Multiple hypothesis tests & False discovery rates

Neo Christopher Chung

Lecture 2, 1000-719bMSB

Hypothesis

A hypothesis is a backbone of science

A hypothesis test is a backbone of statistics

Quick review of a hypothesis testing

Prevalent cases of multiple hypothesis tests

Control the false discovery rates when considering many tests

Hypothesis test

Null hypothesis (H_0) vs. Alternative hypothesis (H_1)

"No changes"
"Equal"

"Changes"
"Not equal"

P-value is the probability to observe cases (statistics) that are as or more extreme than the observation

In a classical sense, we ought to repeat (or imagine repeating) the experiment. In practice, we make assumptions and study designs.

We either **reject or accept** the null hypothesis.

Hypothesis

This treatment may change an expression level of a gene, in some microorganism cells.

Select a random sample of cells for group 1 and group 2, in vitro.
Apply the treatment on a group 1 (case), but not on a group 2 (control).
Carry out the microarray or RNA-seq experiments, followed by summarization.
Check if there is a substantial change.

Null hypothesis (H_0): $\mu_1 = \mu_2 \quad \rightarrow \quad \mu_1 - \mu_2 = 0$
Alternative hypothesis (H_A): $\mu_1 \neq \mu_2 \quad \rightarrow \quad \mu_1 - \mu_2 \neq 0$

Confirmatory data analysis

Compare the means between 2 groups. If two groups came from the same population, they would have the same (population) means.

> The Student's t-test ✅

Analysis of variance (ANOVA) is generalization of t-test, especially for having 3 or more groups to compare.

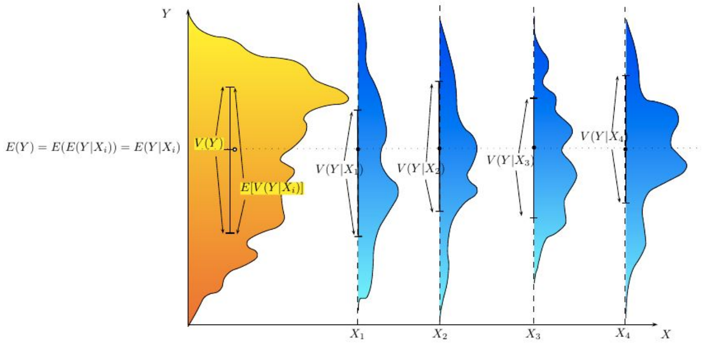


Figure 2: ANOVA : No fit

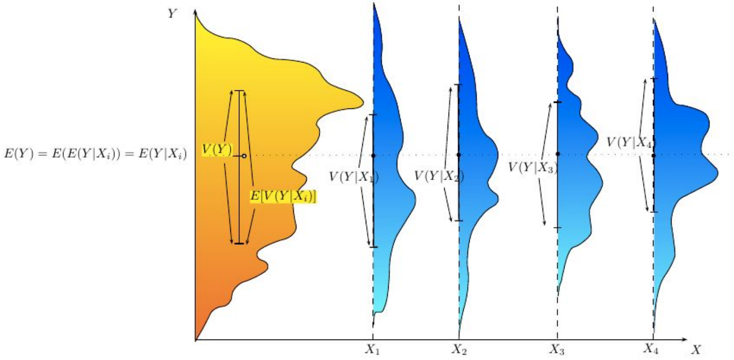


Figure 2: ANOVA : No fit

Random sampling

Samples are randomly chosen from the whole population of cells.

What does it mean to be a whole population of cells? Or can we assume this so easily in an experiment?

The measurements of gene expression on those samples X are random variables (r.v.).

Which samples to apply Treatment or Placebo are randomly assigned.

Such randomizations are required to ensure that the difference (or a lack thereof) we observe would generalize to the population

Student's two-group t-test

Assuming equal sample size, equal variance.

H_0: $\mu_1 - \mu_2 = 0$ vs. H_A: $\mu_1 - \mu_2 \neq 0$

We call the observed data X, labeled with the group 1 and 2.

Then, we form a t-statistics or t-value,

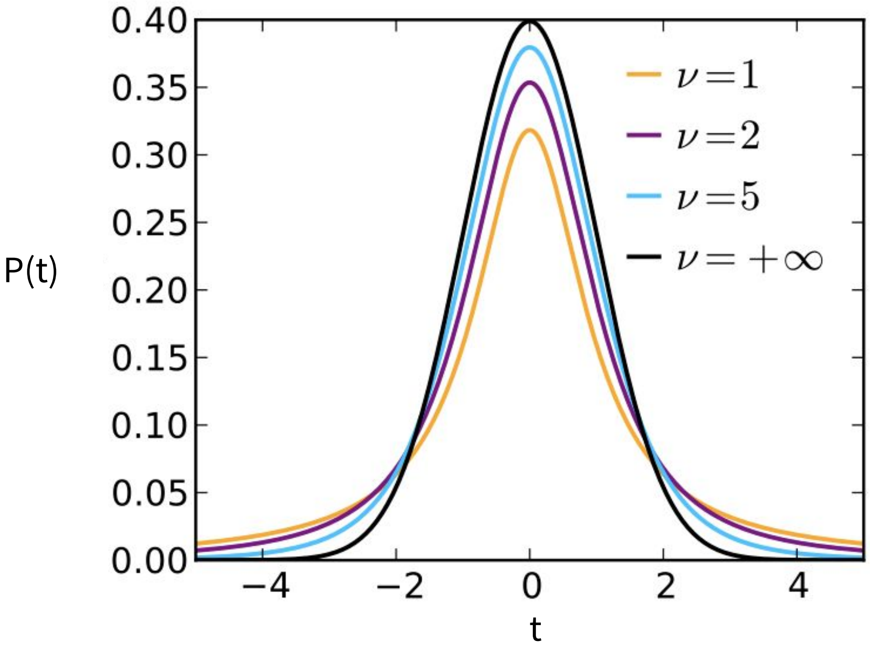
With s = estimated standard deviation of the population:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}}$$, where $$s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}.$$

Student's T-distribution

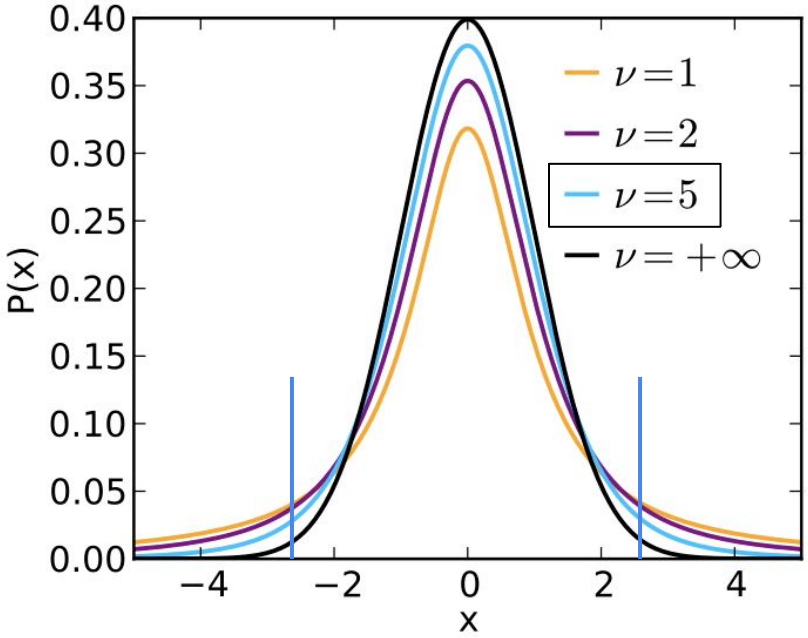
Under the null hypothesis, a t-statistics follows a t-distribution

A degree of freedom v = n-1 where n is a number of observations



Two-sided p-values

Calculate the area under tails; e.g., Critical region for α=0.05



Limitations of frequentist statistics

There are a plenty of issues of **abusing, hacking, or cheating** with p-values

April 16, 2019

Lowering the *P* Value Threshold

Amin Adibi, MSc[1]; Don Sin, MD[2]; Mohsen Sadatsafavi, MD, PhD[1]

» Author Affiliations

JAMA. 2019;321(15):1532-1533. doi:10.1001/jama.2019.0566

To the Editor Mr Wayant and colleagues evaluated the effect of lowering the significance threshold from .05 to .005 on major randomized clinical trials (RCTs) published in 2017.[1] The authors reported that 70.7% of primary end points remained significant and suggested that lowering the threshold might address statistical issues such as *P*-hacking.

nature human behaviour

comment

Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Editorial

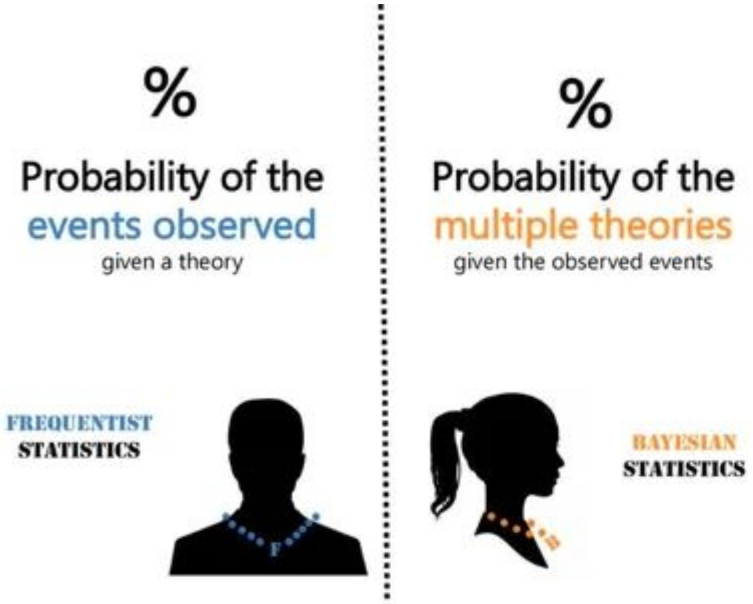
Editorial

David Trafimow & Michael Marks

Pages 1-2 | Published online: 12 Feb 2015

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. <u>From now on, BASP is banning the NHSTP.</u>

Frequentist vs Bayesian Statistics



https://www.analyticsvidhya.com/blog/2023/07/frequentist-vs-bayesian/

Frequentist vs Bayesian Statistics

Aspect	Frequentist Approach	Bayesian Approach
Probability Interpretation	Objective: Probabilities represent long-term frequencies or limiting behavior of repeated experiments.	Subjective: Probabilities represent degrees of belief or uncertainty based on prior knowledge and data.
Treatment of Parameters	Fixed: Parameters are fixed, unknown constants. Estimation involves finding the “best” estimate based on data.	Random: Parameters are treated as random variables with their own probability distributions. They are updated based on prior beliefs and data, resulting in posterior distributions.
Prior Information	N/A: Typically, prior information is not explicitly incorporated into the analysis.	Crucial: Bayesian analysis involves specifying prior distributions representing prior beliefs about parameters before observing data.
Hypothesis Testing	p-values and hypothesis tests are prone to misinterpretation and controversies.	Bayesian hypothesis testing uses Bayes Factors or posterior probabilities for direct comparison.
Interpretation of Results	Focused on the data and observed effects.	Results interpreted in the context of prior beliefs and their update based on data.

https://www.analyticsvidhya.com/blog/2023/07/frequentist-vs-bayesian/

Bayesian hypothesis test

Null hypothesis (H_0) vs. Alternative hypothesis (H_1)

“No changes”
“Equal”

“Changes”
“Not equal”

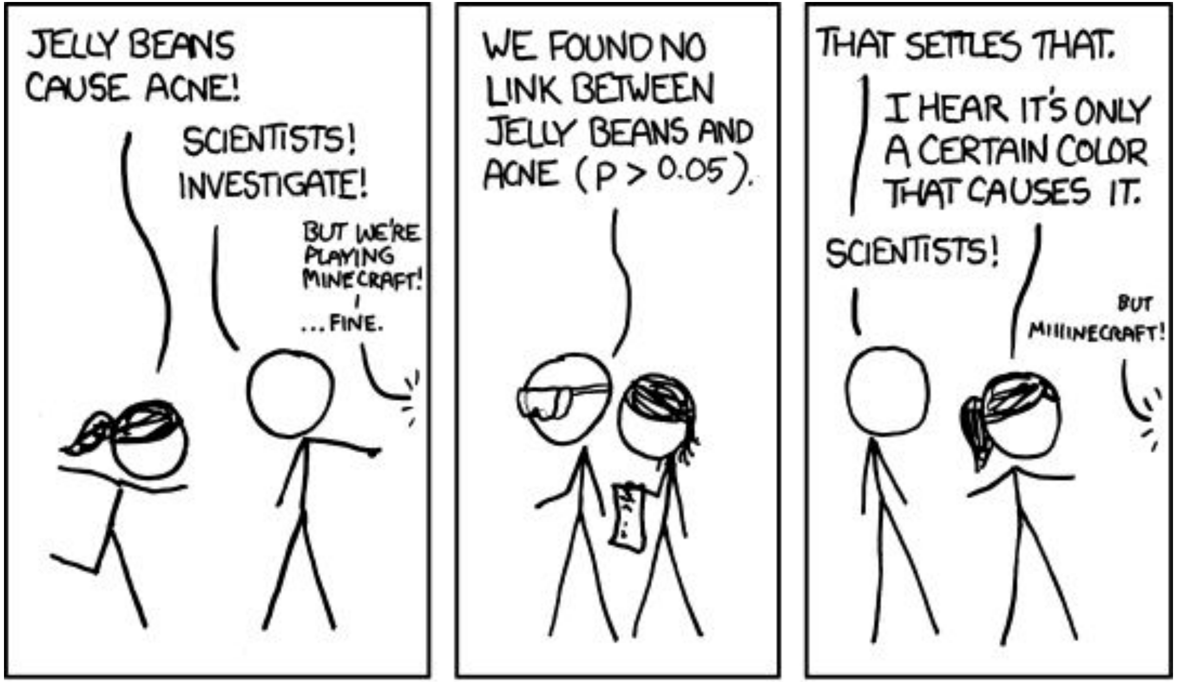
The posterior probabilities of H_0 or H_1 given the observed data.

$$P(H_0|X) \text{ vs } P(H_1|X)$$

Accept the hypothesis with a greater probability

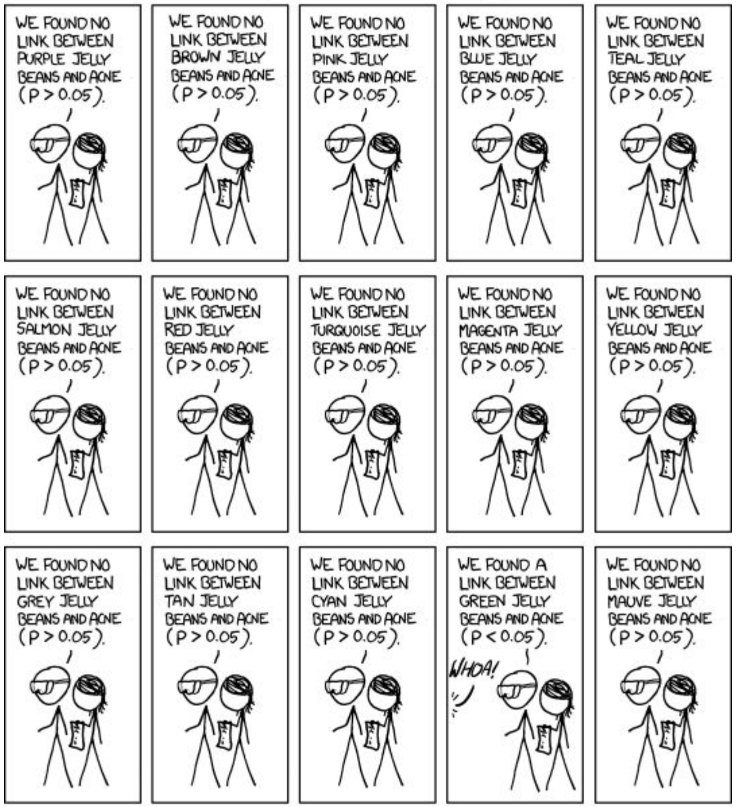
This maximum a posteriori (MAP) test is available in R packages e.g., bayestestR

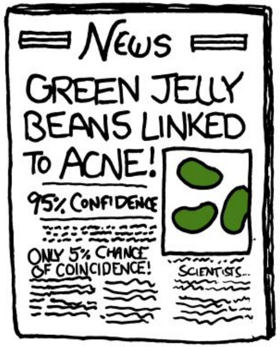
Considering multiple hypotheses



https://xkcd.com/882/

Considering multiple hypotheses





https://xkcd.com/882/

Gene expression in treatment vs. control

Want to see if any gene is related to this treatment factor

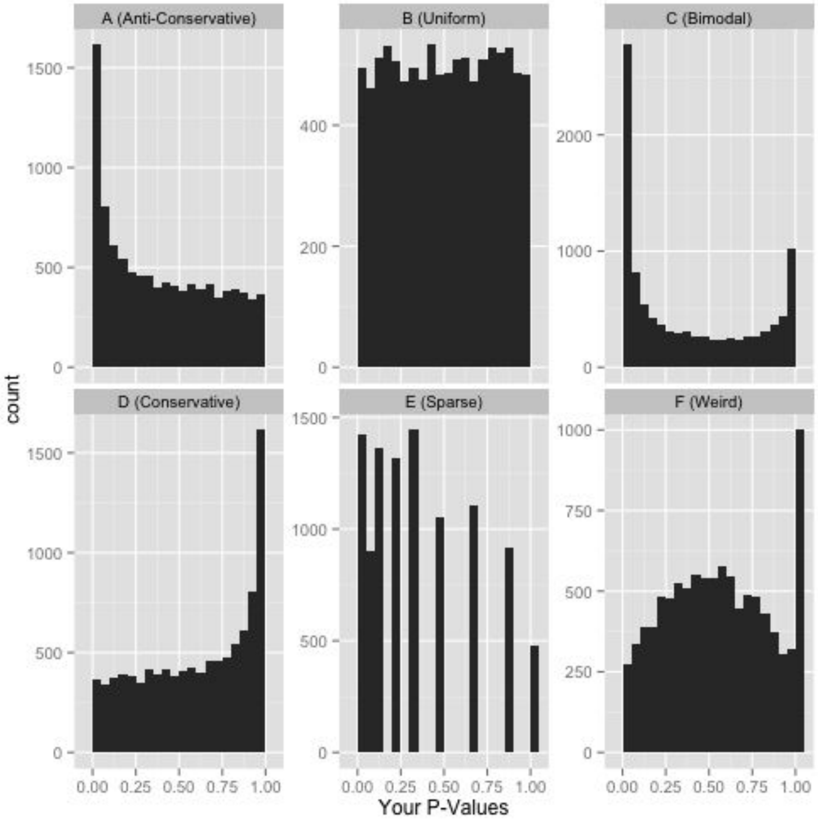
Measure expression levels of 100 genes (or the whole genome)

Calculate 100 p-values

E.g., conduct t-tests on each of 100 genes

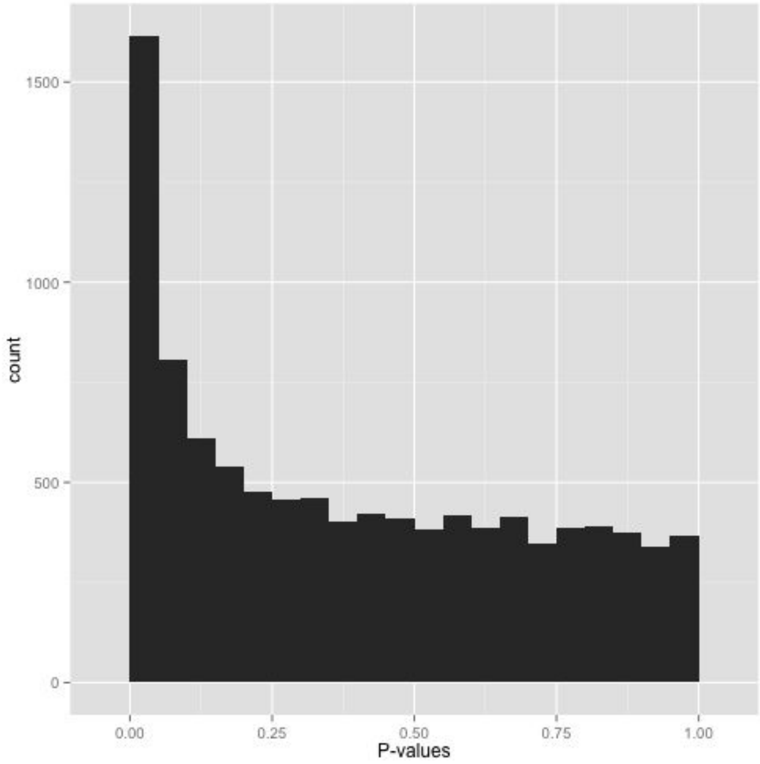
What's the first thing to do with p-values?

Behaviors and Interpretation of p-values

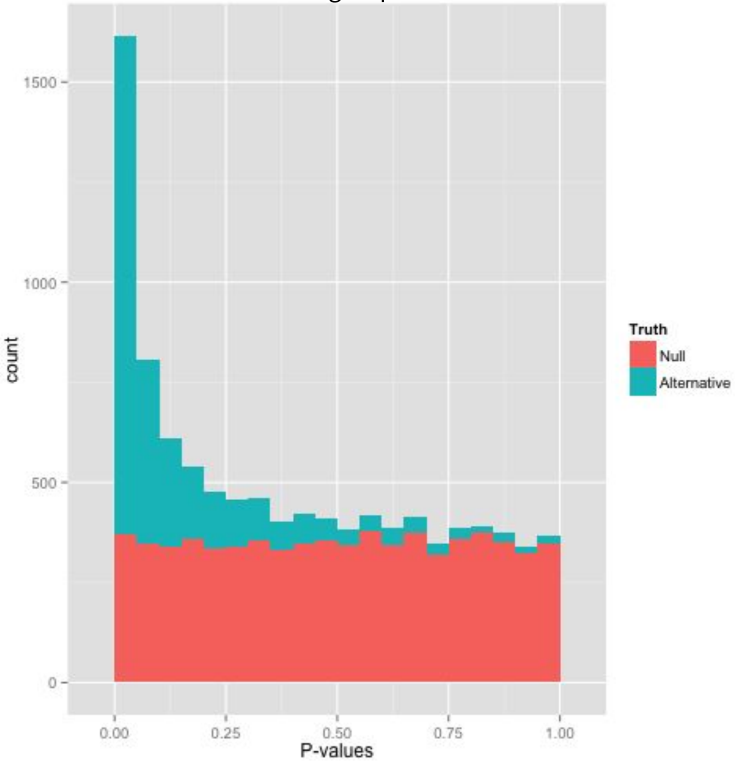


http://varianceexplained.org

Anti-conservative p-values

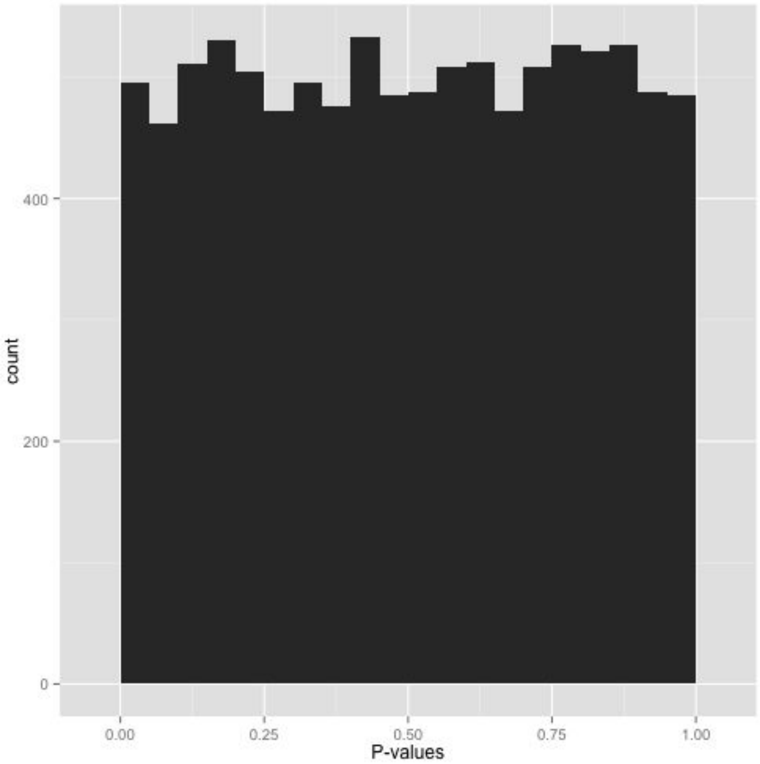


Efron's two-group model



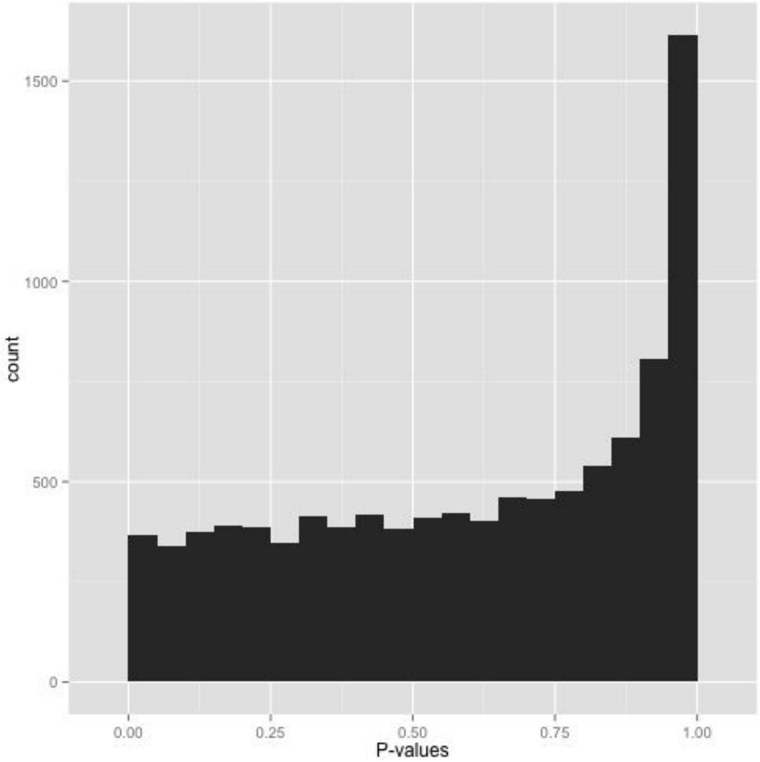
http://varianceexplained.org

Uniform p-values



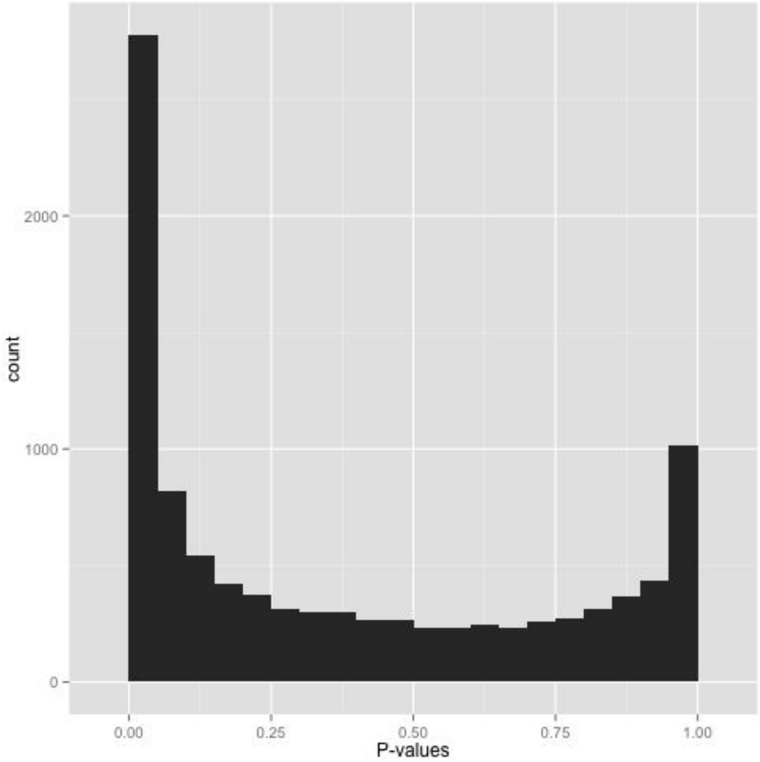
1. A very small number of non-null hypotheses
2. Not enough power
3. Don’t apply thresholding blindly

Conservative p-values



1. Incorrect assumption
2. The distribution doesn’t fit the data
3. P-values have been corrected by some methods

Bimodal p-values



1. One-sided tests applied when two-sided tests are appropriate
2. Look at the characteristics of tests with p-values at/near 1

Gene expression in treatment vs. control

Want to see if any gene is related to this treatment factor

Measure expression levels of 100 genes

Calculate 100 p-values

E.g., conduct t-tests on each of 100 genes

Set $\alpha = 0.05$

Genes with p-value $< \alpha$ are deemed ‘significant’

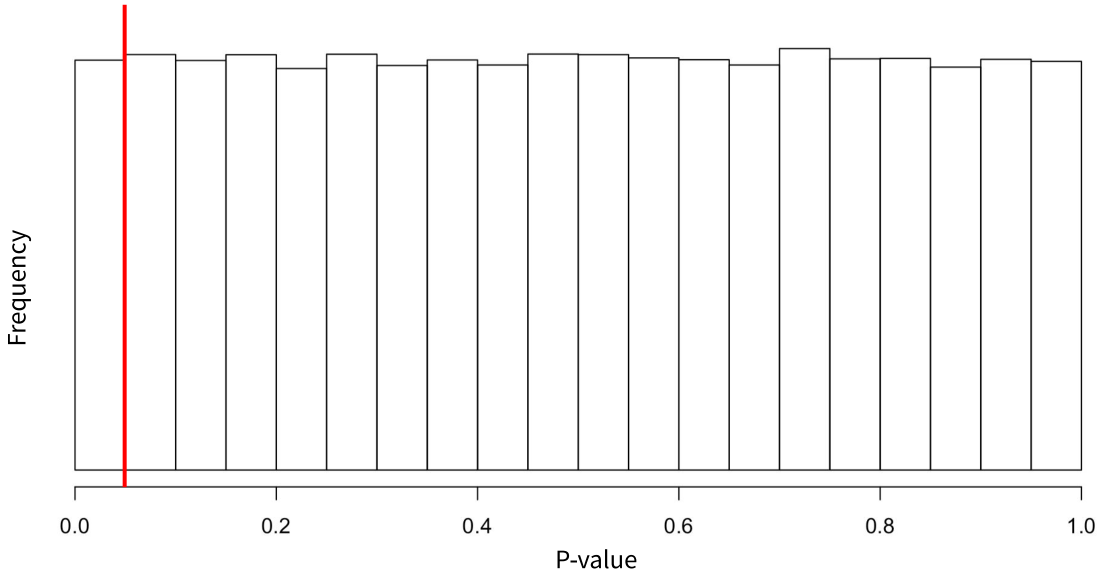
Under the null hypothesis (no difference, truly),

an <u>expected number of false positives = 5</u>

Plot the p-values

Under the null hypothesis, p-values would follow i.i.d. Uniform (0,1) distribution.

See how we get p-values < 0.05 with multiple hypotheses



Classification of multiple hypothesis tests

	Null hypothesis is true (H_0)	Alternative hypothesis is true (H_1)	Total
Significant aka Positive Prediction	**V** (false positive; false discovery)	**S** (true positive; true discovery)	**R** (known with a threshold)
Non-significant aka Negative Prediction	**U** (true negative)	**T** (false negative)	**m - R**
Total	$\mathbf{m_0}$ (must be estimated)	$\mathbf{m - m_0}$	**M** (total tests)

Control a family wise error rate

FWER is the probability of making at least one type 1 error in the family.

$$\text{FWER} = \Pr(\text{ \# false positives } \geq 1) = \Pr(V \geq 1)$$

$$= 1 - \Pr(V = 0)$$

If one control FWER at α, the probability of making one or more type 1 error is controlled at α

Procedures for FWER

Bonferroni procedure (derived from Bonferroni, 1936)

Reject if $p_i \leq \alpha/m$ are significant

Benjamini-Hochberg Correction (proposed by Benjamini and Hochberg 1995)

Order the p-values: $p_1, \ldots, p_m$.

If $p_i \leq \alpha * i/m$, then its hypothesis test is significant

Several other procedures available. See p.adjust() in R

FWER is often too stringent in a high dimensional setting.

E.g., $p_i \leq \alpha/m$ where $m > 10000$ genes

False discovery rates

If the false positive rate is the error measure used, then a simple p-value threshold is used. A p-value threshold of 0.05, for example, guarantees only that the expected number of false positives is $E[V] = 0.05\, m$. ~ Likely too liberal.

The error measure that is typically controlled in genome scans for linkage is the familywise error rate, which can be written as $Pr(V >= 1)$. ~ Likely too conservative

FDR: an error measure in between these, specifically, one that provides a sensible balance between the number of false positive features, V, and the number of true positive features, S.

Storey and Tibshirani (2003)

False discovery rates

FDR = E(Q) where Q = V/R   R > 0

Q = 0   R = 0

Positive FDR is <u>more interpretable and easily estimable</u>: pFDR = E(V/R | R > 0)

	H_0	H_1	Total
Significant	**V**	**S**	**R**
Non-significant	**U**	**T**	**m - R**
Total	$\mathbf{m_0}$	$\mathbf{m - m_0}$	**m**

Estimating pFDR (Storey, 2002)

1. Set a series of rejection regions $[0, \gamma_j]$.
    a. You can/should simply set this to the observed p-values. Then, you get a pFDR estimate for any p

2. For each rejection region (for each p-value), estimate the pFDR

    $\gamma_j m_0 / R(\gamma_j)$, where $R(\gamma_j) = \#(p \leq \gamma_j)$

    There are several methods to estimate m_0 the true number of null hypotheses.

q-value

The <u>minimum</u> false discovery rate at which the test may be called significant

Individual q-values can be calculated and are associated with individual p-values

e.g., get $q_1, \ldots, q_m$ from $p_1, \ldots, p_m$.

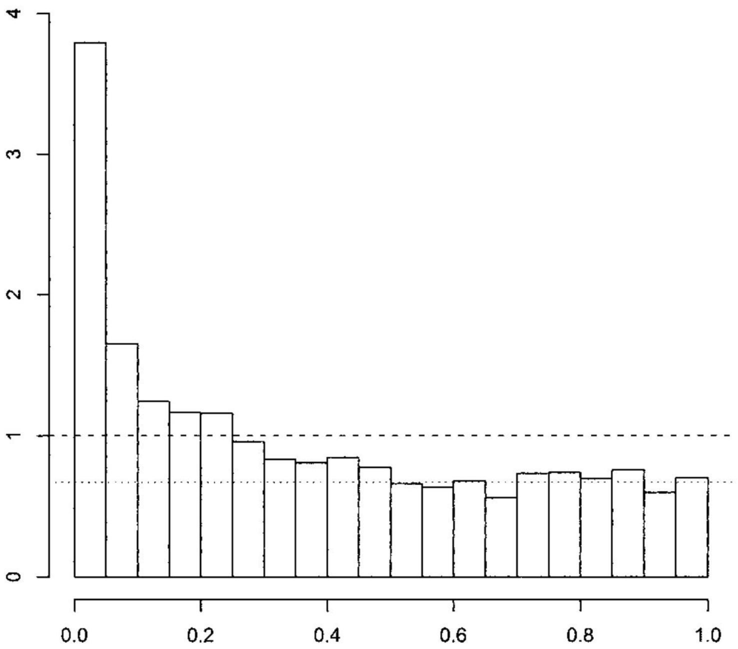
Then, you can threshold q-values at appropriate FDR level

<u>how many false discoveries are you willing to accept?</u>

Given a set of q-values, rejecting the null hypotheses whose $q_i \leq b$ ensures that the

$$E[Q] = b$$

Histogram of p-values



Estimate of proportion of null p-values ~ m0

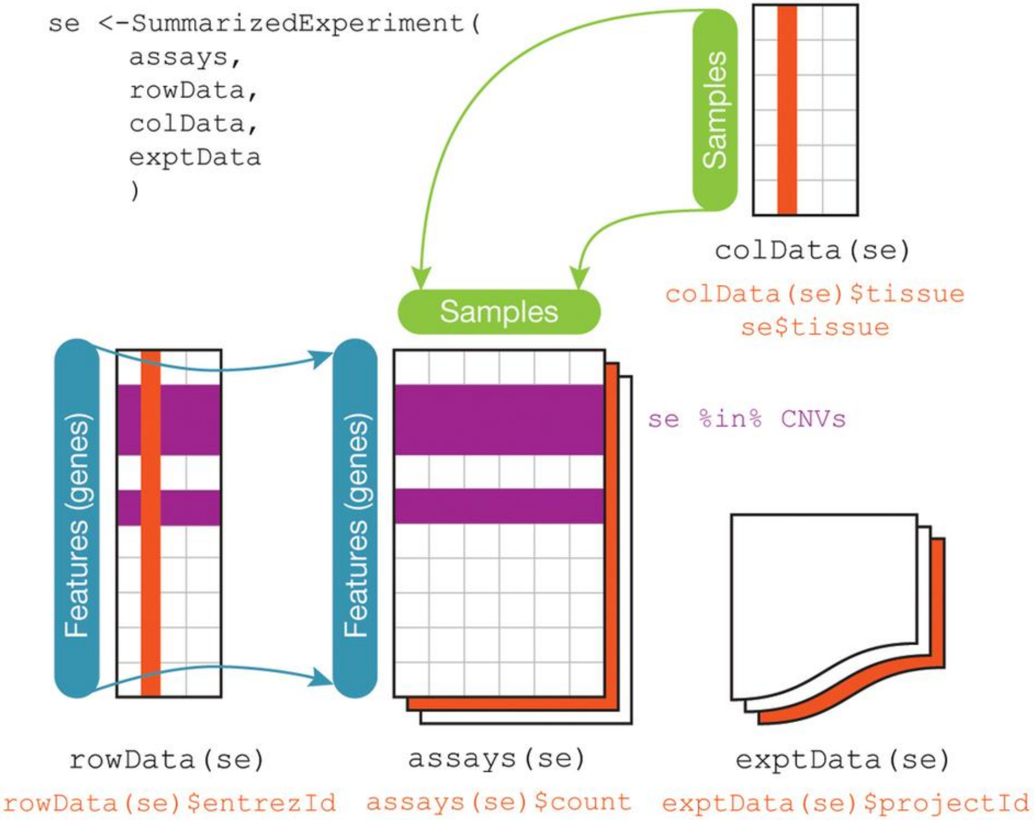
Fig. 1. A density histogram of the 3,170 *p* values from the Hedenfalk *et al.* (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null *p* values.

Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays

C57BL/6J (B6) and DBA/2J (D2) are two of the most commonly used inbred mouse strains in neuroscience research. However, the only currently available mouse genome is based entirely on the B6 strain sequence. Subsequently, oligonucleotide microarray probes are based solely on this B6 reference sequence, making their application for gene expression profiling comparisons across mouse strains dubious due to their allelic sequence differences, including single nucleotide polymorphisms (SNPs).

The emergence of next-generation sequencing (NGS) and the RNA-Seq application provides a clear alternative to oligonucleotide arrays for detecting differential gene expression without the problems inherent to hybridization-based technologies.

Organizing genomic data



Huber et al. 2015

Quality Controls

The most important tool in maintaining high quality gene expression data is a careful study design with both biological and technical replicates, which allow direct assessments of biological and technical variations

biological replicates: a set of samples taken from a set of multiple unique individuals such that each individual contributes a given sample

A technical replicate: a sample that has been partitioned and carried through the sample preparation process from a given point forward

Quality Controls

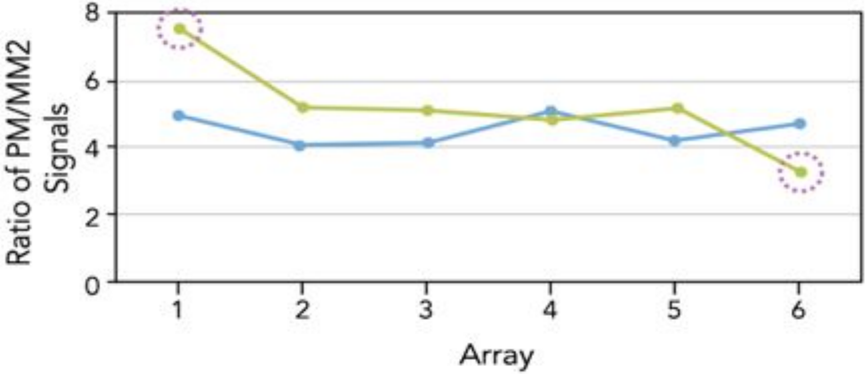
“Internal control features” in many microarray or sequencing technologies could offer another way to maintain high quality

Signal intensity values of hybridization controls

Housekeeping genes should be fairly consistent across arrays when from a similar sample source.

Perfect Match (PM) & Mismatch (MM2). The PM probe signal is expected to be higher than the MM2 probe signal

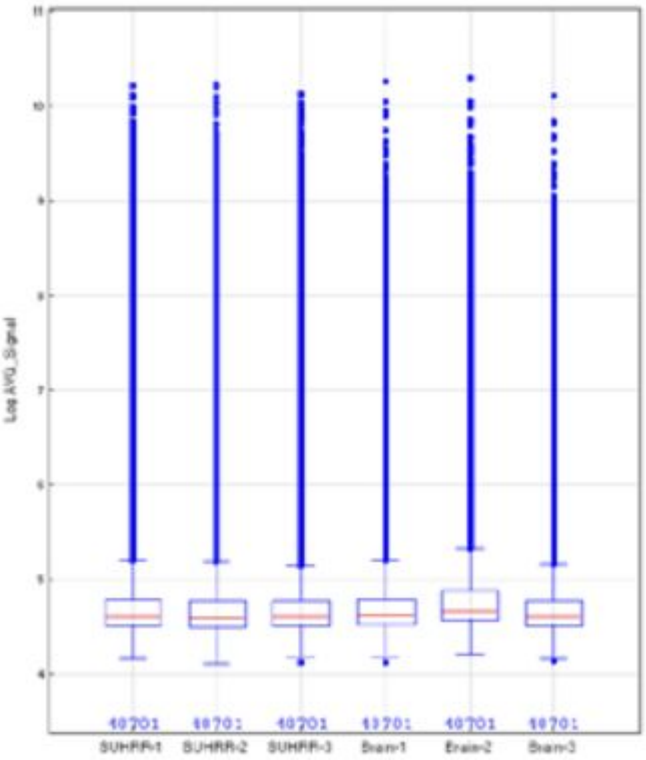
PM/MM2 probe signals



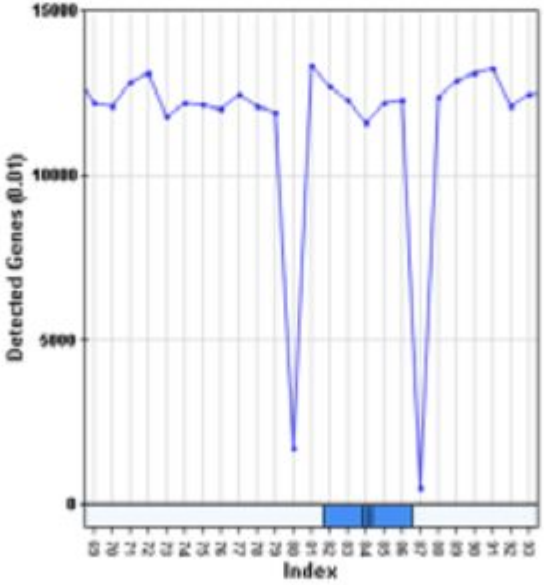
A plot of the ratio of PM/MM2 probe signals across several samples from two different BeadChips (blue and green). In the case of the blue BeadChip, all six samples have similar ratios approximately 4–5 PM/MM2. However, some arrays from the green BeadChip (circled) exhibit deviating ratios, indicating a possible difference in stringency between arrays 1 and 6.

Gene Expression Microarray Data Quality Control by Illumina

Boxplot of (Log) Signal Intensities



of detected genes



This criteria become more important and widely used with some RNA-seq that have a lot of “zeros”

Data wrangling and pre-processing

The process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data.

In genomics, we must convert the outputs from high-throughput genomic technologies into a convenient format. Similarly, we may want to further change the data structure that is more easily analyzed in R or other software

→

Low-level
Images,
Signals,
Spectra

→

FASTA: sequence records
FASTQ: with quality scores
SAM: with mapping info

→

What are we interested in?
Summarize (e.g., genes?)
Filter/select (e.g., coding?)

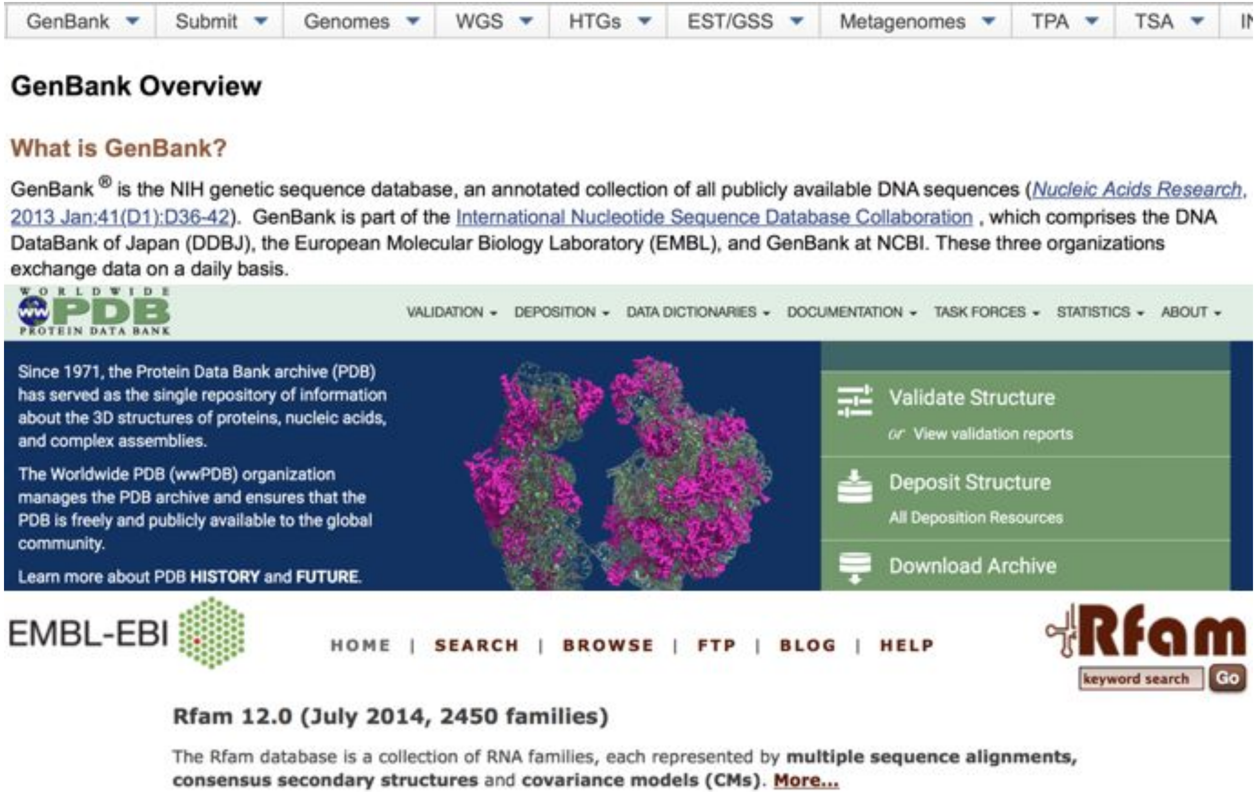
Data curation

Organization and integration of data collected from various sources, annotation of the data, and publication and presentation of the data such that the value of the data is maintained over time, and the data remains available for reuse and preservation

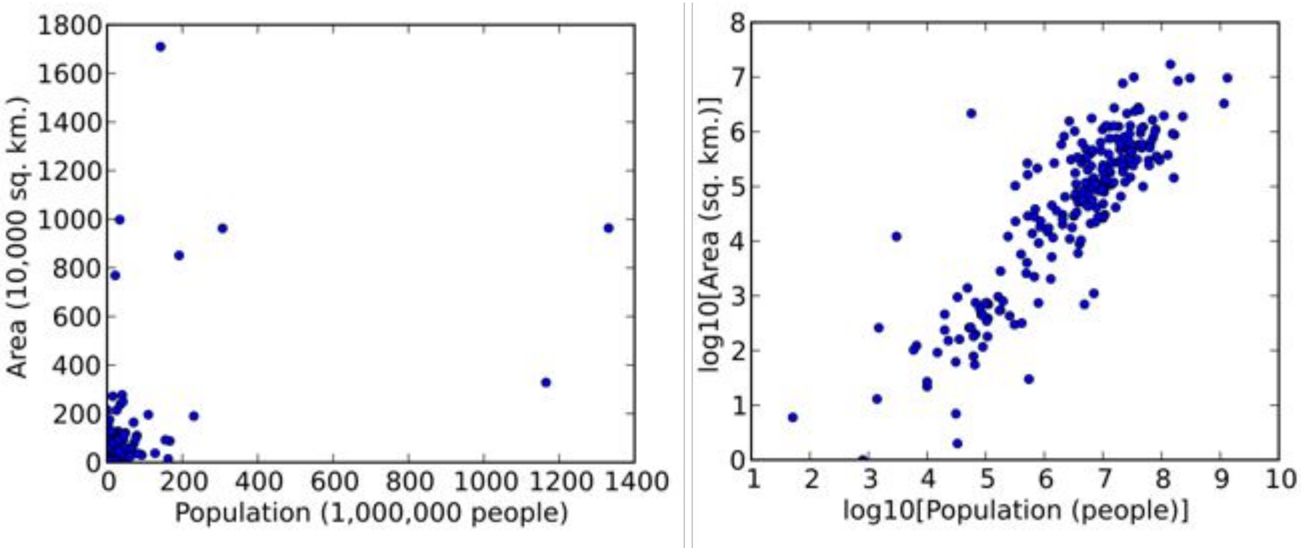
ReCount project

Curated RNA-seq data from many sources

Curated Biological Databases

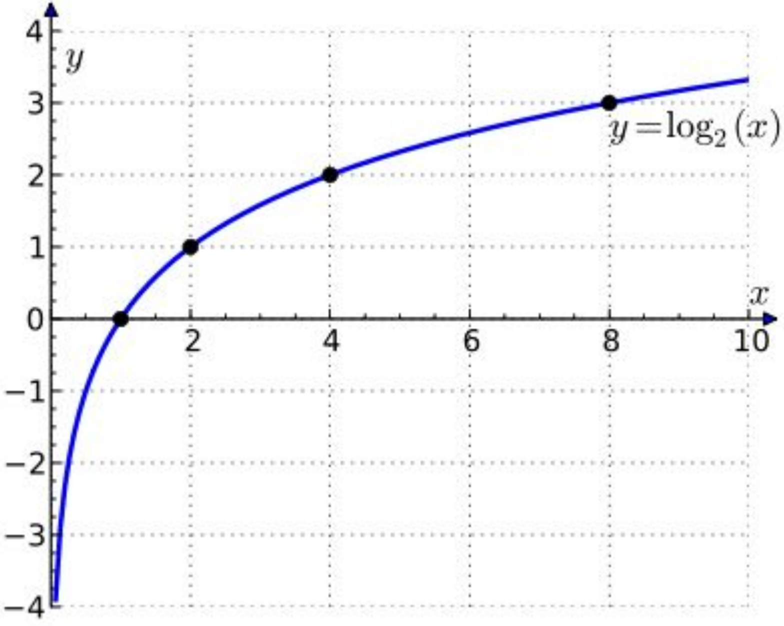


Data transformation



Transform all of the data points using a deterministic function,
to better suit the underlying assumptions of statistical procedures

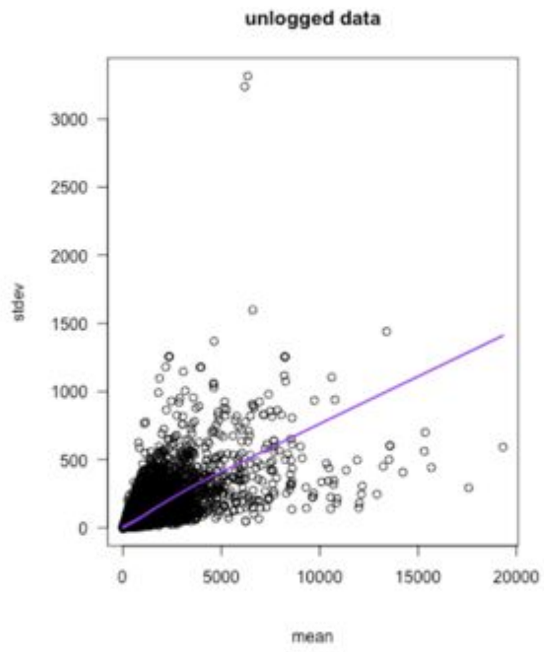
Log transformation

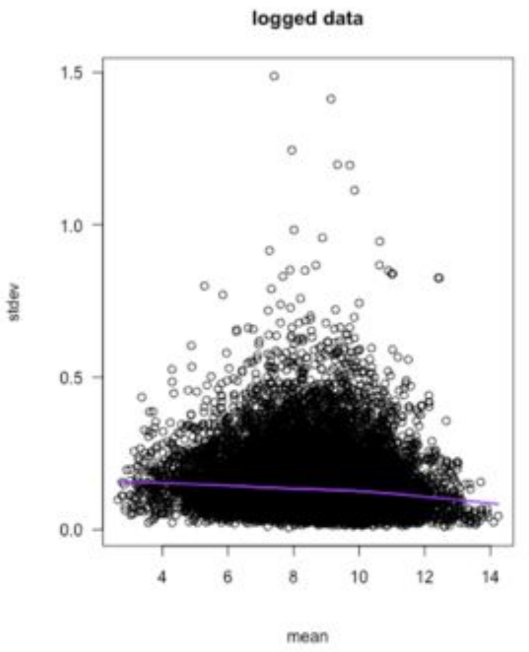


Very popular, due to a need to scale numeric data in [0, inf) to (-inf, inf)

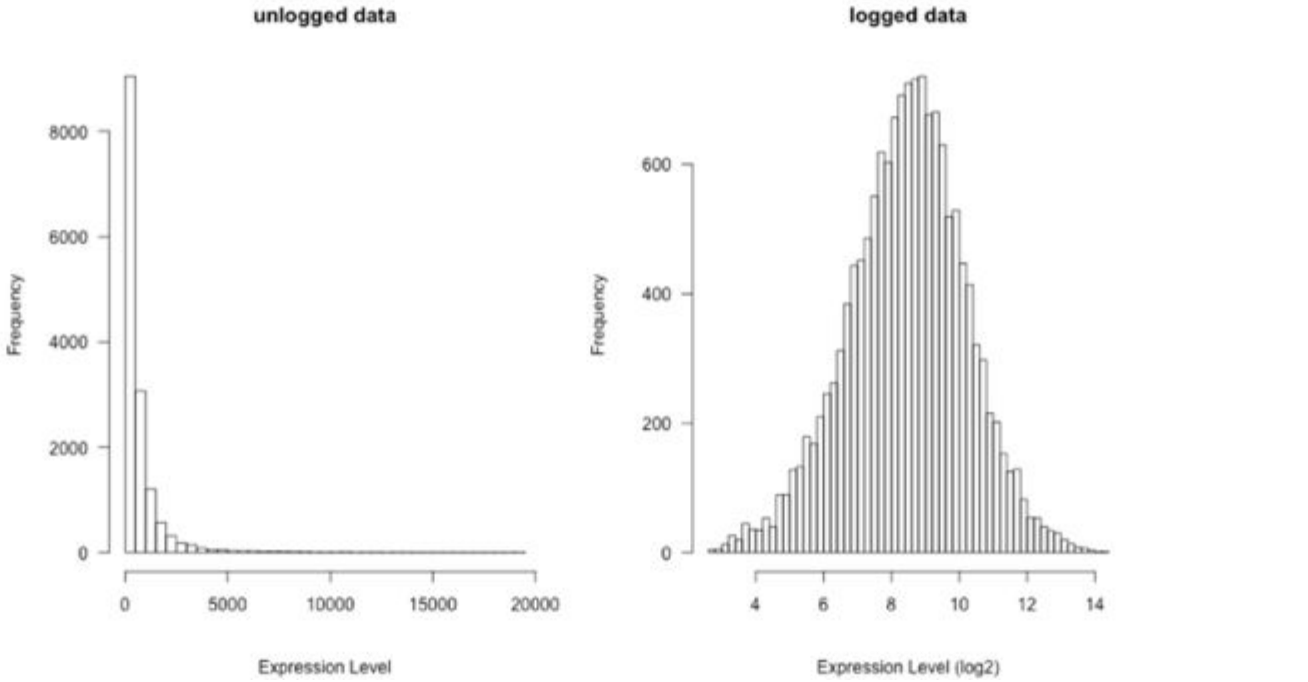
Log transformation

Stabilizes the variance & Compresses the range of data





Log transformation



Perhaps a Normally distributed data may be better for downstream analyses or fit our assumption about the population better