SOC Summer School 2017

A logical introduction to computational biology

Limsoon Wong
About Part II

In Leo Tolstoy’s novel, Anna Karenina, it was famously written: “Happy families are all alike; every unhappy family is unhappy in its own way”.

This gives rise to the 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.

I will use a few examples to illustrate the Anna Karenina Principle. I will also illustrate how this Principle can be used to logically derive better analysis approaches and drive a more insightful testing and analysis of omics data.
Part II

Anna Karenina and the careless null hypothesis in omics data analysis
Anna Karenina Principle

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

*Leo Tolstoy*

Translation

- There are many ways to violate the null hypothesis but only one way that is truly pertinent to the outcome of interest
A Statistician Responds to a Marriage Proposal

GETTING THE NULL HYPOTHESIS RIGHT
Example 1
A seemingly obvious conclusion

- A scientist claims the SNP rs123 is a great biomarker for a disease
  - If rs123 is AA or GG, unlikely to get the disease
  - If rs123 is AG, a 3:1 odd of getting the disease

- A straightforward $\chi^2$ test. Anything more/wrong?
Exercise #1

<table>
<thead>
<tr>
<th>SNP</th>
<th>Genotypes</th>
<th>Controls [n(%)]</th>
<th>Cases [n(%)]</th>
<th>( \chi^2 )</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs123</td>
<td>AA</td>
<td>1</td>
<td>0</td>
<td>0.0%</td>
<td>0.00048</td>
</tr>
<tr>
<td></td>
<td>AG</td>
<td>38</td>
<td>79</td>
<td>97.5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>69</td>
<td>2</td>
<td>2.5%</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: SNP, single nucleotide polymorphism.

- What do you want to conclude from this test?
- What is the null / alternative hypothesis corresponding to this statistical test?
Exercise #2

- “Effective” H0
  - 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 H0
  - rs123 alleles are identically distributed in the two populations
- Apparent H1
  - rs123 alleles are differently distributed in the two populations

- The apparent null / alternative hypothesis is carelessly stated. Why? How to fix this?
Causal inference?

- **Refined H0**
  - Distributions of rs123 alleles in the two samples are identical to the two populations, and
  - rs123 alleles are identically distributed in the two populations

- **Refined H1**
  - Distributions of rs123 alleles in the two samples are different from the two populations, or
  - rs123 alleles are differently distributed in the two populations

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

- **Can we say rs123 mutation causes the disease?**
Three types of reasoning

• Deduction
  – All men are mortal
  – Socrates is a man
  ⇒ Socrates is mortal

• Induction
  – Socrates is a man
  – Socrates is mortal
  ⇒ All men are mortal, provided there is no counter example

• Abduction
  – All men are mortal
  – Socrates is mortal
  ⇒ Socrates is a man, provided there is no other explanation of Socrates' mortality
Abduction in action

- **Hypothesis**
  - If rs123 mutation causes disease, the statistical test is significant

- **Observation**
  - Statistical test is significant

- **Conclusion by abduction**
  - rs123 mutation causes disease
  - provided there is no other explanation for the test to be significant

<table>
<thead>
<tr>
<th>SNP</th>
<th>Genotypes</th>
<th>Controls [n(%)]</th>
<th>Cases [n(%)]</th>
<th>$\chi^2$</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs123</td>
<td>AA</td>
<td>1 0.9%</td>
<td>0 0.0%</td>
<td>4.78E-21</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>AG</td>
<td>38 35.2%</td>
<td>79 97.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>69 63.9%</td>
<td>2 2.5%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: SNP, single nucleotide polymorphism.
Exercise #3

• Hypothesis
  – If rs123 mutation causes disease, the statistical test is significant

• Observation
  – Statistical test is significant

• Conclusion by abduction
  – rs123 mutation causes disease
  – provided there is no other explanation for the test to be significant

• How to incorporate “provided there is no other explanation” into the analysis?
Example 2
A seemingly obvious conclusion

Looks like treatment A is better

What is happening here?

Looks like treatment B is better
Careless null hypothesis

- “Effective” H0
  - Treatments are identically distributed in the two samples

- Assumption
  - All other factors are equalized in the two samples

- Apparent H0
  - Treatments are identically distributed in the two populations

- Apparent H1
  - Treatments are differently distributed in the two populations
Exercise #4

- “Effective” H0
  - Treatments are identically distributed in the two samples

- Assumption
  - All other factors are equalized in the two samples

- Apparent H0
  - Treatments are identically distributed in the two populations

- Apparent H1
  - Treatments are differently distributed in the two populations

- The apparent null / alternative hypothesis is carelessly stated. Why? How to fix this?
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 distribution and p-value

• In statistical hypothesis testing, the null distribution is the probability distribution of the test statistic when the null hypothesis is true.

• The p-value is the probability of a null sample producing a test statistic value as good as or better than the observed test statistic value.
A seemingly obvious conclusion

- A multi-gene signature is claimed as a good biomarker for breast cancer survival
  - Cox’s survival model p-value << 0.05

- A straightforward Cox’s proportional hazard analysis. Anything more/wrong?
Almost all random signatures also have p-value < 0.05

- Theoretical null distribution used in Cox’s proportion hazard analysis does not match the empirical null distribution

- What can we do about this?

Venet et al., PLOS Comput Biol, 2011
Careless null hypothesis

- **“Effective” H0**
  - The biomarker’s values are identically distributed in the two populations

- **Assumption**
  - The null distribution models real world

- **Apparent H0**
  - The biomarker’s values are identically distributed in the two populations

- **Apparent H1**
  - The biomarker’s values are differently distributed in the two populations
Exercise #5

• “Effective” H₀
  – The biomarker’s values are identically distributed in the two populations

• Assumption
  – The null distribution models real world

• Apparent H₀
  – The biomarker’s values are identically distributed in the two populations

• Apparent H₁
  – The biomarker’s values are differently distributed in the two populations

• The apparent null / alternative hypothesis is carelessly stated. Why? How to fix this?
Example 4
Gene-selection methods have poor reproducibility

- Low % of overlapping genes from diff expt in general
  - Prostate cancer
    - Lapointe et al, 2004
    - Singh et al, 2002
  - Lung cancer
    - Garber et al, 2001
    - Bhattacharjee et al, 2001
  - DMD
    - Haslett et al, 2002
    - Pescatori et al, 2007

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEG</th>
<th>POG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Lung Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

Contextualizing based on pathways may help

- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype

- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype
Overlap analysis: ORA

ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (e.g., genes whose t-statistic meets the 5% significance threshold of t-test), and checking the significance of the size of the intersection using the hypergeometric test.

Disappointing performance

upregulated in DMD

DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010
Exercise #6

• What does ORA perform so poorly?
ORA-Paired

- Let $g_i$ be genes in a given pathway $P$
- Let $p_j$ be a patient
- Let $q_k$ be a normal

- Let $\Delta_{i,j,k} = \text{Expr}(g_i, p_j) - \text{Expr}(g_i, q_k)$

- $H_0$: Pathway $P$ is irrelevant to the diff betw patients and normals, so genes in $P$ behave similarly in patients and normals

$\Rightarrow$ t-test whether $\Delta_{i,j,k}$ is a distribution with mean 0

Exercise #7

Which null distribution is appropriate? Why?

ORA-Paired

- Let $g_i$ be genes in a given pathway $P$
- Let $p_j$ be a patient
- Let $q_k$ be a normal

- Let $\Delta_{i,j,k} = \text{Expr}(g_i,p_j) - \text{Expr}(g_i,q_k)$

$\Rightarrow$ t-test whether $\Delta_{i,j,k}$ is a distribution with mean 0

- $H_0$: Pathway $P$ is irrelevant to the diff betw patients and normals, so genes in $P$ behave similarly in patients and normals

- t-distribution with $n \times m$ degrees of freedom
- t-distribution with $n + m$ degrees of freedom
- Generate null distribution by gene-label permutation
- Generate null distribution by class-label permutation
Example 5
Synthetic lethal pairs

**Fact**
- When a pair of genes are synthetic lethal, mutations that affect function 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

**Why interested in synthetic lethality**
- Synthetic-lethal partners of frequently mutated genes in cancer are likely good treatment targets
Exercise #8

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

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

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

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

- Anything wrong with this?
SOMETIMES CHANGING PERSPECTIVE HELPS
Almost all random signatures also have p-value < 0.05

- Instead of asking whether a signature is significant, ask what makes a signature (random or otherwise) significant

Venet et al., PLOS Comput Biol, 2011
• **Proliferation is a hallmark of cancer**

• **Hypothesis:** proliferation-associated genes make a signature significant

Wilson Goh, private communication, 2017

<table>
<thead>
<tr>
<th>Cutoffs</th>
<th>Counts</th>
<th>OddsRatio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above D.05</td>
<td>7043</td>
<td>1.44x</td>
</tr>
<tr>
<td>Below D.05</td>
<td>2766</td>
<td>0.56x</td>
</tr>
<tr>
<td>Marginals</td>
<td>9809</td>
<td></td>
</tr>
</tbody>
</table>
SUMMARY
Anna Karenina Principle

• Careless null / alternative hypothesis due to forgotten assumptions
  – Distributions of the feature of interest in the two samples are identical to the two populations
  – Features not of interest are equalized / controlled for in the two samples
  – No other explanation for significance of the test
  – Null distribution models the real world

• These make it easy to reject the carelessly stated null hypothesis and accept an incorrect alternative hypothesis
Avoiding wrong conclusion, Getting deeper insight

• **Check for sampling bias**
  – Are the distributions of the feature of interest in the two samples same as that in the two populations?

• **Check for exceptions**
  – Are there large subpopulations for which the test outcome is opposite?
  – Are there large subpopulations for which the test outcome becomes much more significant?

• **Check for validity of the null distribution**
  – Can you derive it from the null hypothesis?