Gene expression analysis: Some lessons for statistical hypothesis testing

WONG Limsoon
Plan, 29/1/2024

Statistical hypothesis testing

Onus of proof

Anna Karenina Principle

Null hypothesis & null distribution

Getting them right

Gene expression profiling

Differentially expressed genes (DEG) selection

Poor DEG selection replicability

Addressing the replicability crisis
Statistical hypothesis testing

Formulate null hypothesis $H_0$ 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, reject $H_0$; otherwise, reject $H_1$

Null distribution is the distribution of $t(S_0)$ where $S_0$ ranges over the set of null samples for which $H_0$ holds

This does not mean we accept $H_0$!
Onus of proof: Rejecting H1 ≠ accepting H0

“... a p-value is large doesn’t mean that the null hypothesis is true. All a hypothesis test does is measure the strength of evidence against the null hypothesis. That is, we assume the null hypothesis is true until we have enough evidence to reject. Crucially, we never actually claim that the null hypothesis is true - it is just an assumption!"

A pharmaceutical research team constructs a significant test:

$H_0$ – *Side effects of new drug X are same as standard drug Y*

$H_1$ – *Side effects of new drug X are different from standard drug Y*

Would you be happy to use the new drug based on a large statistically insignificant p-value?

Taken from https://vsni.co.uk/blogs/never-accept-the-null-hypothesis
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 hypothesis incorrectly stated*

*Inappropriate expt design*

And so on
SNP rs123 is good biomarker for a disease:

*If rs123 is AA or GG, low risk for disease*

*If rs123 is AG, high risk for disease*

Straightforward $\chi^2$ test. Anything wrong?
Calculations

AG = 38 + 79 = 117,
Controls + cases = 189
⇒ Population is 117 / 189 = 62% AG

P(AA) = ...
Sample may be biased

Intentionally left blank
Careless null hypothesis

**H0**
rs123 alleles are identically distributed in the two samples

**Assumption**
Distributions of rs123 alleles in the two samples are resp. identical to the two populations

---

**Apparent H1**
rs123 alleles are differently distributed in the two populations

**“Actual” H1**
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
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?
**Meiosis**

In meiosis, the *chromosomes* duplicate (during *interphase*) and *homologous chromosomes* exchange genetic information (*chromosomal crossover*) during the first division, called *meiosis I*. The daughter cells divide again in *meiosis II*, splitting up *sister chromatids* to form haploid *gametes*. Two gametes fuse during *fertilization*, forming a diploid cell (*zygote*) with a complete set of paired chromosomes.

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

<table>
<thead>
<tr>
<th>Gametes made:</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
</tr>
<tr>
<td>48%</td>
</tr>
</tbody>
</table>

Image credit: Khan Academy
Exercise

What is happening here?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lived</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>died</td>
<td>100</td>
<td>165</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lived</td>
<td>40</td>
<td>15</td>
</tr>
<tr>
<td>died</td>
<td>20</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lived</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>died</td>
<td>80</td>
<td>160</td>
</tr>
</tbody>
</table>

Treatment A is better

Treatment B is better
A/B sample not equalized in other attributes, ...
**Careless null hypothesis**

**H0**
Treatment effects are identically distributed in the two samples

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

**Apparent H1**
Treatment effects are differently distributed in the two populations

**“Actual” H1**
Treatment effects are differently distributed in the two populations OR
Some other factors aren’t equalized in the two samples
Exercise

I asked ChatGPT how to test whether my income is atypically low.

What do you think of ChatGPT’s response?

ChatGPT

To test whether your income is atypically low given the mean income ($X$) and standard deviation ($Y$), you can use z-scores to determine how many standard deviations your income is away from the mean. This helps identify whether your income is significantly lower or higher than the average.

The z-score formula is:

$$z = \frac{\text{Your Income} - \text{Mean Income}}{\text{Standard Deviation}} = \frac{\text{Your Income} - X}{Y}$$

Once you calculate the z-score for your income using this formula, you can interpret its value:

- If the z-score is significantly less than -2 (approximately), your income might be considered atypically low.
- A z-score less than -2 indicates that your income is more than two standard deviations below the mean income, which is relatively rare in a normal distribution.
A nation of unusually poor people, really?

Intentionally left blank
Homework

Read this paper

**Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer**

Sriganesh Srihari, Jitin Singla, Limsoon Wong & Mark A. Ragan

*Biology Direct* 10, Article number: 57 (2015)

Write a 1-page review report focusing how it tests for synthetic-lethal gene pairs. Discuss whether their test is a good one.
Plan, 29/1/2024

Statistical hypothesis testing

Onus of proof

Anna Karenina Principle

Null hypothesis & null distribution

Getting them right

Gene expression profiling

Differentially expressed genes (DEG) selection

Poor DEG selection replicability

Addressing the replicability crisis
Microarray

Contain large numbers of DNA molecules spotted on glass slides, nylon membranes, or silicon wafers

Detect what genes are being expressed or found in a cell of a tissue sample

Measure expression of thousands of genes simultaneously
Application: Disease subtype diagnosis

- **Genes**
  - **Samples**
    - **Benign**
    - **Malign**

[Diagram showing genes and samples with benign and malign labels]
Application: Drug-action inference

Which group of genes does the drug affect? Why?
Diagnosis using microarrays & machine learning

Gene expression data collection

Gene selection using e.g., $\chi^2$

Classifier training

Classifier tuning (optional for some machine learning methods)

Apply classifier for diagnosis of future cases
Gene selection basic idea

Class 1 | Class 2
Class 1 | Class 2
Class 1 | Class 2
Gene selection by $\chi^2$

The $\chi^2$ value of a signal is defined as:

$$
\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},
$$

where $m$ is the number of intervals, $k$ the number of classes, $A_{ij}$ the number of samples in the $i$th interval, $j$th class, $R_i$ the number of samples in the $i$th interval, $C_j$ the number of samples in the $j$th class, $N$ the total number of samples, and $E_{ij}$ the expected frequency of $A_{ij}$ ($E_{ij} = R_i * C_j / N$).
### Performance of various classifiers

<table>
<thead>
<tr>
<th>Testing Data</th>
<th>Error rate of different models</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C4.5</td>
<td>SVM</td>
<td>NB</td>
<td>PCL</td>
</tr>
<tr>
<td>T-ALL vs OTHERS1</td>
<td>0:1</td>
<td>0:0</td>
<td>0:0</td>
<td>0:0</td>
</tr>
<tr>
<td>E2A-PBX1 vs OTHERS2</td>
<td>0:0</td>
<td>0:0</td>
<td>0:0</td>
<td>0:0</td>
</tr>
<tr>
<td>TEL-AML1 vs OTHERS3</td>
<td>1:1</td>
<td>0:1</td>
<td>0:1</td>
<td>1:0</td>
</tr>
<tr>
<td>BCR-ABL vs OTHERS4</td>
<td>2:0</td>
<td>3:0</td>
<td>1:4</td>
<td>2:0</td>
</tr>
<tr>
<td>MLL vs OTHERS5</td>
<td>0:1</td>
<td>0:0</td>
<td>0:0</td>
<td>0:0</td>
</tr>
<tr>
<td>Hyperdiploid&gt;50 vs OTHERS</td>
<td>2:6</td>
<td>0:2</td>
<td>0:2</td>
<td>0:1</td>
</tr>
</tbody>
</table>

Total Errors                | 14    | 6    | 8    | 4     

Classifiers based 20 genes selected by $\chi^2$ at each level of the tree.
Multidimensional scaling plot for subtype diagnosis

Obtained by performing PCA on the 20 genes chosen for each level
Plan, 29/1/2024

Statistical hypothesis testing

Onus of proof

Anna Karenina Principle

Null hypothesis & null distribution

Getting them right

Gene expression profiling

Differentially expressed genes (DEG) selection

Poor DEG selection replicability

Addressing the replicability crisis
Poor replicability of gene selection

Low % of overlapping genes from diff expt

Prostate cancer
• Lapointe et al, 2004 vs Singh et al, 2002

Lung cancer
• Garber et al, 2001 vs Bhattacharjee et al, 2001

DMD
• Haslett et al, 2002 vs 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>
Suppose:

*Each gene has 50% chance to be high*

*You have 3 disease and 3 normal samples*

How many genes on a microarray are expected to perfectly correlate to these samples?

\[
\text{Prob(gene is correlated)} = \frac{1}{2^6}
\]

\[
\text{# of genes on array} = 25,000
\]

\[
E(\text{# of correlated genes}) = 390
\]

⇒ Many false positives; these cannot be eliminated based on pure statistics!
Group of genes

Suppose:

*Each gene has 50% chance to be high*

*You have 3 disease & 3 normal samples*

What’s the chance for a group of 5 genes to perfectly correlate to these samples?

When only 1 group is considered, $<< 1/2^6$

$\# \text{ of groups } = \binom{25000}{5}$

$E(\# \text{ of correlated groups}) = \binom{25000}{5} \times (1/2^6)^5$

$= 7.58 \times 10^{10}$

$\Rightarrow$ Even more false positives?

Perhaps no need to consider every group
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 reduced using biological processes.

The unifying biological theme is basis for inferring underlying cause of disease subtype.
Group of Genes

• Suppose
  – Each gene has 50% chance to be high
  – You have 3 disease and 3 normal samples
• What is the chance of a group of 5 genes being perfectly correlated to these samples?

- Prob(group of genes correlated) = \((1/2^6)^5\)
  - Good, \(< < 1/2^6\)
• # of groups = \(\binom{25000}{5}\)
  \(\Rightarrow\) \(E(#\ of\ groups\ of\ genes\ correlated) = \binom{25000}{5} \times (1/2^6)^5 = 7.58 \times 10^{10}\)

- Even more false positives?
• Perhaps no need to consider every group
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.

The bewilderment persists... a crisis?

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

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

Why does ORA perform so poorly in selecting differentially expressed genes?
A biological pathway is a chain of actions of molecules in cell leading to a change in cell

⇒ Behaviour of genes in a pathway is more coordinated than random ones
Issue #2 with ORA

Intentionally left blank
Issue #3 with ORA

Intentionally left blank
<table>
<thead>
<tr>
<th>Plan, 5/2/2024</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistical hypothesis testing</strong></td>
</tr>
<tr>
<td><strong>Onus of proof</strong></td>
</tr>
<tr>
<td><strong>Anna Karenina Principle</strong></td>
</tr>
<tr>
<td><strong>Null hypothesis &amp; null distribution</strong></td>
</tr>
<tr>
<td><strong>Getting them right</strong></td>
</tr>
<tr>
<td><strong>Gene expression profiling</strong></td>
</tr>
<tr>
<td><strong>Differentially expressed genes (DEG) selection</strong></td>
</tr>
<tr>
<td><strong>Poor DEG selection replicability</strong></td>
</tr>
<tr>
<td><strong>Addressing the replicability crisis</strong></td>
</tr>
</tbody>
</table>
Exercise

How to solve the issues identified in ORA?
 ISSUE #1 is solved
Null hypothesis is “Pathway P is irrelevant to the difference between patients and normals, and the genes in P behave similarly in patients and normals”

ISSUE #2 is solved
No need pre-determined list of DE genes

ISSUE #3 is unsolved

Assume absence of batch effects
Let \( g_i \) be a gene in a pathway \( P \)
Let \( p_j \) be a patient
Let \( q_k \) be a normal

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

Test whether \( \Delta_{i,j,k} \) is a distribution with mean 0

How many \( \Delta_{i,j,k} \) are there?
\(|\text{patients}| \times |\text{normals}| \times |\text{genes in } P|\)

Does this mean sample size now larger?
Does this mean more degrees of freedom?
Testing the null hypothesis

“Pathway P is irrelevant to the diff betw patients and normals and so, the genes in P behave similarly in patients and normals”

Method #1

* T-test w/ the right degrees of freedom? 
  # normals + # patients – 1

Method #2

* By the null hypothesis, a dataset & its class-label permutations are exchangeable 
  Get null distribution by class-label permutations 
  Only for large-size sample
Better, but not super-duper good
NEA-Paired: Paired test on subnetworks

Given a pathway P
Let each node and its immediate neighbourhood in P be a subnetwork
Apply ORA-Paired on each subnetwork individually

Issues #1 & #2 are solved as per ORA-Paired
Issue #3 is partly solved

Testing subnetworks instead of whole pathways
Much better performance
Take-home messages

Statistical hypothesis testing needs careful thought

*Right null hypothesis*

*Right null distribution*

*Many nuances wrt confounding factors*

Make effort to understand the domain

*A little domain insight goes a really long way*
References


Lim et al., “A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small”, \textit{Journal of Bioinformatics and Computational Biology}, 13(4):1550018, 2015
Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer

Sriganesh Srihari, Jitin Singla, Limsoon Wong & Mark A. Ragan

Biology Direct 10, Article number: 57 (2015) | Cite this article