P-values and statistical tests

1. Introduction

Marek Gierliński
Division of Computational Biology

Hand-outs available at http://is.gd/statlec
We collaborate on various types of projects

Anything involving data analysis

Marek Gierliński  James Abbott

http://www.compbio.dundee.ac.uk/dag.html
Biology and statistics wishful thinking

Experiment

Statistics

\[ E[pFDR_\gamma] - pFDR(\gamma) \geq E \left[ \frac{\{W(\lambda)/(1 - \lambda)\}_{\gamma} - V(\gamma)}{\{R(\gamma) \lor 1\} \Pr\{R(\gamma) > 0\}} \right]. \]

* 0 \geq 1 - (1 - \gamma)^n under independence. Conditioning on \( R(\gamma) \), it follows that

\[ \frac{W(\lambda)/(1 - \lambda)_{\gamma} - V(\gamma)}{R(\gamma) \lor 1 \Pr\{R(\gamma) > 0\}} \Pr\{R(\gamma) > 0\} \]

Where, \( E\{W(\lambda)|R(\gamma)\} \) is a linear non-increasing function of \( R(\gamma) \), and \( E\{V(\gamma)|R(\gamma)\} \) function of \( R(\gamma) \). Thus, by Jensen’s inequality on \( R(\gamma) \) it follows that

\[ \frac{W(\lambda)/(1 - \lambda)_{\gamma} - V(\gamma)}{R(\gamma) \Pr\{R(\gamma) > 0\}} \Pr\{R(\gamma) > 0\} \geq \frac{E\{W(\lambda)/(1 - \lambda)_{\gamma} - V(\gamma)|R(\gamma) > 0\}}{E\{R(\gamma)|R(\gamma) > 0\} \Pr\{R(\gamma) > 0\}} \]

\[ = E\{R(\gamma)|R(\gamma) > 0\} \Pr\{R(\gamma) > 0\}, \]

it follows that

\[ \frac{W(\lambda)/(1 - \lambda)_{\gamma} - V(\gamma)|R(\gamma) > 0}{\{R(\gamma)|R(\gamma) > 0\} \Pr\{R(\gamma) > 0\}} = \frac{E\{W(\lambda)/(1 - \lambda)_{\gamma} - V(\gamma)|R(\gamma) > 0\}}{E\{R(\gamma)|R(\gamma) > 0\}}. \]

\[ p < 0.05 \]

\[ \text{😊} \]
P-Values: Misunderstood and Misused

Bertie Vidgen and Taha Yasseri *

The fickle P value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

Why Most Published Research Findings Are False

John P.A. Ioannidis
1. Introduction
Null hypothesis, statistical test, p-value
Fisher’s test

2. Contingency tables
Chi-square test
G-test

3. T-test
One- and two-sample
Paired
One-sample variance test

4. ANOVA
One-way
Two-way

5. Non-parametric methods 1
Mann-Whitney
Wilcoxon signed-rank
Kruskal-Wallis

6. Non-parametric methods 2
Kolmogorov-Smirnov
Permutation
Bootstrap

7. Statistical power
Effect size
Power in t-test
Power in ANOVA

8. Multiple test corrections
Family-wise error rate
False discovery rate
Holm-Bonferroni limit
Benjamini-Hochberg limit
Storey method

9. What’s wrong with p-values?
A lot
Null hypothesis
Null hypothesis

Default position
$H_0$: there is no effect

Evidence against $H_0$

Strong evidence?

Position unchanged

Reject $H_0$

Default position
Defendant is innocent

Evidence against

Strong evidence?

Innocent

Guilty
Evidence against $H_0$

- Two samples of mice
  - 12 English mice
  - 9 Scottish mice

- Body mass difference:
  \[ \Delta M = M_S - M_E = 5.0 \text{ g} \]

- Two possibilities
  - real difference
  - fluke

- What are the chances of the fluke?
**Gedankenexperiment** under the null hypothesis

One population of British mice
\[ \mu = 20 \text{ g}, \sigma = 5 \]

Select two samples size 12 and 9

\[ \Delta M = M_E - M_S \]

Build distribution of \( \Delta M \)
**Gedankenexperiment**: result under null hypothesis

Note: in real life we use a statistic with known distribution
Gedankenexperiment: p-value

P-value: probability of getting the observed, or more extreme result, by chance

$n = 10^6$

$\text{observation}$

$n = 12,453$

$p = 0.012$
Null hypothesis and p-value

If both samples were taken from the same population,
then the probability of observing the difference in mean body mass of 5 g, or more, by chance (due to random sampling) would be 1.2%

We observe an effect, but it will occur by chance in 1.2% of repeated experiments (1 in 80)
You have 1.2% chance of making a fool of yourself (if you publish this result)
P-value is the probability of making a fool of yourself
Why “more extreme”? 

Zero probability 

\[ P(X = 5) = 0 \]
Null hypothesis: reject or what?

- Default position: $H_0$
- All other assumptions must be true!

- Reductio ad absurdum

  - Strong evidence?
    - no
      - Nothing
      - absence of evidence is not evidence of absence!
      - evidence too weak?
    - yes
      - Reject $H_0$
      - data are incompatible with $H_0$...
      - ...or any of the other assumptions
      - reject $H_0$ at your own risk

- Absence of evidence is not evidence of absence!
You cannot confirm the null hypothesis

Differential gene expression between WT and a mutant

Genes that are “not different” from 2 replicates...

...are “significantly different” when using 16 replicates

\[ p \geq \alpha \]

Schuch et al. 2016

No effect

Insufficient evidence
You cannot prove the null hypothesis
Statistical testing

Statistical model

Null hypothesis
$H_0$: no effect

All other assumptions

Significance level
$\alpha = 0.05$

$p$-value: probability that the observed effect is random

$p < \alpha$
Reject $H_0$
(at your own risk)
Effect is real

$p \geq \alpha$
Insufficient evidence

Statistical test against $H_0$
Fisher’s exact test
Ronald Fisher

Sir Ronald Aylmer Fisher
(1890-1962)

Rothamsted Experimental Station
(Hertfordshire)
The appreciation of tea

Milk first

Tea first
Null hypothesis:
Ms Bristol has no clue
Let’s draw some balls

Draw $n$ balls without replacement

removing balls changes probability!

What is the probability of finding exactly $k$ white balls?

Urn with $N$ balls
$m$ of them white
Binomial coefficient

- “n chose k”

\[
\binom{n}{k} = \frac{n!}{k! (n - k)!}
\]

- In *combinatorics* it is the number of possible \(k\)-element subsets of an \(n\)-element set

- From a 5-element set there are 10 possible 3-element subsets

\[
\binom{5}{3} = \frac{5!}{3! 2!} = \frac{120}{6 \times 2} = 10
\]
Count all the possibilities

\[
\binom{5}{3} = 10
\]

\[
\begin{bmatrix}
1 & 2 & 3 & 4 & 5
\end{bmatrix}
\rightarrow 
\begin{bmatrix}
\bigcirc & \bigcirc & \bigblackcirc
\end{bmatrix}
\]

\[
\binom{3}{2} = 3 \quad \binom{2}{1} = 2
\]

\[
\frac{\binom{2}{1} \times \binom{3}{2}}{\binom{5}{3}} = \frac{6}{10} = 0.6
\]

Draw 3 balls. What is the probability of finding exactly 2 whites among them?
Hypergeometric probability

- \( N = 36 \) balls
- \( m = 20 \) are white
- \( n = 10 \) balls drawn

What is the probability of finding exactly \( k = 8 \) white balls in the draw?

\[
P(X = 8) = \frac{{\binom{20}{8} \binom{16}{2}}}{{\binom{36}{10}}}
\]

\[
= \frac{125,970 \times 120}{254,186,856} = \frac{15,116,400}{254,186,856} \approx 0.059
\]
### Hypergeometric probability

- **$N$ balls**
- **$m$ are white**
- **$n$ drawn**

What is the probability of finding exactly $k$ white balls in the draw?

$$P(X = k) = \frac{(m)\binom{N-m}{n-k}}{\binom{N}{n}}$$

<table>
<thead>
<tr>
<th></th>
<th>Drawn</th>
<th>Not drawn</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>White</strong></td>
<td>$k$</td>
<td>$m-k$</td>
<td>$m$</td>
</tr>
<tr>
<td><strong>Black</strong></td>
<td>$n-k$</td>
<td>$N+k-n-m$</td>
<td>$N-m$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$n$</td>
<td>$N-n$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Contingency table
Hypergeometric distribution

- If sums are fixed (blue fields), the cells in the table follow hypergeometric distribution

\[
P \left[ \begin{array}{cc} 0 & 20 \\ 10 & 6 \end{array} \right] = 3.2 \times 10^{-5}
\]

\[
P \left[ \begin{array}{cc} 1 & 19 \\ 9 & 7 \end{array} \right] = 0.00090
\]

\[
P \left[ \begin{array}{cc} 2 & 18 \\ 8 & 8 \end{array} \right] = 0.0096
\]

... 

\[
P \left[ \begin{array}{cc} 8 & 12 \\ 2 & 14 \end{array} \right] = 0.059
\]

\[
P \left[ \begin{array}{cc} 9 & 11 \\ 1 & 15 \end{array} \right] = 0.011
\]

\[
P \left[ \begin{array}{cc} 10 & 10 \\ 0 & 16 \end{array} \right] = 0.00073
\]
Hypergeometric distribution

- If sums are fixed (blue fields), the cells in the table follow hypergeometric distribution

\[
P\left[ \begin{array}{c} 0 \\ 10 \\ 6 \end{array} \right] = 3.2 \times 10^{-5}
\]

\[
P\left[ \begin{array}{c} 1 \\ 9 \\ 7 \end{array} \right] = 0.00090
\]

\[
P\left[ \begin{array}{c} 2 \\ 8 \\ 8 \end{array} \right] = 0.0096
\]

...  

\[
P\left[ \begin{array}{c} 8 \\ 2 \\ 14 \end{array} \right] = 0.059
\]

\[
P\left[ \begin{array}{c} 9 \\ 1 \\ 15 \end{array} \right] = 0.011
\]

\[
P\left[ \begin{array}{c} 10 \\ 0 \\ 16 \end{array} \right] = 0.00073
\]
One-sided test

- What is the probability of drawing **8 or more** white balls?

\[ P(X \geq 8) = 0.059 + 0.011 + 0.00073 = 0.071 \]

- **Enrichment**: do we have more than random? (right-sided test)

- **Depletion**: do we have fewer than random? (left-sided test)
Two-sided test

- One-sided test: do we observed too many white balls?
- Two-sided test: do we observe too many or too few white balls?
- Is my result extreme in any way?

- Add all probabilities less or equal \( P(X = 8) \) on both sides

\[
P(X \leq 3 \cup X \geq 8) = 0.13
\]
Tea tasting by Muriel Bristol

Milk first

Tea first
Tea tasting test

- Null hypothesis: Ms Bristol has no ability to tell the difference
- One-sided probability of getting this or more extreme result by chance is
  \[ P(X \geq 3) = 0.229 + 0.014 \approx 0.24 \]
- The null hypothesis cannot be rejected
- Insufficient data!
Contingency table

- Two variables (in columns and rows)
- E.g. treatments vs outcomes
- Contingency = association

<table>
<thead>
<tr>
<th></th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Failure</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

2x2 contingency table
Test of independence

- Two variables (in columns and rows)
- E.g. treatments vs outcomes

- $H_0$: variables are independent

- Ms. Bristol’s answers do not depend on whether she got milk or tea first; they are random

<table>
<thead>
<tr>
<th>Tea served</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>T</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ms. Bristol</td>
<td>T</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>M</td>
<td>T</td>
</tr>
</tbody>
</table>

$p = 0.58$

<table>
<thead>
<tr>
<th>Tea served</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>T</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ms. Bristol</td>
<td>T</td>
<td>T</td>
<td>M</td>
<td>M</td>
<td>T</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>T</td>
<td>T</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

$p = 0.03$
Test of proportion

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>T</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tea served</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ms. Bristol</td>
<td>T</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>M</td>
<td>T</td>
</tr>
</tbody>
</table>

\[
\begin{array}{cc}
4 & 5 \\
2 & 1 \\
\end{array}
\]

4:5 
2:1 
p = 0.58

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>M</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>M</th>
<th>M</th>
<th>T</th>
<th>T</th>
<th>M</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tea served</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ms. Bristol</td>
<td>T</td>
<td>T</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>T</td>
<td>T</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>T</td>
<td>T</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

\[
\begin{array}{cc}
5 & 2 \\
0 & 5 \\
\end{array}
\]

5:2 
0:5 
p = 0.03
Proteomics example

- There are 668 proteins in an experiment
- 7 of them have an associated Gene Ontology term (GO:00301174, regulation of DNA replication initiation)
- We have a cluster of 44 proteins with similar properties
- 6 of them have this GO term
- Is it significantly enriched?

\[ P(X \geq 6) \approx 4 \times 10^{-7} \]
Absolute numbers are important

- A newspaper reports clinical tests on a new cancer drug
- 15% of patients treated with drug A survived
- 30% of patients treated with drug B survived
- So, drug B is 100% better than drug A!
Absolute numbers are important

- A newspaper reports clinical tests on a new cancer drug
- 15% of patients treated with drug A survived
- 30% of patients treated with drug B survived
- So, drug B is 100% better than drug A!
- Actual numbers: 20 and 10 patients
- $p = 0.37$ (two-sided test)

<table>
<thead>
<tr>
<th></th>
<th>Drug A</th>
<th>Drug B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Dead</td>
<td>17</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>10</td>
<td>30</td>
</tr>
</tbody>
</table>

$p = 0.37$
Absolute numbers are important

- A newspaper reports clinical tests on a new cancer drug
- 15% of patients treated with drug A survived
- 30% of patients treated with drug B survived
- So, drug B is 100% better than drug A!
- Actual numbers: 20 and 10 patients
- $p = 0.37$

<table>
<thead>
<tr>
<th></th>
<th>Drug A</th>
<th>Drug B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Dead</td>
<td>17</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>10</td>
<td>30</td>
</tr>
</tbody>
</table>

- If we had 80 and 100 patients and the same proportions
- $p = 0.02$

<table>
<thead>
<tr>
<th></th>
<th>Drug A</th>
<th>Drug B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>12</td>
<td>30</td>
<td>42</td>
</tr>
<tr>
<td>Dead</td>
<td>68</td>
<td>70</td>
<td>138</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>100</td>
<td>180</td>
</tr>
</tbody>
</table>

Moral 1: don’t trust newspapers
Moral 2: estimate the required size of your sample before you do your experiment
Never, ever use percentages in Fisher’s test!

<table>
<thead>
<tr>
<th></th>
<th>Alive</th>
<th>Dead</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug A</td>
<td>15%</td>
<td>85%</td>
<td></td>
</tr>
<tr>
<td>Drug B</td>
<td>30%</td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Fisher’s exact test: summary

<table>
<thead>
<tr>
<th><strong>Input</strong></th>
<th>2×2 contingency table (larger tables possible) typically columns = treatments, rows = outcomes table contains counts counts of subjects falling into categories</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Usage</strong></td>
<td>Examine if there is an association (contingency) between two variables; whether the proportions in one variable depend on the proportions in the other variable; if there is enrichment</td>
</tr>
<tr>
<td><strong>Null hypothesis</strong></td>
<td>The proportions in one variable do not depend on the proportions in the other variable</td>
</tr>
<tr>
<td><strong>Comments</strong></td>
<td>Exact test – count all possible combinations Use when you have small numbers For large numbers (hundreds) use chi-square test Carefully chose between one- and two-sided test</td>
</tr>
</tbody>
</table>
How to do it in R?

# Tea tasting
> fisher.test(rbind(c(3, 1), c(1, 3)), alternative="greater")

Fisher's Exact Test for Count Data

data:  rbind(c(3, 1), c(1, 3))
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
  0.3135693     Inf
sample estimates:
odds ratio
   6.408309

# GO enrichment
> fisher.test(rbind(c(6, 1), c(38, 623)), alternative="greater")

Fisher's Exact Test for Count Data

data:  rbind(c(6, 1), c(38, 623))
p-value = 3.894e-07
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
   14.29724     Inf
sample estimates:
odds ratio
   96.29591
Hand-outs available at http://is.gd/statlec
Two approaches

Fisher

$H_0: \mu_E = \mu_S$

Neyman-Pearson

$H_0: \mu_E = \mu_S$
$H_1: \mu_E < \mu_S$
$\alpha = 0.05$

Critical value

observation

observation

critical region
$\alpha = 0.05$