Anna Karenina Principle

Wong Limsoon

Outline: The Anna Karenina effect is a manifestation of the 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! I will discuss some mechanistic elements with real-life examples, and suggest how they can be logically designed to foil the Anna Karenina effect.
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 $t(S_0)$ where $S_0$ ranges over the set of null samples $S_0$ for which $H_0$ holds
Anna Karenina

Happy families are all alike; every unhappy family is unhappy in its own way.

*Leo Tolstoy*
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*

*And so on*
Biased sample
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

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.
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**.
Synthetic lethality

Why interested in synthetic lethality?

Synthetic-lethal partners of frequently mutated genes in cancer are likely good treatment targets.
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.
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.
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
Propose some possible solutions to this problem

Hypergeometric distribution doesn’t reflect real mutations

- 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

<table>
<thead>
<tr>
<th>Context</th>
<th>Occupation</th>
<th>Income&gt;50K</th>
<th>Income&lt;50K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race = White</td>
<td>Adm-clerical</td>
<td>439 (14%)</td>
<td>2,645 (86%)</td>
</tr>
<tr>
<td></td>
<td>Craft-repair</td>
<td>844 (23%)</td>
<td>2,850 (77%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context</th>
<th>Occupation</th>
<th>Income&gt;50K</th>
<th>Income&lt;50K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race = White,</td>
<td>Adm-clerical</td>
<td>16 (35%)</td>
<td>30 (65%)</td>
</tr>
<tr>
<td>Workclass = Self-emp-not-inc</td>
<td>Craft-repair</td>
<td>90 (18%)</td>
<td>409 (82%)</td>
</tr>
</tbody>
</table>

What is happening here?
Adm/Craft sample not equalized in other attributes, e.g. sex

<table>
<thead>
<tr>
<th>Context</th>
<th>Occupation</th>
<th>Income&gt;50K</th>
<th>Income&lt;50K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race = White,</td>
<td>Adm-clerical</td>
<td>251 (24%)</td>
<td>787 (76%)</td>
</tr>
<tr>
<td>Sex = Male</td>
<td>Craft-repair</td>
<td>829 (24%)</td>
<td>2,695 (76%)</td>
</tr>
</tbody>
</table>

What is happening here?

<table>
<thead>
<tr>
<th>Context</th>
<th>Occupation</th>
<th>Income&gt;50K</th>
<th>Income&lt;50K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race = White, Workclass = Self-emp-not-inc</td>
<td>Adm-clerical</td>
<td>16 (35%)</td>
<td>30 (65%)</td>
</tr>
<tr>
<td></td>
<td>Craft-repair</td>
<td>90 (18%)</td>
<td>409 (82%)</td>
</tr>
</tbody>
</table>

Wong., Stat & Prob Lett, 2018
Careless null hypothesis

“Effective” $H_0$
Income is identically distributed in the two samples

Assumption
All other factors are equalized in the two samples

Apparent $H_1$
Income is differently distributed in the two populations

“Effective” $H_1$
Income is differently distributed in the two populations OR
Some other factors aren’t equalized in the two samples
Confounders abound
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

<table>
<thead>
<tr>
<th>Cutoffs</th>
<th>Counts</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NP</td>
<td>P</td>
</tr>
<tr>
<td>Above 0.05</td>
<td>7043</td>
<td>19 043</td>
</tr>
<tr>
<td>Below 0.05</td>
<td>2766</td>
<td>19 148</td>
</tr>
<tr>
<td>Marginals</td>
<td>9809</td>
<td>38 191</td>
</tr>
</tbody>
</table>

# of random signatures w/ ≥1 prolif gene
Exercise #7

40-50% of random signatures have p-value $\ll 0.05$

How to get rid of them?
An engineer’s solution

<table>
<thead>
<tr>
<th>n</th>
<th>((50%)^n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50.00%</td>
</tr>
<tr>
<td>2</td>
<td>25.00%</td>
</tr>
<tr>
<td>3</td>
<td>12.50%</td>
</tr>
<tr>
<td>4</td>
<td>6.25%</td>
</tr>
<tr>
<td>5</td>
<td>3.13%</td>
</tr>
<tr>
<td>6</td>
<td>1.60%</td>
</tr>
<tr>
<td>7</td>
<td>0.78%</td>
</tr>
</tbody>
</table>

Test using at least 7 independent test sets
Test on many datasets

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


## Projects

**Project 1. Vanderbilt Study: GRE score and PhD performance**
What are the main claims of this study? Can you find some analysis/methodological bugs that might invalidate some of these claims?

**Project 2. Lung cancer and Doppelgangers**
Coudray et al. Nature Medicine 24:1559-1567, 2018
What are the main claims of this study? Can you find some analysis/methodological bugs that might invalidate some of these claims?

**Project 3. Protein function and Twilight Zone**
Seo et al. Bioinformatics 34(13):254-262, 2018
What are the claims of this study? Can you find some analysis/methodological bugs that might cast doubts on these claims?