On p-values and Their Combination

Nicole A. Lazar

How to Think About a *p*-value?

Roughly speaking: Under a given null hypothesis, **and if all assumptions that went into the model hold**, what is the probability of observing a test statistic **as extreme or more extreme** than the one seen in the sample?

Small p-values could arise because the null doesn't hold **or** because other modeling assumptions aren't met.

The latter part is often overlooked!

Some Other Points to Consider

Computing a p-value does not require an alternative hypothesis – this was Fisher's orginal framework.

Neyman-Pearson framework: posits a null and an alternative, but doesn't need a p-value to be computed.

p-value can be thought of as a measure of "compatability" between the model and the data.

What About "Statistical Significance"?

Fisher – used the term to indicate a possibly interesting finding, that requires further research.

Standard for "significance" should be **context specific**.

Over the years, concept *drifted*, plus combination with Neyman-Pearson approach $\rightarrow$ *reject* H_0 *if* $p < 0.05$ *and call the result "statistically significant"* became the dominant paradigm.

Statistical Significance+Big Data

Multiple testing problem – many hypothesis tests conducted simultaneously (e.g. genetics, neuroimaging).

Unadjusted "significance thresholding" leads to many false positives.

Noisy data and small sample size (low power) can lead to false positives or false negatives.

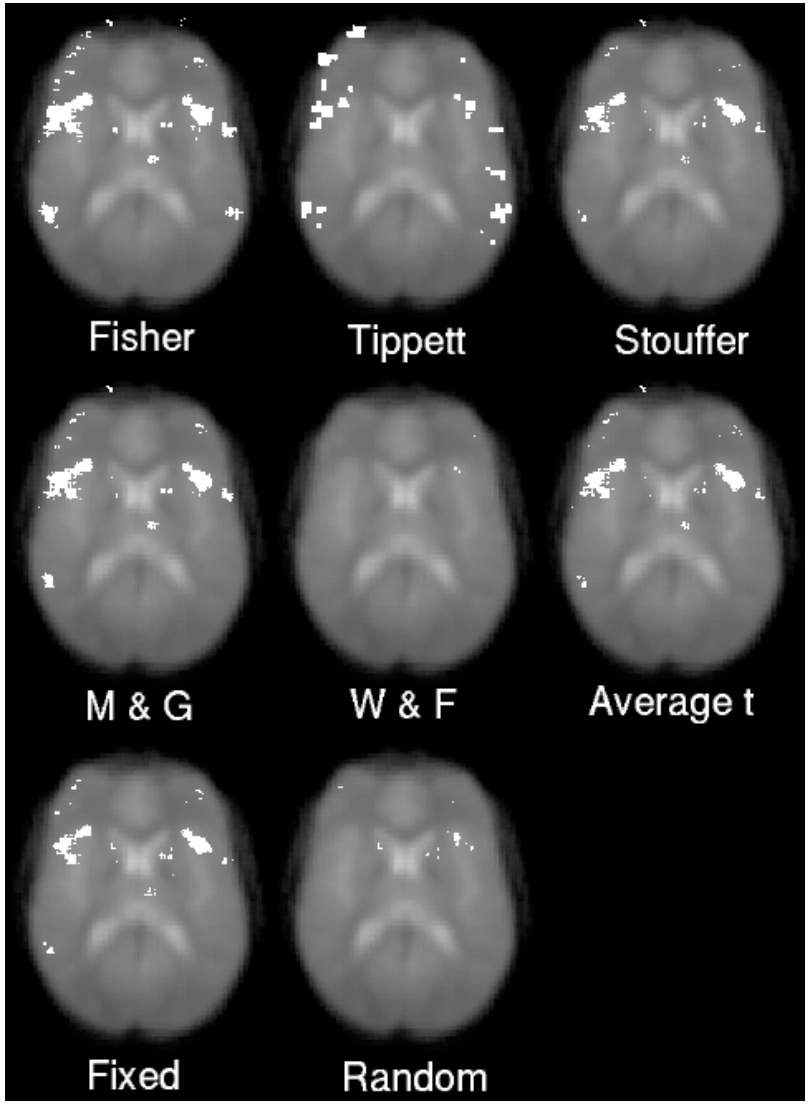
"$p < 0.05$" as "standard" (for publication, grant funding, etc.) leads to significance chasing (p-hacking, etc.)

Why Combine Information?

Convergence of findings from multiple studies.

Indirectly gives larger sample sizes, more reliable (stable, replicable) inference.

An Old Example (Lazar et al., 2002)



(One slice of brain data, 11 subjects, Bonferroni correction)

Combining (Independent) Sources of Information

Old problem, many solutions proposed over the years.

Two main approaches:

1. Combining effect sizes (meta-analysis)
2. Combining p-values

Examples of Combining p-value Methods

Fisher: $T_F = -2\sum_{i=1}^{k} \log p_i$; under (composite) null, $T_F \sim \chi^2_{2k}$.

Tippett: $T_T = \min p_i$; under (composite) null, (transformed) $T_T \sim$ Beta.

Stouffer: $T_S = \sum_{i=1}^{k} \Phi^{-1}(1-p_i)/\sqrt{k}$; under null, $T_S \sim N(0,1)$

M-G: $T_M = -c\sum_{i=1}^{k} \log p_i/(1-p_i)$; under null, $T_M \sim t$ with $c = 5k+4$ df

Some Characteristics

T_F and T_T are sensitive to smallest p-value.

Other methods (not listed above) are sensitive to the largest p-value (these are conservative).

T_M and T_S are compromises between those extremes, and generally similar to each other.

Which Method to Use?

Heard and Rubin-Delanchy (2017) "Choosing between methods of combining p-values"

Establish guidelines, based on likelihood ratio tests, for how to choose a powerful combiner in practice.

Show that the optimal combiner depends on the alternative hypothesis for the distribution of the p-values.

What About Dependent Data?

Methods are for **independent** sources of information.

If they are not independent – more complicated!

Approaches include: appropriate weighting of the individual components; adjusting the distribution under the null.

If not accounted for, test statistic could be "too conservative" **or** "too liberal" (depends on dependence structure).