

Lecture 1 - Introduction

Sta 102 / BME 102

January 13, 2016

Colin Rundel

Syllabus & Policies

General Info

Professor	Dr. Colin Rundel - colin.rundel@stat.duke.edu Old Chemistry 223E
Teaching Assistants	Yanyu Liu - yl304@stat.duke.edu Frank Li - frank.li@duke.edu Nayib Gloria - nayib.gloria@duke.edu
Lecture	Social Sciences 136 Mondays and Wednesdays, 3:05 - 4:20 pm
Labs	Old Chem 101 01L - Tuesdays 10:05 - 11:20 am 02L - Tuesdays 11:45 am - 1:00 pm 03L - Tuesdays 1:25 - 2:50 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.

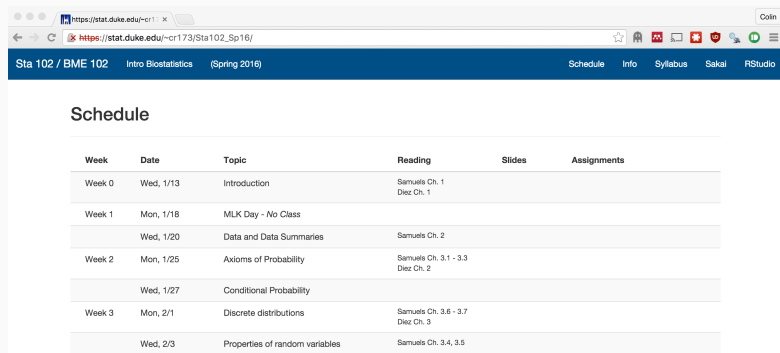
Major topics

- *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*: Univariate and bivariate analyses for numeric and categorical data, decision errors, power.
- *Simple linear regression*: Bivariate correlation and causality, introduction to modeling.
- *Multivariate regression*: Multiple regression, logistic regression.

Course materials

- Statistics for the Life Sciences - Samuels, Witmer, Schaffner
Pearson, 4th Edition, 2012 (ISBN: 9780321652805)
- OpenIntro Statistics - Diez, Barr, Çetinkaya-Rundel
CreateSpace, 3rd Edition, 2015 (ISBN: 194345003X)
- Calculator (\sqrt{x} , $\log(x)$, e^x)

Announcements, slides, assignments, etc. will be posted on course website:



Week	Date	Topic	Reading	Slides	Assignments
Week 0	Wed, 1/13	Introduction	Samuels Ch. 1 Diez Ch. 1		
Week 1	Mon, 1/18	MLK Day - No Class			
	Wed, 1/20	Data and Data Summaries	Samuels Ch. 2		
Week 2	Mon, 1/25	Axioms of Probability	Samuels Ch. 3.1 - 3.3 Diez Ch. 2		
	Wed, 1/27	Conditional Probability			
Week 3	Mon, 2/1	Discrete distributions	Samuels Ch. 3.6 - 3.7 Diez Ch. 3		
	Wed, 2/3	Properties of random variables	Samuels Ch. 3.4, 3.5		

http://stat.duke.edu/~cr173/Sta102_Sp16/

or via Sakai

Office hours

Professor Tuesdays, 3:00 - 5:00 pm or by appointment.

TAs TBD

- 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 attempting the problems beforehand to make the most of OH.

Grading

Homework	-	15%	Midterm 1	-	15%
Labs	-	10%	Midterm 2	-	15%
Project	-	20%	Final	-	25%

- 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

Goal of the homework is for 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 you *must* turn in your own work.
- Excused absences do not excuse homework.

Goal of the labs is for you to have hands on experience with data analysis using statistical software, provide you with tools for the projects.

- 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

The goal of the project is to give you independent applied research experience using real data

- Open ended research project.
- I will provide a large, high quality data set, you choose a research question, select relevant data, analyze it, write up your results.
- Two stages:
 1. Proposal & EDA
 2. Multivariate Analysis

More details after Midterm 1.

- Midterm 1: *Monday, February 22 in class*
- Midterm 2: *Monday, April 4 in class*

- Final: *Monday, May 2, 9:00 am - 12:00 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.

- For the exams you may bring:
 - a calculator (no cell phones, iPods, etc.)
 - a “cheat sheet” - 8.5” × 11” front and back

Special Accommodations

Any students 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 any necessary 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: no credit

For research projects:

- -10% / day late

Excused absences do not excuse assigned work - if you are going to miss class make arrangements ahead of time to turn in your work.

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 an average of at least 30% on the research project 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:

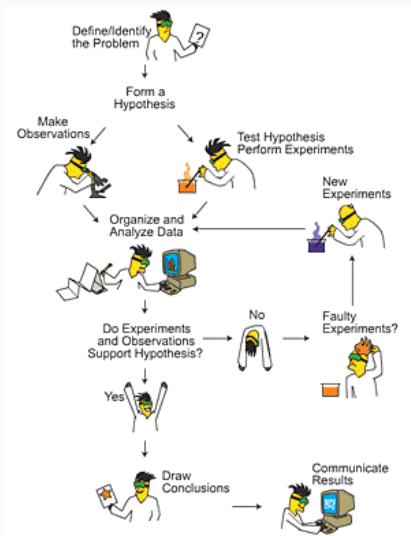
- Use of disallowed materials (including any form of communication with classmates or looking at a classmate's work) during exams.
- Plagiarism of any kind.
- Use of outside answer keys or solution manuals for the homework.

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 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.

Why (Bio)Statistics

Statistics and the Scientific Method



ANNALS OF SCIENCE

DECEMBER 13, 2010 ISSUE

THE TRUTH WEARS OFF

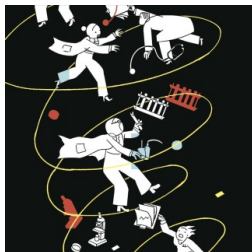
Is there something wrong with the scientific method?

BY JONAH LEHRER

Many results that are rigorously proved and accepted start shrinking in later studies.

ILLUSTRATION BY LAURENT CILLUFFO

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; when there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other

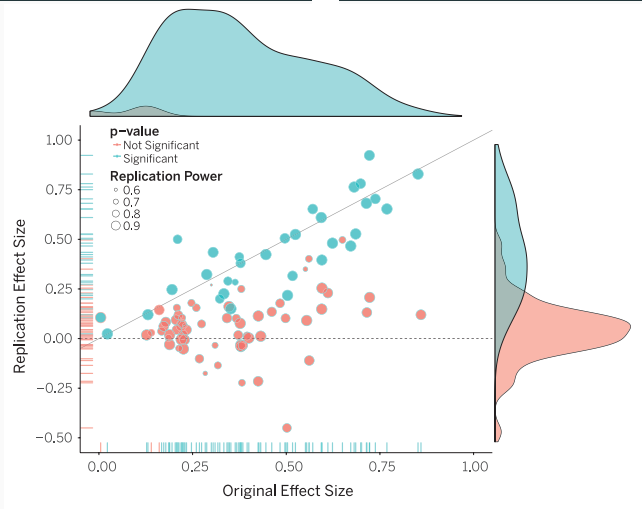
factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p -value less than 0.05. Research is not most appropriately represented and summarized by p -values, but, unfortunately, there is a widespread notion that medical research articles

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none

Reproducibility Project: Psychology



From Science - <http://science.sciencemag.org/content/349/6251/aac4716>

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.'

INTRODUCTION Page 1

CHAPTER I.

VARIATION UNDER DOMESTICATION.

Causes of Variability — Effects of Habit — Correlation of Growth —
Inheritance — Character of Domestic Varieties — Difficulty of
distinguishing between Varieties and Species — Origin of Domestic
Varieties from one or more Species — Domestic Pigeons, their
Differences and Origin — Principle of Selection anciently followed,
its Effects — Methodical and Unconscious Selection — Unknown
Origin of our Domestic Productions — Circumstances favourable
to Man's power of Selection 7-43

CHAPTER II.

VARIATION UNDER NATURE.

Variability — Individual differences — Doubtful species — Wide
ranging, much diffused, and common species vary most — Spe-
cies of the larger genera in any country vary more than the species
of the smaller genera — Many of the species of the larger genera
resemble varieties in being very closely, but unequally, related
to each other, and in having restricted ranges 44-59

CHAPTER III.

STRUGGLE FOR EXISTENCE.

Bears on natural selection—The term used in a wide sense—Geometrical powers of increase — Rapid increase of naturalised animals and plants—Nature of the checks to increase—Competition universal — Effects of climate — Protection from the number of individuals—Complex relations of all animals and plants throughout nature—Struggle for life most severe between individuals and varieties of the same species; often severe between species of the same genus—The relation of organism to organism the most important of all relations .. Page 60-79

CHAPTER IV.

NATURAL SELECTION.

Natural Selection — its power compared with man's selection — its power on characters of trifling importance — its power at all ages and on both sexes — Sexual Selection — On the generality of intercrosses between individuals of the same species — Circumstances favourable and unfavourable to Natural Selection, namely, intercrossing, isolation, number of individuals — Slow action — Extinction caused by Natural Selection — Divergence of Character, related to the diversity of inhabitants of any small area, and to naturalisation — Action of Natural Selection, through Divergence of Character and Extinction, on the descendants from a common parent — Explains the Grouping of all organic beings 80-130

CHAPTER V.

LAWS OF VARIATION.

Effects of external conditions — Use and disuse, combined with natural selection; organs of flight and of vision — Acclimatisation — Correlation of growth — Compensation and economy of growth — False correlations — Multiple, rudimentary, and lowly organised structures variable — Parts developed in an unusual manner are highly variable: specific characters more variable than generic: secondary sexual characters variable — Species of the same genus vary in an analogous manner — Reversions to long-lost characters — Summary 131-170

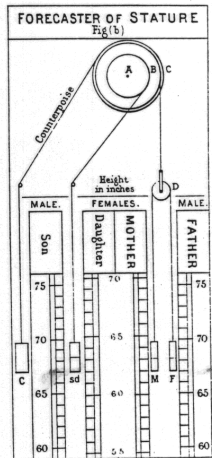
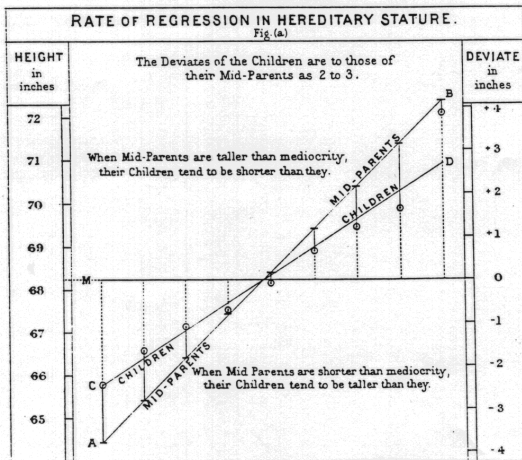
TABLE I.

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.

Plate IX.





“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

Biology:

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

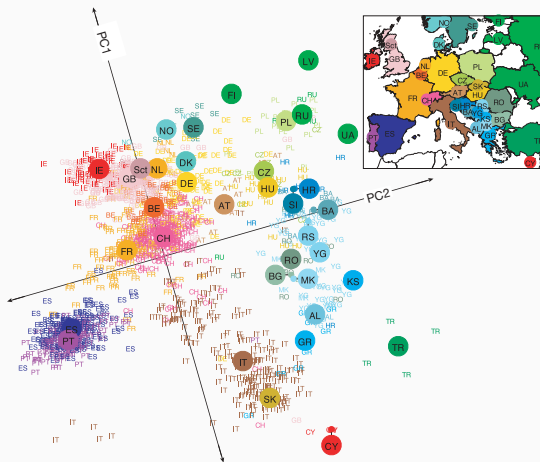
Statistics:

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

Runaway Selection

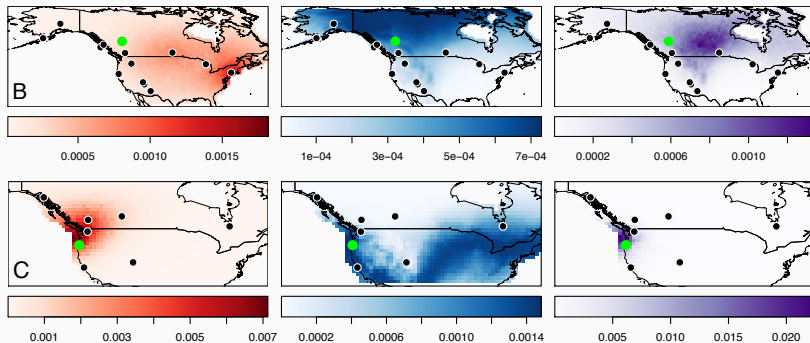


Source: [Irish Elk](#) - [Fiddler Crab](#) - [Peafowl](#)

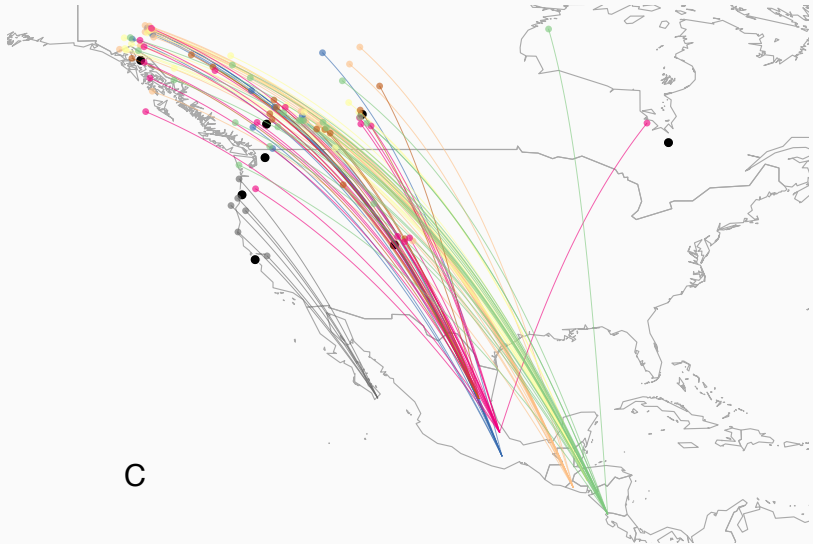


Analysis of 197,146 SNPs in 1,387 Europeans with known family origins

Spatial Mapping



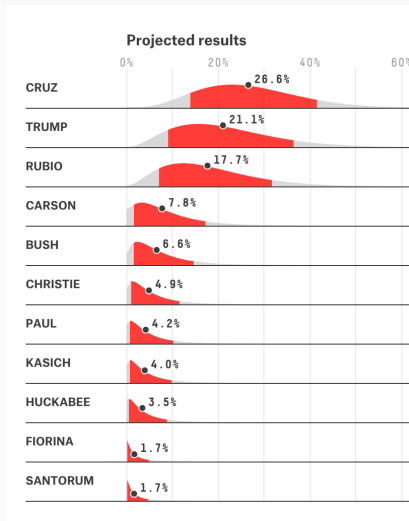
Migratory Connectivity



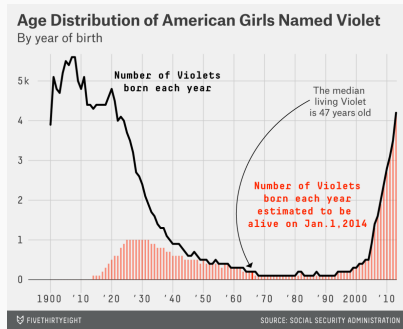
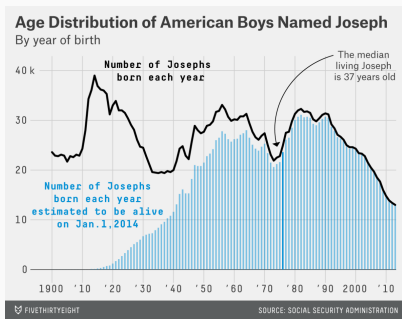
C

Other Applications

The most famous statistician in the world ...



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



<http://fivethirtyeight.com/features/>

[how-to-tell-someones-age-when-all-you-know-is-her-name/](http://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name/)

Why you probably shouldn't be playing ...



<http://graphics.latimes.com/powerball-simulator/>