Statistical testing in the era of big data ($p < 0.05$)

Dimitri Van De Ville
MIP:lab
IBI-STI/CNP (EPFL)
RADIO (UniGE)
http://miplab.epfl.ch/
@dvdevill
The Statistical Crisis in Science

Voodoo Correlations: Have the Results of Some Brain Scanning Experiments Been Overstated?

Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.
SHORT REPORT

The (in)famous GWAS $P$-value threshold revisited and updated for low-frequency variants

João Fadista$^{1,2}$, Alisa K Manning$^{3,4}$, Jose C Florez$^{3,4,5,6}$ and Leif Groop$^{1,7}$
Paradox?

Mount Big Data

Is big data destroying p-values?
Roadmap of the workshop

- Contradictory tendencies
  - Many (emotive) reports about p-value crisis
  - Reviewers even more picky on statistical significance
    - Sufficient power, multiple comparisons, replication,…
  - Adage: Never enough data
  - Big data has arrived, and will become bigger
- Is classical hypothesis testing doomed?
  Should we all go into Bayesian statistics?
  Machine-learning approaches will be the only solution?
- Here, revisit the basic statistical hypothesis testing
  - to understand the core issue
  - to solve it within the conventional framework
One-sample t-test in a nutshell

- Consider $N$ samples modeled to reflect a true effect $\mu$ with a random Gaussian* deviation $\mathcal{N}(0, \sigma^2)$: $x_n = \mu + e_n$, $n = 1, \ldots, N$
- Estimator of $\mu$ is average $\hat{\mu}$
- Estimator of uncertainty on $\hat{\mu}$ is standard deviation $\hat{\sigma}$
- We define $t = \frac{\hat{\mu}}{\hat{\sigma}} \sqrt{N}$
- Question: is there evidence from the data that the underlying $\mu \neq 0$

* Popularity of Gaussian hypothesis? Central limit theorem!
One-sample t-test in a nutshell

- **Null hypothesis** $\mathcal{H}_0$: no effect, $\mu = 0$
- (Implicit) alternative hypothesis $\mathcal{H}_1: \mu \neq 0$
  - Under the null, $t$ follows a known distribution (Student t-distribution with $N - 1$ degrees of freedom)
- $p$-value is probability to mistakenly reject $\mathcal{H}_0$: $p = P(|t| > T | \mathcal{H}_0)$
- Result is considered significant if $p < 0.05$

“If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty or one in a hundred. Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fails to reach this level.”

One-sample t-test in a nutshell

- Thus, $p$-value indicates probability of *false positive* (FP)
- Typically, no explicit $H_1$:
  - No control on false negatives; i.e., $P( | t | \leq T | H_1$ not true)
  - One can only control specificity (1-FP rate), not sensitivity (1-FN rate)
  - No proof of no effect because no point of comparison
Fallacy of statistical testing

- Any true effect $\mu_0 \neq 0$ can become significant for $N$ sufficiently large

\[ \frac{\mu_0}{\sigma}\sqrt{N} > T \quad \quad N > T^2 \frac{\sigma^2}{\mu_0^2} \]

- “[N] must be big enough that an effect of such magnitude as to be of scientific significance will also be statistically significant. It is just as important, however, that the study not be too big, where an effect of little scientific importance is nevertheless statistically detectable”

- As $N$ increases, **discriminability**, as measured by classification accuracy, of individual samples becomes very small

- As $N$ increases, **consistency**, as measured by population prevalence, of the effect becomes very small

[Lenth, 2001]
Effect size

- Bottomline: $p$-values are relevant if effect size is non-trivial!
- Standardized effect size: Cohen’s $d = t/\sqrt{N}$; $R^2 = \mu^2/(\mu^2 + \sigma^2)$; $\rho = \sqrt{R^2}$

<table>
<thead>
<tr>
<th>Effect size</th>
<th>Cohen’s $d$</th>
<th>Coefficient of determination $R^2$</th>
<th>Correlation</th>
<th>Classification accuracy</th>
<th>Population prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>~1</td>
<td>~1/2=0.50</td>
<td>~0.71</td>
<td>~70%</td>
<td>~50%</td>
</tr>
<tr>
<td>Medium</td>
<td>~1/2=0.50</td>
<td>~1/5=0.20</td>
<td>~0.45</td>
<td>~60%</td>
<td>~20%</td>
</tr>
<tr>
<td>Small</td>
<td>~1/4=0.25</td>
<td>~1/17=0.06</td>
<td>~0.24</td>
<td>~55%</td>
<td>~6%</td>
</tr>
<tr>
<td>Trivial</td>
<td>~1/8=0.13</td>
<td>~1/65=0.02</td>
<td>~0.12</td>
<td>~52.5%</td>
<td>~1%</td>
</tr>
<tr>
<td>None</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50%</td>
<td>0%</td>
</tr>
</tbody>
</table>

“… one should be cautious that extremely large studies may be more likely to find a formally statistical significant difference for a trivial effect that is not really meaningfully different from the null.” (Ioannidis, 2005)

[Friston, NeuroImage, 2012]
Sample size and sensitivity

Consider now fixed specificity $\alpha = 0.05$, then we have

$$\alpha = \int_{u(\alpha)}^{\infty} T(t; N - 1) dt$$

Sample size and sensitivity

- Consider now fixed specificity $\alpha = 0.05$, then we have
  \[ \alpha = \int_{u(\alpha)}^{\infty} T(t; N - 1) dt \]

- Under the assumption of a true effect size $d$, we can compute sensitivity as
  \[ 1 - \beta(d) = \int_{u(\alpha)}^{\infty} T(t; N - 1, d\sqrt{N}) dt \]
  where $T(t; K, \delta)$ is the non-central t-distribution with $K$ degrees of freedom and non-centrality parameter $\delta$

- Sensitivity depends on sample size ($N$) and effect size ($d$)

[Friston, NeuroImage, 2012]
**Under-powered?**

- Sensitivity depends on sample size ($N$) and effect size ($d$)
  - Significant effect with small sample size is likely to be caused by large effect size!
  - If you are criticized in this way:

  "The fact that we have demonstrated a significant result in a relatively under-powered study suggests that the effect size is large. This means, quantitatively, our result is stronger than if we had used a larger sample-size."

  = conflation of significance and power

Over-powered?

- Sensitivity depends on sample size ($N$) and effect size ($d$)
  - Sensitivity to trivial effect sizes increases with sample size!
  - Ultimately, with very large sample sizes, sensitivity will reach 100% for every non-null effect size
  - Explains a lot about the crisis!
  - More is not better

Loss-function analysis

- Let us define a simple loss function $l$:
  - Cost $+1$ for detecting trivial effect size of $1/8$ [bad]
  - Cost $-1$ for detecting large effect size of $1$ [good]
- Expected loss:
  $$l = (1 - \beta(1/8)) - (1 - \beta(1))$$
  $$= \beta(1) - \beta(1/8)$$
- Optimal sample size at minimal loss
- Does not increase dramatically even if significance needs to be (much) stronger (e.g., due to multiple comparisons)
Protected inference

- Inference is based on controlling FP rate under $H_0$, which translates in a flat sensitivity at $\alpha$ for no effect:
  - specificity = sensitivity to null effects
- So let us suppress sensitivity to trivial effects instead!

$$1 - \beta(d) = \int_0^\infty T(t; N - 1, d\sqrt{N}) dt$$

where this time we use

$$\alpha(d') = \int_0^\infty T(t; N - 1, d'\sqrt{N}) dt$$

with $d' = 1/8$

Protected inference

- Protection fixes $\beta(1/8) = 0.05$ and thus increasing $N$ becomes harmless
- Concretely, threshold to be applied to $t$-values is penalized

A note on non-parametric testing

- Consider $N$ samples modeled to reflect a true effect $\mu$ with a random deviation of unknown, but symmetric distribution: $x_n = \mu + e_n$, $n = 1, \ldots, N$
- Estimator of $\mu$ is average $\hat{\mu}$ (could also be median etc)
- Null hypothesis $\mathcal{H}_0$: no effect, $\mu = 0$
  - In that case, we can randomly flip or permute the signs of $x_n$ and recompute our measure of interest under the null as $\hat{\mu}_k^{(0)}$, $k = 1, \ldots, K$
  - If $\hat{\mu} > \max \hat{\mu}_k^{(0)}$ or $\hat{\mu} < \min \hat{\mu}_k^{(0)}$, then $\mathcal{H}_0$ is rejected with $p = 2/(K + 1)$
- Use $K = 39$ randomizations to be able to assess 0.05 significance
- Less assumptions about distribution, but essentially same problem that trivial effects will be picked up as $N$ increases
Bias-variance trade-off of effect size

- Inferential statistics; e.g., presence of treatment effect
  - In-sample effect size is about data at hand
  - In-sample effect size overestimates true effect size because some large test statistics can also be obtained by chance
- Estimation; e.g., predicting treatment effect
  - Out-of-sample effect size is an unbiased estimate of true effect size
  - However, test is less efficient
- *Le beurre et l’argent du beurre*

Reproducible science

- More data allows you to do more things
- Terminology becomes important!

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>Same</td>
</tr>
<tr>
<td>Different</td>
<td>Different</td>
</tr>
<tr>
<td></td>
<td>Reproducible</td>
</tr>
<tr>
<td></td>
<td>Generalisable</td>
</tr>
</tbody>
</table>
And some resources

Statistics for Biologists

There is no disputing the importance of statistical analysis in biological research, but too often it is considered only after an experiment is completed, when it may be too late.

This collection highlights important statistical issues that biologists should be... show more

Since September 2013 Nature Methods has been publishing a monthly column on statistics called "Points of Significance." This column is intended to provide researchers in biology with a basic introduction to core statistical concepts and methods, including experimental design. Although targeted at biologists, the articles are useful guides for researchers in other disciplines as well. A continuously updated list of these articles is provided below.

Importance of being uncertain - How samples are used to estimate population statistics and what this means in terms of uncertainty.

Error Bars - The use of error bars to represent uncertainty and advice on how to interpret them.

https://www.nature.com/collections/qghhqm/pointsofsignificance
Good luck... and stay out of hell!

The nine circles of scientific hell

I Limbo
II Lust
III Gluttony
IV Greed
V Anger
VI Heresy
VII Violence
VIII Fraud
IX Treachery

[Neuroskeptic, Perspectives on Psychological Science, 2012]