Data-Driven Reasoning and Study Design

Data 100: Principles and Techniques of Data Science

Sandrine Dudoit

Department of Statistics and Division of Biostatistics, UC Berkeley

Spring 2019

Outline

Data-Driven Reasoning
Case Study: How to Find Housing in Berkeley?
Case Study: How Big is a Crowd?
Getting the Question Right
Data-Driven Study Design
Workflow Design
Getting the Data Right
Bad Data
What are Bad Data?
Sampling Bias in Political Polls
Duke Personalized Medicine Scandal
Probabilistic

Outline

4.4 A/B Testing

Learning Objectives

- How to approach a data-enabled inquiry, i.e., identify, reason about, and answer data-enabled questions.
- How to design a workflow for a particular data-enabled inquiry, i.e., lay out precisely each step in this workflow, from framing the question to translating, interpreting, and implementing (into actions) the results.
- How to use existing data.
- How to envisage what new types of questions could be addressed if we could collect certain data (i.e., measure certain variables) and how we might collect such data – "Futurism".

Learning Objectives

- How to communicate with and provide input to others to frame questions, collect data, and interpret and make decisions based on data analysis results.
- How to approach, properly use, and assess probabilistic designs.
- How to develop good research practice, cf. research responsible conduct and integrity, computational reproducibility and verifiability.
- How to avoid bad data!

Data-Driven Reasoning

- Let us first revisit a few of the data-enabled inquiries introduced in the first lecture to reason through the process of addressing them.
- We start with purposely vague questions, concerning vastly different topics and requiring different types of answers, to illustrate the process of identifying and framing data-enabled questions and identifying relevant data (i.e., what to measure).
- These examples illustrate common themes in approaching a data-driven question.
  - Framing questions is non-trivial. It is an iterative process.
  - Different data are relevant depending on the type of question and required answer.

Data-Driven Reasoning

- These aspect of Data Science are often glossed over in Computer Science (CS) and Statistics, where the questions and data are typically given.

How to Find Housing in Berkeley?

"I have just arrived in Berkeley, how do I find a place?"

Let's try to frame the question, i.e., make it more precise.

What do you want/need to know, exactly?

- "How much does it cost to rent an apartment?"
- "What's the cheapest apartment I can get?"
- "Can I afford to live in Berkeley or am I better off living in another city?"
- "I have a lead on an apartment, is the rent too expensive?"
- "I'll only be in Berkeley for 6 months, are there short term apartments?"
- "Should I have roommates or can I get a place by myself?"
- "My parents think it might be better to buy an apartment because I will be in Berkeley for 4 years. Are they right?"

How to Find Housing in Berkeley?

- "Which are the most expensive areas in the East Bay?"
- "What are the differences in housing costs between the 10 UC campuses?"

How to Find Housing in Berkeley?

- These are all relevant questions when it comes to finding housing in Berkeley. However, some are more precise and easier to translate into a question about data than others.
- Some questions are also more general (e.g., last two).
- Our next task is to map these questions into data-driven approaches and, in particular, specify relevant data to collect.
- Let's focus on a subset of questions and reason through them further.

How to Find Housing in Berkeley?

"I have a lead on an apartment, is the rent too expensive?"
Alternate, related questions: "Can you find a cheaper apartment to rent?" and, if so, "Would you live in it?".

- The answer depends of course on the rent relative to features of the apartment, e.g., how big, in what neighborhood.
- However, "big" and "neighborhood" are vague and can be described in various ways.
  - Relevant variables for assessing how big a housing unit is are, e.g., square footage, number of bedrooms, number of bathrooms.
  - Relevant variables for describing a neighborhood are, e.g., its name, safety rating, walking score, distance to campus.

How to Find Housing in Berkeley?

- The value placed on each of these variables will vary from person to person, i.e., we each have different loss functions.

How to Find Housing in Berkeley?

Now, how do you find more information, i.e., data, on rents for apartment meeting your criteria.

- Ask friends. This is limited and likely biased data.
- Google.
  - Search "typical rent in Berkeley". But whats "typical"? The median?
  - Search "average rent in Berkeley". One answer is $3,800, another $3,123.
  - Comparing your rent to search results gives you a quick-and-dirty answer, but it is over all rentals and too general (i.e., doesn't account for your housing criteria). Also, we may each get different answers depending on our search history.
- Collect rental data. Compare your rent to the distribution of rents for apartments meeting your criteria.

How to Find Housing in Berkeley?

- Relevant variables. In order to obtain this distribution, you will need to systematically collect the following types of variables for apartments in Berkeley: Rent, square footage, number of bedrooms, number of bathrooms, neighborhood name/safety rating/walking score/distance to campus, parking availability, washer/dryer availability, utilities included in rent or not.
- Where can you find such data?
  E.g. Craigslist, HotPads. Not necessarily exhaustive, perhaps selection bias, but probably good enough for the type of answer you need.
- How can you programmatically acquire this data (vs. click on and read through each listing)?
  Webscraping.

How to Find Housing in Berkeley?

- How can you analyze the data to get your answer?
  Some data cleaning inevitable (e.g., multiple listings with same ID, missing values), exploratory data analysis, graphical summaries, numerical summaries.

How Big is a Crowd?

Figure 1: *Crowd size.* Trump (left) and Obama (right) presidential inaugurations.

How Big is a Crowd?

Trump 2016 Presidential Inauguration.
(`https://en.wikipedia.org/wiki/Inauguration_of_Donald_Trump#Crowd_size`)
Donald Trump: *"an unbelievable, perhaps record-setting turnout"*.
Sean Spicer, White House Press Secretary: *"largest audience ever to witness an inauguration, period, both in person and around the globe"*.
Kellyanne Conway, Trump Counselor and Spokesperson: *"alternative facts"*.

How Big is a Crowd?

How do we check these claims, or "alternative facts"?

- Before we spend time doing this: Why does this matter? To whom? Maybe it's a stupid question?
- Also, where do you think crowd size estimates reported in the media or by politicians come from? These number can vary greatly depending on the source, e.g., event organizer, law enforcement.
  E.g. For the Trump inauguration, estimates for attendance on the Mall range from 300,000 to over a million.

How Big is a Crowd?

- Back to the question: "Did Trump have the biggest inauguration ever?"
  What is "big"? Do we count only people that were on the Mall? Or also television/radio/Internet audiences? Only US or worldwide audiences?
  This is a comparative question. We need comparable data on previous inaugurations, but previous inaugurations didn't have the same technologies, both for viewing the event and for estimating attendance.
- Other related, more manageable question: "How many attended the Trump inauguration in person on the Mall?"
- How accurate do we have to be? Depends on the purpose of the question, i.e., how the answer will be used.
  Actually, how does one measure accuracy?

How Big is a Crowd?

- There were no mechanisms in place for doing a census, i.e., directly counting the persons attending the inauguration on the Mall (e.g., ticket-only attendance, controlled entry and counting at checkpoints).
- Proxies for crowd size are: Public transportation ridership, Twitter feeds, Facebook and Instagram check-ins, crowd pictures, surveys. Each have their pros and cons. There is also the added difficulty that attendance varies over time and that some of the proxies are only snapshots in time.
- Image data.
  - First, what's an image?
  - How should we collect images of the inauguration? There could be selection bias. Some of the 2016 images were cropped to make the crowd appear larger.
  - How do we combine data from multiple images?

How Big is a Crowd?

  - Should we use many high-resolution images to capture the entire crowd?
  - Given an image, how do we count people? We can turn to the well-established field of image processing (segmentation, background correction, etc.).
  - How do we quantify accuracy? Nothing close to probability sampling here, so can't rely on usual statistical inference machinery.
- Prospective question: How can we predict the crowd size for an upcoming event?
  - Why is such information helpful? How accurate do we need to be, e.g., within 1% of true count or 2 persons from true count?

How Big is a Crowd?

  - Do a census: Try to count everyone. Easiest would be ticket-only attendance or controlled entry and counting at checkpoints. If this is not possible, partition the venue and count persons in each area (again, only snapshot in time).
  - Use proxies, e.g., collect images using helicopter or drones.
  - Carry out benchmarking experiments using "training data" to compare methods and assess accuracy.
- Events for which crowd size estimates are reported and debated: Women's March, March for Science, March for Life, Gilets Jaunes demonstrations.
- There is actually an area of research called Crowd Science (e.g., `http://www.gkstill.com`).

Getting the Question Right

- Recall: *"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."* (Tukey).
- Being a little off in formulating a question can lead to being a long way off at the end, as errors are propagated and can be amplified during the workflow with which we answer the question.
- One of the hardest and underestimated aspects of Applied Statistics, as well as Data Science, is to translate, when appropriate, a possibly vague domain question into a statistical inference question, i.e., a parameter to be estimated or for which to test hypotheses.

Getting the Question Right

- If this step is not done properly, we may be estimating the wrong parameter and be completely off in our answer.
- However, not all Data Science questions are about "formal" statistical inference (i.e., estimation or testing), far from it!
- Some questions are best answered by collecting appropriate data and providing effective numerical and graphical summaries of these data.
  E.g. Suppose we are interested in studying the demographics of UC Berkeley undergraduates over time. In this case, we have a census, i.e., we get to observe the entire population and there is no sampling. Hence, standard errors and p-values are meaningless. The challenge is in finding relevant variables to compare and effective numerical and graphical summaries of the data.

Getting the Question Right

- By definition, numerical and graphical data summaries are functions of the data, i.e., statistics.
  - When computing these summaries on data from a sample drawn from a population, we are therefore implicitly performing inference on the population.
  - However, this can be done quite effectively without making strong probabilistic statements about the distributions of these statistics, e.g., reporting p-values for a t-test.
  - Such statements can problematic in many situations, as they are only valid under certain assumptions, e.g., the data have a Gaussian distribution.
  - Distributional assumptions, are often hard to verify or unrealistic.
  - As a result, reported quantities such as p-values can be plain wrong, in addition to not being particular informative or easy to interpret.

Getting the Question Right

- The literature is full of articles providing precise, but not necessarily right, answers (i.e., p-values, risk) without formulating a question.
- Sophisticated methods (e.g., logistic regression, neural networks, deep learning, hidden Markov models, cross-validation) are applied without considering their appropriateness (i.e., scope, assumptions, limitations) or whether they actually answer the question of interest.
- It is often easier to focus on technical mathematical or computational details of a question (not necessarily the right question!), at the expense of losing the "soft" big picture questions and issues that are actually remarkably difficult to be precise about.

About the Answer

- An appropriate answer for a data-driven inquiry necessarily depends on the question and purpose of the study, i.e., what actions will be taken based on the answer.
  - Pilot study with software prototype, where answer is a new question leading to a new inquiry.
  - Study requiring polished, reusable, and extensible pipeline with reliable and efficient software implementation of analysis methods.
  - Study for which results will be used to set up clinical trials for a new drug.
- In particular, in term of statistical inference, the following issues should be considered.

About the Answer

- Loss function. What are the costs of the different types of errors? This is connected to ethics.
  E.g. In the case of the COMPAS algorithm, both false positives (jailing an individual that wouldn't have recidivated to jail) and false negatives (not jailing an individual that would have recidivated) have serious real-life implications.
- Domain significance vs. statistical significance. One can have a statistically highly significant result for an insignificant difference in the domain.
  E.g. A tiny p-value for a drop of 0.01 mmHg in systolic blood pressure with a new drug.
- Accuracy, precision, and bias.
  - Accuracy measures how close on average an estimator is to the true value of the parameter (i.e., what we want to estimate or learn).

About the Answer

- Precision measures how variable an estimator is around its average (not the parameter!).
- Bias measures how close the average of an estimator is to the parameter.
- Here, averages and variability are over repeated sampling from the population. We will discuss these issues in detail in upcoming lectures.
- One can be very precise about a completely wrong answer!

Accuracy, Precision, and Bias

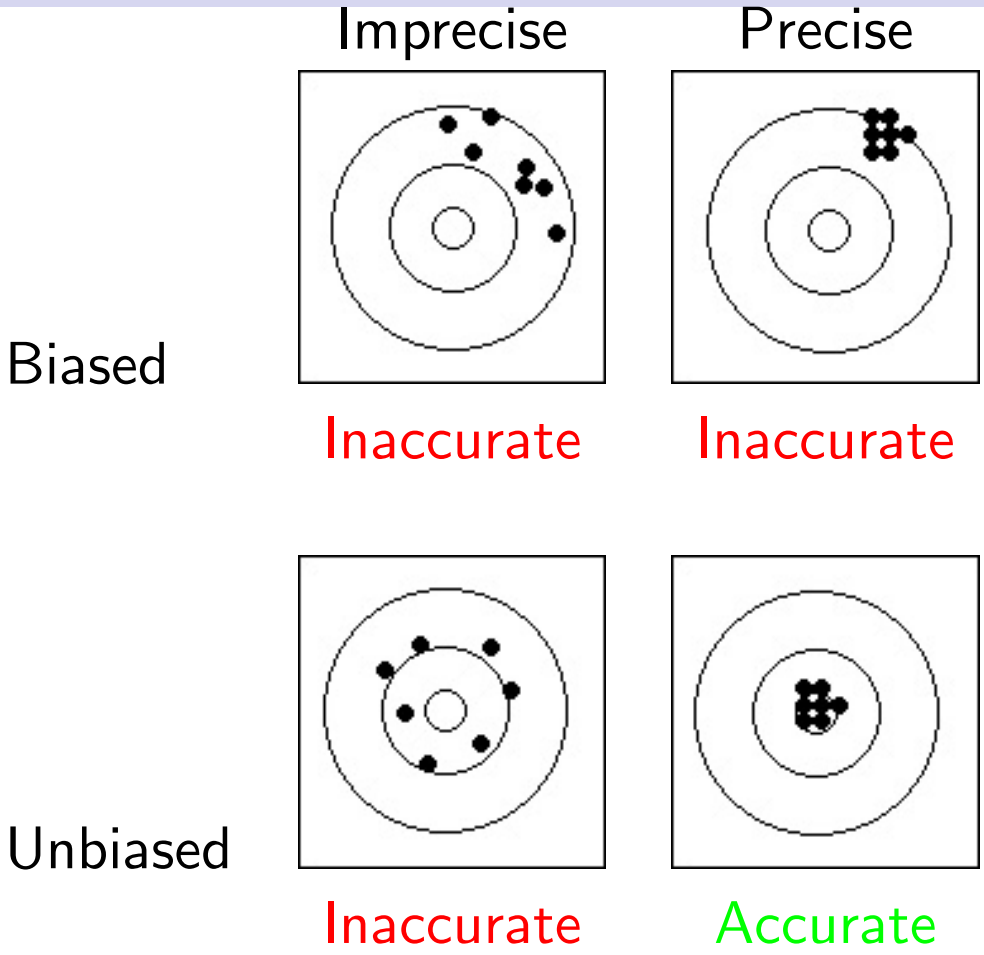


Figure 2: *Accuracy, precision, and bias.*
`https://cals.arizona.edu/classes/rnr613/accuracy.html`.

Data-Driven Study Design

- Data-driven study design is broadly concerned with designing the workflow for a particular data-enabled inquiry (cf. first lecture), i.e., laying out precisely each step (sequential and transversal) in this workflow.
- Study design is more about **what** to do, rather than **how** to do it.
- In particular, when it comes to statistical inference, you should be an informed and cautious user of methodology. That is, it is more important to know which methods are appropriate (in terms of their scope, assumptions, interpretation of results, and pros and cons compared to other methods) than focus on their mathematical or computational details (i.e., what's under the hood). Leave that to the theoretical statisticians!

Data-Driven Study Design

- Study design is essential to avoid or mitigate problems such as the collection of irrelevant, biased, or erroneous data and ensure that the question of interest can be answered as accurately as possible given available resources (e.g., biological specimen, time, money).
- The examples in Section 3, below, are all studies that led to "bad data" for a variety of reasons falling under the broad purview of study design:
  - Framing questions,
  - identifying what data to collect,
  - survey sampling design,
  - managing data,
  - computational reproducibility and verifiability,
  - research integrity.

Team Science

- Some data-driven projects can be carried out by a single investigator.
  E.g. Class project with existing data (e.g., Craigslist rental listings, Gapminder data); pilot study on specific aspects of the workflow or with software prototype; following up on a news report that piqued your attention by collecting and examining data.
- However, because of their scope, scale, and the type of output required, many projects require the breadth of depth of a team of investigators.
  E.g. Investigating the effectiveness of a malaria vaccine; cosmology project involving data collection from ground- and space-based telescopes.

Team Science

- An initial and key aspect of study design is therefore to assemble the right team to design and carry through the Data Science workflow, from A to Z, i.e., from framing the question to translating, interpreting, and implementing (into actions) the results of the study.
- The team and design of course depend on the type of question and answer needed. It typically is interdisciplinary and includes computer scientists, statisticians, and domain experts.
- Different members of the team will focus on different aspects of the workflow, to reflect differences in expertise emphasis, but they should work in a coordinated manner, i.e., communicate and provide feedback to each other, to allow proper flow and iteration of the sequential aspects and integration of the transversal aspects.

Workflow Design

The following transversal aspects of the Data Science workflow should be addressed at the study design stage.

- Making decisions regarding computing and data technologies.
- Ensuring computational reproducibility and verifiability of the analysis workflow and results.
- Handling matters related to research responsible conduct and integrity, ethics, privacy, security, and governance.

Workflow Design

- Some of the "softer" study design issues (e.g., framing questions, project management) are typically neither addressed by Statistics, nor CS, nor domain disciplines. However, they are very much part of Data Science and good research practice.
- This reflects the transdisciplinary nature of Data Science and the fact that it is fundamentally distinct from CS, Statistics, and domain disciplines.

Workflow Design

- Note that Statistics traditionally adopts a much narrower view of study design, which is concerned primarily with procedures for data collection and how these relate to the subsequent optimal inference step (e.g., survey sampling, randomized controlled trials). That is, the data and parameters of interest are already defined and one seeks optimal designs, i.e., designs that minimize variance, minimize risk, or maximize power given available resources (e.g., sample size).

Workflow Design

How would you design the workflow for some of the data-driven projects we've discussed? What expertise is required? What sort of team would you build?

Getting the Data Right

As data are inherently at the core of Data Science, it is essential to carefully consider what data to collect, how to collect these data, and whether the data are dependable, in order to ensure that the question of interest can be answered properly. That is, we should make sure we "know our data". Study design therefore concerns, among other things,

- data relevance, i.e., are the data pertinent for addressing the question;
- data provenance, i.e., where do the data come from;
- data reliability, i.e., can the data be used and trusted.

As we'll see below, these three notions are closely related and often hard to separate.

Getting the Data Right

- What data to collect, i.e., what to measure.
  - What is the unit of observation?
  - What are the variables to be measured/recorded on each unit?
- What is the population of interest, i.e., the scope of the inquiry.
  E.g. If we want a new blood pressure drug to be effective and safe for both men and women and also underrepresented minorities (URM), then the clinical trial shouldn't enroll only white men.

Getting the Data Right

There are ethical, practical, financial, and scientific constraints on the data that may be collected.

- The variable(s) of interest may not be measurable or easily measurable.
  E.g. Administering experimental treatment to human subjects, destructive measurement process, subjective property (e.g., "intelligence").
- Instead, one may use proxy variables or proxies, i.e., an easily measurable variable related to the variable of interest.
  E.g. Work with "model organisms" (e.g., yeast, mice) instead of humans; use per-capita gross domestic product (GDP) as a proxy for standard of living or quality of life;

Getting the Data Right

use years of education, GPA, or IQ as proxies for cognitive ability; use images for crowd size.

- Most of the time, one cannot collect data for the entire population of interest.
  - Instead, one obtains data for a sample (i.e., subset) drawn from this population.
  - The sample is, in some sense, a proxy for the population.
  - This is where statistical inference comes into play: How to use the sample to learn about the population.

Getting the Data Right

- Census. One can collect data on the entire population of interest.
  - In this case, there is no variability due to sampling, i.e., standard errors and p-values are meaningless.
  - However, there are still many non-trivial and important issues, that are very much within the scope of Data Science.
    E.g. Determining relevant variables, appropriate numerical and graphical data summaries.
- Sample. The sample should be representative of the population and selected according to well-defined probabilistic procedures to allow assessment of the accuracy of the answer, cf. estimator bias, standard errors.

Getting the Data Right

- Found data. Found data are data that were collected for some other purpose and are available to the investigators. E.g. Open-access Web databases, administrative datasets.
  - Using found data can be tricky.
  - It is often unclear why and how the data were collected and how reliable they are.
  - The data are not necessarily what we need, i.e., the right variables or from the right population.
  - The data are not necessarily a sample from a population of interest and, if a sample, they were not necessarily obtained according to well-defined probabilistic procedures, thus making statistical inference problematic.
  - Worthwhile and non-trivial analyses can still be performed, by focusing on relevant aspects of the data and using appropriate numerical and graphical data summaries.

Getting the Data Right

How to collect the data. Having specified the relevant data, one must determine the process for obtaining these data. There are a broad range of issues and approaches, including:

- Designing the data collection procedure, e.g., survey/questionnaire, sampling scheme, or randomized controlled experiment.
- Collecting the data. Acquiring available data/found data (e.g., Webscraping), generating new data (e.g., sensors), and fusing/merging data sources (e.g., record linkage).

Entire courses could be devoted to each of these topics, which involve Statistics, data technologies, and domain expertise.

Getting the Data Right

Data on data.

- Problems with the data are propagated and can be amplified during the workflow: Garbage in, garbage out.
- Study design should therefore involve the inclusion of quality and sanity checks throughout the workflow.
- We should collect "data on data" or metadata.

Getting the Data Right

- Controls.
  - Controls (a.k.a., standards) are observations or variables with known values/behavior.
  - Controls are useful for quality assessment/control (QA/QC), instrument calibration, normalization of measures across observations or variables, benchmarking methods, and validating results.
  - Controls can be positive or negative, according to whether they vary in value or not.
  - Controls can be internal or external, depending on whether they are obtained along with the main data or from another source.

Getting the Data Right

- E.g. Control observations. When assessing the effect of a "treatment" on an outcome (e.g., new drug for blood pressure), a control group is typically included to minimize the effects of variables other than the treatment (cf. confounding[1]). The control group should be similar to the treatment group with respect to all relevant variables except the treatment.
- E.g. Control variables. When seeking to identify genes that are differentially expressed (DE) between different types of cells (e.g., disease vs. healthy cells, cells in embryogenesis) using high-throughput transcriptome sequencing (RNA-Seq), positive and negative controls can be obtained by spiking in synthetic RNA sequences at known concentrations in the samples to be sequenced. The controls can then be used to compare DE methods or validate the results of a particular method using receiver operating characteristic (ROC) curves (i.e., plots of true

Getting the Data Right

positive rate vs. false positive rate, https://en.wikipedia.org/wiki/Receiver_operating_characteristic).

- Variables describing the data.
  - Variables describing how the data were collected and processed are useful for the purpose of QA/QC, data normalization, and validation and interpretation of results.
  - For instance, one should record variables such as batch (i.e., day of data collection, run of instrument, interviewer/lab technician), in order to detect and adjust for nuisance/unwanted effects on the measures of interest and examine possible confounding.

Getting the Data Right

  - E.g. In genomics, the biological effects of interest are often smaller than nuisance technical effects, e.g., from experimental protocol, instrument (cf. "technical noise"). If "treatment" and "control" samples are run in separate batches, then biological and technical effects are confounded, i.e., there is no way to tell whether a difference between the two groups is due to biology (interesting!) or just technical effects (uninteresting!), no matter how precise your answer.

- Model diagnosis. As part of exploratory data analysis (EDA) and optimal statistical inference, we should assess whether a model is appropriate for the question and data using, e.g., residual plots, simulation.
- "Look at data", at each step of the workflow!

Getting the Data Right

- The above steps are highly valuable for detecting gross errors, as well as more subtle problems with data.
  E.g. Mislabeled samples (species labels reversed), failed run of an instrument (microarray printer crashing), bugs in code – I have had to deal with all of these issues in my research!
- They are also helpful in separating signal from noise.
- Think ahead, it will make your life easier later on.

[1] A confounding variable is a variable that has an effect on both variables of interest and causes a spurious association between them. E.g. association between murder rate and sale of ice cream, with weather as confounding variable.

Bad Data

- No matter how "big" or costly the data, a poorly designed study can lead to "bad" data and hence inaccurate/plain wrong answers or no answer at all.
  Big bad data $\longrightarrow$ Garbage in, garbage out (GIGO).
- The data can be "bad" or "dirty" for a variety of reasons, including, falsification (cf. integrity), recording errors, missing values, difficulty of access and manipulation, being outdated (cf. versioning and reproducibility issues), wrong variables measured, poor proxies, confounding, sampling bias.

Bad Data

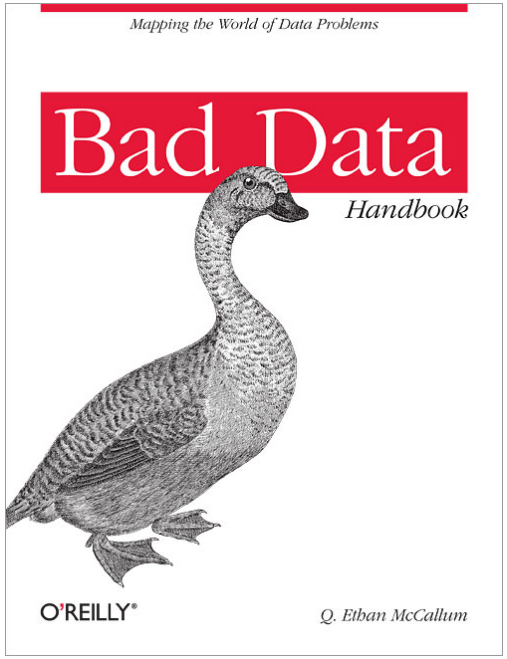


Figure 3: *Bad data.* Q. E. McCallum (2012). *Bad Data Handbook*, O'Reilly (`https://www.oreilly.com/library/view/bad-data-handbook/9781449324957/`). *"Bad Data is data that gets in the way." "... missing values, malformed records, and cranky file formats ..." "... data that you can't access, data that you had and then lost, data that's not the same today as it was yesterday ..."*

Bad Data

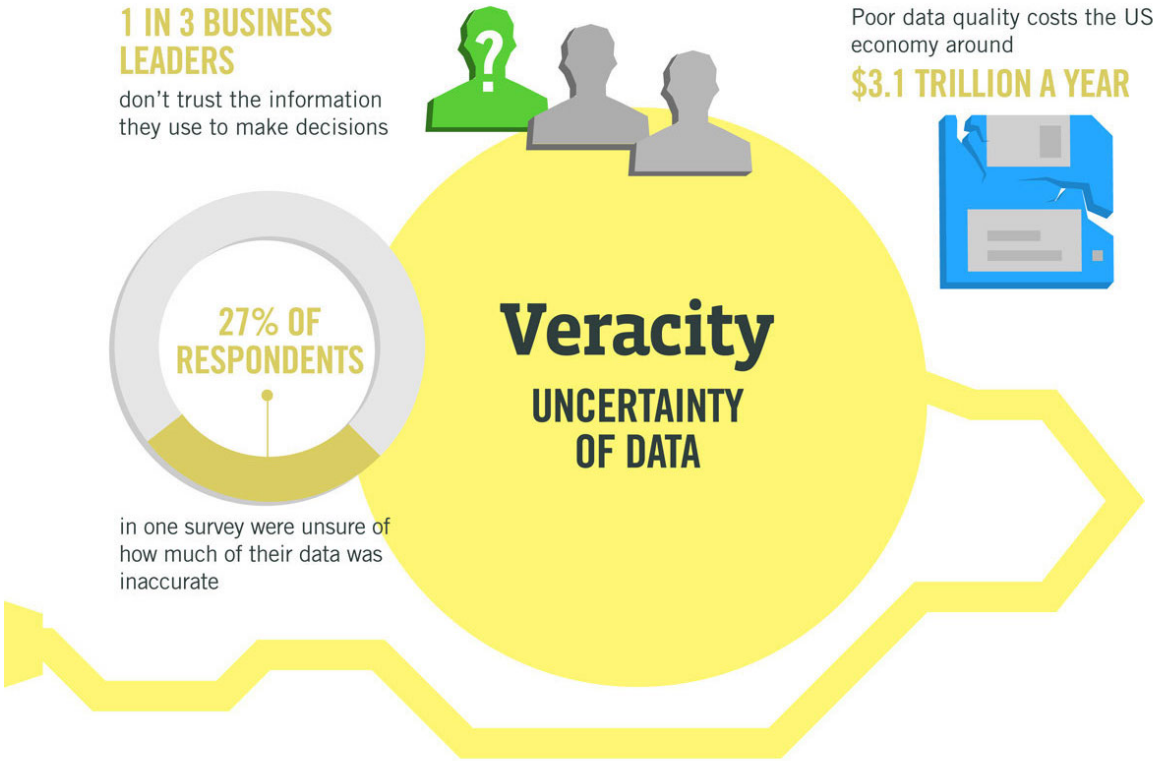


Figure 4: *Four V's of Big Data: Veracity.* `https://www.ibmbigdatahub.com/infographic/four-vs-big-data`.

Bad Data

- "Bad Data Costs the U.S. $3 Trillion Per Year" (`https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year`), based on IBM figure (`https://www.ibmbigdatahub.com/infographic/four-vs-big-data`).
  "The reason bad data costs so much is that decision makers, managers, knowledge workers, data scientists, and others must accommodate it in their everyday work. And doing so is both time-consuming and expensive. The data they need has plenty of errors, and in the face of a critical deadline, many individuals simply make corrections themselves to complete the task at hand. They dont think to reach out to the data creator, explain their requirements, and help eliminate root causes."

Bad Data

- It has been estimated that 60% of data scientists spend most of their time cleaning and organizing data (CrowdFlower; `https://visit.figure-eight.com/data-science-report?utm_source=internal%20referral&utm_medium=email&utm_campaign=data%2520science%2520report`).
- Data Science on Data Science: Actually, how reliable are the above figures? ☺

Study Design and Bad Data

- Proper study design is essential to avoid bad data and the resulting costs, in terms of erroneous and potentially dangerous conclusions (e.g., putting patients at risk) and wasted time and money.
- Although some amount of data cleaning is unavoidable, proper study design can substantially reduce the burden. More about data cleaning in upcoming lectures.

Sampling Bias in Political Polls: 1936 Roosevelt vs. Landon

1936 Roosevelt vs. Landon Literary Digest Poll.

- About 40 million voters expected for the 1936 presidential election.
- The Literary Digest magazine sent out 10 million mock ballots to poll voters and received back 2.4 million.
- The poll predicted Alfred Landon's victory, but Franklin Roosevelt ended up winning the election with 24% more of the popular vote.
- The 10 million voters who received mock ballots were not representative of the electorate; they were drawn from wealthier voters.
- The 2.4 million who responded were not representative of those who received mock ballots; they were more passionate voters and hence more likely to respond.

Sampling Bias in Political Polls: 1936 Roosevelt vs. Landon

Table 1: *1936 Roosevelt vs. Landon presidential election: Popular vote.*

	Landon (Rep)	Roosevelt (Dem)
Predicted	57%	43%
Actual	38%	62%

Sampling Bias in Political Polls: 1948 Truman vs. Dewey

1948 Truman vs. Dewey Gallup Poll.

- In an attempt to avoid previous mistakes, the Gallup Poll used quota sampling to predict the results of the 1948 presidential election.
- Demographic strata (by gender, age, ethnicity, and income level) were defined based on the US census.
- Each interviewer polled a set number of people or quota from each stratum. Interviewers were told that they could additionally interview whomever they wished, as long as they fulfilled their quotas. However, Republicans were overrepresented among additional polled individuals, as they were easier to interview (e.g., lived in nicer neighborhoods).

Sampling Bias in Political Polls: 1948 Truman vs. Dewey

- The Gallup Poll predicted that Thomas Dewey would earn at least 5% more of the popular vote than Harry Truman would. Truman ended up winning by more that 4%.

Table 2: *1948 Truman vs. Dewey presidential election: Popular vote.*

	Dewey (Rep)	Truman (Dem)
Predicted	49.5%	44.5%
Actual	45.1%	49.6%

Sampling Bias in Political Polls: 1948 Truman vs. Dewey

Figure 5: *1948 Truman vs. Dewey presidential election: Truman with Chicago Daily Tribune front page.*

Sampling Bias in Political Polls

- The fundamental problem: Sampling/selection bias, i.e., samples that are not representative of the population in that certain types of individuals are overrepresented and others underrepresented.
- Both the Literary Digest and Gallup polls made the mistake of assuming that their samples were representative of the US voting population. However, their sampling schemes based on human judgment lead to selection bias in favor of Republicans.
- Polls now typically rely on probability sampling, i.e., methods that assign a precise probability to the event that each particular sample is drawn, to reduce bias as much as possible in the data collection process.
- However, probability sampling is not foolproof.

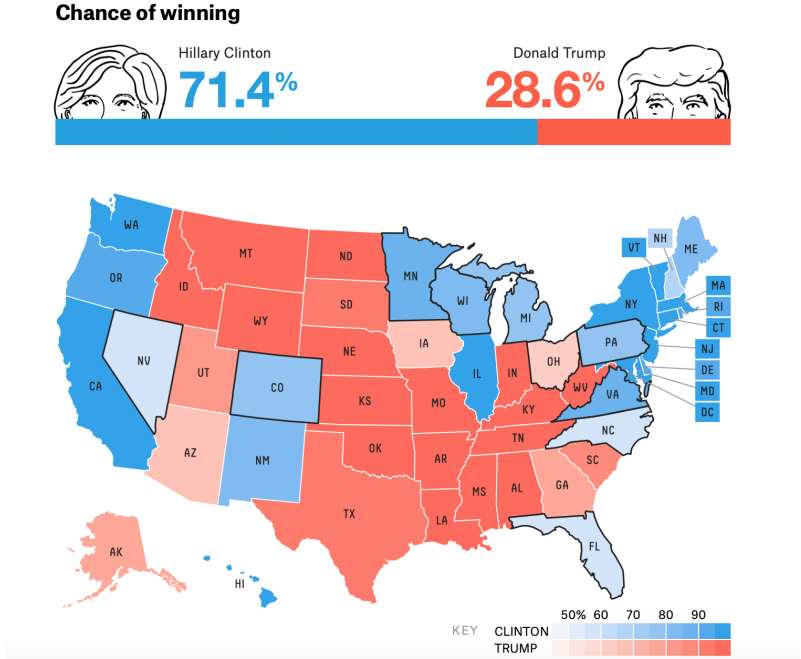
Sampling Bias in Political Polls: 2016 Trump vs. Clinton

2016 Trump vs. Clinton polls.

- Donald Trump's 2016 election victory took many by surprise, as most of the polling had suggested a victory for Hillary Clinton.

Sampling Bias in Political Polls: 2016 Trump vs. Clinton

- Five Thirty Eight election forecast: `https://projects.fivethirtyeight.com/2016-election-forecast/`.



Sampling Bias in Political Polls: 2016 Trump vs. Clinton

- A report from the American Association for Public Opinion Research (AAPOR; `https://www.aapor.org/education-resources/reports/an-evaluation-of-2016-election-polls-in-the-u-s.aspx`) notes three main reasons as to why the polls underestimated support for Trump.
  - Many polls did not adjust for overrepresentation of college graduates, i.e., sampling bias.
  - Real change in vote preference during the final weeks of the campaign.
  - Late-revealing trump voters.

Sampling Bias in Political Polls: 2016 Trump vs. Clinton

- We revisit the following analysis concerning sampling bias in the 2016 Election polls: `http://bit.ly/2nxIHbn`, `https://www.ru.nl/sociology/mt/sig/downloads/`.
- How can we use the actual election results to check whether polls could have predicted the right outcome?
- For simplicity, let's assume polls are done through simple random sampling (SRS), i.e., sampling at random without replacement from the population of interest.
- Actually, what is the population of interest when it comes to predicting the outcome of the presidential election?

Sampling Bias in Political Polls: 2016 Trump vs. Clinton

Table 3: *2016 Election cerfified results.* Percentage of votes in swing states won by Trump.

State	Trump %	Clinton %	Other %
Florida	47.8	49.0	3.2
Michigan	47.3	47.5	5.2
Pensylvania	47.9	48.6	3.6
Wisconsin	46.5	47.2	6.3

Sampling Bias in Political Polls: 2016 Trump vs. Clinton

- Polls based on simple random sampling from the population of voters in swing states can be simulated by sampling from a multinomial distribution with probabilities set to the true actual proportions of votes for each candidate from certified election results.
- How do we aggregate the poll results at the state level to predict the outcome of election?
- For samples sizes n ranging from 100 to 5,000, we perform 1 million such simulated polls and record the proportion of polls declaring each candidate a winner.

Sampling Bias in Political Polls: 2016 Trump vs. Clinton

- In order to investigate the effect of small sampling bias on the polling results, we perform the same simulation with a 1% bias in favor of Clinton (i.e., add 0.01 and subtract 0.01 from the proportion of Clinton and Trump votes, respectively y).
- The multinomial distribution is a generalization of the binomial distribution. In terms of a box model, it corresponds to drawing n tickets at random with replacement from a box with K types of tickets ($K = 2$ for binomial). If we let π_k denote the proportion of tickets of type k in the box, then the chance of drawing x_k tickets of type k in the sample ($\sum_k x_k = n$), for each $k = 1, \ldots, K$, is

$$\frac{n!}{x_1! \ldots x_K!} \pi_1^{x_1} \ldots \pi_K^{x_K}. \quad (1)$$

Sampling Bias in Political Polls: 2016 Trump vs. Clinton

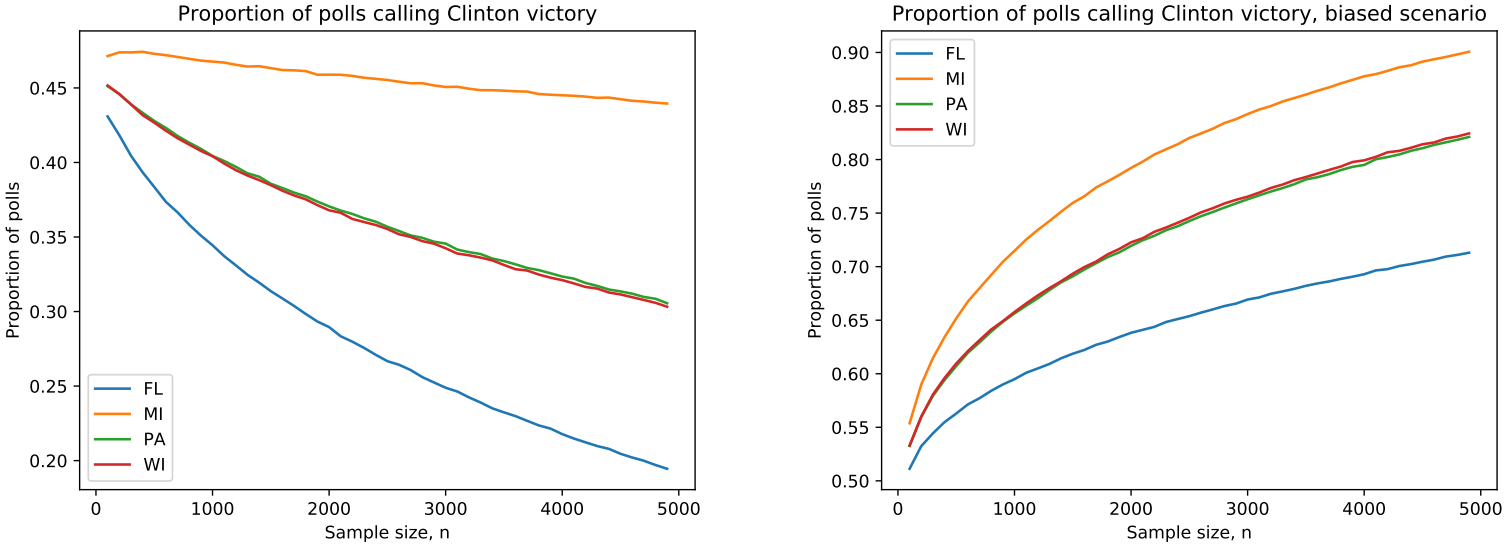


Figure 6: *2016 Election polls: Simulation study.* Proportion of polls declaring Clinton a winner, by state. Left: True vote proportions. Right: Biased vote proportions in favor of Clinton.

Sampling Bias in Political Polls: 2016 Trump vs. Clinton

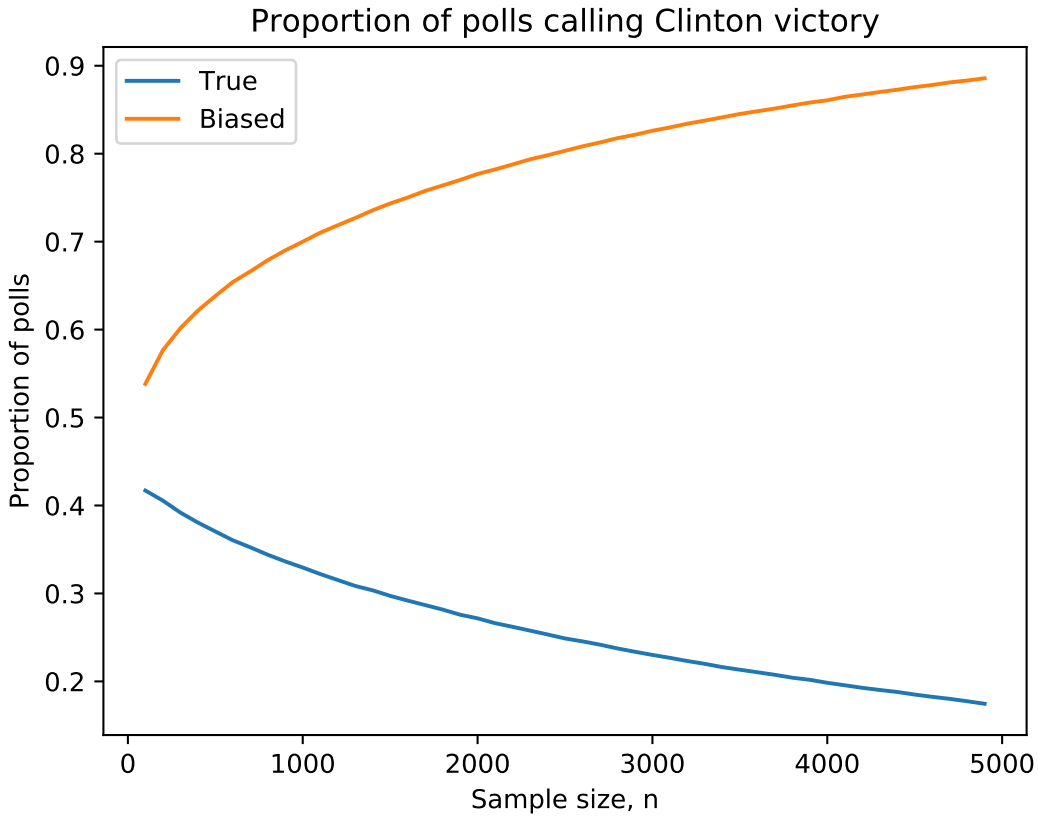


Figure 7: *2016 Election polls: Simulation study.* Proportion of polls declaring Clinton a winner.

Sampling Bias in Political Polls: 2016 Trump vs. Clinton

- With the true actual vote proportions (i.e., representative samples), SRS predicts a Trump victory with samples sizes as low as $n = 100$.
- With a 1% bias in favor of Clinton (i.e., sampling bias), SRS predicts a Clinton victory.
- Sampling bias is not corrected by getting more data, on the contrary! The polls are more and more precise about a wrong answer.
- In general, as sample size increases, precision increases, but not necessarily accuracy. This is because of bias.

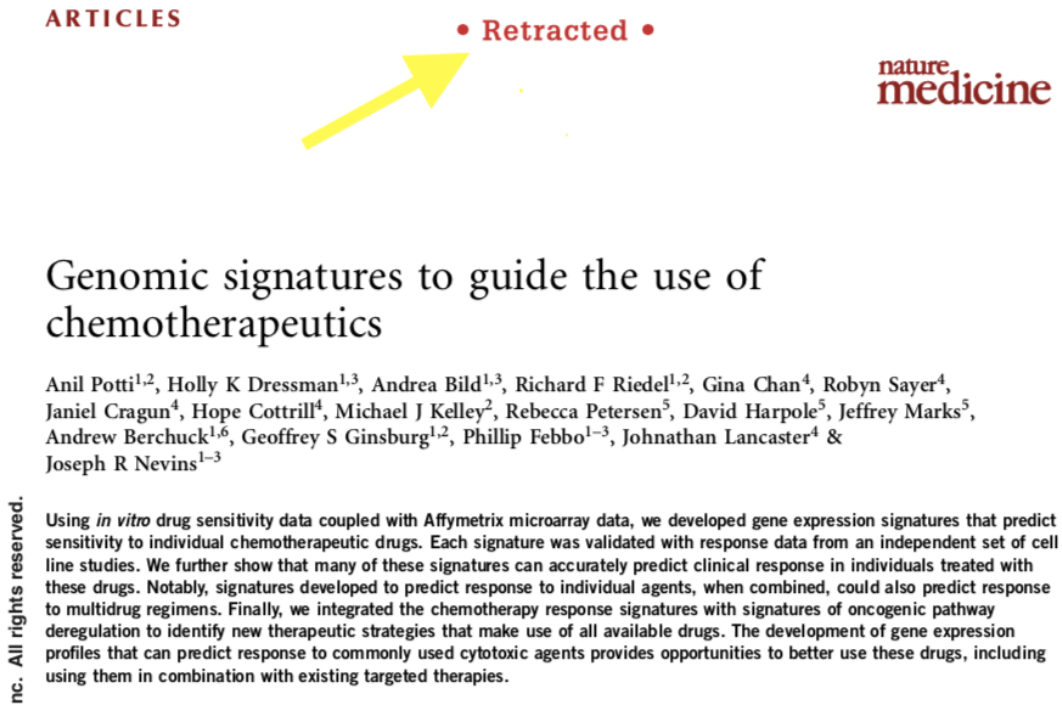
After the Election: Election Audits

- There is still much to be done after an election.
- Election audits are reviews conducted after the polls are closed to determine whether the electoral outcome is correct.
- Voting practices vary greatly by country and even by state within the US. They include paper and electronic ballots. Likewise for methods for reading and counting the ballots.
- Accordingly, there are a variety of considerations and approaches for designing audits.
  E.g. Choosing the number of ballots to audit, selecting/finding these ballots, determining when the audit can stop.

After the Election: Election Audits

- A risk-limiting audit (RLA) is a procedure that is guaranteed to have a large chance of progressing to a full hand count of the votes if the electoral outcome is wrong. The outcome according to the hand count then replaces the outcome being audited. (`https://www.stat.berkeley.edu/~stark/Vote/auditTools.html`).

Duke Personalized Medicine Scandal



ARTICLES

• Retracted •

nature medicine

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

nc. All rights reserved.

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic strategies that make use of all available drugs. The development of gene expression profiles that can predict response to commonly used cytotoxic agents provides opportunities to better use these drugs, including using them in combination with existing targeted therapies.

Figure 8: *Duke personalized medicine scandal.* https://www.nature.com/articles/nm1491.

Duke Personalized Medicine Scandal

- Potti et al. (2006) proposed an approach for personalized medicine, whereby patient sensitivity to chemotherapeutic drugs is predicted based on *in vitro* drug sensitivity and microarray[2] gene expression measures.
- This high-profile study led to follow-up studies and, more critically, the enrollment of patients into clinical trials.
- Coombes et al. (2007) and Baggerly and Coombes (2009) discuss their failure to reproduce the results in Potti et al. (2006), despite using the same data and software (`https://bioinformatics.mdanderson.org/supplements/reprorsch-all/`). They report a variety of simple, but serious errors in data that invalidate the conclusions of the study. More generally, they advocate the use of software

Duke Personalized Medicine Scandal

such as R's Sweave to facilitate computationally reproducible research.

- *"We do not believe that any of the errors we found were intentional. We believe that the paper demonstrates a breakdown that results from the complexity of many bioinformatics analyses. This complexity requires extensive double-checking and documentation to ensure both data validity and analysis reproducibility. We believe that this situation may be improved by an approach that allows a complete, auditable trail of data handling and statistical analysis. We use Sweave, a package that allows analysts to combine source code (in R) and documentation (in LaTeX) in the same file. Our Sweave files are available at (`http://bioinformatics.mdanderson.org/`*

Duke Personalized Medicine Scandal

supplements/reprorsch-chemo). *Running them reproduces our results and generates figures, tables and a complete PDF manuscript."* (Coombes et al., 2007)

- Lack of proper documentation of the study workflow lead Baggerly and Coombes to painstaking exercises in *"*"*forensic bioinformatics*" *where aspects of raw data and reported results are used to infer what methods must have been employed"* (Baggerly and Coombes, 2009).
- In their investigation, they noted a variety of errors in data, including mislabeled samples (drug sensitivity status reversed!), duplicated or triplicated samples, and several incompatible versions of the same data.
  "One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is

Duke Personalized Medicine Scandal

our experience that the most simple errors are common." (Baggerly and Coombes, 2009)

- No matter how appropriate or sophisticated the downstream inference methods (here, principal component analysis, Bayesian probit regression, cross-validation), these errors in data invalidate the conclusions of the study and put patients at risk.
- Potti was eventually found guilty in 2015 of "research misconduct", including data falsification, by the Office of Research Integrity (ORI) (`https://www.federalregister.gov/documents/2015/11/09/2015-28437/findings-of-research-misconduct`).

Duke Personalized Medicine Scandal

- This mediatized scientific saga led to inappropriate enrollment of patients in clinical trials, premature launch of companies, and retraction of dozens of research papers (original 2006 article retracted in 2011).
- In addition to the dire real-life consequences of putting patients at risk, this scandal also raised key general issues about the conduct of research, i.e., the Data Science workflow, including, scientific ethics and integrity, data reliability, and computational reproducibility. Addressing these issues is part of study design.

Duke Personalized Medicine Scandal

- The Duke scandal and Baggerly and Coombes' work gave impetus for the adoption of computationally reproducible research practices and, in particular, tools such as Make, Git, Sweave (Leisch, 2002), knitr (Xie, 2013), and Jupyter (`https://jupyter.org/`).

[2] Microarrays are high-throughput biological assays that allow the simultaneous measurement of gene expression/transcription levels for entire genomes.

When Contact Changes Minds

When contact changes minds: An experiment on transmission of support for gay equality.

- In their highly-publicized *Science* article, LaCour and Green (2014) conclude that a single conversation with canvassers could change the minds of voters on divisive social issues such as same-sex marriage.
- They base this conclusion on a "randomized placebo-controlled trial" ... which turned out never happened!
- The data were found to have been fabricated by LaCour. They appear to have been simulated by adding Gaussian noise to data from a previous study.

When Contact Changes Minds

- Details on story: `https://en.wikipedia.org/wiki/When_contact_changes_minds`.
  Original article, now retracted: `http://science.sciencemag.org/content/346/6215/1366`.

Probability Review

- Many fundamental aspects of Data Science, including study design, involve chance/randomness and uncertainty.
- Probability Theory allows us to characterize randomness and quantify uncertainty.
- In this section, we survey probabilistic designs for data collection.
- For a review of Probability Theory:
  - Freedman et al. (2007), Part IV: `https://books.wwnorton.com/books/webad.aspx?id=11597`;
  - Lau et al. (2019): `https://www.textbook.ds100.org/ch/02/design_prob_overview.html`;
  - Adhikari and Pitman (2019): `http://prob140.org/textbook/`.

Probability Review

- For foundations on probabilistic designs: Freedman et al. (2007), Chapters 1 and 2 and Part VI: `https://books.wwnorton.com/books/webad.aspx?id=11597`.

Survey Sampling

Survey sampling. The purpose of a survey is to measure characteristics and/or attitudes of a population, e.g., via a questionnaire. Survey sampling is the process of selecting a sample of elements from a target population and recording variables of interest on these elements.
E.g. Election polls: Sample voters and record preferred candidate.

- Self-selected sample. Sample is whoever chooses to answer.
- Convenience sample. Sample is whomever/whatever is convenient for investigator.
- Judgment sample. Sample is whomever/whatever investigator deliberately selects.

Survey Sampling

- Probability sample. Sample is selected based on probabilistic procedure.

Survey Sampling

- Unlike the other three forms of sampling, probability sampling allows assigning a precise probability to the event that each particular sample is drawn from the population.
- This allows to quantify uncertainty/confidence about an estimator, prediction, or hypothesis test.
- Be suspicious whenever standard errors, p-values, or confidence levels are reported without a proper explanation of the sampling procedure. They could be meaningless or seriously wrong.
- Entire courses are devoted to survey sampling (e.g., STAT 152). Below is a very brief review of basic sampling approaches.

Simple Random Sampling

- A useful representation for sampling is a box model, where the population of interest is represented by a box of N tickets, each with values written on them (the data!).
- A simple random sample (SRS) of size n is obtained by drawing n tickets at random **without** replacement from the box.
- For a small sample compared to the population, SRS is very close to sampling at random **with** replacement.

Simple Random Sampling

- How many ways are there so select an SRS of size n from a population of size N?

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

- What is the chance that a particular element of the population is selected by an SRS?

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}}.$$

Cluster Sampling

- In cluster sampling, the population is divided into clusters of individuals. One then uses SRS to select entire clusters instead of individuals.
- Cluster sampling makes data collection easier. For example, it is much easier to poll entire towns of a few hundred people each than to poll thousands of people distributed across the entire US. This is why many polling agencies use forms of cluster sampling to conduct surveys.
- The main downside of cluster sampling is that it tends to produce greater variation in estimation. This typically means that we need to take larger samples than with SRS.

Stratified Sampling

- In stratified sampling, the population is divided into strata of individuals, e.g., based on demographics. We then select SRS of individuals in each stratum.
- In both cluster sampling and stratified sampling the population is split into groups of individuals; in cluster sampling we use a single SRS to select groups, whereas in stratified sampling we use one SRS per group to select individuals.
- Stratified sampling results in increased precision and representation.

Stratified Sampling

- Stratified sampling could be viewed as a proper way to conduct quota sampling. It allows the investigator to ensure that subgroups of the population are well-represented in the sample without using human judgment to select the individuals in the sample.
- Stratified sampling can be difficult to conduct in practice because we may not know how large each stratum is. In some cases, we can take advantage of US census data to define the strata.

Survey Sampling

Figure 9: *Cluster and stratified sampling.*
`https://keydifferences.com/difference-between-stratified-and-cluster-sampling.html`.

Designed Experiments

- Designed experiments are used to examine the association/effect (causal?) of a treatment[3] on an outcome when the variable(s) of interest is(are) under the control of the investigator, i.e., when the investigator can determine who/whet gets the treatment. E.g. Clinical trial to test effect of new drug on patients with Alzheimer's disease, A/B test for two versions of a website.
- A randomized controlled trial (RCT) is a type of designed experiment in which participants in the trial are randomly allocated to either the group receiving the treatment under investigation or a control group receiving standard treatment, no treatment, or a placebo.
- A RCT is often considered the gold standard for many types of investigations, e.g., clinical trials.

Designed Experiments

- Randomization is used to avoid sampling bias. It allows assessing the effect of the treatment compared to the control, while other variables are kept constant.
- There are different types of RCT designs (e.g., crossover, cluster, factorial) and random allocation in real trials can be complex.
- RCTs are often used to test the efficacy or effectiveness of various types of medical interventions and may provide information about adverse effects, such as drug reactions.

[3] Here, the term "treatment" is used broadly, to refer to the variable whose effect on an outcome is to be examined. The treatment could be a new drug in a clinical trial, an new type of fertilizer in an agricultural experiment, or a new marketing strategy in A/B testing.

Observational Studies

Observational studies are used to examine the association/effect (causal?) of a treatment on an outcome when the variable(s) of interest is(are) not under the control of the investigator.

E.g. Study effect of smoking on health.

- Case-control study. Two existing groups differing in outcome ("case" or "control") are identified/sampled and compared on the basis of variables potentially associated with the outcome.
- Cross-sectional study. Data are obtained at a specific point in time for each subject.
- Longitudinal study. Data are obtained at multiple timepoints for each subject.

Observational Studies

- Cohort study or panel study. A particular form of longitudinal study where a group of subjects is closely monitored over a span of time.
  E.g. Framingham heart study (`https://en.wikipedia.org/wiki/framingham_heart_study`).

A/B Testing

- A/B testing (a.k.a., split testing or bucket testing) is concerned with determining whether two samples were drawn from the same population, i.e., have the same data generating distribution.
- A/B testing is widely used in industry for marketing and website and mobile app design purposes, e.g., comparing two different types of subject headers in e-mailing campaigns, two different landing pages for websites.
- A/B tests are workhorses of conversion rate optimization (CRO), an Internet marketing practice whose goal is to increase the percentage of website visitors that convert into customers or, more generally, take any desired action on a webpage.

A/B Testing

- Google engineers ran their first A/B test in 2000 in an attempt to determine the optimum number of results to display on the search engine's results page.
- Although the applications (and the name) are new, there is really nothing new methodologically about A/B testing. The t-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guiness brewery and whose pseudonym was "Student".
- A/B tests typically focus on comparing means. In principle, one could test for any difference between the two distributions or a difference in terms of any parameter, e.g., mean, median, or variance. Different test statistics are appropriate for different purposes.

A/B Testing

- The two samples could be obtained in various ways, through randomization of subjects to the two treatment groups ("A" and "B") or sampling from two different populations.

A/B Testing

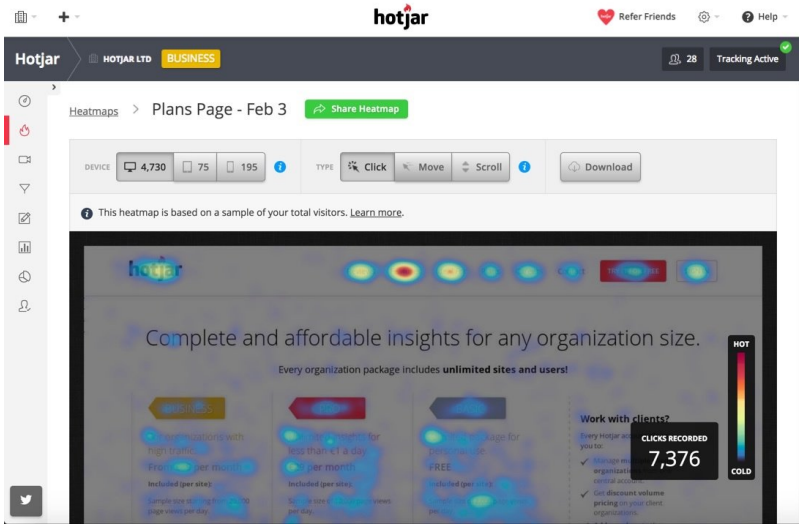
A/B test design issues.

- Variants to be compared.
  E.g. Website landing page, check-out page, product page. Can use click, move, and scroll heatmaps, which identify areas of visitor activity, to guide the identification of relevant variables.
- Relevant outcome.
  E.g. Revenue increase, customer conversion rate.
- Timing of the test.
- Number of tests to conduct. Cf. Multiple testing.
- How to select the two samples. A/B tests are not immune to sampling bias!

A/B Testing

Figure 10: *A/B testing and CRO.* `https://vwo.com/ab-testing/`.

A/B Testing



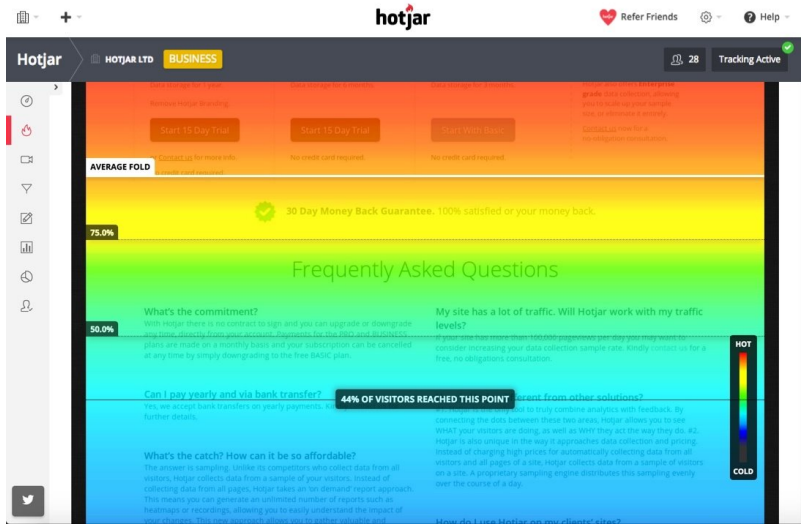


Figure 11: *A/B testing and CRO: Click and scroll heatmaps.* `https://www.hotjar.com/tour`.

A/B Testing

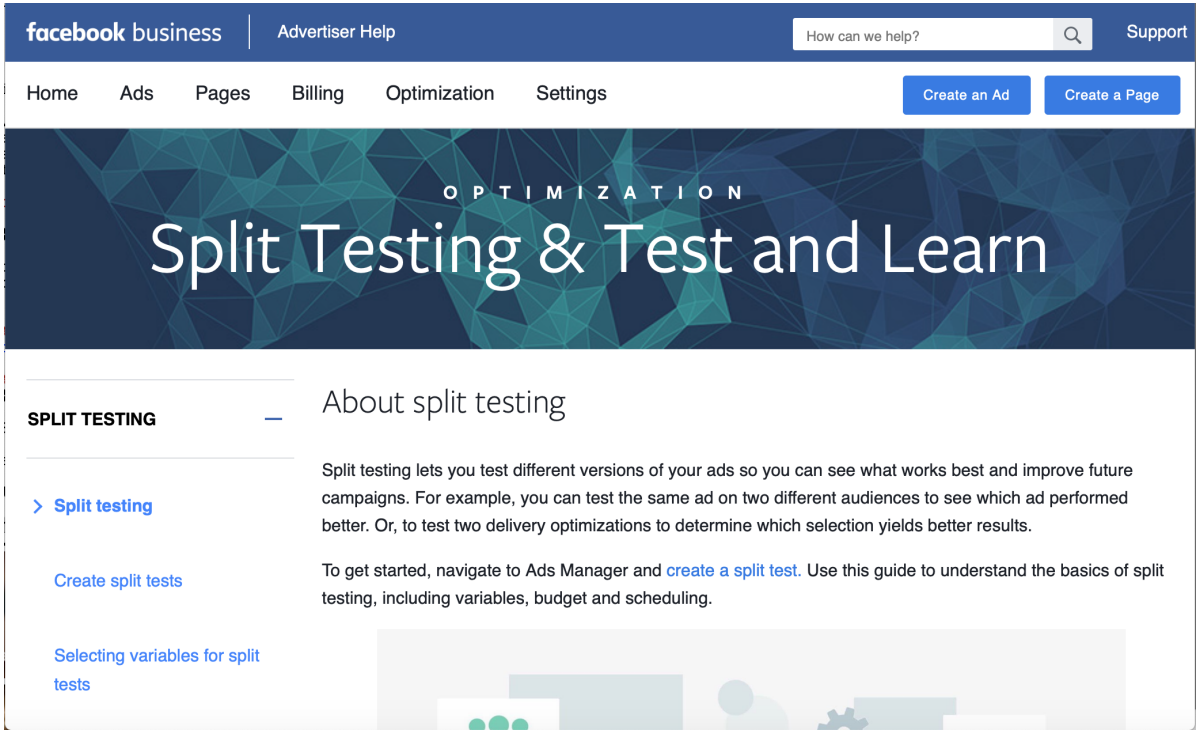


Figure 12: *A/B testing on Facebook.* https://www.facebook.com/business/help/1738164643098669.

A/B Testing

Caveat.

- Note that while two-sample t or z-statistics can be useful as descriptive summaries of the differences between the two samples, any probabilistic statements related to these as part of a t or z-test procedure (e.g., statistical significance based on p-values, Bayesian posterior probabilities) are only valid and meaningful to the extent that their underlying assumptions are satisfied.
- In other words, there is an important distinction between a test statistic and a full testing procedure which makes probabilistic statements about this statistic.

A/B Testing

- For instance, it makes no sense to go through the motions of hypothesis testing and focus on the probabilistic interpretation of p-values for a t-test when one has data for the two entire populations, when one has convenience samples, or when the data are far from Gaussian and the sample size is tiny.

References

A. Adhikari and J. Pitman. *Probability for Data Science*. 2019. URL http://prob140.org/textbook/.

K. A. Baggerly and K. R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, 3(4):1309–1334, 2009.

K. R. Coombes, J. Wang, and K. A. Baggerly. Microarrays: retracing steps. *Nature Medicine*, 13:1276–1277, 2007.

D. Freedman, R. Pisani, and R. Purves. *Statistics*. Norton, 4th edition, 2007.

M. J. LaCour and D. P. Green. When contact changes minds: An experiment on transmission of support for gay equality. *Science*, 364 (6215):1366–1369, 2014.

S. Lau, J. Gonzalez, and D. Nolan. *Principles and Techniques of Data Science*. 2019. URL https://www.textbook.ds100.org.

References

F. Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, Heidelberg, Germany, 2002. Physika Verlag. URL `www.ci.tuwien.ac.at/~leisch/Sweave`. ISBN 3-7908-1517-9.

A. Potti, H. K. Dressman, A. Bild, R. F. Riedel, G. Chan, R. Sayer, J. Cragun, H. Cottrill, M. J. Kelley, R. Petersen, D. Harpole, J. Marks, A. Berchuck, G. S. Ginsburg, P. Febbo, J. Lancaster, and J. R. Nevins. Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*, 12:1294–1300, 2006.

Y. Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2013. URL `http://yihui.name/knitr/`. ISBN 978-1482203530.