Lecture 1 - Introduction

Sta 102 / BME 102

May 16, 2016

Colin Rundel & Mine Çetinkaya-Rundel

Course Details

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.
- *Multiple regression:* Multiple regression, logistic regression.

Course materials

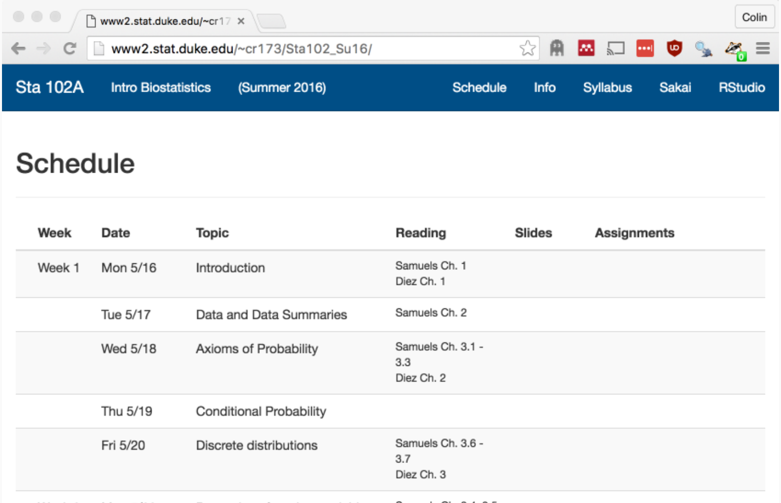
- Statistics for the Life Sciences - Samuels, Witmer, Schaffner
  Pearson, 4^{th} Edition, 2012 (ISBN: 9780321652805)

- OpenIntro Statistics - Diez, Barr, Çetinkaya-Rundel
  CreateSpace, 3^{rd} Edition, 2015 (ISBN: 194345003X)

- Calculator ($\sqrt{x}$, $\log(x)$, e^x)

Webpage

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



Schedule

Week	Date	Topic	Reading	Slides	Assignments
Week 1	Mon 5/16	Introduction	Samuels Ch. 1 Diez Ch. 1		
	Tue 5/17	Data and Data Summaries	Samuels Ch. 2		
	Wed 5/18	Axioms of Probability	Samuels Ch. 3.1 - 3.3 Diez Ch. 2		
	Thu 5/19	Conditional Probability			
	Fri 5/20	Discrete distributions	Samuels Ch. 3.6 - 3.7 Diez Ch. 3		

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.
- 8 homeworks planned - lowest score will be dropped.
- Show all your work to receive credit.

Labs

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.

- 8 labs planned - lowest score will be dropped.
- Write ups due the following lab session - majority of each lab can be completed in class, turned in via Sakai.
- Submit both Rmd and HTML files.

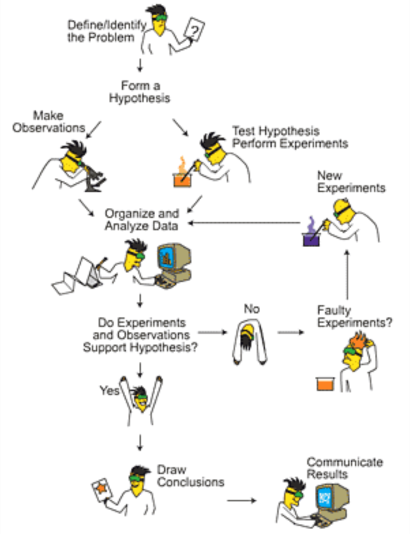
Research Projects

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

- Open ended research project.
- You find a data set, choose a research question, select relevant data, analyze it, write up your results.
- Multiple stages: proposal, EDA, analysis.

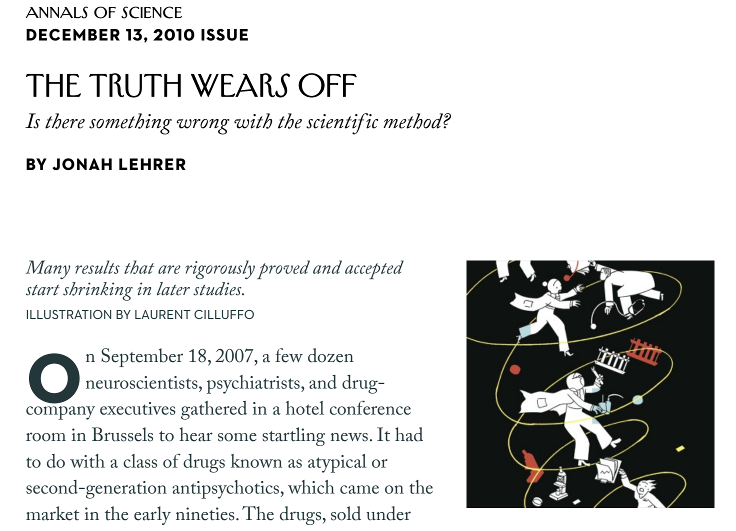
Why (Bio)Statistics

Statistics and the Scientific Method


Define/Identify the Problem
Form a Hypothesis
Make Observations
Test Hypothesis Perform Experiments
New Experiments
Organize and Analyze Data
Do Experiments and Observations Support Hypothesis?
No
Faulty Experiments?
Yes
Draw Conclusions
Communicate Results

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

The Decline Effect



ANNALS OF SCIENCE

DECEMBER 13, 2010 ISSUE

THE TRUTH WEARS OFF

Is there something wrong with the scientific method?

BY JONAH LEHRER

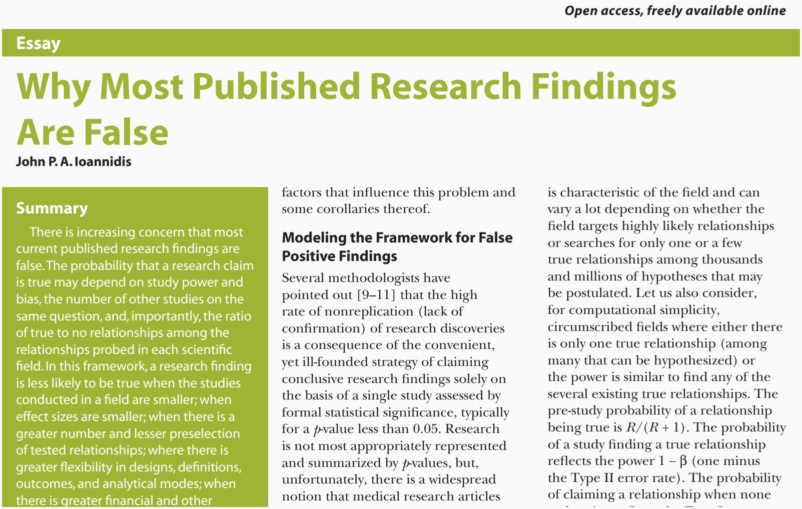
Many results that are rigorously proved and accepted start shrinking in later studies.

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

From The New Yorker - *http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off*

Nonreplication



Open access, freely available online

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; 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.

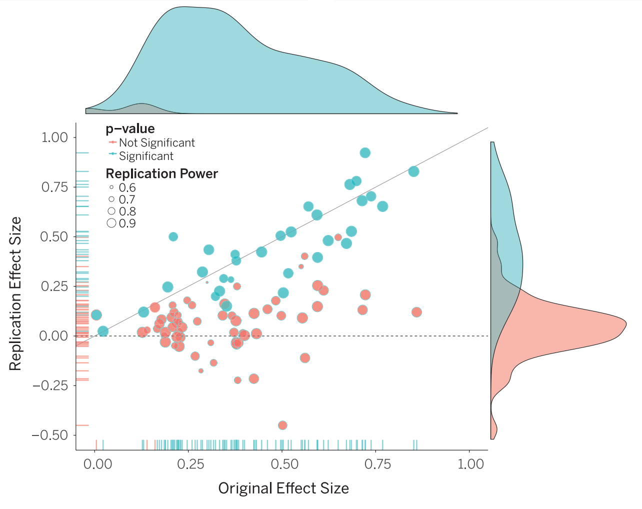
Modeling the Framework for False Positive Findings

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

From PLOS Medicine - *http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124*

Reproducibility Project: Psychology

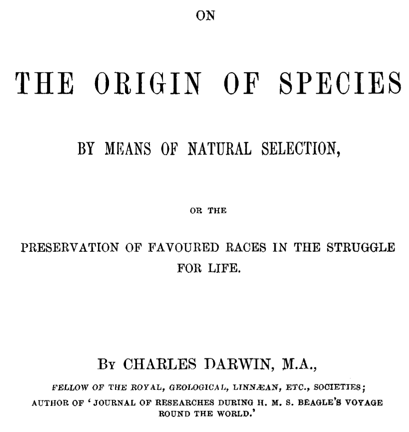


From Science - *http://science.sciencemag.org/content/349/6251/aac4716*

ASA Statement of p-values

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Charles Darwin



ON

THE ORIGIN OF SPECIES

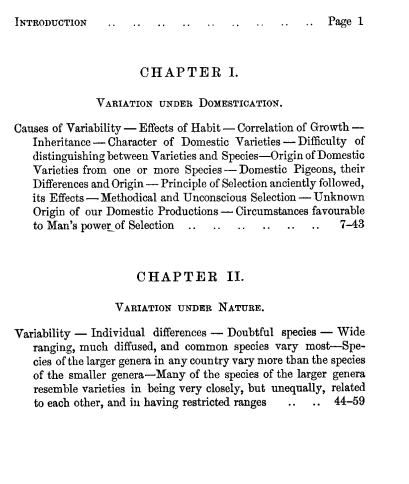
BY MEANS OF NATURAL SELECTION,

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.'



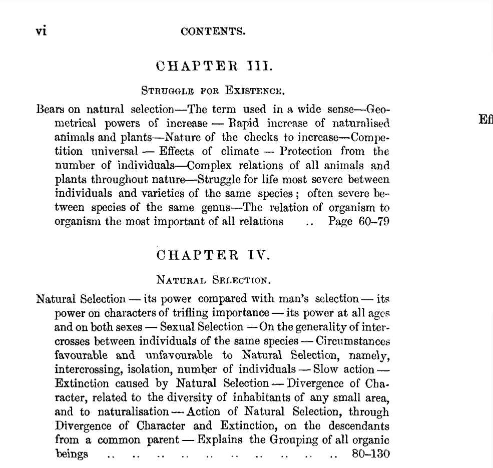
CHAPTER I.

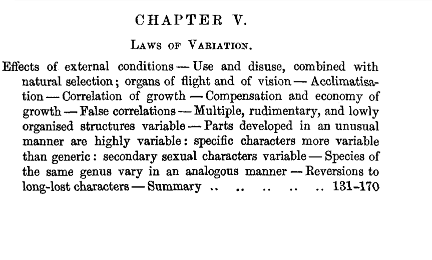
VARIATION UNDER DOMESTICATION.

CHAPTER II.

VARIATION UNDER NATURE.

Charles Darwin





Francis Galton

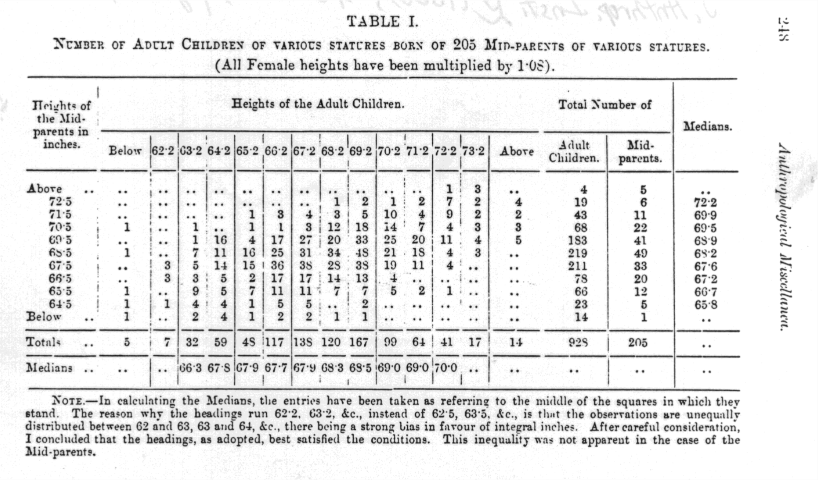


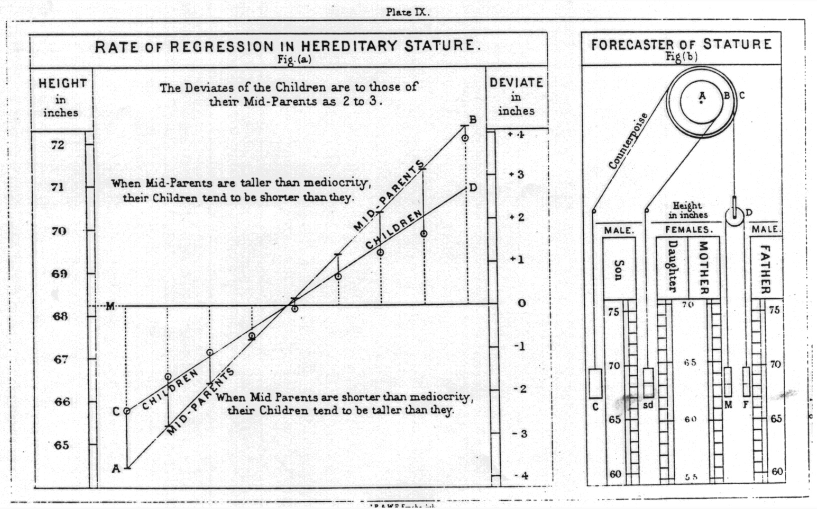
TABLE I.

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.

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
- ...

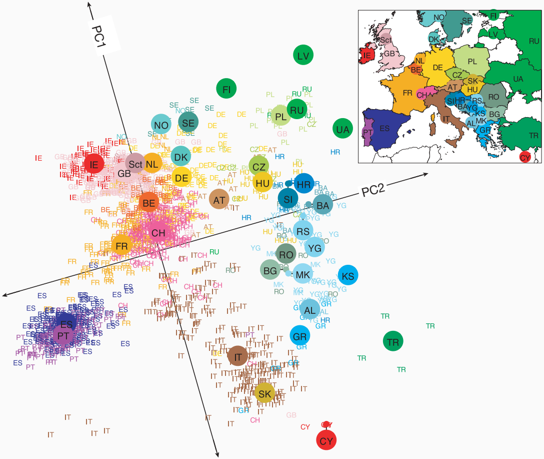
Statistics:

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

Runaway Selection

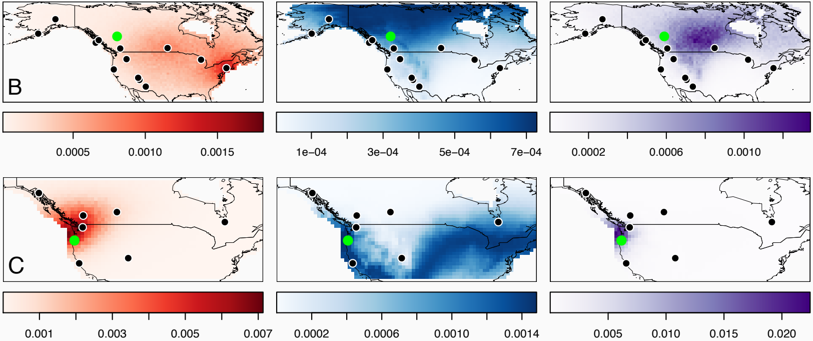
Source: Irish Elk - Fiddler Crab - Peafowl

Novembre et al. - Nature 2008

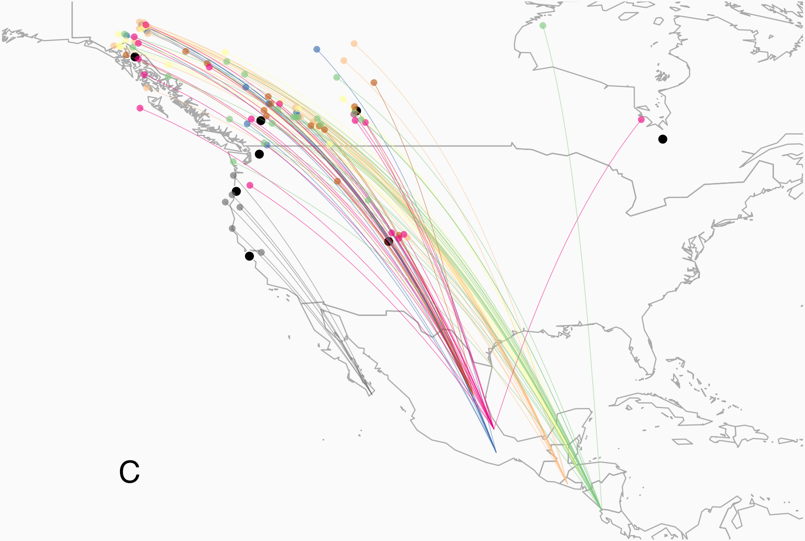


Analysis of 197,146 SNPs in 1,387 Europeans with known family origins

Spatial Mapping

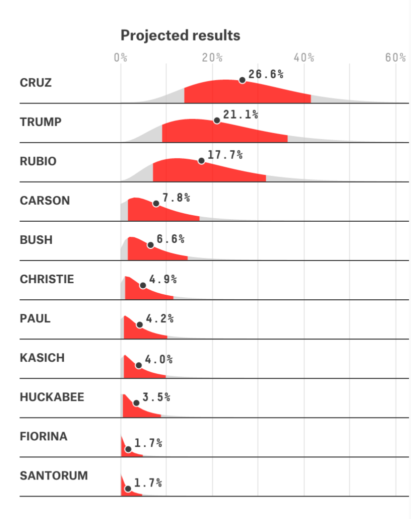


Migratory Connectivity



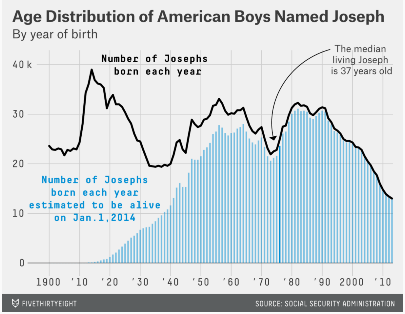
Other Applications

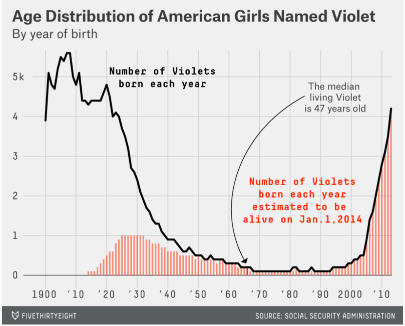
The most famous statistician in the world ...



http://projects.fivethirtyeight.com/election-2016/primary-forecast/iowa-republican/

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/

Why you probably shouldn't be playing ...

MAIL IN ANY NON-WINNING
CASH CASH CASH SCRATCHER AND YOU
COULD WIN A $500 VISA GIFT CARD

:162932
.50048539136-7

558144
20160108 14:15
$10.00 - 1 Draw
01/09/16

PowerPlay No

A 07 27 35 49 61 Powerball: 06 EP
B 23 25 26 46 49 Powerball: 15 EP
C 02 22 31 34 47 Powerball: 21 EP
D 25 48 53 58 68 Powerball: 25 EP
E 06 17 41 48 69 Powerball: 25 EP

POWERBALL EST
JACKPOT $800 MILLION

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

Data collection and study design

Using a sample to make inferences about the population

- Ultimate goal: make inferences about populations

Using a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access

Using a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access
- Solution: use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*

Using a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access
- Solution: use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Using a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access
- Solution: use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Using a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access
- Solution: use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Suppose we want to know how many offspring female lemurs have, on average. It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center. We use the sample mean from these data as an estimate for the unknown population mean. Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

Sampling is natural

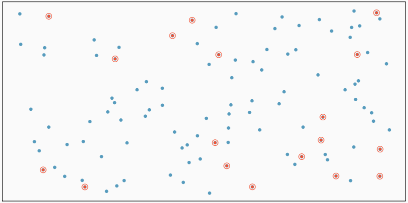
- When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's *exploratory analysis*
- If you generalize and conclude that your entire soup needs salt, that's an *inference*
- For your inference to be valid, the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population)

Sampling methods

Sampling methods

Simple random:

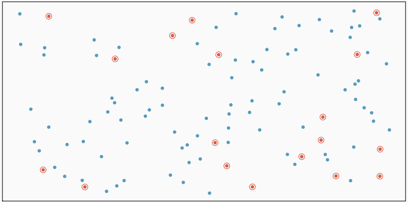
Drawing names from a hat



Sampling methods

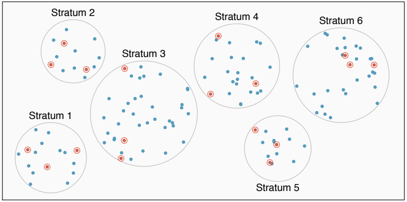
Simple random:

Drawing names from a hat



Stratified: homogenous strata

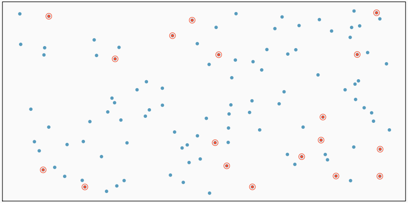
Stratify to control for SES



Sampling methods

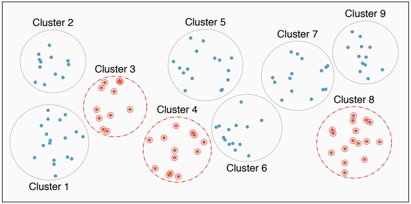
Simple random:

Drawing names from a hat



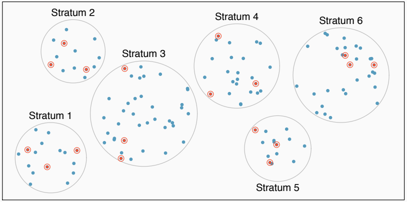
Cluster: heterogenous clusters

Sample all chosen clusters



Stratified: homogenous strata

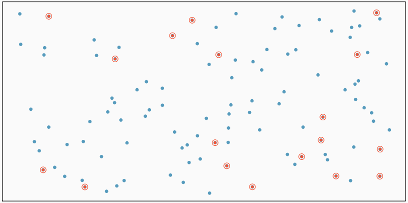
Stratify to control for SES



Sampling methods

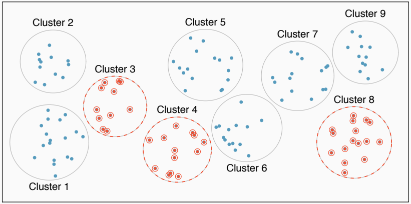
Simple random:

Drawing names from a hat



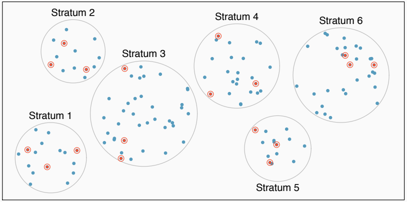
Cluster: heterogenous clusters

Sample all chosen clusters



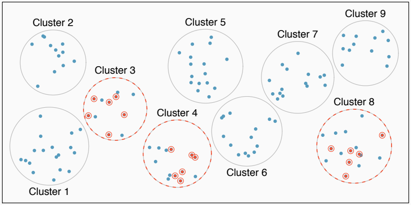
Stratified: homogenous strata

Stratify to control for SES



Multistage:

Random sample in chosen clusters



Clicker question

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. Which approach would likely be the *least* effective?

(a) Simple random sampling

(b) Stratified sampling, where each stratum is a neighborhood

(c) Cluster sampling, where each cluster is a neighborhood

Biases in study design

- *Non-response:* If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population

Biases in study design

- *Non-response:* If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue since such a sample will also not be representative of the population

Biases in study design

- *Non-response:* If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue since such a sample will also not be representative of the population
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample

Clicker question

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.

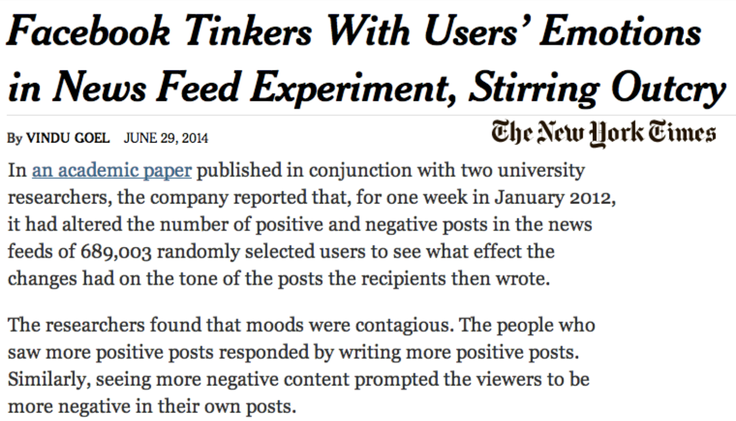
II. Overall, the school district has strong support from parents to move forward with the policy approval.

III. It is possible that majority of the parents of high school students disagree with the policy change.

IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) I and III (d) III and IV (e) Only IV

What type of study is this? What is the scope of inference (causality / generalizability)?



Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

By VINDU GOEL JUNE 29, 2014

The New York Times

In an academic paper published in conjunction with two university researchers, the company reported that, for one week in January 2012, it had altered the number of positive and negative posts in the news feeds of 689,003 randomly selected users to see what effect the changes had on the tone of the posts the recipients then wrote.

The researchers found that moods were contagious. The people who saw more positive posts responded by writing more positive posts. Similarly, seeing more negative content prompted the viewers to be more negative in their own posts.

http://www.nytimes.com/2014/06/30/technology/

facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed. What type of study is this?

What is the conclusion of the study?

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

Four principles of experimental design

- We would like to design an experiment to investigate if increased stress causes muscle cramps:

Four principles of experimental design

- We would like to design an experiment to investigate if increased stress causes muscle cramps:
    - Treatment: increased stress
    - Control: no or baseline stress

Four principles of experimental design

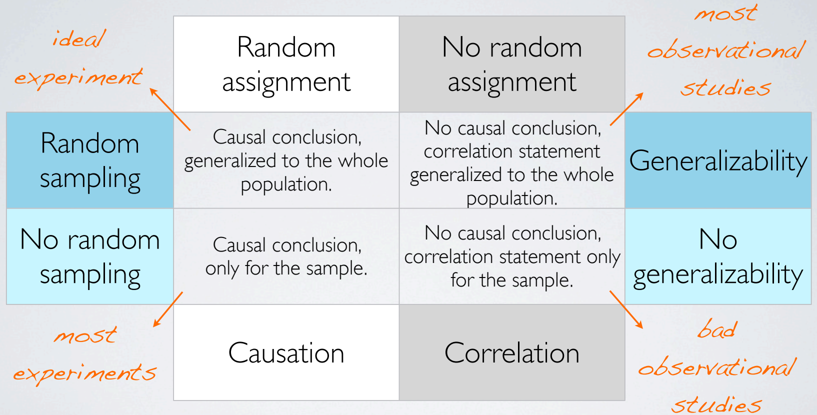
- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - Treatment: increased stress
  - Control: no or baseline stress
- It is suspected that the effect of stress might be different on younger and older people: *block* for age.

Four principles of experimental design

- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - Treatment: increased stress
  - Control: no or baseline stress
- It is suspected that the effect of stress might be different on younger and older people: *block* for age.

Why is this important? Can you think of other variables to block for?

Random sampling helps generalizability, random assignment helps causality



	Random assignment	No random assignment	
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
	Causation	Correlation	

ideal experiment

most observational studies

most experiments

bad observational studies

Summary

Summary of main ideas

1. Use a sample to make inferences about the population
2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
3. Sampling schemes can suffer from a variety of biases
4. Experiments use random assignment to treatment groups, observational studies do not
5. Four principles of experimental design: randomize, control, block, replicate
6. Random sampling helps generalizability, random assignment helps causality