Combining p-values: Fisher versus Stouffer and the discovery (?) of the odderon

Roger Barlow
The University of Huddersfield
with thanks to Wlodek Guryn who got me interested

LHCb Statistics group

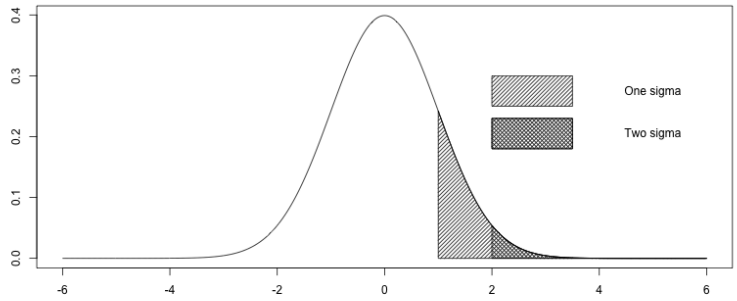
3^{rd} May 2021

The problem

Reminder

p: the probability of getting a result this extreme under the null hypothesis

Z: the equivalent number of standard deviations in the tail of a Gaussian



$Z = 1$ corresponds to $p = 0.159$

$Z = 2$ corresponds to $p = 0.0228$

$Z = 3$ corresponds to $p = 0.0013$

$Z = 5$ corresponds to $p = 2.9 \times 10^{-7}$

Note

Using one-tailed p-values, appropriate when looking for a positive signal

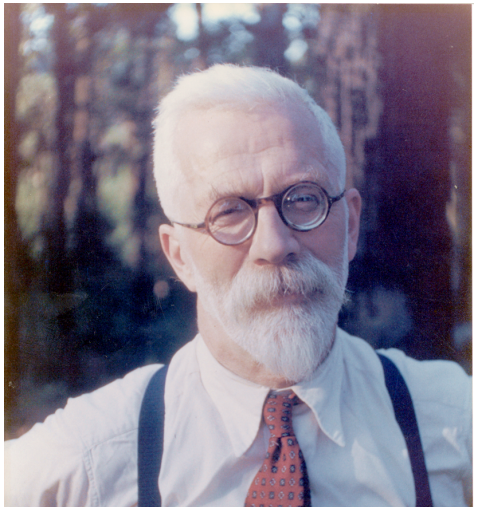
`p=1-pnorm(Z)` in R or `TMath::Prob(Z*Z,1)/2` in Root

`Z=qnorm(1-p)` in R or `TMath::ErfcInverse(p)` in Root

The Question

With 2 (or more) results, how can you combine their p (or Z) values?

Two methods

Fisher 

Stouffer 

$$p = P_{\chi^2}(-2\sum_{i=1}^{N} \log p_i, 2N)$$

$$Z = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_i$$

Actually several other methods - see R Cousins arXiv:0705.2209v2 (2008) for list and discussion. Other methods are mostly extensions of these two.

An example

Experiment A sees an effect at the 2.2 sigma level.
Experiment B sees it at the 3.3 sigma level

p-values are 0.01390345 and 0.0004834241

Let's combine them!

Fisher says:
$-2(\log p_1 + \log p_2) = 23.82$
and $P_{\chi^2}(23.82, 4) = 8.677 \times 10^{-5}$
Corresponding Z is 3.755

Stouffer says:
$Z = (2.2 + 3.3)/\sqrt{2} = 3.889$
Corresponding p-value 5.031×10^{-5}

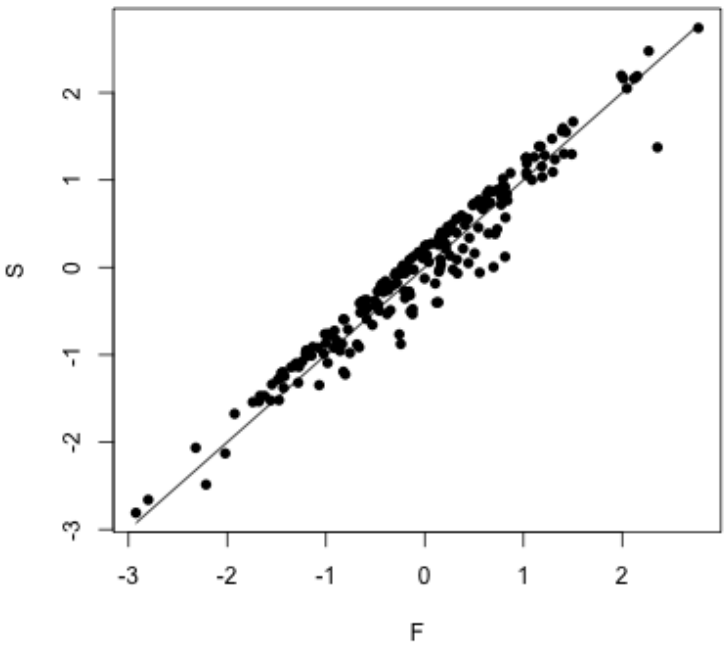
Two methods give similar but not identical results. Stouffer's Z value is a bit higher.

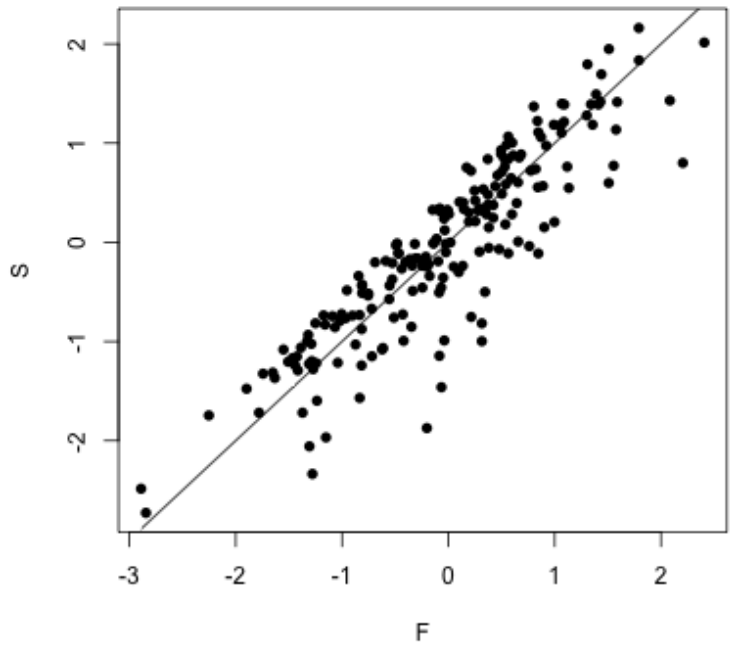
This case is typical - but not universal

What's happening?

Take random p-values and look at the combined Z from the two methods

Combining 2 experiments and 5 experiments





Lots of Stouffer values slightly above the line of equality - but tail below

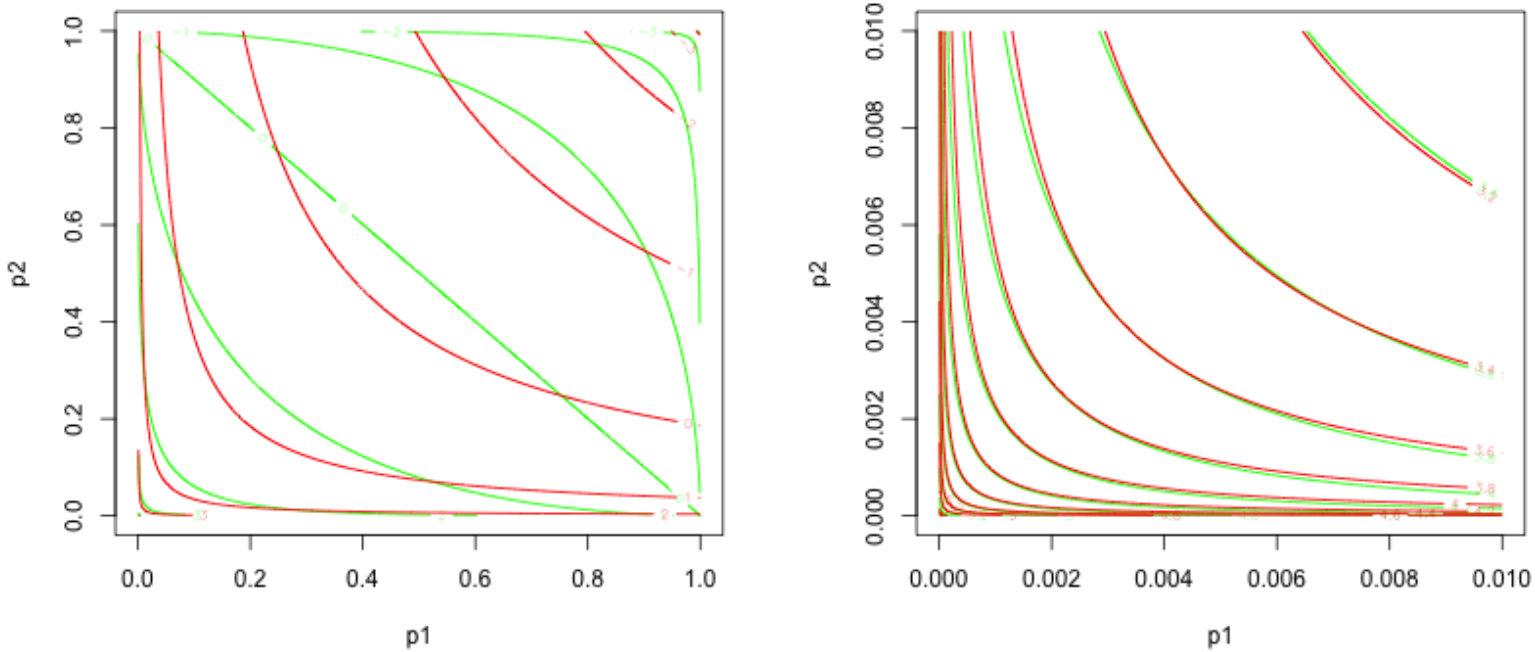
What's going on?

Fisher says:
Under H_0, each p is uniformly distributed
Then $y = -2 \log p$ has exponential distribution $\propto e^{-y/2}$
This happens to be the χ^2 distribution for 2 degrees of freedom
The sum of N of these is a χ^2 distributions with $2N$ d.o.f.

Stouffer says:
Under H_0, each Z is distributed according to a Gaussian with mean 0 and sigma=1
The sum is distributed according to a Gaussian with mean 0 and sigma $\sqrt{N}$

Compare and Contrast - 2 results

Contours of constant combined p (or Z) according to Fisher and Stouffer.
Contours labelled with Z
Overall plots differ - though bottom left corner (interesting part) similar



Both sets of contours are 'correct': under H_0 this plane is uniformly populated and contours demarcate appropriate area.
Although significance is the same, power will not be. Difference depends on alternative hypothesis H_1.

Compare and Contrast - 2 results

continued

Consider green (Stouffer) $Z = 0$ contour, straight line at 45°

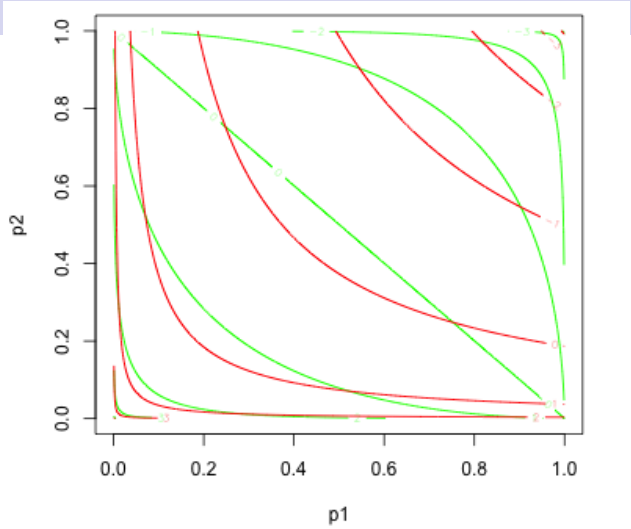
Cases where $Z_2 = -Z_1$. Cancel to give 0

$p_2 = 1 - p_1$. Taking logs emphasises small values

.5 and .5 $\rightarrow$ -0.69 - 0.69 = -1.38

.1 and .9 $\rightarrow$ -2.30 -0.11 =- 2.41

.05 and .95 $\rightarrow$ -3.00 -0.05 = -3.05



With Stouffer, opposite Z values cancel.

With Fisher, one low p-value is not nullfied by the other p value being large

Which applies? You need to decide

H_0 is: *All the experiments see nothing*

Is the alternative *All the experiments see something*

or *Some of the experiments see something*

Test cases

- A new fertiliser is applied to 10 similar fields.
- A new fertiliser is applied to 10 different crops.
- ATLAS and CMS search for the same particle
- A particle search is done using muons and using electrons.

Test cases

- A new fertiliser is applied to 10 similar fields.
  If one shows a somewhat improved crop yield but the rest have little change, there is probably no effect. Combine data samples if possible? Otherwise use Stouffer
- A new fertiliser is applied to 10 different crops.

- ATLAS and CMS search for the same particle

- A particle search is done using muons and using electrons.

Test cases

- A new fertiliser is applied to 10 similar fields.
  If one shows a somewhat improved crop yield but the rest have little change, there is probably no effect. Combine data samples if possible? Otherwise use Stouffer
- A new fertiliser is applied to 10 different crops.
  If broccoli yields go up, but the rest have little change, that suggests it could be a good fertiliser for broccoli. Fisher is appropriate
- ATLAS and CMS search for the same particle

- A particle search is done using muons and using electrons.

Test cases

- A new fertiliser is applied to 10 similar fields.
  If one shows a somewhat improved crop yield but the rest have little change, there is probably no effect. Combine data samples if possible? Otherwise use Stouffer
- A new fertiliser is applied to 10 different crops.
  If broccoli yields go up, but the rest have little change, that suggests it could be a good fertiliser for broccoli. Fisher is appropriate
- ATLAS and CMS search for the same particle
  If one sees an excess but the other sees a deficit then they cancel. Nothing to see here, move along. Stouffer.
- A particle search is done using muons and using electrons.

Test cases

- A new fertiliser is applied to 10 similar fields.
  If one shows a somewhat improved crop yield but the rest have little change, there is probably no effect. Combine data samples if possible? Otherwise use Stouffer
- A new fertiliser is applied to 10 different crops.
  If broccoli yields go up, but the rest have little change, that suggests it could be a good fertiliser for broccoli. Fisher is appropriate
- ATLAS and CMS search for the same particle
  If one sees an excess but the other sees a deficit then they cancel. Nothing to see here, move along. Stouffer.
- A particle search is done using muons and using electrons.
  An excess in one but not the other could be an indication of breakdown of lepton universality. Fisher.

Test cases

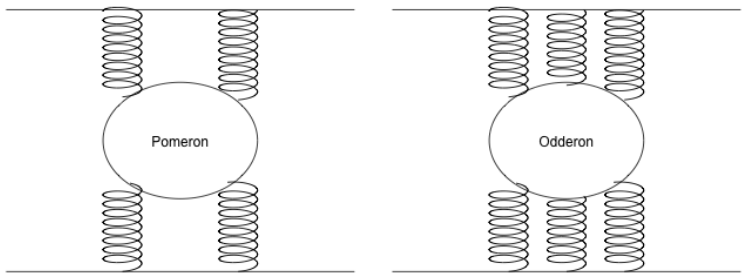
- A new fertiliser is applied to 10 similar fields.
  If one shows a somewhat improved crop yield but the rest have little change, there is probably no effect. Combine data samples if possible? Otherwise use Stouffer
- A new fertiliser is applied to 10 different crops.
  If broccoli yields go up, but the rest have little change, that suggests it could be a good fertiliser for broccoli. Fisher is appropriate
- ATLAS and CMS search for the same particle
  If one sees an excess but the other sees a deficit then they cancel. Nothing to see here, move along. Stouffer.
- A particle search is done using muons and using electrons.
  An excess in one but not the other could be an indication of breakdown of lepton universality. Fisher.

Stouffer imposes the requirement that the experiments are equivalent. If this is true, it adds to the power of the test.

Example: the odderon

V M Abazov *et al.*, (DØ and TOTEM). arXiv:2012.03981v1 (2020)

C-odd 3-gluon exchange, contributing to elastic high-energy hadron scattering, like the Pomeron (which is C-even 2-gluon exchange)



TOTEM found a C-odd term at the 4.7 sigma level using pp scattering, from the real to imaginary ratio of the forward scattering amplitude

DØ used $p\overline{p}$ scattering and establish the need for a C-odd term (difference between pp and $p\overline{p}$) at the 3.4 sigma level

4.7 sigma + 3.4 sigma =?

The odderon

continued

4.7 sigma + 3.4 sigma =?

Stouffer says $(4.7 + 3.4)/\sqrt{2} = 5.7$ sigma

Fisher says effective $\chi^2 = -2(\log p_1 + \log p_2) = 42.1$, which for 4 dof corresponds to $p = 1.57 \times 10^{-8}, Z = 5.53$

Which is appropriate? Probably Fisher.

But both values are more than 5, so 'discovery' is justified

More information helps

Large p-value (small Z) on its own is uninformative: no information on quality of experiment

If Z_i are measurements of a quantity which is zero under H_0: $Z_i = x_i/\sigma_i$

Best estimate of x is $\sum_i w_i x_i$. with $w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2}$

which has error $\frac{1}{\sqrt{\sum_i 1/\sigma_i^2}}$

Overall $Z = \frac{\sum_i Z_i/\sigma_i}{\sqrt{\sum_i 1/\sigma_i^2}}$

If the σ_i are all the same, this reduces to Stouffer's method

Conclusion

Fisher's and Stouffer's methods for combining significances both work

Fisher is easy and Stouffer is very easy

Use Stouffer if the experiments are measuring the same thing, so deficits can cancel excesses

Use Fisher if the experiments are separate, and the impact of an excess is not reduced by deficits elsewhere.

Combining Totem and DØ results to 'discover' the odderon is statistically valid