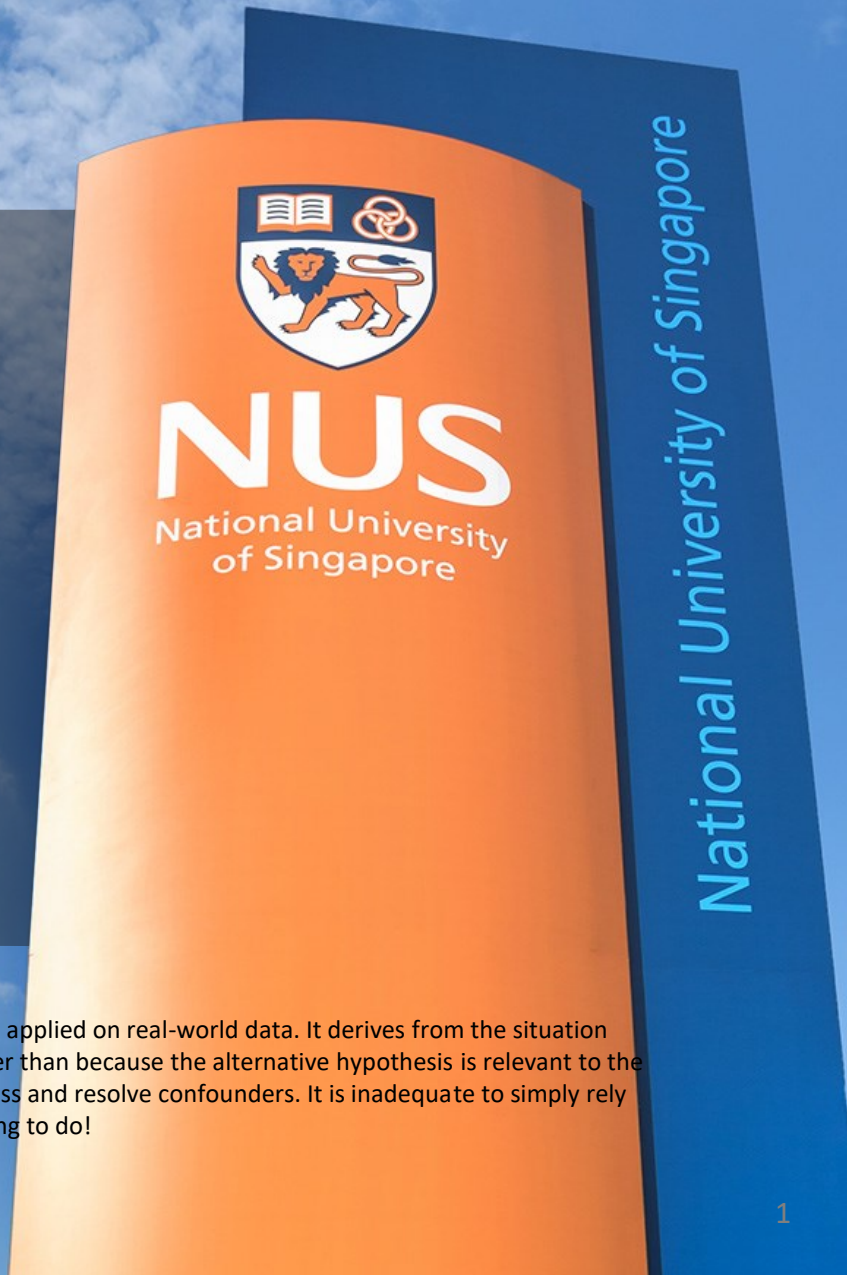
The reluctant data scientist

Wong Limsoon



Outline: There is a big theory–practice gap that exists when theoretical statistics are applied on real-world data. It derives from the situation where the null hypothesis is rejected for extraneous reasons (or confounders), rather than because the alternative hypothesis is relevant to the disease phenotype. The mechanics of applying statistical tests therefore must address and resolve confounders. It is inadequate to simply rely on manipulating the P-value; indeed, I will show how/why this can be the wrong thing to do!

Hypothesis testing

Steps of hypothesis testing

Formulate null H_0 and alternate hypothesis H_1

Devise a test statistic, $t(\cdot)$

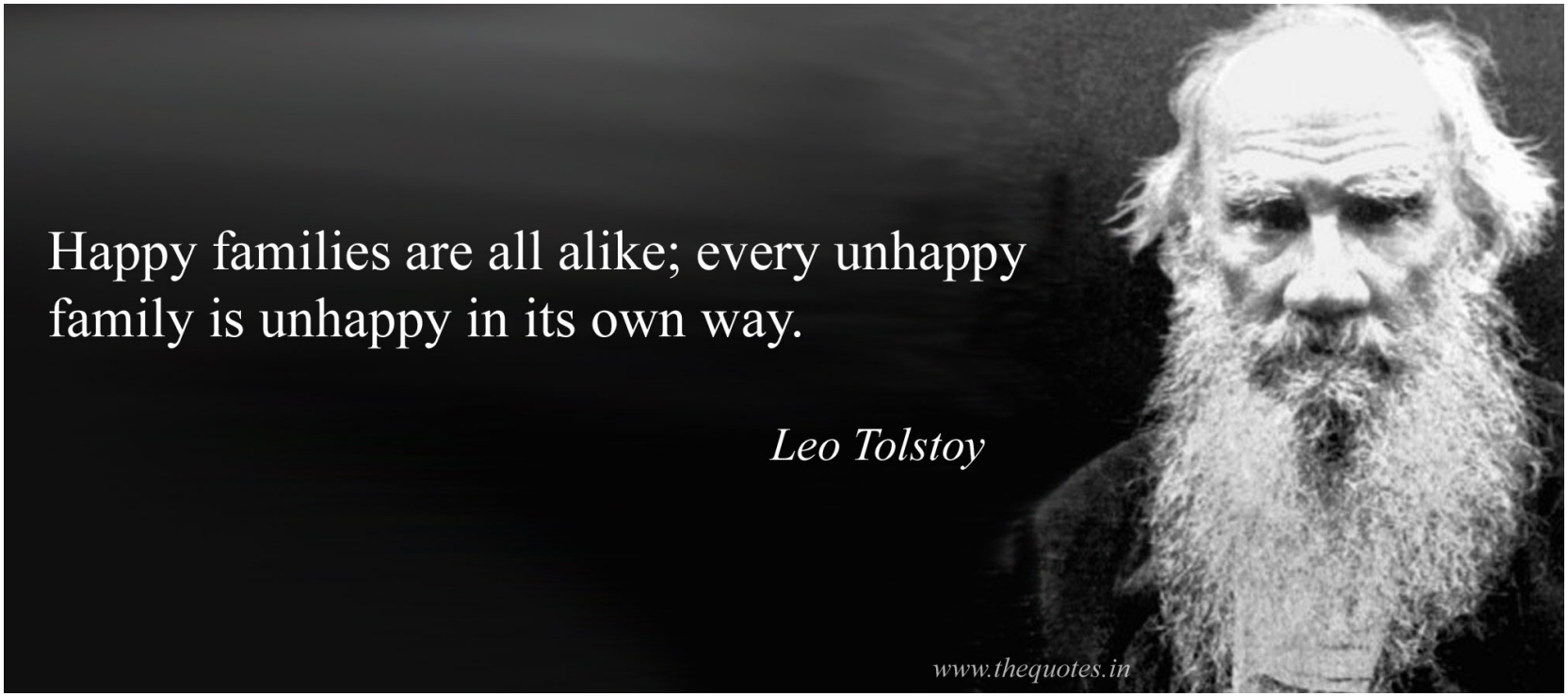
Evaluate $t(S)$ on a sample S

Compare $t(S)$ to the null distribution

If significant, accept H_1; otherwise, accept H_0

Null distribution is the distribution of $t(\cdot)$ over the set of null samples for which H_0 holds

Anna Karenina



Anna Karenina Principle

There are many ways to violate the null hypothesis but only one way that is truly pertinent to the outcome of interest

Sample is biased

Null distribution used is inappropriate

Null / alternative hypothesis incorrectly stated

Inappropriate expt design

Biased sample

Exercise #1

SNP	Genotypes	Group: Controls [n(%)]		Group: Cases [n(%)]		χ^2	*P* value
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

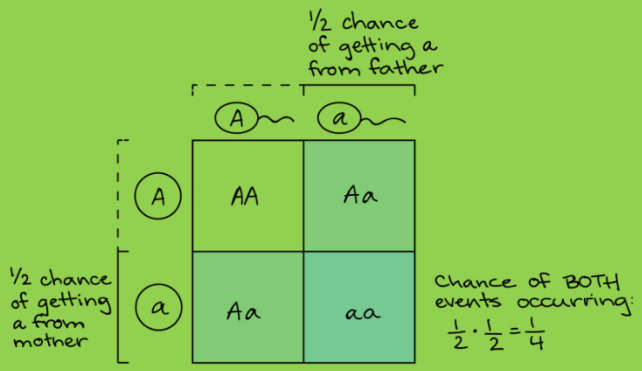
SNP rs123 is a great biomarker for a disease, based on a prospective study

If rs123 is AA or GG, unlikely to get the disease

If rs123 is AG, ~3x higher risk of disease

A straightforward $\chi 2$ test. Anything wrong?

There may be sample bias



Basic rule of human genetics

SNP	Genotypes	Group: Controls [n(%)]		Group: Cases [n(%)]		χ^2	*P* value
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

AG = 38 + 79 = 117,

Controls + cases = 189

⇒ Population ~62% AG

⇒ Population >9% AA, unless AA is lethal

"Big data check" shows AA is non-lethal for this SNP ⇒ sample is biased

Careless null hypothesis

"Effective" H_0

rs123 alleles are identically distributed in the two samples

Assumption

Distributions of rs123 alleles in the two samples are identical to the two populations

Apparent H_1

rs123 alleles are differently distributed in the two populations

"Effective" H_1

rs123 alleles are differently distributed in the two populations OR

Distribution of rs123 alleles in the two samples are not identical to the two populations

Exercise #2

Suppose distributions of rs123 alleles in the two samples are identical to the corresponding populations and the test is significant

Can we say rs123 mutation causes the disease?

When two genes are close together, this is what happens during meiosis

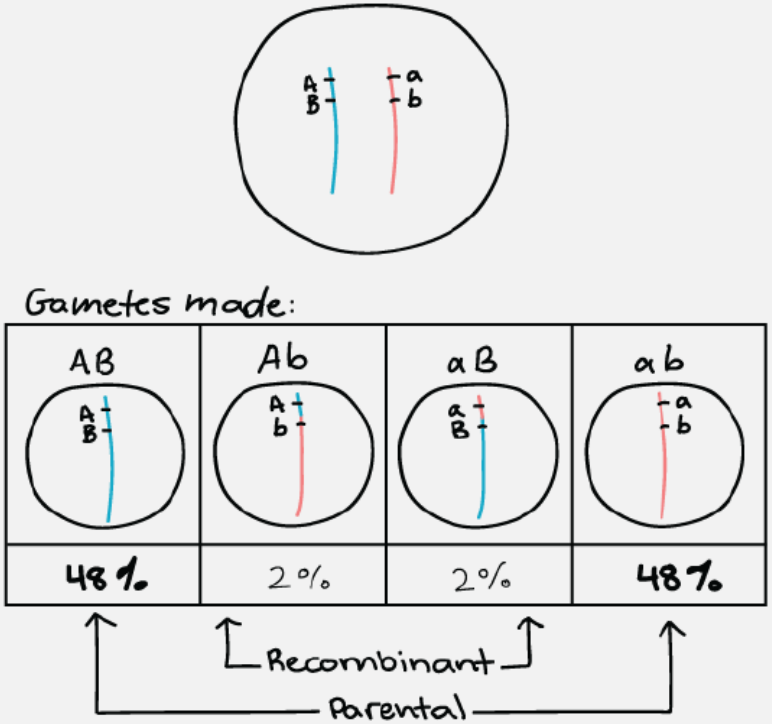
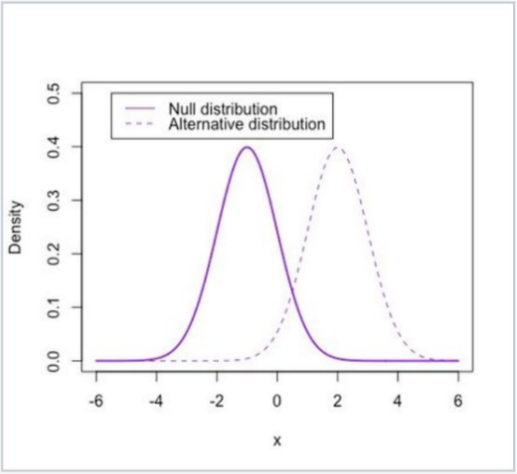


Image credit: Khan Academy

In statistical hypothesis testing, the **null distribution** is the probability **distribution** of the test statistic when the **null** hypothesis is true. For example, in an F-test, the **null distribution** is an F-**distribution**.



Null and alternative distribution

Inappropriate null distribution

Synthetic lethality

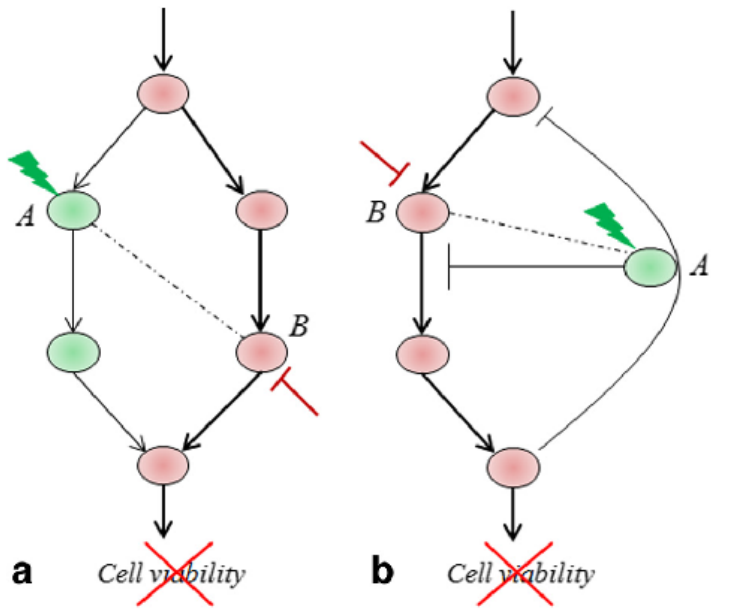


Fig. 7 Two models for pathway-based targeting of synthetic lethal genes *B* in conjunction with deleted/downregulated genes *A*: **a** parallel pathways model where targeting *B* results in disruption of both survival pathways, and **b** negative feedback-loop model where targeting *B* shunts of (forward) signals for cell survival

Why interested in synthetic lethality?

Synthetic-lethal partners of frequently mutated genes in cancer are likely good treatment targets

Srihari et al. Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. *Biology Direct*, 10:57, 2015.

Synthetic lethal pairs

Fact:

When a pair of genes is synthetic lethal, mutations of these two genes avoid each other

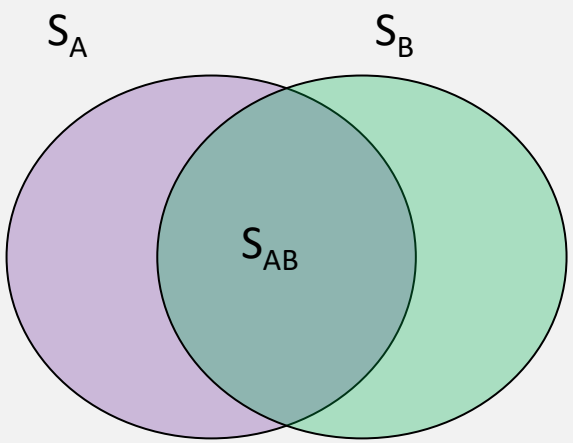
Observation:

Mutations in genes (A,B) are seldom observed in the same subjects

Conclusion by abduction:

Genes (A,B) are synthetic lethal

Exercise #3



$$P[X \leq |S_{AB}|] = 1 - P[X > |S_{AB}|], \quad (1)$$

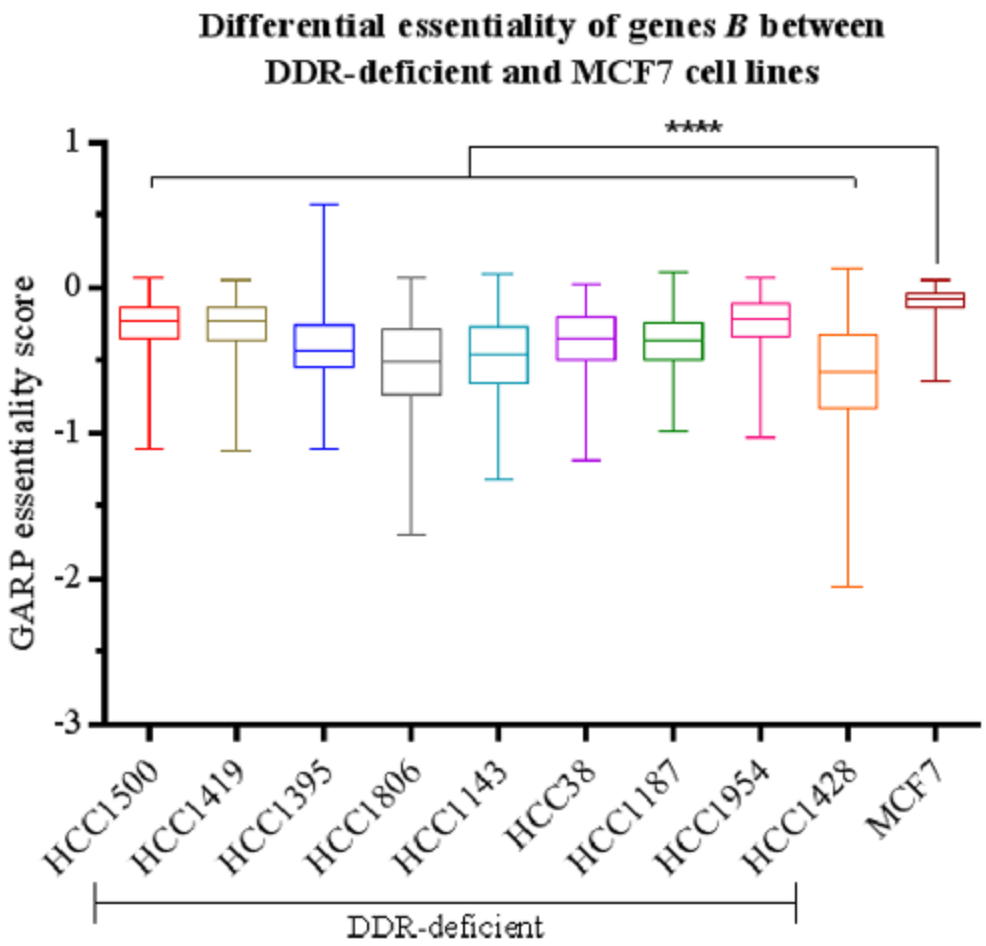
where $P\ [X > |S_{AB}|]$ is computed using the hypergeometric probability mass function for $X = k > |S_{AB}|$:

$$P[X > |S_{AB}|] = \sum_{k=|S_{AB}|+1}^{|S_B|} \frac{\binom{|S_A|}{k}\binom{|S|-|S_A|}{|S_B|-k}}{\binom{|S|}{|S_B|}}$$

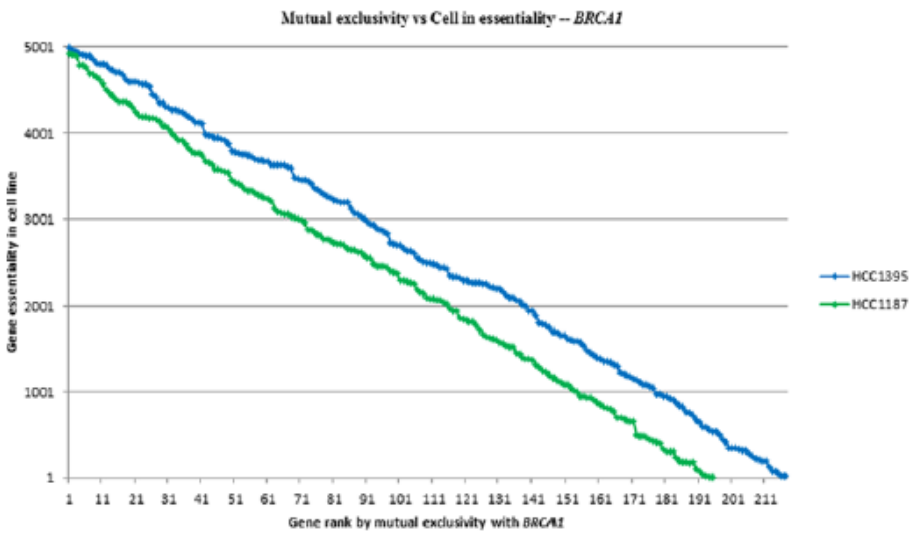
Mutations of genes (A,B) avoid each other if $P[X \leq S_{AB}] \leq 0.05$

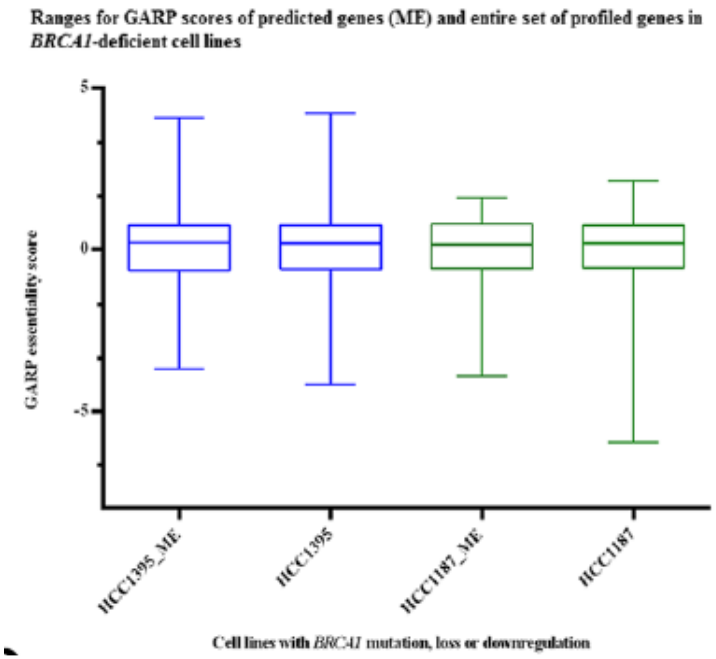
Anything wrong with this?

Seems to work fine



What is happening?

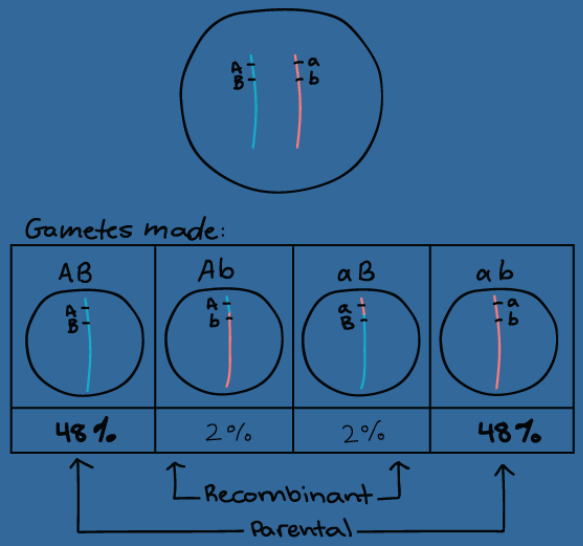




Among top ME-genes, GARP score ranks correlate with mutual exclusion ranks

But GARP scores of ME-genes (i.e. have mutually exclusive mutations to BRCA1) are like other genes

Hyper-geometric distribution doesn't reflect real mutations



Hypergeometric distribution

Mutations are independent

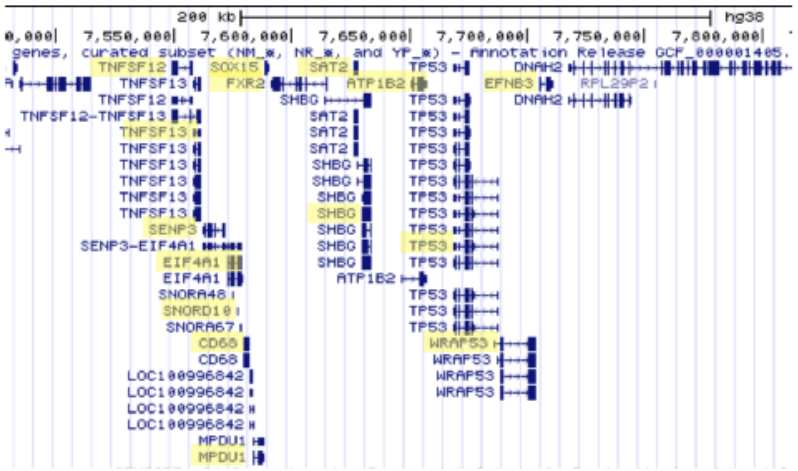
Mutations have equal chance to appear in a subject

Real-life mutations

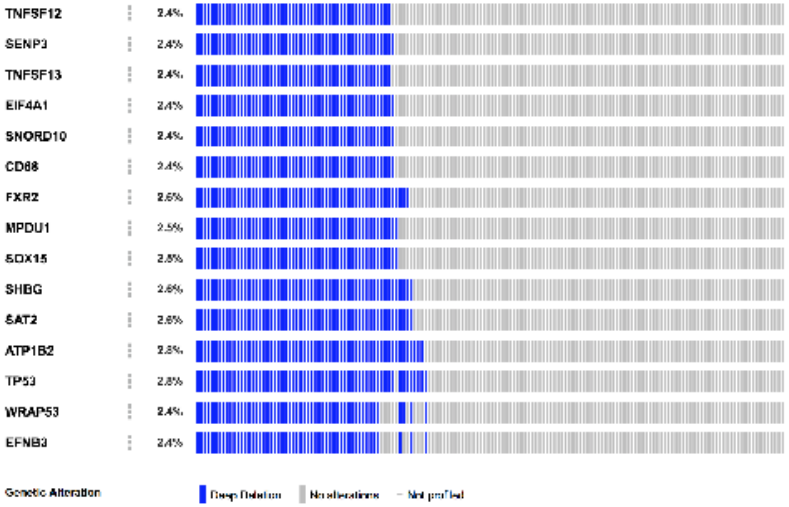
Inherited in blocks; those close to each other are correlated

Some subjects have more mutations than others, e.g. those w/ defective DNA-repair genes

Real-life example: Mutations of TP53 and its neighbours



(a) Genomic location of genes close to TP53

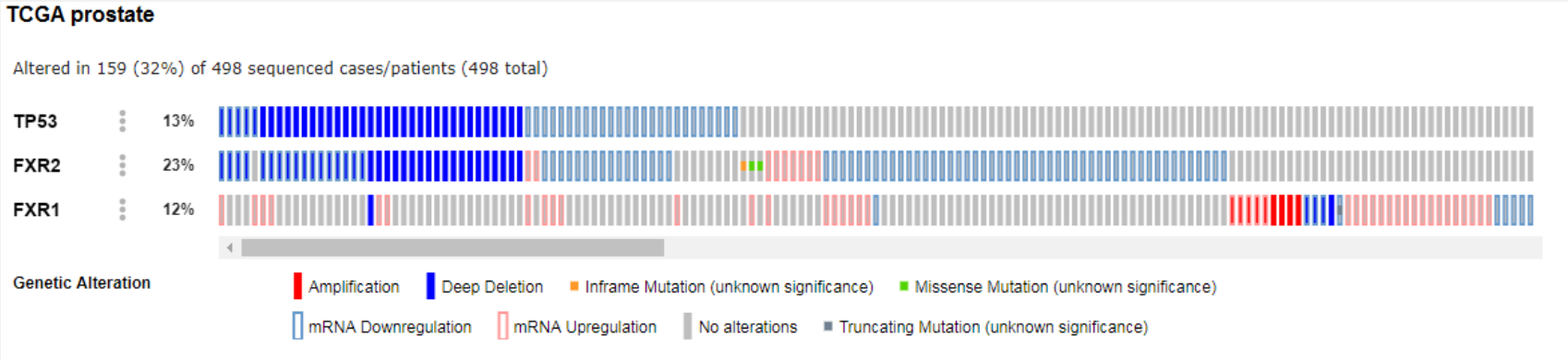


(b) CNA profile of genes close to TP53

Exercise #4

FXR2 is located near TP53

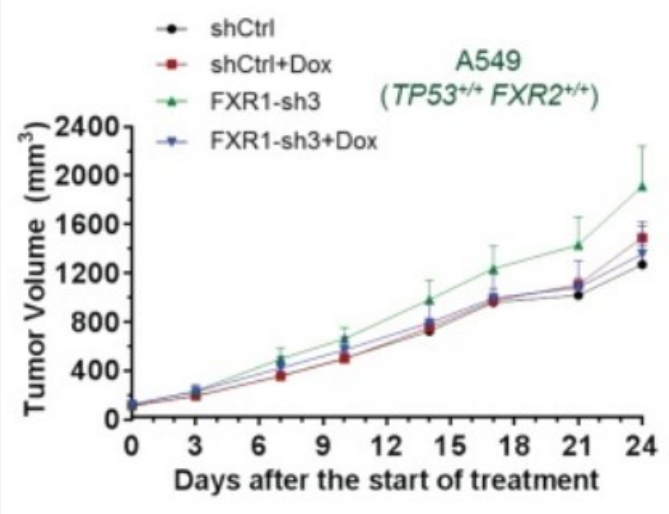
FXR1 and FXR2 buffer each other's function

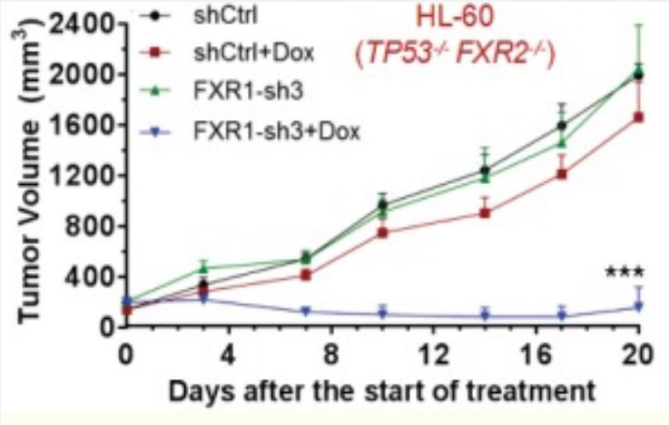


Is FXR1 synthetic lethal to TP53?

Does inhibiting FXR1 lead to cell death for TP53-deleted cell lines?

Tumour bearing homozygous TP53/FXR2 co-deletion shrinks upon doxycycline-induced FXR1 knock down



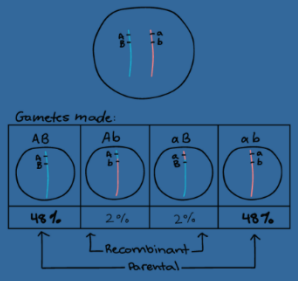


Fan et al., eLife, 6:e26129, 2017

Exercise #5

Propose some possible solutions to this problem

Hyper-geometric distribution doesn't reflect real mutations



Hypergeometric distribution

Mutations are independent

Mutations equal chance to appear in a subject

Real-life mutations

Inherited in blocks; those close to each other are correlated

Some subjects have more mutations than others, e.g. those with defective DNA-repair genes

Inappropriate experiment design

Exercise #6

Overall

	A	B
lived	60	65
died	100	165

Treatment A is better

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

Treatment B is better

What is happening here?

A/B sample not equalized in other attributes, e.g. sex

Overall

	A	B
lived	60	65
died	100	165

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

Taking A

Men = 100 (63%)

Women = 60 (37%)

Taking B

Men = 210 (91%)

Women= 20 (9%)

Careless null hypothesis

"Effective" H_0

Treatment effects are identically distributed in the two samples

Assumption

All other factors are equalized in the two samples

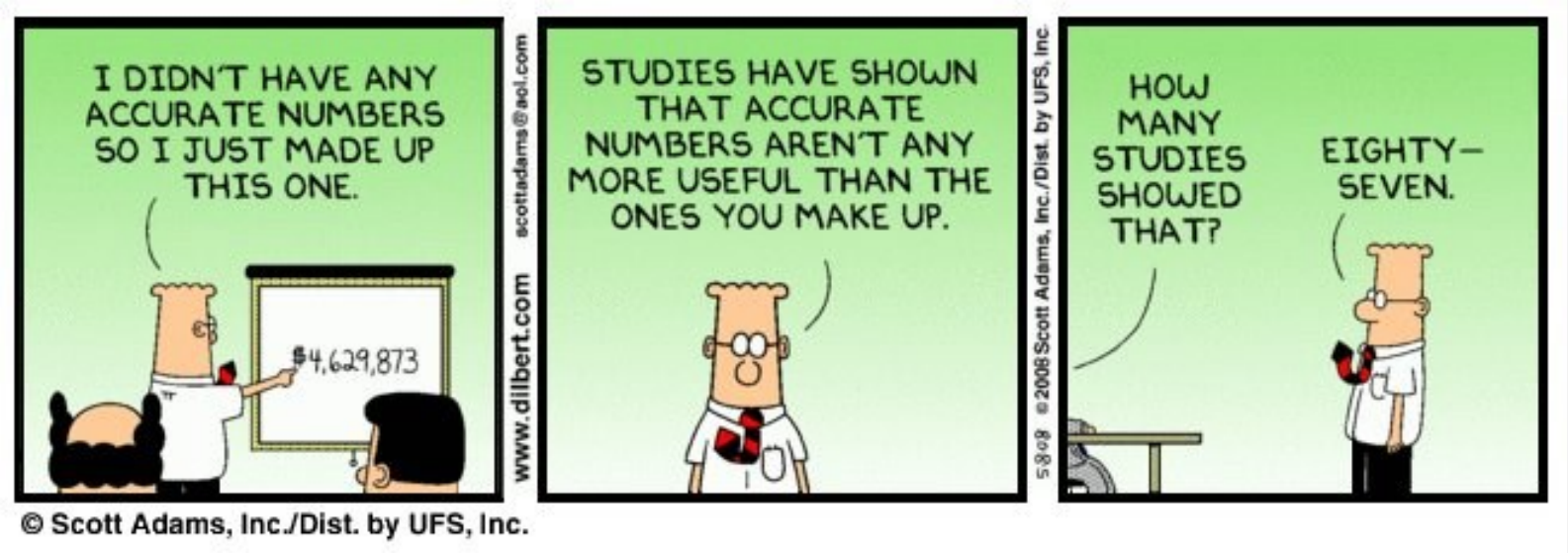
Apparent H_1

Treatment effects are differently distributed in the two populations

"Effective" H_1

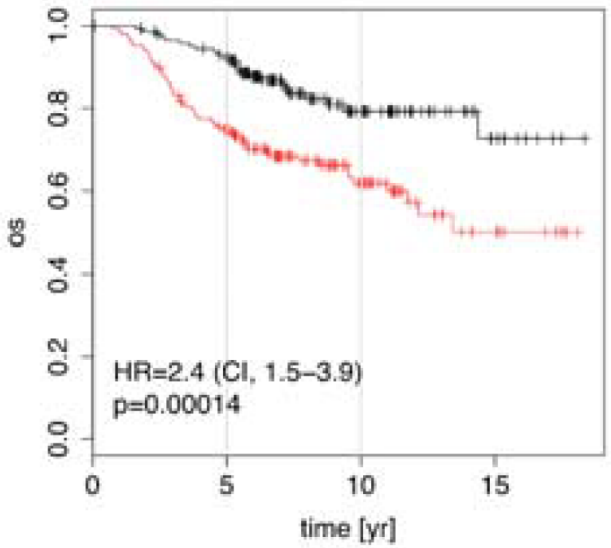
Treatment effects are differently distributed in the two populations OR

Some other factors aren't equalized in the two samples



Confounders abound

Venet et al., *PLOS Comput Biol*, 2011



A seemingly obvious conclusion

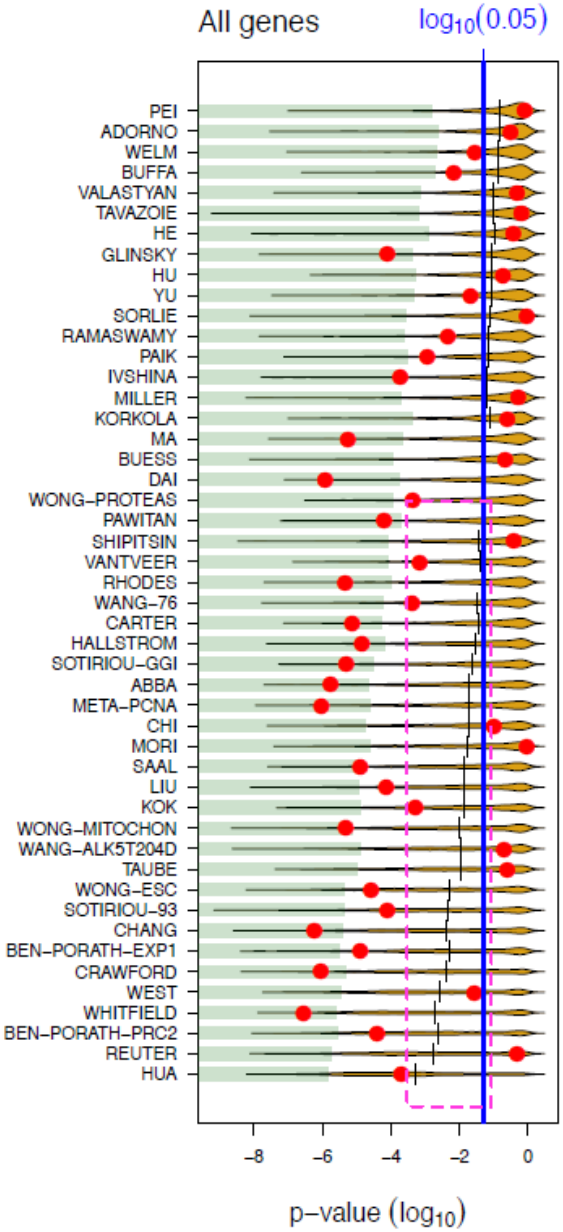
A multi-gene signature (social defeat in mice) good as a biomarker for breast cancer survival

Cox's survival model p-value << 0.05

A straightforward Cox's analysis. Anything wrong?

Almost all random signatures also have p-value < 0.05

Venet et al., *PLOS Comput Biol*, 2011



What makes random signatures significant?

Proliferation is a hallmark of cancer

Hypothesis: Proliferation-associated genes make a signature significant

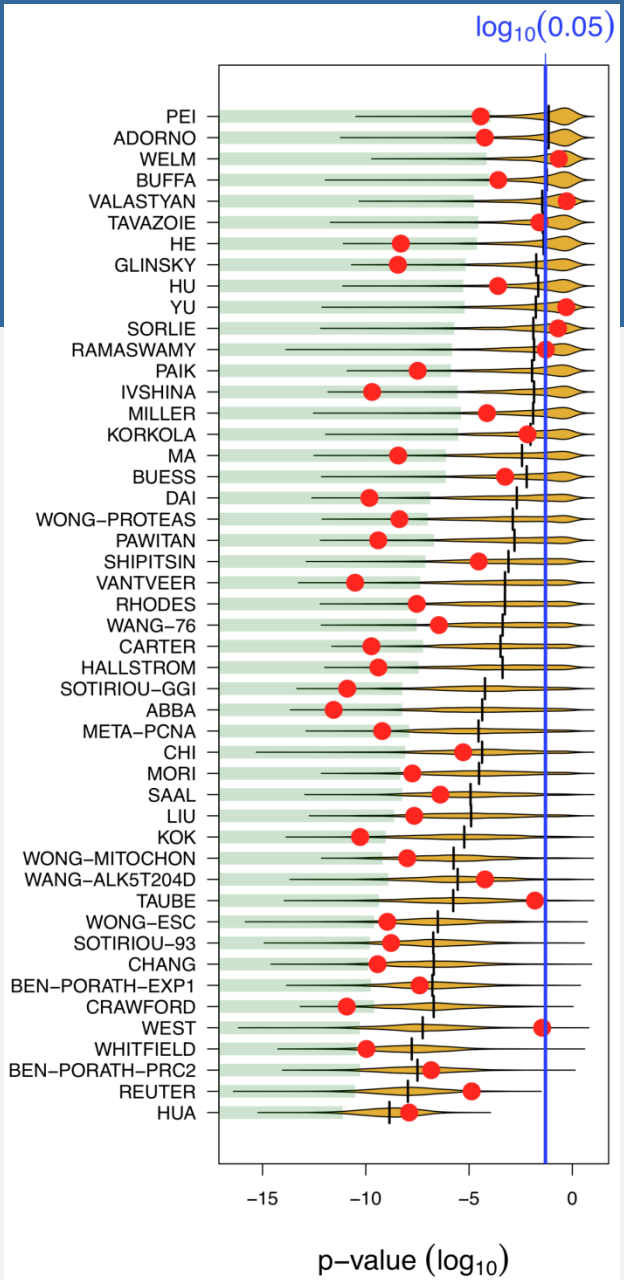
# of random signatures w/ ≥1 prolif gene

Cutoffs	Counts		
	NP	P	Marginals
Above 0.05	7043	19 043	26 086
Below 0.05	2766	19 148	21 914
Marginals	9809	38 191	48 000

Exercise #7

40-50% of random signatures have p-value << 0.05

How to get rid of them?



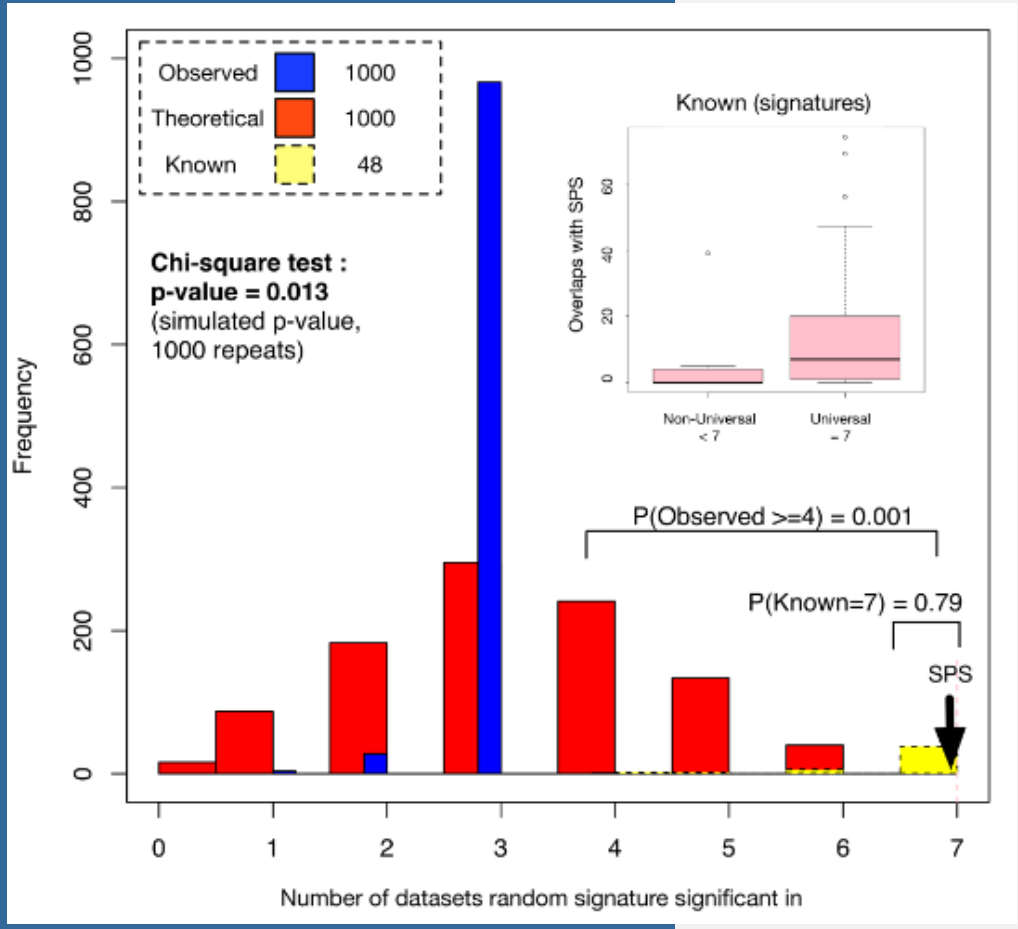
An engineer's solution

n	$(50\%)^n$
1	50.00%
2	25.00%
3	12.50%
4	6.25%
5	3.13%
6	1.60%
7	0.78%

Test using at least 7 independent test sets

Test on many datasets

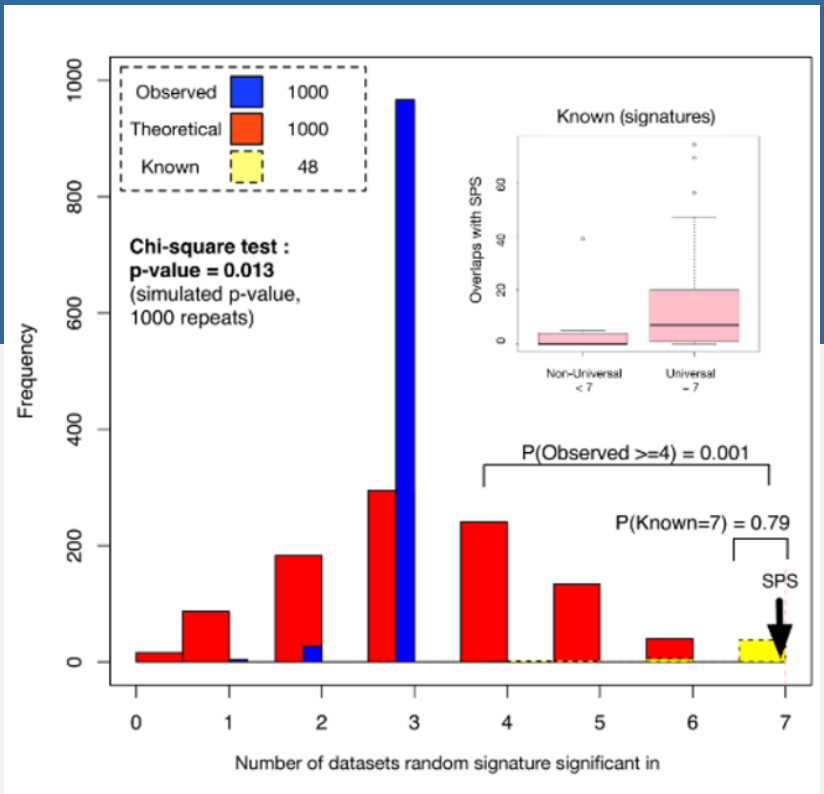
Goh & Wong. Turning straw into gold: Building robustness into gene signature inference. *Drug Discovery Today*, 24(1):31-36, 2019.



Validated signatures are universally significant

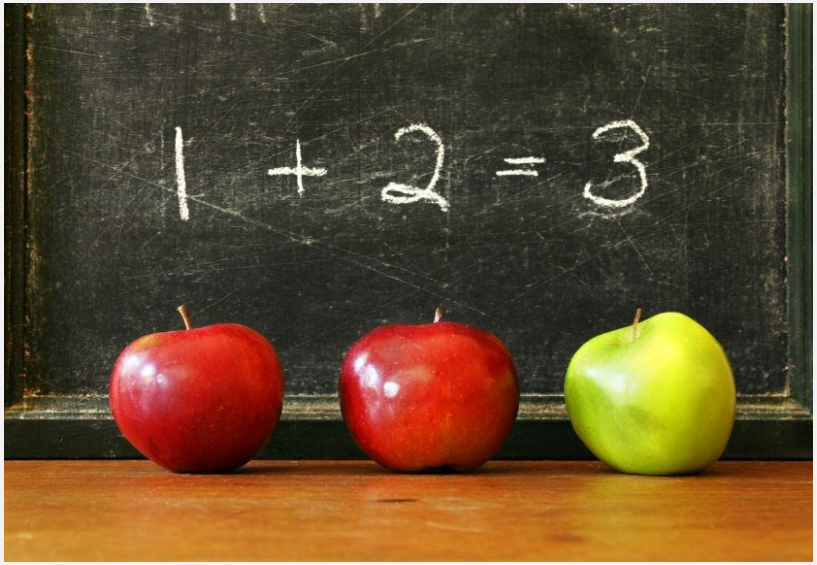
Random signatures are not universal, even though they get better p-values than known signatures on some datasets

Exercise #8



The red bars show the theoretical binomial distribution on expected # of random signatures that should be significant on n datasets

What do you think is happening here?


1 + 2 = 3

What have we learned?

When a statistical test is significant, think again!

Sample is biased

Null distribution used is inappropriate

Null / alternative hypothesis incorrectly stated

Inappropriate expt design

Confounders are aplenty

“Independent” test data are not as independent as you think

References

Goh & Wong. Dealing with confounders in –omics analysis. *TIBTECH*, 36(5):488-498, 2018

Srihari et al. Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. *Biology Direct*, 10:57, 2015.

Pinoli et al. Identifying collateral and synthetic lethal vulnerabilities within the DNA-damage response. *BMC Bioinformatics*, 22:250, 2021

Goh & Wong. Why breast cancer signatures are no better than random signatures explained. *Drug Discovery Today*, 23(11):1818-1823, 2018

Goh & Wong. Turning straw into gold: Building robustness into gene signature inference. *Drug Discovery Today*, 24(1):31-36, 2019

Ho et al. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns*, 1(8):100129, 2020