).
- Nov. 26: final project report (15% of final grade).
- Dec. 4 - 8: presentation (10% of final grade).

The KAUST grading system is a 0 to 4 scale using letter grades, according to the following equivalence.

A	4.0
A-	3.67
B+	3.33
B	3.0
B-	2.67
C+	2.33
C	2.0
C-	1.67
D +	1.33
D	1.00
D-	.67
F	0

Link

KAUST
King Abdullah University of Science and Technology

Code of Conduct

Take some time to review Kaust's Code of Conduct in the Student Handbook.

▶ Link

KAUST
King Abdullah University of Science and Technology

Office Hours (JO).

Tuesday 3:00 - 4:00 pm

Wednesdays 3:00 - 4:00 pm

Other times by appointment

Teaching Assistants

TBA

TAs will take turns to cover the office hours.
The roster will appear in BB.

Tentative list of topics to be covered:

1. Introduction to data analysis. Introduction to R computing
2. Introduction to R computing. Data Summaries.
3. Graphs in R. Quantile Plots.
4. One sample problems. Estimation.
5. Hypothesis Tests. Comparing two populations.
6. Contingency tables. Proportions.
7. Comparing many populations (ANOVA).
8. Introduction to experimental design. Simple linear regression.
9. Simple Linear Regression. Model formulation. Inference. Diagnostics.
10. Multiple Linear Regression
11. Multiple Linear Regression
12. Nonparametric function estimation. Kernel smoothing
13. Binary regression models. Logistic regression.
14. Regression models for count data. Bootstrap for regression
15. Resampling Methods and the Bootstrap

Tutorials

- Tutorials will be held on Saturdays at 10 am.
- They will be online, using Zoom.
- Announcements and invitations on the BB page for the course.
- There will be tutorials as long as there is a reasonable level of attendance.

Sessions

- Saturday, Sep. 10, Probability refresher.

Communication

Outline

KAUST
King Abdullah University of Science and Technology

Description

- The course will be a face-to-face hybrid class. Some activities will be synchronous, and others asynchronous.
- It will be a 'flipped' course.
- Lectures have been divided into smaller topics and recorded. There will be no live lectures. All are pre-recorded for you to watch when it is convenient for you.
- You also have access to the presentations and the scripts with the R code used in the recorded lecture. For some topics, there are also class notes.
- You have an amount of work to cover each week on your own.
- The material is on the BB page for the course, organized by weeks.

Class Meetings

- This has been a large class in recent years
- At this moment we have 51 regular students, and 3 auditing.
- The course has been divided into three sections.
- The idea is to have balanced sections with approximately the same number of students

The three groups are

- Group A: Tuesdays, 9:30 - 11:00 am
- Group B: Mondays, 1:00 - 2:30 pm
- Group C: Sundays, 8:00 - 9:30 am

Class Meetings

- Groups will meet with me once a week, according to their schedules
- Group size should not exceed 22 students.
- These meetings will not be lectures.
- You are expected to watch the videos, review the material, run the scripts, and solve the problems on the folder named 'Week n' before meeting with me on week n (i.e., do this on week $n-1$).
- I will do a quick summary of the topics covered in the videos and answer any doubts you may have. You are expected to take an active part in these discussions.
- Sessions will be recorded and posted on BB.
- Attendance is required

Weekly Problem Lists

- There will be weekly lists of problems, usually 3-4 problems in each one.
- These problems are not the homework.
- All students must solve these problems before the weekly meetings.
- The solution to these problems will be the main topic for the meetings.
- Solutions will be presented by students selected at random the previous week.
- Failing to show up when you are due to present a problem without proper justification leads to losing 2 pts/100 in the final grade.
- For next week, I will present the solution to problems 1 and 2.
- I need two volunteers to do problems 3 and 4.

Every week you have to:

1. Watch the videos, run the scripts, review the material, understand the topic that is being covered.
2. Do the homework
3. Solve the problems in the weekly list
4. Attend the weekly session

For **next week** you have to:

1. Watch the following videos
   - V1 A quick tour
   - V2 Using R
   - V3 Basic Functions
   - V4 Objects and Data 1
   - V5 Objects and Data 2
   - V6 Objects and Data 3
   - V7 Some useful functions
2. Solve Problem List 1 (BB/Course Content/Week 2/Problem List 1)

NO PHONES IN CLASS

NO TABLETS OR COMPUTERS

Unless used for class-related work

Outline

Blackboard

Link

Outline

- R is an open-source software environment for statistics freely distributed by CRAN (Comprehensive R Archive Network) at the following address http: //cran.r-project.org/.
- CRAN features precompiled binaries as well as source code for R, add-on packages, documentation (including manuals, frequently asked questions, and the R newsletter) as well as general background information.
- You can also go to the home page for the R project, located at http://r-project.org, which includes a link to CRAN.
- The installation of R varies according to the operating system (Windows, Mac OS X, or Linux) but the functions are exactly the same, and most of the programs are portable from one system to another.
- Installing R is very simple; just follow the instructions.

- RStudio facilitates the use of R by integrating R help and documentation, providing a workspace browser and data viewer, and supporting syntax highlighting, code completion, and smart indentation.
- It integrates reproducible analysis with Sweave, knitr, and R Markdown, supports the creation of slide presentations, and includes a debugging environment.
- It facilitates the creation of dynamic web applications using Shiny.
- RStudio for MacOS, Windows, or Linux can be downloaded from http://www.rstudio.com/ide.
- RStudio requires R to be installed on the local machine. It is also useful to have LaTeX installed (required for certain documents in RMarkdown).
- Documentation of the advanced features in the system are available on the RStudio website.

S

- The *S* language was developed in the 70s and 80s at Bell Labs, by a team led by John M. Chambers.
- The aim of the language, as expressed by Chambers, is *'to turn ideas into software, quickly and faithfully'*.
- John M. Chambers was awarded the ACM Software System Award in 1998. *'The S system ... has forever altered how people analyze, visualize, and manipulate data'*

R

- R is a free implementation of the *S* language (a dialect).
- The first versions of `R` were written by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand.
- Current development is coordinated by the R Development Core Team, a group of international volunteers.
- Today `R` is widely used by the Statistics academic community, and there are thousands of contributed packages for all kinds of statistical procedures.

Some Useful Cheatsheets

- https://www.rstudio.com/resources/cheatsheets/
- https://ugoproto.github.io/ugo_r_doc/R_CS/
- https://www.datacamp.com/community/data-science-cheatsheets
- https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/colorPaletteCheatsheet.pdf

Also visit

- https://www.r-graph-gallery.com/

RStudio