STAT021 Statistical Methods II

Lecture 1 Overview

Lu Chen
Swarthmore College
9/4/2018

STAT021

What we should know already

- summary statistics, correlation, probability, (random) variables, distributions (Normal), Central Limit Theorem, statistical inference (significance test and confidence interval), two-sample t test, least squares regression, etc.

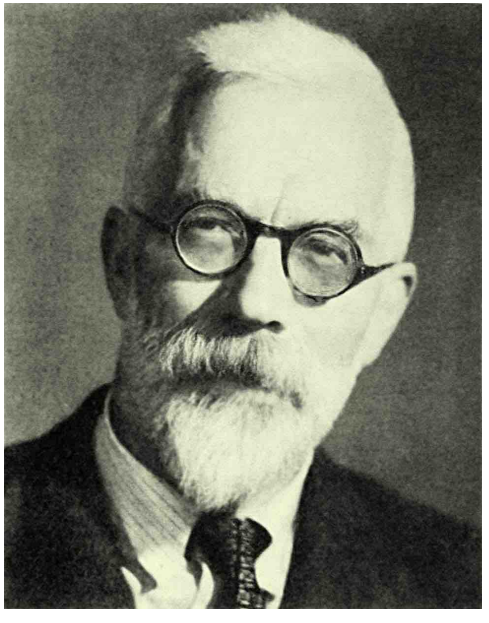
What we will learn

- **An**alysis **o**f **va**riance (**ANOVA**): compare several group means
- **S**imple **l**inear **r**egression (**SLR**): linear relationship between two quantitative factors
- **M**ultiple **l**inear **r**egression (**MLR**): linear relationship between a quantitative factor and some other factors
- **Logistic regression**: relationship between a binary factor and some other factor(s)

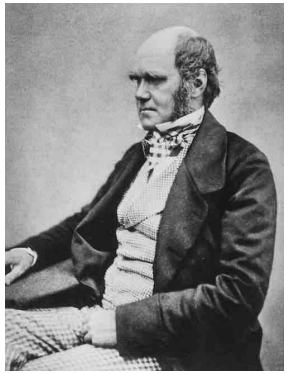
STAT021 is about ...

Life of Ronald Fisher

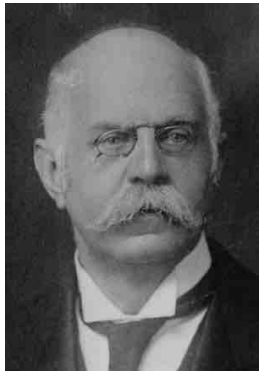


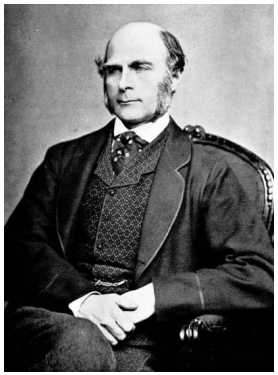
Life of Ronald Fisher



Charles Darwin

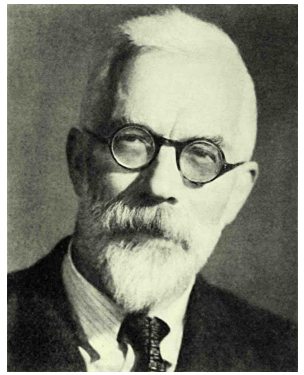


Leonard Darwin



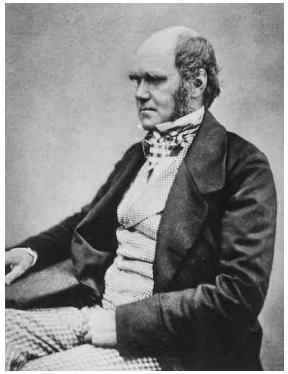
Francis Galton

Karl Pearson



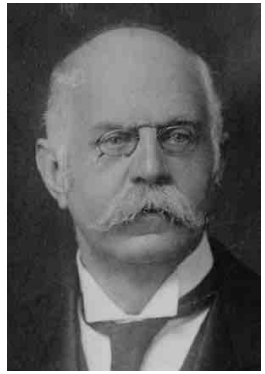
Ronald Fisher

Life of Ronald Fisher

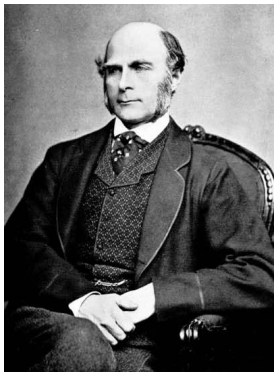


Charles Darwin

Father

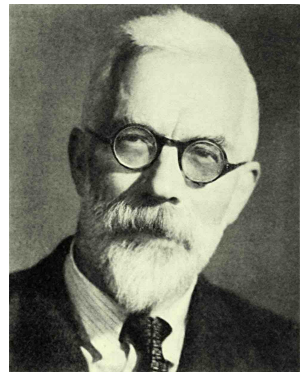


Leonard Darwin



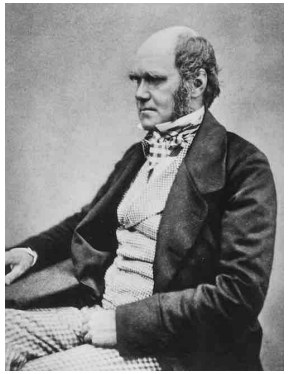
Francis Galton

Karl Pearson



Ronald Fisher

Life of Ronald Fisher

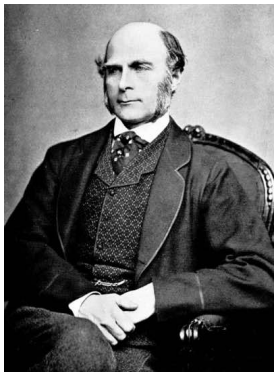


Charles Darwin

Father

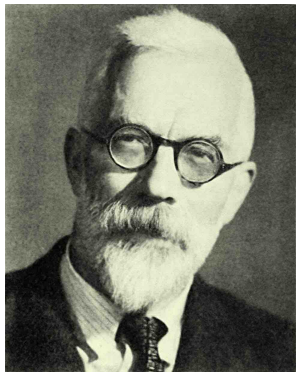
Leonard Darwin

Half-cousin



Francis Galton

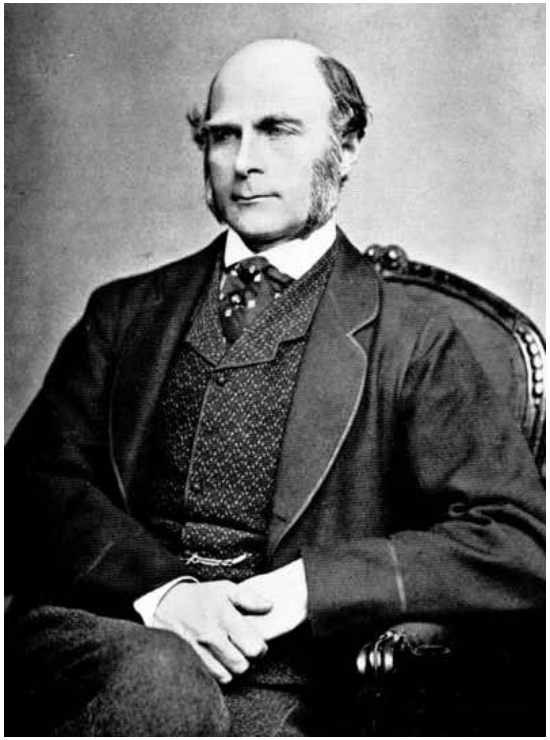
Karl Pearson



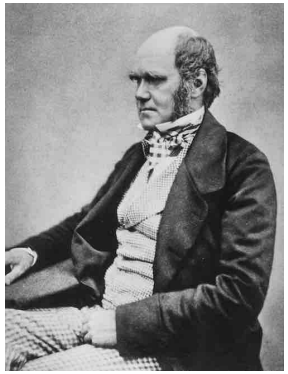
Ronald Fisher

Francis Galton

- 2/16/1822 - 1/17/1911
- Half-cousin of Charles Darwin
- Questionnaires and surveys
- Concept of variation - standard deviation
- Median
- Concept of correlation
- Regression to the mean



Life of Ronald Fisher

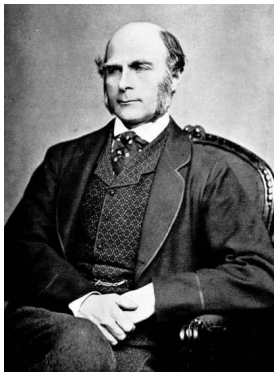


Charles Darwin

Father

Leonard Darwin

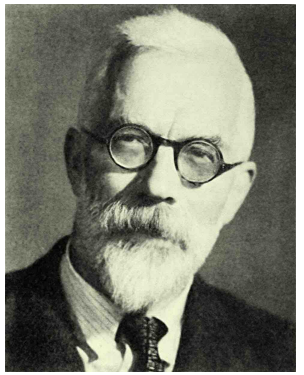
Half-cousin



Francis Galton

Mentor

Karl Pearson

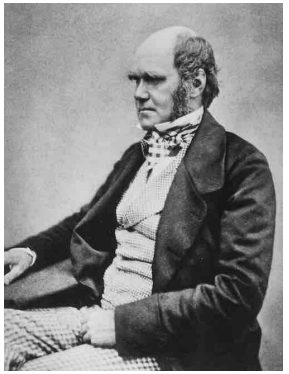


Ronald Fisher

Karl Pearson

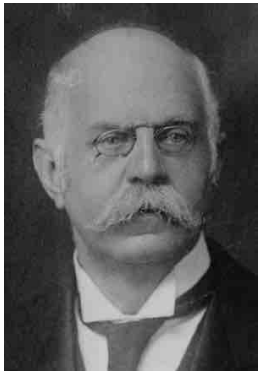
- 3/27/1857 - 4/27/1936
- Galton's protege
- Correlation coefficient
- and its relationship to linear regression
- *P*-value
- Foundation of significance test
- Chi-square test

Life of Ronald Fisher



Charles Darwin

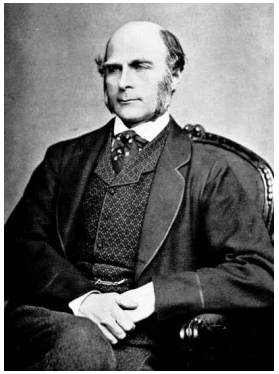
Father



Leonard Darwin

Half-cousin

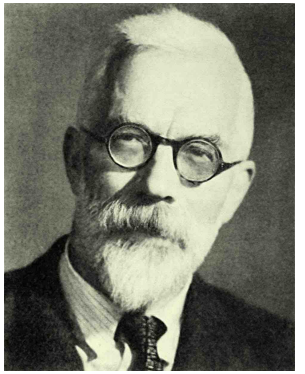
Mentor



Francis Galton

Mentor

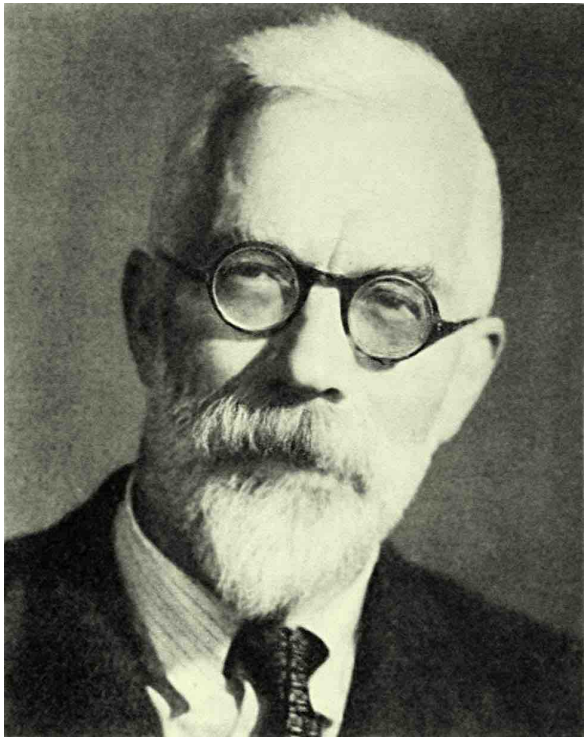
Karl Pearson



Ronald Fisher

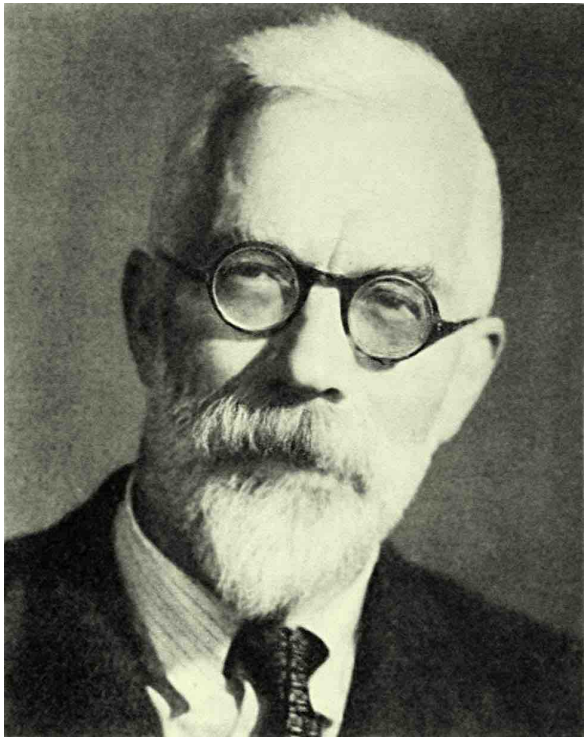
Ronald Fisher

> "a genius who almost single-handedly created the foundations for modern statistical science"
>
> – Anders Hald



Ronald Fisher

- 2/17/1890 - 7/29/1962
- Student of Leonard Darwin
- Design of experiments
- Variance
- Analysis of Variance (ANOVA)
- F-distribution
- Popularized P-value and proposed $\alpha = 0.05, z = 1.96$ as the cutoff
- Null hypothesis
- Popularized Student's t distribution
- Popularized maximum likelihood estimation (MLE)



Statistics is a great tool

Francis Galton

- Statistician, geneticist, polymath, sociologist, psychologist, anthropologist, geographer, inventor, meteorologist, etc.

Karl Pearson

- Mathematician, biostatistician, meteorologist, etc.

Ronald Fisher

- Biologist, statistician, founder of population genetics, etc.

Statisticians are curious about the world

You've Been Cutting Cake The Wrong Way Your Whole Life

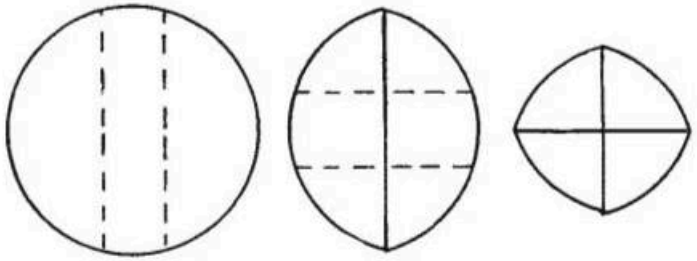
Good-bye, stale cake.

British author and broadcaster Alex Bellos has revealed how we've been wrong when cutting cake until now.

Statisticians are curious about the world

Cutting a Round Cake on Scientific Principles.

CHRISTMAS suggests cakes, and these the wish on my part to describe a method of cutting them that I have recently devised to my own amusement and satisfaction. The problem to be solved was, "given a round tea-cake of some 5 inches across, and two persons of moderate appetite to eat it, in what way should it be cut so as to leave a minimum of exposed surface to become dry?" The ordinary method of cutting out a wedge is very faulty in this respect. The results to be aimed at are so to cut the cake that the remaining portions shall fit together. Consequently the chords (or the arcs) of the circumferences



Broken straight lines show intended cuts. Ordinary straight lines show the cuts that have been made. The segments are kept in apposition by a common elastic band that encloses the whole. In the above figures about one-third of the area of the original disc is removed by each of the two successive operations.

of these portions must be equal. The direction of the first two vertical planes of section is unimportant; they may be parallel, as in the first figure, or they may enclose a wedge. The cuts shown on the figures represent those made with the intention of letting the cake last for three days, each successive operation having removed about one-third of the area of the original disc. A common india-rubber band embraces the whole and keeps its segments together.

F. G.

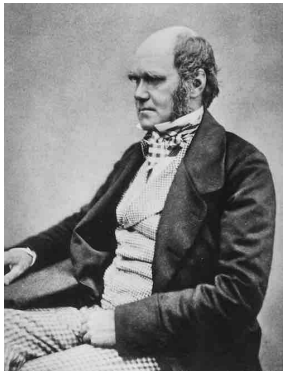
- *Nature*, December 20, 1906
- **Cutting a Round Cake on Scientific Principles**
- Question of interest: Given a round tea-cake of some 5 inches across, and two persons of moderate appetite to eat it, in what way should it be cut so as to leave a minimum of exposed surface to become dry?
- Author: F. G. - Francis Galton, 1822-1911

Statisticians are curious about the world

- Rothamsted Experimental Station, UK, 1919
- Muriel Bristol, phycologist, Fisher's colleague
- Milk tea: milk or tea first?
- 8 cups: 4 tea+milk, 4 milk+tea
- The "Lady tasting tea"
- Small sample problem: Fisher's exact test
- Null hypothesis: the lady cannot tell the difference
- Reject the null hypothesis only if the lady correctly categorize ALL 8 cups

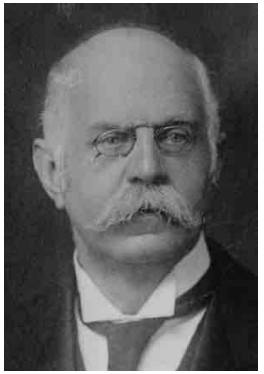
Statistics can be dangerous

EUGENICISTS



Charles Darwin

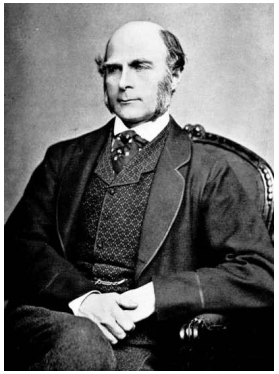
Father



Leonard Darwin

Half-cousin

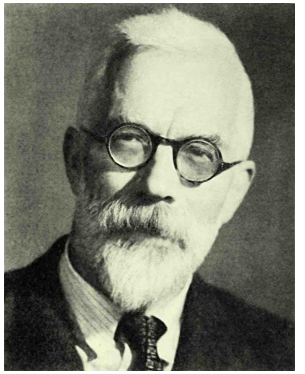
Mentor



Francis Galton

Mentor

Karl Pearson



Ronald Fisher

Statistics can be dangerous

- **Francis Galton** defined ***Eugenics*** in 1883: the study of all agencies under human control which can improve or impair the racial quality of future generations.
  - *Hereditary Genius*. Based on his biographical studies, Galton believed that desirable human qualities were hereditary traits, and unaffected by education or living conditions.
- **Karl Pearson** was a eugenicist who applied his social Darwinism to entire nations.
  - He saw war against "inferior races" as a logical implication of his scientific work on human measurement.
  - Regarding the Jewish immigration into Britain, Pearson alleged that these immigrants "will develop into a parasitic race. [...] Taken on the average, and regarding both sexes, this alien Jewish population is somewhat inferior physically and mentally to the native population".

Statistics can be dangerous

- **Fisher** publicly spoke out against the 1950 study showing that smoking tobacco causes lung cancer, arguing that correlation does not imply causation.
- His biographers Yates and Mather said that "the reason for his interest was undoubtedly his dislike and mistrust of puritanical tendencies of all kinds; and perhaps also the personal solace he had always found in tobacco."

FDA Center for Tabacco Products

The course

- STAT021 Statistical Methods II
- About me
  - Office: SC 139
  - Email: lchen6@swarthmore.edu
  - Tel: (610)690-5764
- Office hours
  - Tuesdays 2:40-4:10 pm
  - Thursdays after class by appointment

The course - Evaluation

- Participation (5%): in class, Moodle discussion forum; Attendance will be taken at three randomly picked classes.
- Homework (40%): assigned weekly on Thursday; due on the following Wednesday 11:55pm on Moodle. No late homework will be accepted.
  - Late homework will not be accepted except for ***documented*** illness, family emergencies or other excuses.
- Midterm exam (20%): closed-book. One two-sided cheat-sheet is allowed.
- Take-home data analysis project (15%).
- Final exam (20%): in class, closed book. One two-sided cheat-sheet is allowed.

The course - Schedule

Week	Lectures	Contents	Homework
1~2	Lecture 1~4	**Introduction**	HW 1~2
3~4	Lecture 5~8	**Analysis of Variance**	HW 3~4
5~8	Lecture 9~13	**Simple Linear Regression**	HW 5~6
7	**Fall break**	None	None
8	**Midterm**	Lecture 1~13	None
9~12	Lecture 14~21	**Multiple Linear Regression**	HW 7~10
13	**Project**	Lecture 14~21	None
14~15	Lecture 22~24	**Logistic Regression**	HW 11
	Final exam	Lecture 14~24	

The course - How to get help

1. Moodle discussion forums.
2. Stat clinics: SC 158; Sundays, Mondays and Wednesdays at 7:00-10:00 pm. For more information, visit website.
3. Office hours: SC 139; Tuesdays 2:40-4:10 pm, Thursdays after class by appointment.
4. Email: lchen6@swarthmore.edu; asking questions or scheduling meetings with me.
5. **If none of the above is helpful**, you may apply for a tutor.

Lecture slides will be uploaded to Moodle on Tuesday/Thursday afternoon.

R

https://cran.r-project.org.

Wikipedia:

> "R is an open source programming language and software environment for **statistical computing** and **graphics** that is supported by the R Foundation for Statistical Computing."

RStudio

https://www.rstudio.com.

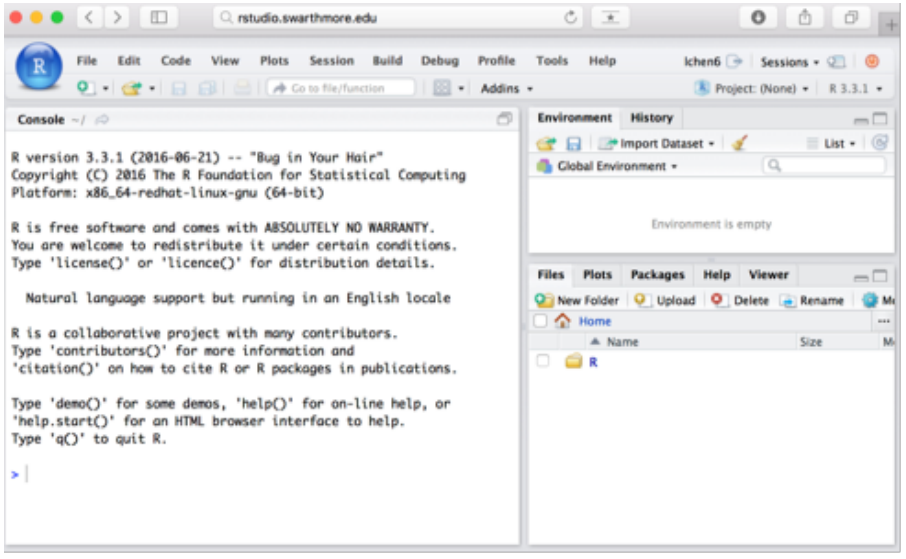
Wikipedia:

> "RStudio is a free and open-source **integrated development environment** (IDE) for **R**, a programming language for statistical computing and graphics."

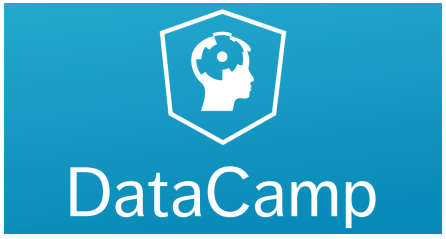
R is needed for intalling RStudio.

RStudio server

- Swarthmore holds an RStudio server at http://rstudio.swarthmore.edu.
- To access it, log in with your Swarthmore ID and password.
- RStudio server functions the same as RStudio desktop version but can be accessed anywhere internet is available.
- Using RStudio server is **strongly** encouraged for STAT021.



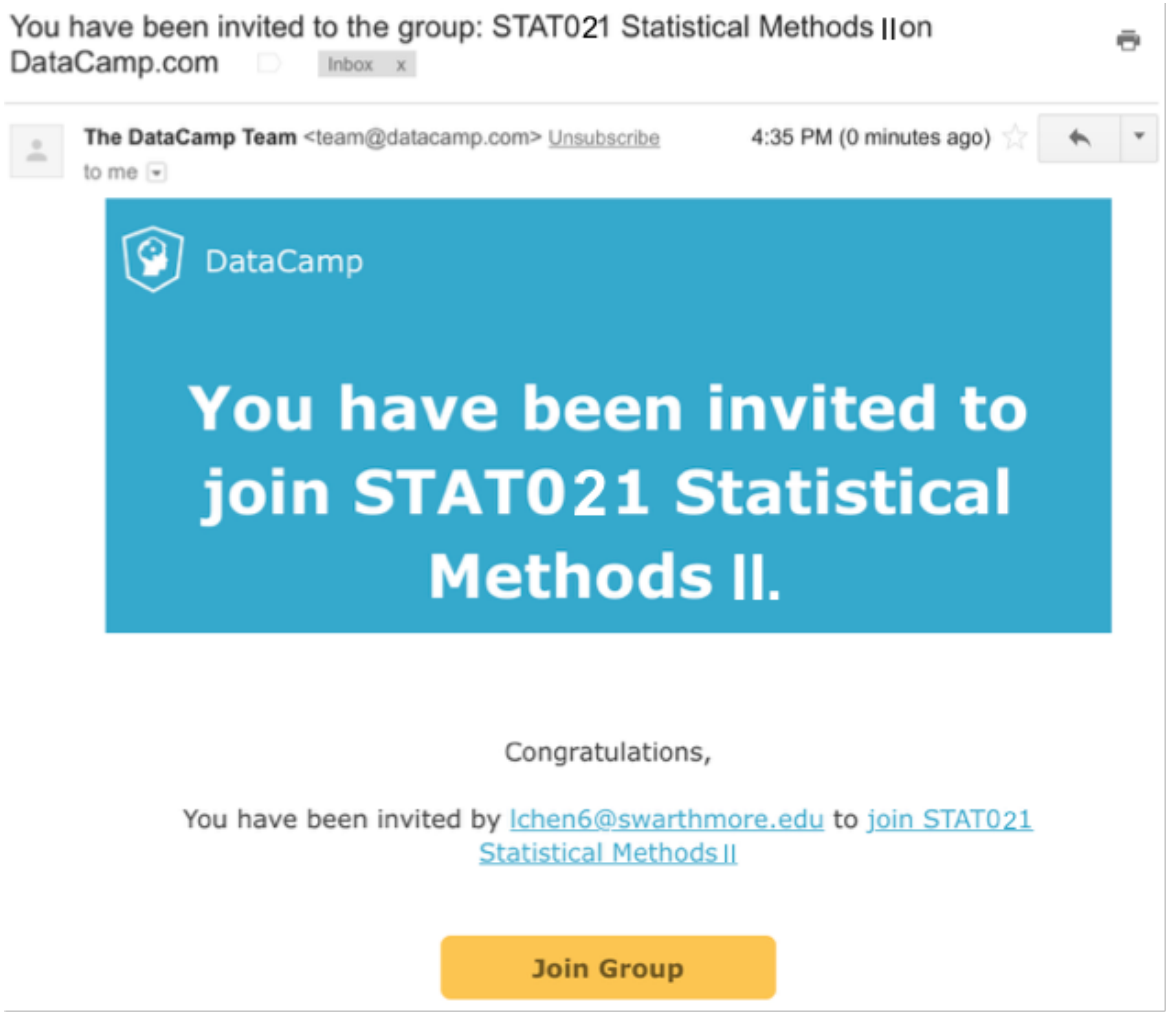
Learning R programming language



https://www.datacamp.com.

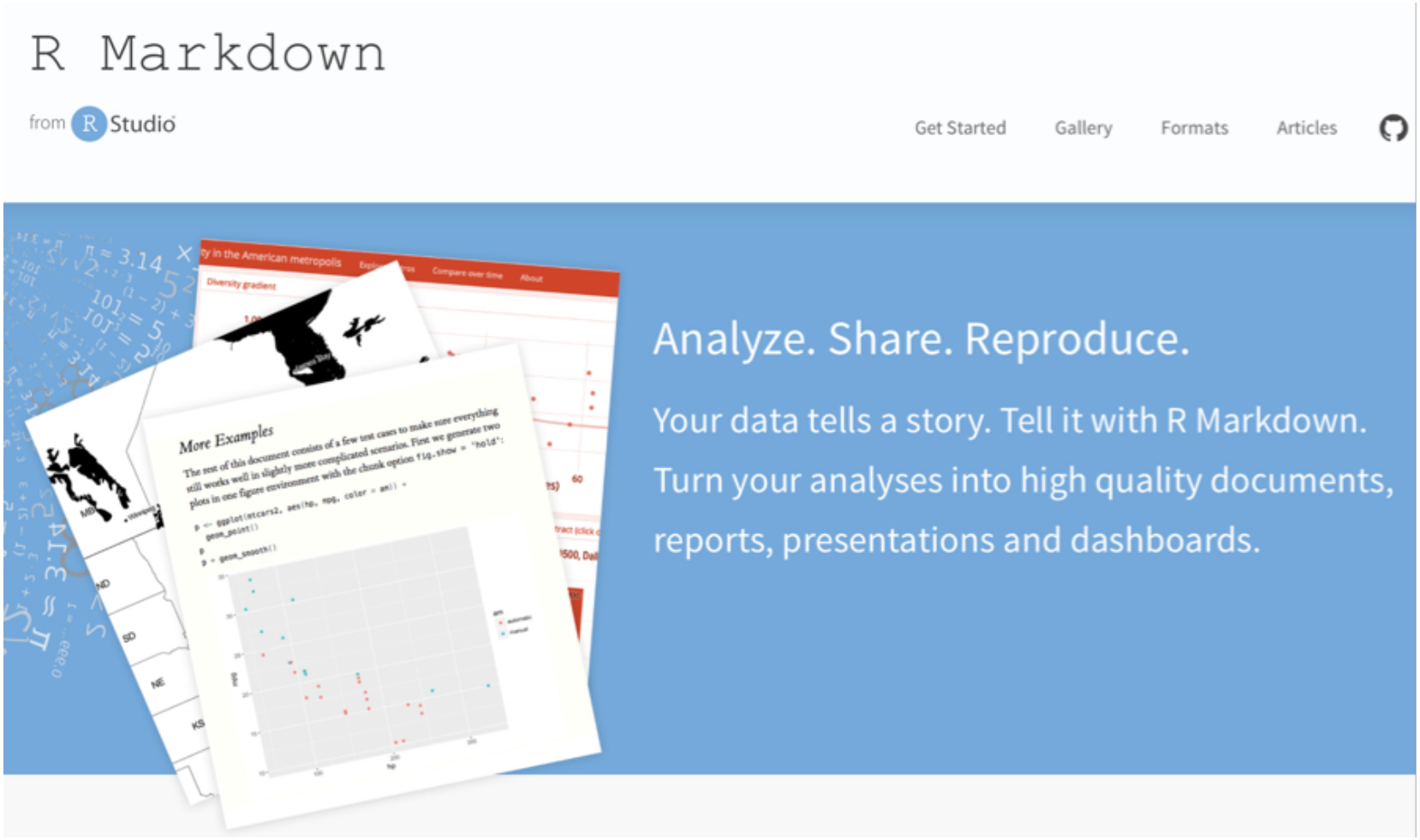
- DataCamp is an interactive way to learning R and other programming languages online.
- Its introductory courses are free. But it costs you $25 per month to access all other courses.
- Thanks to DataCamp, it will support STAT021 by giving us **FREE** access to all the courses for 6 months.

Sign up for DataCamp



- Wait for my invitation email to sign up for DataCamp and join the STAT021 group.
- Sign up an account using your **Swarthmore email**.

RMarkdown


R Markdown
from RStudio
Get Started
Gallery
Formats
Articles
Analyze. Share. Reproduce.
Your data tells a story. Tell it with R Markdown.
Turn your analyses into high quality documents,
reports, presentations and dashboards.
More Examples

Paperless homework

- Homework will be given in an .Rmd file on Moodle. Download this .Rmd file and upload it to RStudio server.
- You conduct analyses and answer questions in this single file on RStudio server.
- RStudio generates a .pdf file from the .Rmd file. Submit the .pdf file to Moodle.
- All homework will be graded on Moodle.

Picture from http://www.websigmas.com/wp-content/uploads/2014/11/Go-Paperless.png

Lecture 1 Glossary

- **R**: statistical programming language
- **RStudio**: a software based on R
- **RStudio server**: the online version of RStudio (recommended)
- **DataCamp**: a website to learn R programming language
- **RMarkdown**: a tool that generates reports with text, R codes and output in one place
  - This is one of the great features of RStudio

Learning R: How to get help

- R programming
  - Learn and practice: DataCamp
  - Ask questions: Google, StackOverflow
  - Other sources https://www.RStudio.com/Online-Learning/
- RStudio software
  - Working with the RStudio IDE (Part 1)
  - RStudio IDE Cheatsheet
- RMarkdown
  - RMarkdown Cheatsheet
  - RMarkdown Reference Guide

Homework 1 (20%)

- (5%) Log in to Swarthmore RStudio Server by 9/5 Wednesday 11:55pm.
  - If you do not have access to it, email me.
- (5%) Schedule a 10-minute meeting with me. Sign up in class or here by 9/5 Wednesday 11:55pm.
- (10%) Learn R on DataCamp
  - You will receive an email for DataCamp sign-up.
  - Complete the first three chapters of **Introduction to R**
    - *Chapter 1 Intro to basics*
    - *Chapter 2 Vectors*
    - *Chapter 3 Matrices*
  - *Chapter 1* is due on 9/5 Wednesday 11:55pm.