We are all statisticians now
I DON'T KNOW HOW TO DO STATISTICS BUT IT DOESN'T MATTER BECAUSE I DIDN'T HAVE DATA.
N = SAMPLE SIZE
N = \frac{(\$ \text{ YOU HAVE})}{(\$ \text{ PER SAMPLE})}
RNA-seq

2008 $N \approx 2$

2010 $N \approx 70$

2013 $N \approx 900$

PMIDS: 19056941, 20220758, 24092820
Discover your family tree, together

Start My Family Tree!
FamiLinx

Crowd-sourced genealogy for human genetics

FamiLinx provided by the Erlich Lab
Increased ciprofloxacin resistance in gonococci isolated in Scotland.

Forsyth A, Moyes A, Young H.

Abstract
A review of the susceptibility of Neisseria gonorrhoeae isolated from 4415 episodes of infection in Scotland between 1991 and 1999 showed that the proportion of isolates with lowered susceptibility (ciprofloxacin minimum inhibitory concentration [MIC] > or = 0.05 mg/L) increased from 0.5% in 1991 to 5% in 1999 (p<0.001), whereas the proportion of isolates with clinical resistance (ciprofloxacin MIC > or = 1 mg/L) was significantly higher in 1999 than the average for the preceding 4 years (2.2% vs 0.9%; p=0.02). Ciprofloxacin is a recommended treatment for gonococcal infection in the UK but if resistance continues to increase at the present rate it might not be suitable as a first-line treatment of gonorrhoea for much longer.

Comment in
Empirical estimates suggest most published medical research is true

Leah R. Jager, Jeffrey T. Leek

(Submitted on 16 Jan 2013)

The accuracy of published medical research is critical both for scientists, physicians and patients who rely on these results. But the fundamental belief in the medical literature was called into serious question by a paper suggesting most published medical research is false. Here we adapt estimation methods from the genomics community to the problem of estimating the rate of false positives in the medical literature using reported P-values as the data. We then collect P-values from the abstracts of all 77,430 papers published in The Lancet, The Journal of the American Medical Association, The New England Journal of Medicine, The British Medical Journal, and The American Journal of Epidemiology between 2000 and 2010. We estimate that the overall rate of false positives among reported results is 14% (s.d. 1%), contrary to previous claims. We also find there is not a significant increase in the estimated rate of reported false positive results over time (0.5% more FP per year, P = 0.18) or with respect to journal submissions (0.1% more FP per 100 submissions, P = 0.48). Statistical analysis must allow for false positives in order to make claims on the basis of noisy data. But our analysis suggests that the medical literature remains a reliable record of scientific progress.

Comments: 11 pages, 4 figures, Correspondance to J. Leek
Subjects: Applications (stat.AP)
Cite as: arXiv:1301.3718 [stat.AP]
(or arXiv:1301.3718v1 [stat.AP] for this version)

https://github.com/jtleek/swfdr
about me

research

blogging

teaching
about me

ttleek.com

simplystatistics.org

jhudatascience.org
July 17, 2012

Johns Hopkins University Partners with Coursera to Offer Free Online Classes

The Johns Hopkins University today announced that it has joined Coursera, an upstart education venture formed to offer high-quality college-level university courses online for free, creating new opportunities for learning worldwide.

Johns Hopkins is one of 17 top-tier institutions, including Princeton University, Stanford University, the University of Pennsylvania and the University of Michigan, that have signed agreements with Coursera to make some of their Web-based courses available to a wider student audience without charging tuition.

The universities are offering undergraduate and graduate courses taught by their professors in the arts, sciences, engineering, law, medicine, nursing and public health for the disciplines. The
from: jtleek@gmail.com

Roger let me know you gave him a ballpark figure for the number of students registered for his course "Computing for Data Analysis". Could you give me an idea of how many have registered for my course "Data Analysis?"
from: pangwei@coursera.org

Hi Jeff,

7,000 students! It's pretty awesome. (You'll be able to check this out yourself next week, once the class sites are up.)
from: rdpeng@gmail.com

You are f**ed.

-roger
9 classes
1 month long
Every month
Cumulative Enrollment

- Sep 2012
- Jul 2013
- Apr 2014
Dude, we get it data/statistics is everywhere

*(what is the big deal)*
Genomic signatures to guide the use of chemotherapeutics

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

Using \textit{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
DERIVING CHEMOSensitivity FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY* AND KEVIN R. COOMBES†

U.T. M.D. Anderson Cancer Center

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in “forensic bioinformatics” where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report, we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.
NORTH CAROLINA

DURHAM COUNTY

IN THE GENERAL COURT OF JUSTICE
SUPERIOR COURT DIVISION I
CIVIL CASE NO. DUR 11 CVS 04131

Richard Aiken, Jean K. Carroll, Executrix of the Estate of Harold G. Carroll, Jean K. Carroll, Individually, Peggy Cox, as Administratrix of the Estate of Paul F. Cox, Peggy Cox, Individually, Helene L. Fligel, Jason Gannon, as Personal Representative of the Estate of Jennifer L. Gannon, John Haddock, as Executor of the Estate of Karen Heath, Walter Jacobs, as Executor of the Estate of Juliet J. Jacobs, Walter Jacobs, Individually, Polly Johnson, as Executor of the Estate of Malcom W. Johnson, and Polly Johnson, Individually,
Plaintiffs

VS.

DEFENDANTS

COMPLAINT
(JURY TRIAL DEMANDED)
what went wrong?

2 things
what went wrong? transparency

The data/code weren’t reproducible
what went wrong? transparency

There was a lack of cooperation
what went wrong? expertise

They used silly prediction rules

\[ \Pr(\text{FEC}) = \frac{5}{8}[\Pr(\text{F}) + \Pr(\text{E}) + \Pr(\text{C})] - \frac{1}{4} \]
What went wrong? 

They had study design problems

(Batch effects)
what went wrong? expertise

Their predictions weren’t locked down

Today: $\text{Pr}(\text{FEC}) = 0.8$
Tomorrow: $\text{Pr}(\text{FEC}) = 0.1$
At the end of the day the Potti analysis was fully reproducible

The problem is that the analysis was wrong
Financial Times: Piketty’s Data Is Full of Errors

By Jordan Weissmann
Disclose all data in publications

Keith Baggerly

Published online 22 September 2010
How to share data with a statistician

This is a guide for anyone who needs to share data with a statistician. The target audiences I have in mind are:

- Scientific collaborators who need statisticians to analyze data for them
- Students or postdocs in scientific disciplines looking for consulting advice
- Junior statistics students whose job it is to collate/clean data sets

For more information and updates, please visit: https://github.com/jtleek/datasharing
Reproducibility = Solved

(You’re welcome)
E-GEUV-6 - Re-analysis of 667 assays from the GEUVADIS study

<table>
<thead>
<tr>
<th>Status</th>
<th>Released on 17 April 2014, last updated on 27 April 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Samples (463)</td>
<td>[Click for detailed sample information and links to data]</td>
</tr>
<tr>
<td>Protocols (5)</td>
<td>[Click for detailed protocol information]</td>
</tr>
<tr>
<td>Description</td>
<td>Re-analysis of 667 assays (463 samples) from Geuvadis study has been performed to demonstrate the flexibility of the Ballgown package to identify transcript-level eQTLs and identify non-linear artifacts in transcript data. Our package Ballgown is freely available from: <a href="https://github.com/alyssafrazee/ballgown">https://github.com/alyssafrazee/ballgown</a>. Geuvadis RNA sequencing data set of 465 human lymphoblastoid cell line samples from the CEU, FIN, GBR, TSI and YRI populations from the 1000 Genomes sample collection was created by the Geuvadis consortium (<a href="http://www.geuvadis.org">www.geuvadis.org</a>, <a href="http://www.geuvadis.org/web/geuvadis/our-rnaseq-project">http://www.geuvadis.org/web/geuvadis/our-rnaseq-project</a>). Original Geuvadis mRNA and small RNA sequencing data, clean data that passed QC and other filters, processed files and analysis results are available under accession E-GEUV-1, E-GEUV-2, E-GEUV-3.</td>
</tr>
<tr>
<td>Experiment types</td>
<td>RNA-seq of coding RNA, operator variation design, population based design</td>
</tr>
<tr>
<td>Contacts</td>
<td>Alyssa C. Frazee <a href="mailto:acfrazee@gmail.com">acfrazee@gmail.com</a>, Jeff Leek <a href="mailto:jtleek@gmail.com">jtleek@gmail.com</a></td>
</tr>
<tr>
<td>MINSEQE</td>
<td>* * * * *</td>
</tr>
</tbody>
</table>

**Files**
- Investigation description
- Sample and data relationship
- Processed data (667)
- [Click to browse all available files](#)

- E-GEUV-6.idf.txt
- E-GEUV-6.sdrf.txt
- [Click to browse processed data](#)
1st Discussion Point:

Statistical thinking is (often) an afterthought
Abstract

Background
Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation plays during normal cellular differentiation and tissue specification. We also need thorough, genome level studies to determine the meaning of methylation of individual CpG dinucleotides in terms of gene expression.

Results
In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arrays and compared those to gene expression by RNA sequencing. We have identified unique signatures of global DNA methylation for human heart, kidney and liver, and showed that DNA methylation data can be used to correctly classify various tissues. It indicates that DNA methylation reflects tissue specificity and may play an important role in tissue differentiation. The integrative analysis of methylation and RNA-Seq data showed that gene methylation and its transcriptional levels were comprehensively correlated. The location of methylation markers in terms of distance to transcription start site and CpG island showed no effects on the regulation of gene expression by DNA methylation in normal tissues.

http://bit.ly/OgW3xv
Synthesis, Structure, and Catalytic Studies of Palladium and Platinum Bis-Sulfoxide Complexes

Emma, please insert NMR data here! where are they? and for this compound, just make up an elemental analysis...
<table>
<thead>
<tr>
<th>FIELD</th>
<th>NOTEWORTHY APPLICATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial Intelligence</td>
<td>machine learning, natural language processing, vision, mathematical models of cognition and learning</td>
</tr>
<tr>
<td>Chemistry</td>
<td>chemical and biomolecular engineering</td>
</tr>
<tr>
<td>Computational Science</td>
<td>computational fluid mechanics, computational materials sciences</td>
</tr>
<tr>
<td>Earth and Planetary Science</td>
<td>climate modeling, seismology, geographic information systems</td>
</tr>
<tr>
<td>Marketing</td>
<td>online advertising, consumer behavior</td>
</tr>
<tr>
<td>Physical Sciences</td>
<td>astronomy, particle physics, geophysics, space sciences</td>
</tr>
<tr>
<td>Signal Processing</td>
<td>compressed sensing, inverse imagining</td>
</tr>
<tr>
<td>Statistics</td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>genomics, proteomics, ecoinformatics, computational cell biology</td>
</tr>
<tr>
<td>Economics</td>
<td>macroeconomic policy, taxation, labor economics, microeconomics, finance, real estate</td>
</tr>
<tr>
<td>Engineering</td>
<td>sensor networks (traffic control, energy-efficient buildings, brain-machine interface)</td>
</tr>
<tr>
<td>Environmental Sciences</td>
<td>deforestation, climate change, impacts of pollution</td>
</tr>
<tr>
<td>Humanities</td>
<td>digital humanities, archeology, land use, cultural geography, cultural heritage</td>
</tr>
<tr>
<td>Law</td>
<td>privacy, security, forensics, drug/human/CRBN trafficking, criminal justice, incarceration, judicial decision making, corporate law</td>
</tr>
<tr>
<td>Linguistics</td>
<td>historical linguistics, corpus linguistics, psycholinguistics, neurolinguistics</td>
</tr>
</tbody>
</table>
A guide for the lonely bioinformatician

45 Replies

This may be a uniquely UK centric blog post but I suspect not. Let me start with a brief story. Sat with a coffee in our canteen a few weeks ago, I overheard a conversation between a few PIs about a grant application. “Don’t worry”, the lead PI said, “we’ve put money on the application to fund a bioinformatician”. Good planning I hear you say, and I agree; however, note that none of the PIs in that discussion were themselves bioinformaticians; none of them can code; put them in front of a Linux terminal and they wouldn’t know what to do.

Yes – we were witnessing the birth of yet another “pet bioinformatician”. What I mean by this term is a single bioinformatician employed within laboratory based group. These guys are becoming more and more common in UK academic groups, and it concerns me because it is possible they will become isolated and pick up bad practices as they don’t have a senior bioinformatician to guide them. It also concerns me that their career and professional development might suffer.

http://biomickwatson.wordpress.com/2013/04/23/a-guide-for-the-lonely-bioinformatician/
Yes, we are witnessing the birth of Yet another “pet bioinformatician”. What I mean by this term is a single bioinformatician employed within a laboratory based group.

http://biomickwatson.wordpress.com/2013/04/23/a-guide-for-the-lonely-bioinformatician/
2nd Discussion Point:

Most data analysts are untrained
med school entrance requirements

One year of biology
One year of physics
One year of English
Two years of chemistry
(through organic chemistry)

https://www.aamc.org/students/applying/requirements/
To: jtleek@gmail.com

I am a postdoctoral fellow in redacted group. I collected data on redacted...

Preliminary analysis has pulled out some interesting things but we need some professional assistance...

We want to submit at the end of next month.
3rd Discussion Point:

Statistics is not math
(and data analysis isn’t statistics)
association between shoe size and literacy
Fast track — Mechanisms of Disease

Use of proteomic patterns in serum to identify ovarian cancer
Fast track — Mechanisms of Disease

Use of proteomic patterns in serum to identify ovarian cancer

How Bright Promise in Cancer Testing Fell Apart

Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors.
Common genetic variants account for differences in gene expression among ethnic groups

Richard S Spielman¹, Laurel A Bastone², Joshua T Burdick³, Michael Morley³, Warren J Ewens⁴ & Vivian G Cheung¹,³,⁵
Common genetic variants account for differences in gene expression among ethnic groups
Richard S Spielman¹, Laurel A Bastone², Joshua T Burdick³, Michael Morley³, Warren J Ewens⁴ & Vivian G Cheung¹,³,⁵

Correspondence

On the design and analysis of gene expression studies in human populations
Joshua M Akey¹, Shameek Biswas¹, Jeffrey T Leek² & John D Storey¹,²
Published Online July 1 2010

Science DOI: 10.1126/science.1190532

REPORT

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani¹, Nadia Solovieff¹, Annibale Puca², Stephen W. Hartley¹, Efthymia Melista³, Stacy Andersen⁴, Daniel A. Dworkis³, Jemma B. Wilk⁵, Richard H. Myers⁵, Martin H. Steinberg⁶, Monty Montano³, Clinton T. Baldwin⁶,⁷ and Thomas T. Perls⁴,*

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.H.P.)
This article has been retracted

Published Online July 1 2010

Science DOI: 10.1126/science.1190532

REPORT

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani1,*, Nadia Solovieff1, Annibale Puca2, Stephen W. Hartley1, Efthymia Melista3, Stacy Andersen4, Daniel A. Dworkis3, Jemma B. Wilk5, Richard H. Myers5, Martin H. Steinberg6, Monty Montano3, Clinton T. Baldwin6,7 and Thomas T. Perls4,*

*To whom correspondence should be addressed. E-mail: sebas@bu.edu (P.S.); thperls@bu.edu (T.H.P.)
Discussion Point:

How do we balance skepticism & excitement?
Doctors cannot compete with machines says @vkhosla bit.ly/1ndScXa #bigdatamed
B I G  D A T A

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,1,2* Ryan Kennedy,1,3,4 Gary King,3 Alessandro Vespignani3,5,6

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (10, 15).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week’s errors predict this week’s errors (temporal autocorrelation), and the direction and

1. Statistical thinking is (often) an afterthought

2. Most data analysts are untrained

3. Statistics is not math (and data analysis isn’t statistics)

4. How do we balance skepticism & excitement?
jtleek.com/talks