P-hacking, HARKing, and Multiple Testing

Many scientific investigations will at times hinge on a small set of specific questions:

- subjects receiving a treatment have less pain than subjects receiving a placebo
- average yearly income in population A is larger than the average yearly income in population B
- Variables X and Y are correlated
- Variables X_1 and X_2 have different probability distributions.

Hypothesis Testing can help address these questions (NHST)

1. Turn question into a pair of hypotheses, null and alternative

   Ex null: treat. and placebo lead to same level of pain
   alt: treatment less pain than placebo.

   null: $\text{corr}(X,Y) = 0$ alt: $\text{corr}(X,Y) \neq 0$

2. Devise a test statistic S, a numeric quantity you can (i) measure in the data you have, (ii) know what its value will be if null is true.

   $\Rightarrow P_{null}(s)$ - distribution of S under the null (determine mathematically or computationally)

3. Ask prob. of seeing S { as large or larger than / as small or smaller than / at least as extreme as } S_{real} if null holds.

   $\Rightarrow$ This is the "p-value" of the test

   $\Rightarrow$ large p-value = data cannot rule out null hyp.
   small p-value = null hyp unlikely to hold ($\neq$ alt. hyp holding)

When is p small? When $p < \alpha$, α: <u>significance level</u>

$\Rightarrow \alpha = 0.05$ is an arbitrary cutoff.

Hypothesis testing is very common in science but it also has <u>many problems.</u>

Some criticisms
- null hypotheses are almost always false
- small p means data unlikely under null, but data could be just as unlikely under alternative
- statistical signif. $\neq$ scientific signif.
- only considers statistical errors, other errors not considered
- small p does not mean small prob. of being wrong

  $\Rightarrow$ Ioannidis: 74% of studies w/ $p < 0.05$ were found to be wrong.

- p-values are often misunderstood or abused.

Hypothesis testing can be made more reliable by taking a nuanced view of the data and admitting when answers are not clear-cut, by pre-registering research studies, by independently replicating research, and so forth, but a perfect solution does not exist.

$\Rightarrow$ Actions you take as a researcher can seem reasonable but make things <u>worse</u>.

<u>Multiple Testing and P-hacking</u>

Suppose you have two groups of people. You suspect they are different in some way but are not sure how, so you measure many attributes of each person, then do a hypothesis test on the difference for each attribute.

Testing an attribute at level $\alpha \Rightarrow$ A <u>false positive</u> (concluding there is a difference when null is true) will occur with prob. α

Suppose there is no difference for any of n attributes. What is the prob. of <u>at least one</u> false positive?

$$\begin{pmatrix}\text{Prob of at}\\ \text{least one}\\ \text{false positive}\end{pmatrix} = 1 - \begin{pmatrix}\text{prob of}\\ 0 \text{ false}\\ \text{positives}\end{pmatrix} = 1 - (1-\alpha)^n$$ (Assuming tests are independent)

Ex: $n = 20$, $\alpha = 0.05 \Rightarrow$ prob $= 0.6415$. wow!

$\Rightarrow$ As you keep testing it becomes more and more likely you will observe a significant result due only to chance.

- See also: birthday paradox
- see also: big data $\rightarrow$ many attributes to examine?

<u>P-hacking (data dredging)</u>

Performing multiple tests within a given data set to <u>search for</u> a statistically significant result.

Ex. Run test on data $\Rightarrow$ not significant
then Collect more data $\Rightarrow$ not signif.
then Remove outliers $\Rightarrow$ significant

} multiple tests increased chance of false. pos.

Filtering data, adding more variables/attributes, collecting more data, transforming the data (ex: take log), all can lead to p-hacking when performed during hypothesis testing.

Another side to p-hacking: instead of changing the data until the test is significant

change the <u>test</u> until a test is significant.

$\Rightarrow$ Can also do both: change data and sample many tests $\Rightarrow$ risk of false pos. even higher.

Pre-registering a study and replicating a study can help w/ this higher risk of false positives.

HARKing hypothesizing after results are known

describing a hypothesis made <u>after</u> analyzing the data as if it were a prediction made <u>before</u> using the data. Disingenuous.

⇒ Make a scientific paper appear more impactful by making the hypothesis test appear stronger.

⇒ Increases chance that a false positive becomes embedded in the theories of a field, making the entire field less reliable over time.

Correcting for Multiple Testing

Suppose you conduct m hypothesis tests w/ p-values $P_1, P_2, \ldots, P_m$

For i^{th} test, reject null if $P_i < \frac{\alpha}{m}$ (Bonferroni Method)

Why?

Let R = event that at least one null hyp. incorrectly rejected
R_i = event that null hyp i is incorrectly rejected

Recall: if $E_1, E_2, \ldots E_k$ are events then $Pr\left(\bigcup_{i=1}^{k} E_i\right) \leq \sum_{i=1}^{k} Pr(E_i)$ "union bound"

$$\Rightarrow Pr(R) = Pr\left(\bigcup_{i=1}^{m} R_i\right) \leq \sum_{i=1}^{m} Pr(R_i) = \sum_{i=1}^{m} \frac{\alpha}{m} = \alpha$$

Thus, if each test is held at the stricter level of α/m then the overall prob for at least one false pos is brought back to α, and we have prevented false positives from occurring more often than we want.

Bonferroni Method is very conservative → tries to make even one false rejection unlikely.

It is sometimes more reasonable to control the <u>false discovery rate</u> (FDR)

$$FDR = \frac{\#\text{ rejections of null that are false}}{\#\text{ rejections of null}}$$

<u>Benjamini-Hochberg</u> (BH) controls FDR at level α. (see, e.g. Wasserman (2004) pgs 166-167 for full details)