Syllabus & Policies
<table>
<thead>
<tr>
<th>General Info</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Professor</strong></td>
</tr>
<tr>
<td>Dr. Colin Rundel - <a href="mailto:colin.rundel@stat.duke.edu">colin.rundel@stat.duke.edu</a></td>
</tr>
<tr>
<td>Old Chemistry 223E</td>
</tr>
<tr>
<td><strong>Teaching Assistants</strong></td>
</tr>
<tr>
<td>Yanyu Liu - <a href="mailto:yl304@stat.duke.edu">yl304@stat.duke.edu</a></td>
</tr>
<tr>
<td>Frank Li - <a href="mailto:frank.li@duke.edu">frank.li@duke.edu</a></td>
</tr>
<tr>
<td>Nayib Gloria - <a href="mailto:nayib.gloria@duke.edu">nayib.gloria@duke.edu</a></td>
</tr>
<tr>
<td><strong>Lecture</strong></td>
</tr>
<tr>
<td>Social Sciences 136</td>
</tr>
<tr>
<td>Mondays and Wednesdays, 3:05 - 4:20 pm</td>
</tr>
<tr>
<td><strong>Labs</strong></td>
</tr>
<tr>
<td>Old Chem 101</td>
</tr>
<tr>
<td>01L - Tuesdays 10:05 - 11:20 am</td>
</tr>
<tr>
<td>02L - Tuesdays 11:45 am - 1:00 pm</td>
</tr>
<tr>
<td>03L - Tuesdays 1:25 - 2:50 pm</td>
</tr>
</tbody>
</table>
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


- Calculator ($\sqrt{x}$, $\log(x)$, $e^x$)
Announcements, slides, assignments, etc. will be posted on course website:

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.
Goal of the homework is for you to 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 the 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.
Exams

• 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.
• 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 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

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
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; where 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 actually exists
Rationale:

Introduction:

Open Science Collaboration

psychological science

Estimating the reproducibility of psychological science

RATIONALE:

INTRODUCTION:

Open Science Collaboration

In this article, we examine the reproducibility of findings in psychological science. We focus on the extent to which previously published findings can be replicated using the same methods and conditions. We also consider the role of the characteristics of the original and replication teams in predicting the success or failure of replication efforts.

RESULTS:

The mean effect size (r) of the replication effects was 0.197 (SD=0.257) for a significant effect and 0.403 (SD=0.188) for a nonsignificant effect. These results are consistent with other studies that have found a substantial decline in the strength of evidence for previously published findings. Ninety-seven percent of original studies had significant results (P < 0.05), and 95% confidence intervals were in the same direction. Thirty-six percent of replications had significant results, and 68% of replications were in the same direction of the original. These results suggest that the reproducibility of psychological science is low.

CONCLUSION:

The reproducibility of psychological science is a problem that requires attention. Scientific claims should not gain credibility because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error. Reproducibility is not well understood because of the complexity of the factors that influence it. However, innovative ideas become old news fast. Journal reviewers and editors may discount new tests of a published idea as unimportant for many findings in psychological research. This project provides accumulating evidence to increase certainty when facing uncertainty. We believe that replication points out paths that are likely; innovation points out paths that are possible; and prioritizing novelty over replication belies the uncertainty of scientific evidence. However, innovative ideas become old news fast. Journal reviewers and editors may discount new tests of a published idea as unimportant for many findings in psychological research. This project provides accumulating evidence to increase certainty when facing uncertainty. We believe that replication points out paths that are likely; innovation points out paths that are possible; and prioritizing novelty over replication belies the uncertainty of scientific evidence.
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, LINNEAN, 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 — Species 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
CONTENTS.

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. Page 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. Page 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 1.05).

<table>
<thead>
<tr>
<th>Heights of the Mid-parents in Inches</th>
<th>Below 62.2</th>
<th>63.2</th>
<th>64.2</th>
<th>65.2</th>
<th>66.2</th>
<th>67.2</th>
<th>68.2</th>
<th>69.2</th>
<th>70.2</th>
<th>71.2</th>
<th>72.2</th>
<th>73.2</th>
<th>Above</th>
<th>Total Number of Adult Children</th>
<th>Medians</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>71.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>70.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>43</td>
<td>11</td>
</tr>
<tr>
<td>69.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>68</td>
<td>22</td>
</tr>
<tr>
<td>68.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>183</td>
<td>41</td>
</tr>
<tr>
<td>67.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>219</td>
<td>40</td>
</tr>
<tr>
<td>66.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>211</td>
<td>33</td>
</tr>
<tr>
<td>65.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>78</td>
<td>20</td>
</tr>
<tr>
<td>64.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>66</td>
<td>12</td>
</tr>
<tr>
<td>Below</td>
<td></td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>5</td>
<td>7</td>
<td>32</td>
<td>59</td>
<td>48</td>
<td>117</td>
<td>138</td>
<td>120</td>
<td>167</td>
<td>90</td>
<td>64</td>
<td></td>
<td>928</td>
<td>205</td>
</tr>
<tr>
<td>Medians</td>
<td></td>
<td>66.3</td>
<td>67.8</td>
<td>67.9</td>
<td>67.7</td>
<td>67.9</td>
<td>68.3</td>
<td>68.5</td>
<td>69.0</td>
<td>69.0</td>
<td>70.0</td>
<td>69.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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.
The Deviates of the Children are to those of their Mid-Parents as 2 to 3.

When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they.

When Mid Parents are shorter than mediocrity, their Children tend to be taller than they.
“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)

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
- ...

29
Runaway Selection

Source: Irish Elk - Fiddler Crab - Peafowl
The direction of the PC1 axis and its relative strength may reflect a special role for this geographic axis in the demographic history of Europeans (as first suggested in ref. 10). PC1 aligns north-northwest/south-southeast (NNW/SSE, 216°) and accounts for approximately twice the amount of variation as PC2 (0.30% versus 0.15%, first eigenvalue 54.09, second eigenvalue 52.04). However, caution is required because the direction and relative strength of the PC axes are affected by factors such as the spatial distribution of samples (results not shown, also see ref. 9). More robust evidence for the importance of a roughly NNW/SSE axis in Europe is that, in these same data, haplotype diversity decreases from south to north (A.A. et al., submitted). As the fine-scale spatial structure evident in Fig. 1 suggests, European DNA samples can be very informative about the geographical origins of their donors. Using a multiple-regression-based assignment approach, one can place 50% of individuals within 310 km of their reported origin and 90% within 700 km of their origin (Fig. 2 and Supplementary Table 4, results based on populations with \( n \)). Across all populations, 50% of individuals are placed within 540 km of their reported origin, and 90% of individuals within 840 km (Supplementary Fig. 3 and Supplementary Table 4). These numbers exclude individuals who reported mixed grandparental ancestry, who are typically assigned to locations between those expected from their grandparental origins (results not shown). Note that distances of assignments from reported origin may be reduced if finer-scale information on origin were available for each individual.

Population structure poses a well-recognized challenge for disease-association studies (for example, refs 11–13). The results obtained here reinforce that the geographic distribution of a sample is important to consider when evaluating genome-wide association studies.

---

**Figure 1**

Population structure within Europe. (a) A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small colored labels represent individuals and large colored points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The PC axes are rotated to emphasize the similarity to the geographic map of Europe. AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia; SC, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia.

(b) A magnification of the area around Switzerland from (a) showing differentiation within Switzerland by language.

(c) Genetic similarity versus geographic distance. Median genetic correlation between pairs of individuals as a function of geographic distance between their respective populations.

Analysis of 197,146 SNPs in 1,387 Europeans with known family origins.
Figure 2: Posterior probability maps, from left to right, of the genetic, isotopic and combined model output. Rows (A) and (B) reflect the results for the same exemplar Hermit Thrush under cross validation by individual and location respectively, while (C) and (D) are the same Wilson’s Warbler. ● indicates the true origin of the sample and • indicate all other sampling locations.
Figure 5: Results of localization to breeding grounds of samples taken from wintering Wilson’s Warblers using Genetic (A), Isotopic (B) or Combined (C) approaches. Assignment is based on location with the greatest posterior probability.
Other Applications
The most famous statistician in the world ...


<table>
<thead>
<tr>
<th>Candidate</th>
<th>Projected Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cruz</td>
<td>26.6%</td>
</tr>
<tr>
<td>Trump</td>
<td>21.1%</td>
</tr>
<tr>
<td>Rubio</td>
<td>17.7%</td>
</tr>
<tr>
<td>Carson</td>
<td>7.8%</td>
</tr>
<tr>
<td>Bush</td>
<td>6.6%</td>
</tr>
<tr>
<td>Christie</td>
<td>4.9%</td>
</tr>
<tr>
<td>Paul</td>
<td>4.2%</td>
</tr>
<tr>
<td>Kasich</td>
<td>4.0%</td>
</tr>
<tr>
<td>Huckabee</td>
<td>3.5%</td>
</tr>
<tr>
<td>Fiorina</td>
<td>1.7%</td>
</tr>
<tr>
<td>Santorum</td>
<td>1.7%</td>
</tr>
</tbody>
</table>
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/