INTRODUCTION TO EXPERIMENTAL DESIGN AND STATISTICAL ANALYSIS

Olga Vitek

College of Science College of Computer and Information Science



WHY STATISTICS?

- Variation and uncertainty are unavoidable
 - Technical variation: sampling handling, storage, processing
 - Instrumental variation: elution time, ion suppression
 - Signal processing: peak boundaries, identity, intensity
 - Biological variation: variation in protein abundance
- Overall goal: effective, reproducible research
 - Experimental design: unbiased and efficient experiments
 - Data analysis: objective conclusions in presence of uncertainty
 - Statistical tools: re-analysis, peer review, reproducibility

"Statistics: a body of methods for making wise decisions in the face of uncertainty." (W. A. Wallis)

WHY STATISTICS?

nature biotechnology

nature.com > journal home > archive > issue > opinion and comment > correspondence > full text

NATURE BIOTECHNOLOGY | OPINION AND COMMENT | CORRESPONDENCE

Sequencing technology does not eliminate biological variability

Kasper D Hansen, Zhijin Wu, Rafael A Irizarry & Jeffrey T Leek

Affiliations | Corresponding authors

Nature Biotechnology 29, 572-573 (2011) | doi:10.1038/nbt.1910

OUTLINE

- Translate scientific question into statistics
 - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
 - Replication, randomization, blocking
- Basic data analysis
 - Simple summaries and models
- Words of caution
 - Instability, multiplicity, reproducible research

STATISTICAL GOAL I: CLASS DISCOVERY Discover proteins or subjects with similar patterns

- No known class labels
 - E.g., no 'healthy' or 'disease'
 - All variation treated equally
 - No error rates
- Can't find something meaningful if unsure what we look for
 - Best used for visualization



Gehlenborg et al, Nature Methods, 2010

STATISTICAL GOAL 2: CLASS COMPARISON Compare mean abundances in subject groups

- Known class labels
 - Compare group averages
 - Report p-values, posterior probabilities etc
- Useful when compare groups of subjects
 - Best used for basic biology
 - Initial (Tier III) biomarker discovery screen



DIFFERENTIALLY ABUNDANT PROTEINS ARE NOT ALWAYS BIOMARKERS



BIOMARKER PROTEINS ARE NOT ALWAYS DIFFERENTIALLY ABUNDANT



STATISTICAL GOAL 3: CLASS PREDICTION Classify each subject into a known group

- Known class labels
 - Predict individual subjects
 - Report misclassification error (sensitivity, specificity, predictive value etc)
- Useful when focus on an individual
 - Tier I or Tier II biomarker discovery studies



OUTLINE

- Translate scientific question into statistics
 - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
 - Replication, randomization, blocking
- Basic data analysis
 - Simple summaries and models
- Words of caution
 - Instability, multiplicity, reproducible research

A STATISTICIAN'S VIEW OF THE EXPERIMENT



DEFINITION OF BIAS AND INEFFICIENCY



Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ systematically different from $\mu_{1k} - \mu_{2k}$ **Inefficiency:** Large $Var(\bar{Y}_{.1..} - \bar{Y}_{.2..})$

DEFINITION OF BIAS AND INEFFICIENCY



Bias: $\bar{Y}_{.1..} - \bar{Y}_{.2..}$ systematically different from $\mu_{1k} - \mu_{2k}$ **Inefficiency:** Large $Var(\bar{Y}_{.1..} - \bar{Y}_{.2..})$

PRINCIPLE I: REPLICATION (1) carries out the inference and (2) minimizes inefficiencies



Two levels of randomness imply two types of replication:

- *Biological replicates:* selecting multiple subjects from the population
- *Technical replicates:* multiple runs per subject

Oberg and Vitek, J. Proteome Research, 8, 2009

PRINCIPLE 2: RANDOMIZATION Prevents bias



Two levels of randomness imply two types of randomization:

- *Biological replicates:* random selection of subjects from the population
- *Technical replicates:* random allocation of samples to all processing steps

Oberg and Vitek, J. Proteome Research, 8, 2009

EXAMPLE: LACK OF RANDOMIZATION

Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

- Serum samples with five types of cancer
- SELDI-TOF MS
 - normalized, peak picked

Hierarchical clustering of samples



BEWARE OF BIG EFFECTS THEY LIKELY REFLECT FLAWS OF THE DESIGN

- Study of gene expression between Asians and Europeans
- Found that 78% of genes are differentially
 - Asians were profiles in one year, and Europeans in another
 - The difference therefore likely reflects a batch effect



Journal home > Archive > Letter > Full Text

Journal content	Letter
 Journal home 	Nature Genetics 39, 226 - 231 (2007)
Advance online publication	Published online: 7 January 2007 doi:10.1038/ng1955
Current issue	expression among ethnic groups
Archive	Richard S Spielman ^{\pm} , Laurel A Bastone ^{2} , Joshua T Burdick ^{3} , Michael Morley ^{3} ,
Focuses and Supplements	Warren J Ewens ⁴ & Vivian G Cheung ^{1,3,5}
Source: a bl	og hy leff Leek, Biostatistics, John



Т	hi	is	ī	e	s		0
		0		0	a	u	c

- Table of contents

Source: a blog by Jeff Leek, Biostatistics, John Hopkins University http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/

BEWARE OF BIG EFFECTS THEY LIKELY REFLECT FLAWS OF THE DESIGN

• Study of gene expression between Asians and Europeans

'To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of'

--- Ronald Fisher

 Journal nome 	Nature Genetics 39, 226 - 231 (2007)	
Advance online	Published online: 7 January 2007 doi:10.1038/ng1955	
Current issue	Common genetic variants account for differences in gene expression among ethnic groups	This issue
Archive	Richard S Spielman ¹ , Laurel A Bastone ² , Joshua T Burdick ³ , Michael Morley ³ ,	Table of contents
Focuses and Supplements	Warren J Ewens ⁴ & Vivian G Cheung ^{1,3,5}	

Source: a blog by Jeff Leek, Biostatistics, John Hopkins University http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/



= systematic allocation

Two levels of randomness imply two types of blocks:

- *Biological replicates:* subjects having similar characteristics (e.g. age)
- *Technical replicates:* samples processed together (e.g. in a same day)

Oberg and Vitek, J. Proteome Research, 8, 2009

EXAMPLE: LACK OF BLOCKING

Hu, Coombes, Morris, Baggerly, Briefings in Functional Genomics, 2005

- Serum samples with two types of cancer
- SELDI-TOF MS, 3 fractions
 - normalized, peak picked



MATCHING Blocking with respect to biological risk factors



Complete randomization = inflated variance



Block-randomization = restriction on randomization = systematic allocation

Käll and Vitek, PLoS Computational Biology, 7, 2011

EXAMPLE

Block-randomized selection of subjects from repository

				Disease group		
		Control	Stable angina	Unstable angina	NSTEMI	STEMI
	≥ 58 y.o; Female	354	300	49	39	29
Stratification	≥ 58 y.o; Male	701	843	143	86	54
Stratification	< 58 y.o; Female	80	56	5	5	8
	< 58 y.o; Male	264	190	34	23	27

Counts in the initial repository of samples

				Disease group		
		Control	Stable angina	Unstable angina	NSTEMI	STEMI
	≥ 58 y.o; Female	3	3	3	3	3
Stratification	≥ 58 y.o; Male	3	3	3	3	3
Stratification	< 58 y.o; Female	2	2	2	2	2
	< 58 y.o; Male	2	2	2	2	2

Counts of subjects included in the study

Mass spectra acquired without technical replication

OUTLINE

- Translate scientific question into statistics
 - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
 - Replication, randomization, blocking
- Basic data analysis
 - Simple summaries and models
- Words of caution
 - Instability, multiplicity, reproducible research

COMPARE DESIGNS In terms of bias and (in)-efficiency



TWO-SAMPLET-TEST Simple example: label-free experiment, one feature/protein



FoldChange = <u>'typical' value in group 1</u> <u>'typical' value in group 2</u>

log2(FoldChange) =
= log2('typical' value in group 1)
-log2('typical' value in group 2)

•
$$\bar{Y}_{1.} - \bar{Y}_{2.} = \text{estimates log-fold change}$$

 $\frac{1}{n_1} \sum_j Y_{1j} - \frac{1}{n_2} \sum_j Y_{2j} = \frac{1}{n_1} \sum_j \log_2 X_{1j} - \frac{1}{n_2} \sum_j \log_2 X_{2j} =$
 $\log_2 \left(\prod_j X_{1j} \right)^{\frac{1}{n_1}} - \log_2 \left(\prod_j X_{2j} \right)^{\frac{1}{n_2}} = \log_2 \frac{\left(\prod_j X_{1j} \right)^{\frac{1}{n_1}}}{\left(\prod_j X_{2j} \right)^{\frac{1}{n_2}}}$

Conclusion:

On log scale, estimates of FC are ratios of geometric means

TWO-SAMPLET-TEST

Simple example: label-free experiment, one feature/protein



TWO-SAMPLET-TEST Simple example: label-free experiment, one feature/protein



H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$ Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$

observed t =
$$\frac{\text{difference of group means}}{\text{estimate of variation}} = \frac{\bar{Y}_{1.} - \bar{Y}_{2.}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

 $s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_{1.})$



i=1

ASSUMPTION: NORMAL DISTRIBUTION The Central Limit Theorem



Conclusion: As n increases the n

As n increases, the mean is less variable and more Normal

EFFECT OF SAMPLE SIZE As n increases, the estimates stabilize



Simulated example Krzywinski and Altman, Points of Significance Collection, *Nature Methods*

FINDING DIFFERENTIALLY ABUNDANT PROTEINS False positive rate



H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$ Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$





Distribution of the

 α = False

Positive Rate

FINDING DIFFERENTIALLY ABUNDANT PROTEINS P-value



H0: 'status quo', no change in abundance, $\mu_1 - \mu_2 = 0$ Ha: change in abundance, $\mu_1 - \mu_2 \neq 0$





p = p-value

Distribution of the

ALTERNATIVE TO TESTING: CONFIDENCE INTERVALS Not all error bars are made equal



A 95% CI: if we repeatedly collect data and draw confidence intervals, then 95% of them will contain the true mean

$$(\bar{Y}_{1.} - \bar{Y}_{2.}) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \ (\bar{Y}_{1.} - \bar{Y}_{2.}) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

CI are wider than bars indicating standard error of the mean!

Width of the intervals depends on the sample size

Simulated example Krzywinski and Altman, Points of Significance Collection, Nature Methods

ERROR BARS PROVIDE DIFFERENT INSIGHT Absence of overlap does not always mean stat. significance



Simulated example Krzywinski and Altman, Points of Significance Collection, Nature Methods

STATISTICAL POWER Probability to detect a difference when it exists



Correct inference

- Specificity, 1α
- **Power**, sensitivity, β

Incorrect inference

- Type I error, α
- Type II error, 1β

STATISTICAL POWER Probability to detect a difference when it exists



Correct inference

- Specificity, 1α
- **Power**, sensitivity, β

Incorrect inference

- Type I error, α
- Type II error, 1β





 α trades off the sensitivity and the specificity of the test

HOW TO GAIN POWER? Better choice: increase signal/noise



Correct inference

- Specificity, 1α
- **Power**, sensitivity, β

Incorrect inference

- Type I error, α
- Type II error, 1β

sample size increases statistical power

OUTLINE

- Translate scientific question into statistics
 - Statistical terms for 'biomarker' (or 'signature')
- Experimental design
 - Replication, randomization, blocking
- Basic data analysis
 - Simple summaries and models
- Words of caution
 - Instability, multiplicity, reproducible research

AMERICAN STATISTICAL ASSOCIATION (ASA) STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES The American Statistician, February 2016

- 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
- Scientific conclusions and business policy decisions should not be based only on whether a p-value passes a specific threshold

AMERICAN STATISTICAL ASSOCIATION (ASA) STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES The American Statistician, February 2016

- 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 a hypothesis

WITH SMALL SAMPLE SIZE, P-VALUES ARE UNSTABLE



- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
 - Larger sample size
 - Adjustment for multiple testing



Simulated example

Halsey, Curran-Everett, Volwer and Drummond, Nature Methods, 2015

WITH SMALL SAMPLE SIZE, P-VALUES ARE UNSTABLE



- Repeatedly sampling data leads to different results
- The problem worsens when testing many proteins
- Partial solutions:
 - Larger sample size
 - Adjustment for multiple testing



Simulated example

Halsey, Curran-Everett, Volwer and Drummond, Nature Methods, 2015

WITH SMALL SAMPLE SIZE, CONCLUSIONS ARE BIASED



CONFIDENCE INTERVALS PROVIDE COMPLEMENTARY INSIGHT





Simulated example Halsey, Curran-Everett, Volwer and Drummond, *Nature Methods*, 2015

PITFALL: MULTIPLE TESTING

- An fMRI on dead fish
- Found many active brain regions
 - All background noise and random variation



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett¹, Abigail A. Baird², Michael B. Miller¹, and George L. Wolford³ ¹ Psychology Department, University of California Santa Barbara, Santa Barbara, CA; ² Department of Psychology, Vassar College, Poughkeepsie, NY; ³ Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

<u>Subject.</u> One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at

GLM RESULTS



Source: a blog by Jeff Leek, Biostatistics, John Hopkins University http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/









MULTIVARIATE TYPE I ERROR How many false positives can we tolerate?

	# of proteins with	# of proteins with	Total
	no detected difference	detected difference	
# true non-diff. proteins	U	\mathbf{V}	m ₀
# true diff. proteins	\mathbf{T}	\mathbf{S}	$\mathbf{m_1} = \mathbf{m} - \mathbf{m_0}$
Total	m - R	R	m

- Type I error rate
 - $\alpha = \Pr\{\text{Type I error}\} \\ = \mathbf{E}\left[\frac{\mathbf{V}}{\mathbf{m}_{\mathbf{0}}}\right]$
- Family-wise error rate (FWER)

 α^{*} = Pr{at least one Type I error}
 = P[V > 0]
- False discovery rate (FDR)

$$q = E\{\text{proportion of Type I error}\}$$
$$= \mathbf{E}\left[\frac{\mathbf{V}}{\max(\mathbf{R}, \mathbf{1})}\right]$$





Statistics: P values are just the tip of the iceberg

Jeffrey T. Leek & Roger D. Peng

28 April 2015

Statistical considerations are key at every step



PITFALL: OUTCOME SWITCHING

- Anti-depressant Paxil was studied for several main outcomes
 - None showed an effect
 - Some secondary outcomes dis
- Switched the outcome of the trial and used to market the drug



SCIENCE & HEALTH 🔛

۵0 ۹

How researchers dupe the public with a sneaky practice called "outcome switching"

Updated by Julia Belluz on December 29, 2015, 8:10 a.m. ET ™ julia.belluz@voxmedia.com



Source: a blog by Jeff Leek, Biostatistics, John Hopkins University <u>http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/</u>

PITFALL: NOT PRE-SPECIFIED DATA SELECTION AND ANALYSIS

- Compare 2 groups: women at peak and off peak fertility cycle
 - A series of choices of which women to include in which comparison group
- Conclude that at peak fertility women are more likely to wear red or pink shirts

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*

> Andrew Gelman^{\dagger} and Eric Loken^{\ddagger} 14 Nov 2013

Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of *potential* comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on

Source: a blog by Jeff Leek, Biostatistics, John Hopkins University

http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/

RESEARCHER DEGREE OF FREEDOM

SCIENCE Association for Psychological Science
Home OnlineFirst All Issues Subscribe RSS Semail Alerts

False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons1,

Leif D. Nelson2 and

Uri Simonsohn1



Author Affiliations

Joseph P. Simmons, The Wharton School, University of Pennsylvania, 551 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104 E-mail: jsimmo@wharton.upenn.edu Leif D. Nelson, Haas School of Business, University of California, Berkeley, Berkeley, CA 94720-1900 E-mail: leif_nelson@haas.berkeley.edu Uri Simonsohn, The Wharton School, University of Pennsylvania, 548 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104 E-mail: uws@wharton.upenn.edu « Previous | Next Article » Table of Contents

This Article

Published online before print October 17, 2011, doi: 10.1177/0956797611417632

Psychological Science November 2011 vol. 22 no. 11 1359-1366

» Abstract Free Full Text Free



All Versions of this Article: >>> Version of Record - Nov 7, 2011 0956797611417632v1 - Oct 17, 2011

What's this?

- Services

- Email this article to a colleague
- Alert me when this article is cited



- Define the problem
 - translate biological/clinical goal into statistical goal
- Experimental design: avoid bias and inefficiency
 - randomization, replication, blocking
- Follow a pre-defined design and analysis protocol
 - do not alter the design
 - do not cherry-pick data/parameters
 - understand and state limitations
- Document all the steps
 - in form of the executable code

Q 🖧

MS statistical Tool For Quantitative Mass Spectrom etry-Based Proteomics

HOME INSTALLATION WORKFLOWS DATASETS INSSTATSQC NEWS CONTACT



MS Statistical Tool For Quantitative Mass Spectrometry-Based Proteomics

HOME INSTALLATION WORKFLOWS DATASETS MSSTATSQC NEWS CONTACT

MSSTATSQC

Longitudinal system suitability monitoring tools for quantitative mass spectrometry based proteomic experiments

Statistical process control (SPC) is a general and well-established method of quality control (QC) which can be used to monitor and improve the quality of a process such as LC MS/MS. 'MSstatsQC' is an opensource R-based web application for statistical analysis and monitoring of quality control and system suitability testing (SST) samples produced by spectrometry-based proteomic experiments. Our framework termed 'MSstatsQC' is available through http://www.msstats.org/msstatsqc. It uses SPC tools to track ID free system suitability metrics including total peak area, retention time, full width at half maximum (FWHM) and peak asymmetry for selection reaction monitoring (SRM) based proteomic



Martin Krzywinski & Naomi Altman

57

