Lecture 1 - Introduction

Sta 102 / BME 102

Colin Rundel

January 9, 2014

General Info

Professor:	Dr. Colin Rundel - *colin.rundel@stat.duke.edu* Old Chemistry 223E
Teaching Assistants:	Stephen Desilets - *stephen.desilets@duke.edu* Xi Yang - *xi.yang@duke.edu* Tongrong Wang - *tongrong.wang@duke.edu*
Lecture:	Old Chemistry 116 Wednesdays and Fridays, 11:45 am - 1:00 pm
Labs:	Old Chem 101 01L - Mondays 10:05 - 11:20 pm 02L - Mondays 11:45 - 1:00 pm

Course goals & objectives

1. Recognize the importance of data collection, identify limitations in data collection methods, and determine how they affect the scope of inference.
2. Use statistical software to summarize data numerically and visually, and to perform data analysis.
3. Have a conceptual understanding of the unified nature of statistical inference.
4. Apply estimation and testing methods to analyze single variables or the relationship between two variables in order to understand natural phenomena and make data-based decisions.
5. Model numerical response variables using a single explanatory variable or multiple explanatory variables in order to investigate relationships between variables.
6. Interpret results correctly, effectively, and in context without relying on statistical jargon.
7. Critique data-based claims and evaluate data-based decisions.
8. Complete an independent research project employing what you learn in this class.

Major topics

- *Introduction to data:* Observational studies and non-causal inference, principles of experimental design and causal inference, exploratory data analysis: description, summary and visualization.
- *Probability and distributions:* The basics of probability and chance processes, Bayesian perspective in statistical inference, the normal distribution.
- *Framework for inference:* Central Limit Theorem and sampling distributions
- *Statistical inference for numerical variables*
- *Statistical inference for categorical variables*
- *Simple linear regression:* Bivariate correlation and causality, introduction to modeling
- *Multivariate regression:* Multiple regression, logistic regression

Course materials

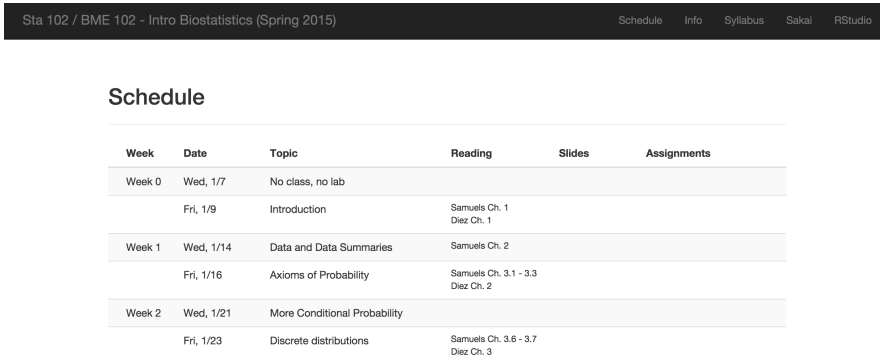
Optional Materials: Statistics for the Life Sciences - Samuels, Witmer, Schaffner Pearson, 4^{th} Edition, 2012 (ISBN: 9780321652805)

OpenIntro Statistics - Diez, Barr, Çetinkaya-Rundel CreateSpace, 2^{nd} Edition, 2012 (ISBN: 1478217200)

Four function calculator

Webpage

Announcements, slides, assignments, etc. posted on the website.



http://stat.duke.edu/~cr173/Sta102_Sp15/ or via Sakai

Support

Office hours Tuesdays 1:00 - 3:00 pm or by appointment.

SECC Sundays - Thursdays 4 - 9 pm (Old Chemistry 211A)

The statistics education center has upper level statistics students available to help you. For more information and a schedule see *http://stat.duke.edu/courses/resources-students* .

- You are highly encouraged to stop by with any questions or comments about the class, or just to say hi and introduce yourself.
- Homework will be due on Wednesdays - I strongly recommend at least attempting all problems to make the most of OH.

Grading

Homework	-	10%	Midterm 1	-	15%
Labs	-	10%	Midterm 2	-	15%
Project 1	-	10%	Final	-	30%
Project 2	-	10%			

- Grades will be curved at the end of the course after overall averages have been calculated.
  - Average of > 90 guaranteed A-.
  - Average of > 80 guaranteed B-.
  - Average of > 70 guaranteed C-.
- The more evidence there is that the class has mastered the material, the more generous the curve will be.
- Letter midterm grades will be assigned after Midterm 1

Homework

Objective: Help you develop a more in-depth understanding of the material and help you prepare for exams and the project.

- Questions from the textbooks and outside sources. (Full questions will be downloadable as a PDF from course website)
- Due at the beginning of class on the due date.
- 11 homeworks planned - lowest score will be dropped.
- Show all your work to receive credit.
- You are encouraged to work with others, but turn in your own work.
- Excused absences do not excuse homework.

Labs

Objective: Give you hands on experience with data analysis using statistical software, provide you with tools for the projects.

https://vm-manage.oit.duke.edu/containers/rstudio

- 12 labs planned - lowest score will be dropped.
- Write ups due the following week - most can be completed in class, turned in via Sakai.
- You must attend the lab you are enrolled in, if you do not attend in a given week you are eligible for at most 50% credit on that lab.

Research Projects

Objective: Give you independent applied research experience using real data

Project 1:

- Pick data and research question(s) from a curated list.
- Analyze the data, write up your results.

Project 2:

- Open ended research project.
- You choose a research question, find relevant data, analyze it, write up your results.

More details on both in the weeks to come.

Exams

- Midterm 1: *Friday, February* 13^{th}
- Midterm 2: *Friday, March* 27^{th}

- Final: *Tuesday, April* 28^{th} *from 2 - 5 pm* (Cumulative)

- Exam dates cannot be changed. No make-up exams will be given. If you cannot take the exams on these dates you should drop this class.

- You may bring a calculator to the exams (no cell phones, iPods, etc.) and you may also bring one sheet of notes ("cheat sheet"). This sheet must be no larger than $8\frac{1}{2}" \times 11"$ and must be prepared by you (no photocopies). You may use both sides of the sheet.

Students with disabilities

Students with disabilities who believe they may need accommodations in this class are encouraged to contact the Student Disability Access Office at (919) 668-1267 as soon as possible to better ensure that such accommodations can be made.

Late Work Policy

- For homework and lab write ups:
  - late but during class: -10%
  - after class on due date: -20%
  - next day or later: -100%
- For research projects: -10% / day late

Other Policies

- The final exam must be taken at the stated time and you cannot pass this class if you do not take the final exam.
- You must score at least >30% on both research projects to pass this class.
- Regrade requests must be made within one week of when the assignment is returned, and must be submitted in writing.

Academic Dishonesty

Any form of academic dishonesty will result in an immediate 0 on the given assignment and will be reported to the Office of Student Conduct. Additional penalties may also be assessed if deemed appropriate. If you have any questions about whether something is or is not allowed, ask me beforehand.

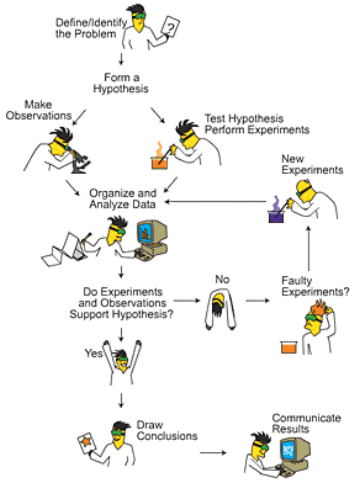
Some examples:

- Use of disallowed materials (including any form of communication with classmates or looking at a classmate's work) during exams.
- Plagiarism of any kind.
- Use of outside answer keys or solution manuals for the homework.

Tips for success

1. Complete the reading before each lecture, and review again at the end of each chapter.
2. Be an active participant during lectures and labs.
3. Ask questions - during class or office hours, or by email. Ask me, the TAs, and your classmates.
4. Do the problem sets - start early and make sure you attempt and understand all questions.
5. Start your project early and allow adequate time to complete the necessary components.
6. Give yourself plenty of time time to prepare a good cheat sheet for exams. This requires going through the material and taking the time to review the concepts that you're not comfortable with.
7. Do not procrastinate - don't let a week go by with unanswered questions as it will just make the following week's material even more difficult to follow.

Statistics and the Scientific Method



From Universe Today - *http://www.universetoday.com/74036/what-are-the-steps-of-the-scientific-method/*

Charles Darwin

ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

Charles Darwin

Francis Galton

TABLE I.

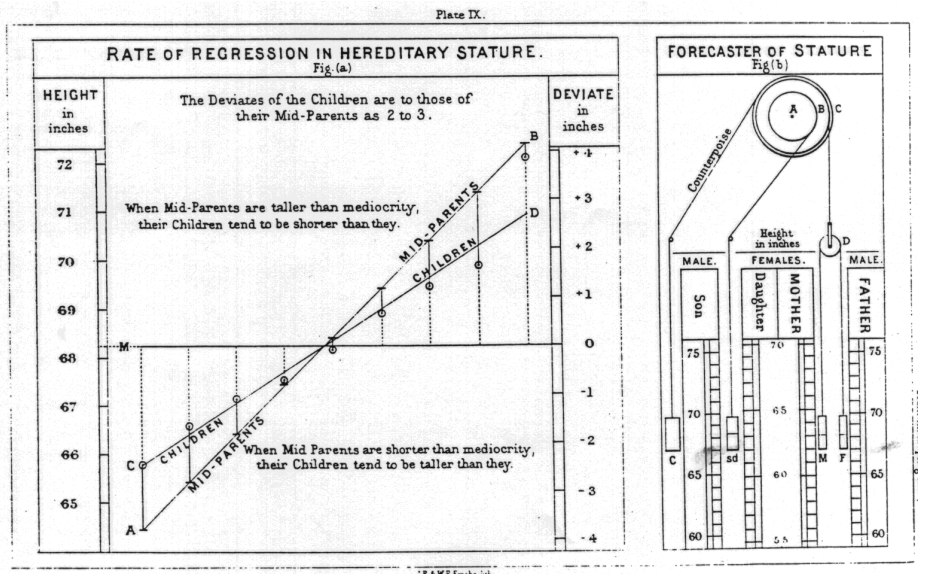
NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.													Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Anthropological Miscellanea.

Francis Galton



R.A. Fisher

"I occasionally meet geneticists who ask me whether it is true that the great geneticist R.A. Fisher was also an important statistician." - L.J. Savage (Annals of Statistics, 1976)

Source: *http://www.swlearning.com/quant/kohler/stat/biographical_sketches/Fisher_3.jpeg*

R.A. Fisher cont.

Biology:

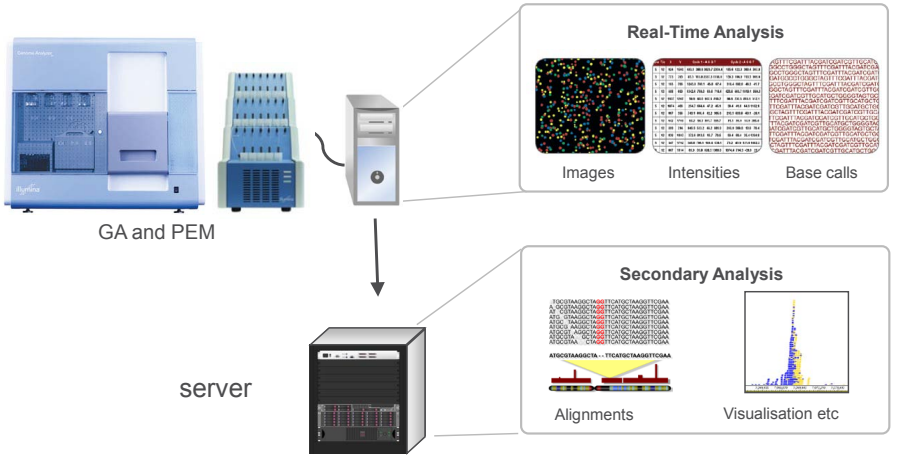
- Heterozygote advantage
- Population genetics (Modern evolutionary synthesis)
- Fisherian runaway selection
- ...

Statistics:

- Analysis of Variance
- Null hypothesis
- Maximum Likelihood
- F distribution
- Fisher's Exact test
- Fisher Information
- Randomization testing
- ...

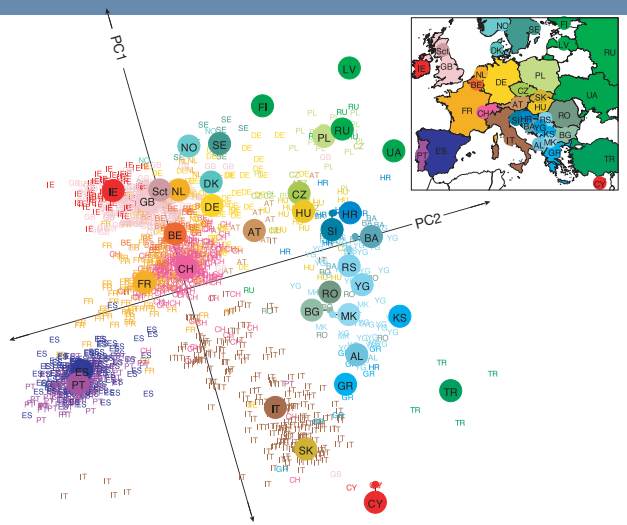
Source: Irish Elk - Fiddler Crab - Peafowl

Next Gen Sequencing



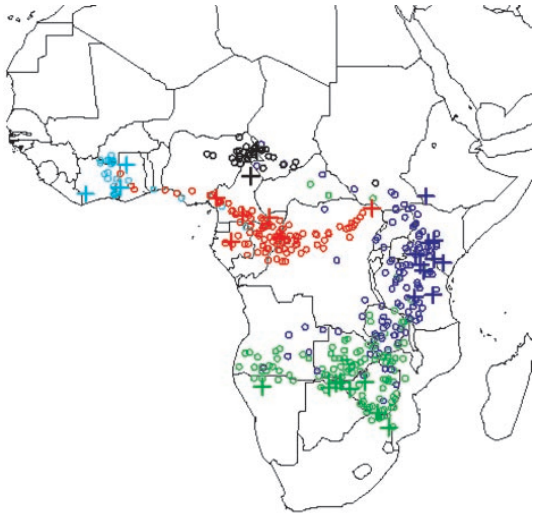
http://www.ebi.ac.uk/industry/Documents/workshop-materials/newsequence291009/Basecalling-Klaus_Maisinger.pdf

Novembre et al. - Nature 2008

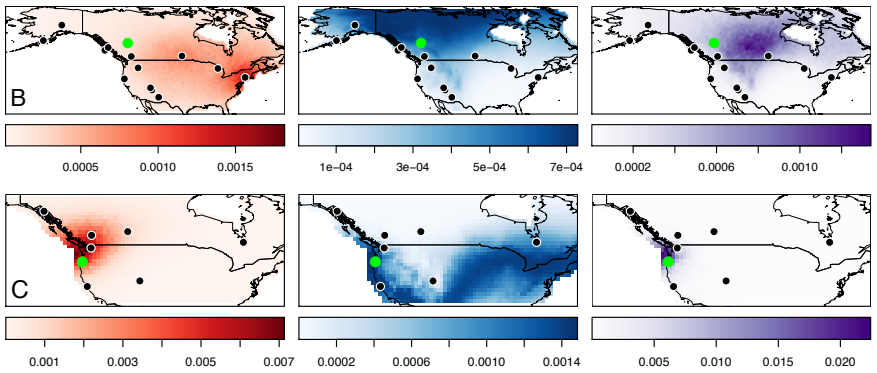


Analysis of 197,146 SNPs in 1,387 Europeans with known family origins

Wasser et al. - PNAS 2004

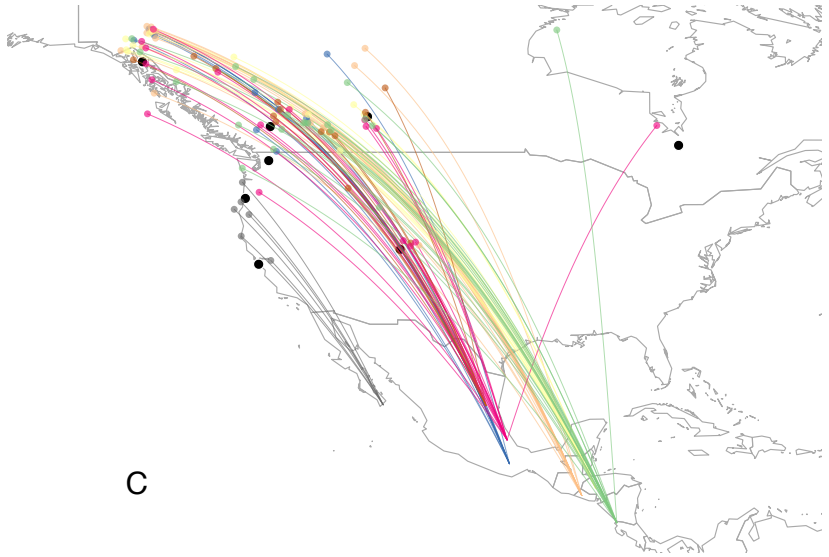


Migratory Connectivity



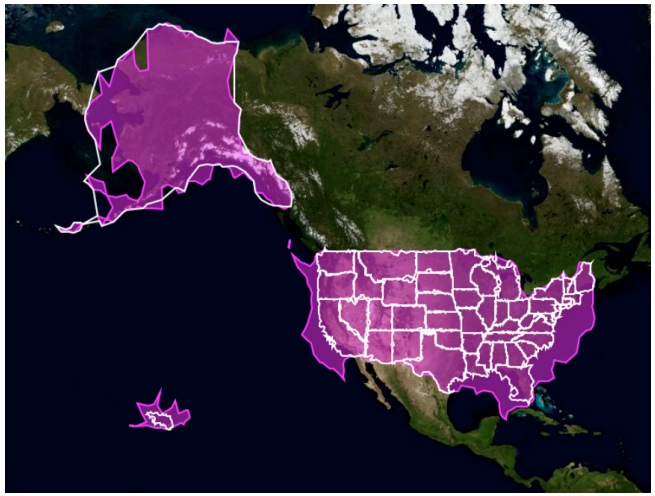
Rundel et al. - Molecular Ecology 2013

Migratory Connectivity



Rundel et al. - Molecular Ecology 2013

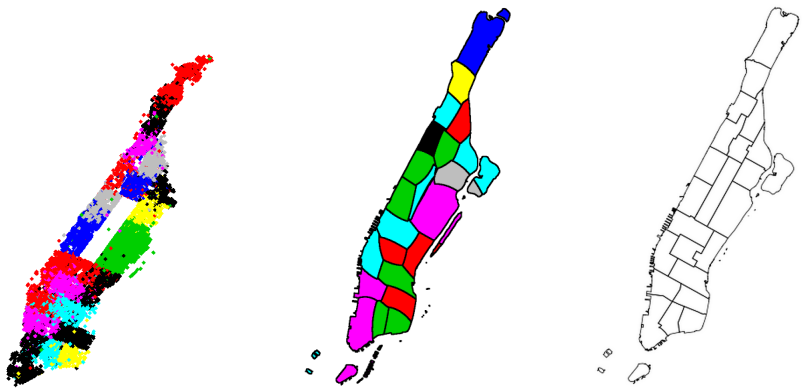
Map based on Flickr tags



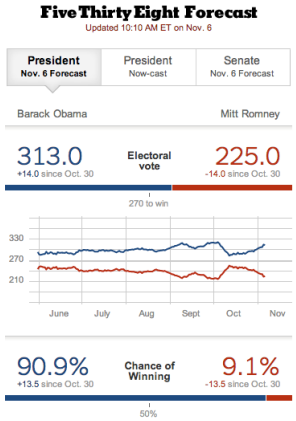
http://aaronstraupcope.com

Manhattan - Parking Tickets and Police Precincts

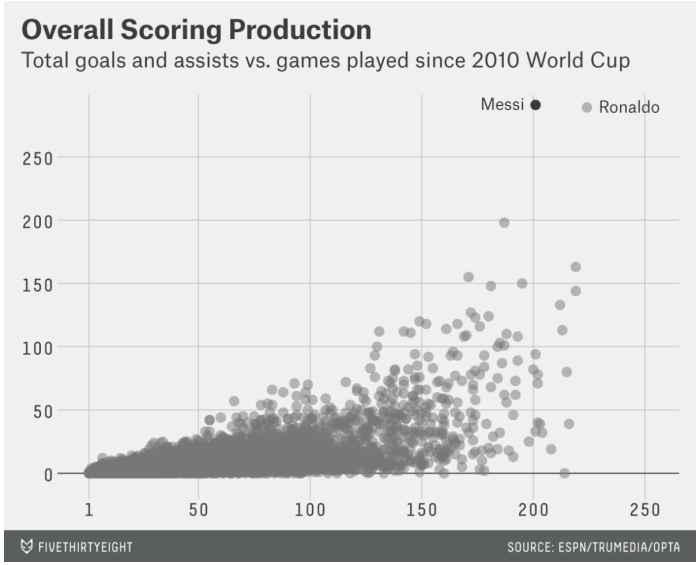
raw data | prediction | truth



The most famous statistician in the world ...



538 - Lionel Messi Is Impossible



538 - How to Tell Someone's Age When All You Know Is Her Name