The Look Elsewhere Effect

Discussion for ODSL Journal Club

Zoom / 25 June 2021

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

Multiple testing

The Look Elsewhere Effect

An example and practical solution (Gross and Vitells)

Brief comments on:

- Multidimensional LEE
- Bayesian approach to LEE

Frequentist Hypothesis Test

For a frequentist hypothesis test of a null (no-signal) hypothesis H_0 define a "critical region" w in the data space $\boldsymbol{x}$, which has probability content assuming H_0 not greater than a prespecified small constant α (the "size" of the test):

$$P(\mathbf{x} \in w | H_0) \leq \alpha$$

Inequality needed for discrete data; here suppose equality.

Choose w to maximize power wrt alternative $= P(\boldsymbol{x} \text{ in } w \mid H_1)$

If $\boldsymbol{x}$ is found in w, reject H_0 and announce discovery of new signal.

Equivalently, reject H_0 if its p-value is less than α:

$$p_0 = P(\text{data equally or more incompatible than what was observed} | H_0)$$

The probability to reject H_0 if it is true is equal to α (the type-I error rate).

Multiple Testing, Bonferroni Correction

If we carry out N tests, the "Family Wise Error Rate" is

$$P(\text{reject } H_0 \text{ in any of } N \text{ tests}|H_0) \equiv \overline{\alpha}$$

If the tests are independent and each of size α, then

$$\overline{\alpha} = 1 - (1 - \alpha)^N$$

For N large, FWER $\longrightarrow$ 1 and one will surely discover a signal.

Even if the tests are not independent, can show $\overline{\alpha} \leq \alpha N$

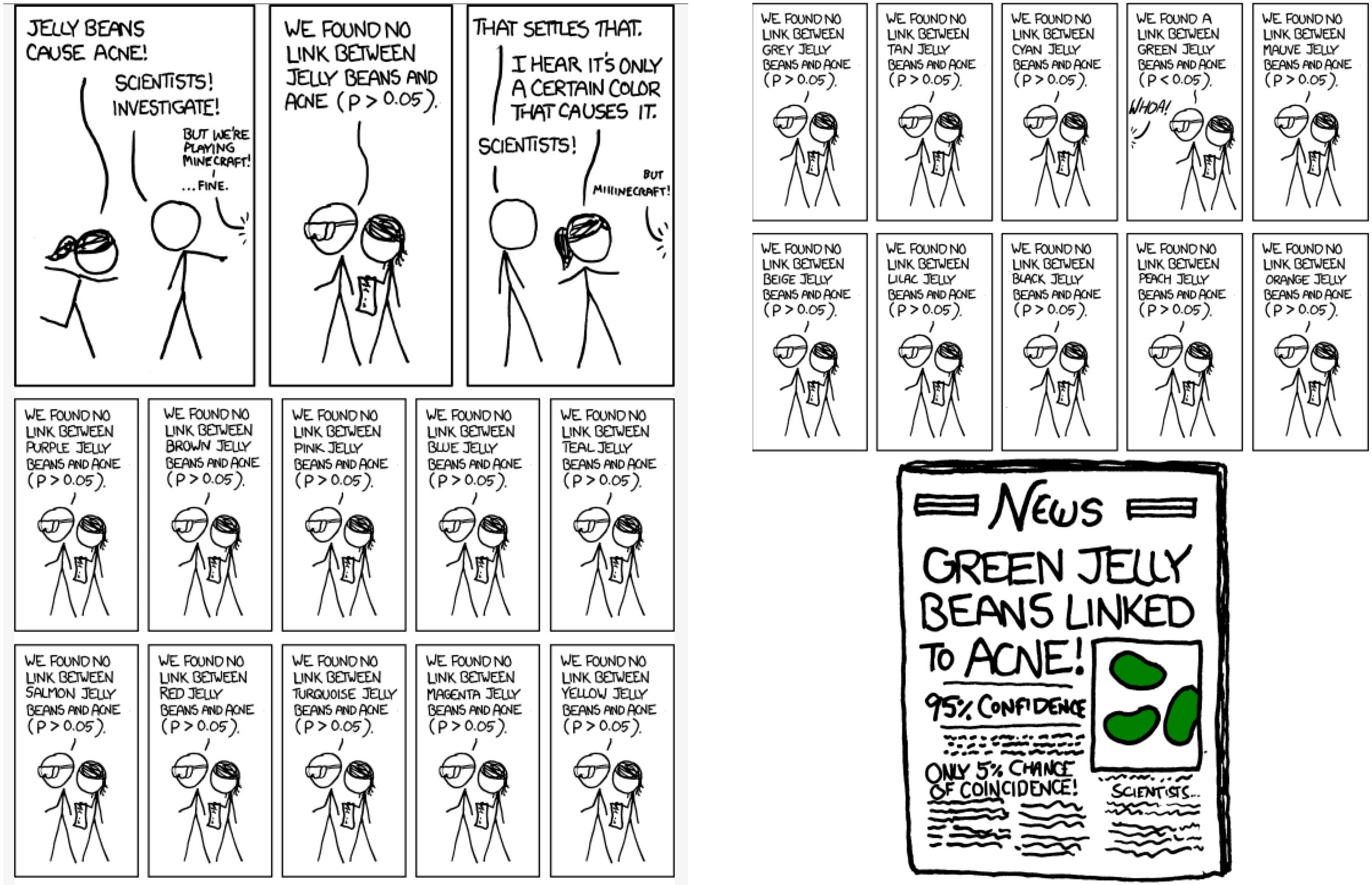
So we can ensure the FWER does not exceed α if we reject H_0 if we reject H_0 when

$$p_0 < \frac{\alpha}{N}$$

Bonferroni, C. E., Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 1936

Other corrections less conservative (Sidak, Holm-Bonferroni,...)

https://imgs.xkcd.com/comics/significant.png



For α = 0.05, N = 20, FWER = 0.64

Multiple Testing $\longrightarrow$ LEE

For discrete tests this is called "multiple testing" or "multiple comparisons".

In particle physics we often carry out a test of H_0 (no-signal) designed to have high power with respect to an alternative $H_1(\theta)$ indexed by a continuous parameter (e.g., mass of a new particle).

There is a test for each θ, so $N \to \infty$ but the tests are not independent (e.g., two masses close to each other).

This is the Look Elsewhere Effect (~ continuous multiple testing).

Out of the tests carried out, the let the smallest p-value $= p_{\mathrm{local}}$.

We want $p_{\mathrm{global}} = P(p_{\mathrm{local}} \leq p_{\mathrm{local,obs}} \mid H_0)$

For N independent tests, $p_{\mathrm{global}} = 1 - (1 - p_{\mathrm{local}})^N$ (not useful here).

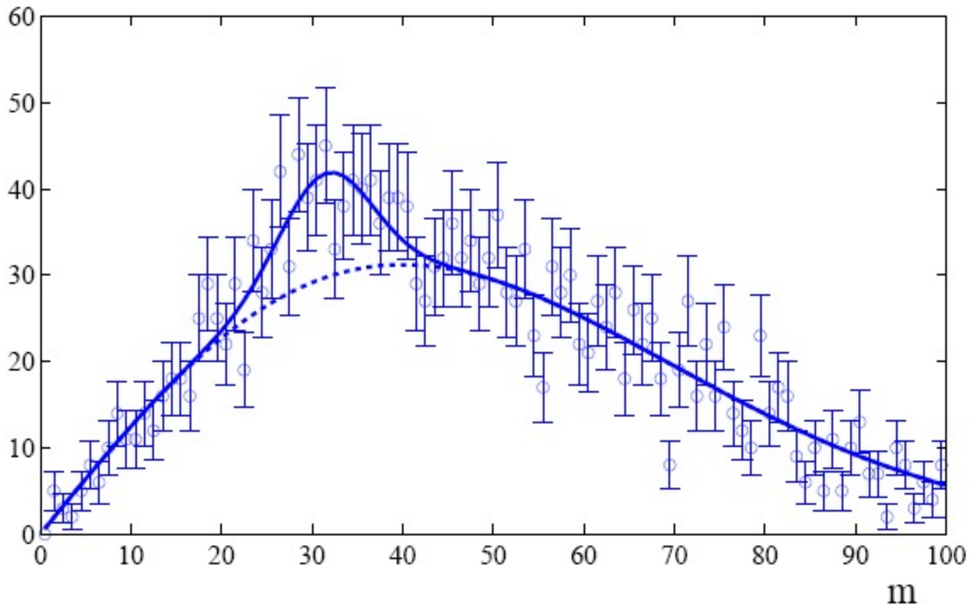
Protype example of LEE and a solution

Eilam Gross and Ofer Vitells. Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C - Particles and Fields*, 70:525–530, 2010.

R. B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika 74 (1987), 33-43.

Suppose a model for a mass distribution allows for a peak at a mass m with amplitude μ.

The data show a bump at a mass m_0.



How consistent is this with the no-bump ($\mu = 0$) hypothesis?

Local p-value

First, suppose the mass m_0 of the peak was specified a priori.

Test consistency of bump with the no-signal ($\mu = 0$) hypothesis with e.g. likelihood ratio

$$t_{\text{fix}} = -2 \ln \frac{L(0, m_0)}{L(\hat{\mu}, m_0)}$$

where "fix" indicates that the mass of the peak is fixed to m_0.

The resulting p-value

$$p_{\text{local}} = \int_{t_{\text{fix,obs}}}^{\infty} f(t_{\text{fix}}|0)\, dt_{\text{fix}}$$

gives the probability to find a value of t_{fix} at least as great as observed at the specific mass m_0 and is called the local p-value.

Global p-value

But suppose we did not know where in the distribution to expect a peak.

What we want is the probability to find a peak at least as significant as the one observed anywhere in the distribution.

Include the mass as an adjustable parameter in the fit, test significance of peak using

$$t_{\mathrm{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{m})}$$

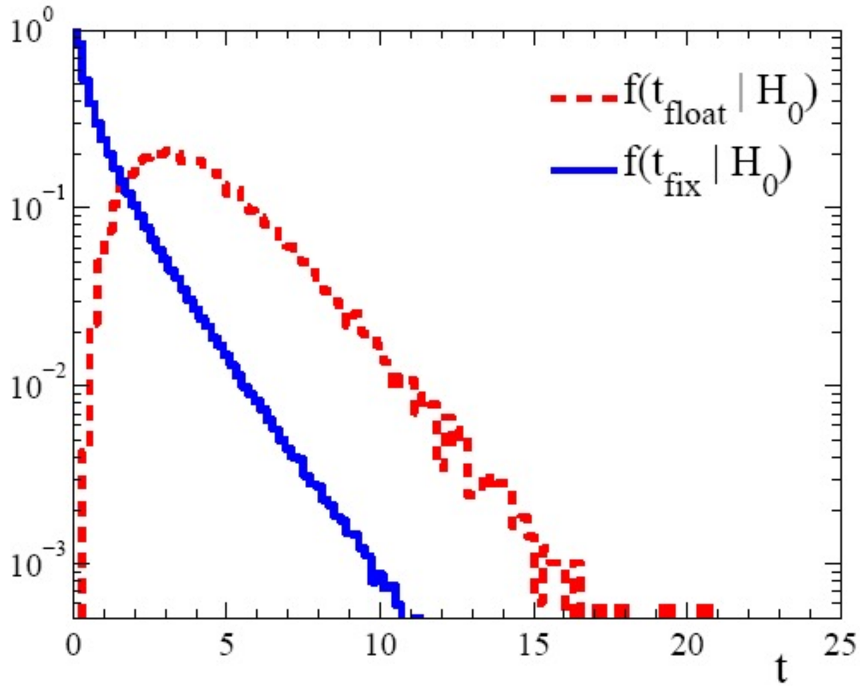
(Note m does not appear in the $\mu = 0$ model.)

$$p_{\mathrm{global}} = \int_{t_{\mathrm{float,obs}}}^{\infty} f(t_{\mathrm{float}}|0)\, dt_{\mathrm{float}}$$

Distributions of t_{fix}, t_{float}

For a sufficiently large data sample, $t_{\text{fix}} \sim$ chi-square for 1 degree of freedom (Wilks' theorem), significance $Z_{\text{fix}} = \Phi^{-1}(1 - p_{\text{local}}) = \sqrt{t_{\text{fix}}}$.

For t_{float} there are two adjustable parameters, μ and m, and naively Wilks theorem says $t_{\text{float}} \sim$ chi-square for 2 d.o.f.



In fact Wilks' theorem does not hold in the floating mass case because on of the parameters (m) is not-defined in the $\mu = 0$ model.

So getting t_{float} distribution is more difficult.

Approximate correction for LEE

We would like to be able to relate the p-values for the fixed and floating mass analyses (at least approximately).

Gross and Vitells (using result from Davies) show the p-values are approximately related by

$$p_{\mathrm{global}} \approx p_{\mathrm{local}} + \langle N(c) \rangle$$

where $\langle N(c) \rangle$ is the mean number "upcrossings" of $t_{\mathrm{fix}} = -2\ln\lambda$ in the fit range based on a threshold

$$c = t_{\mathrm{fix,obs}} = Z_{\mathrm{local}}^2$$

and where $Z_{\mathrm{local}} = \Phi^{-1}(1 - p_{\mathrm{local}})$ is the local significance.

So we can either carry out the full floating-mass analysis (e.g. use MC to get p-value), or do fixed mass analysis and apply a correction factor (much faster than MC).

E. Gross and O. Vitells, EPJC 70:525–530, 2010.

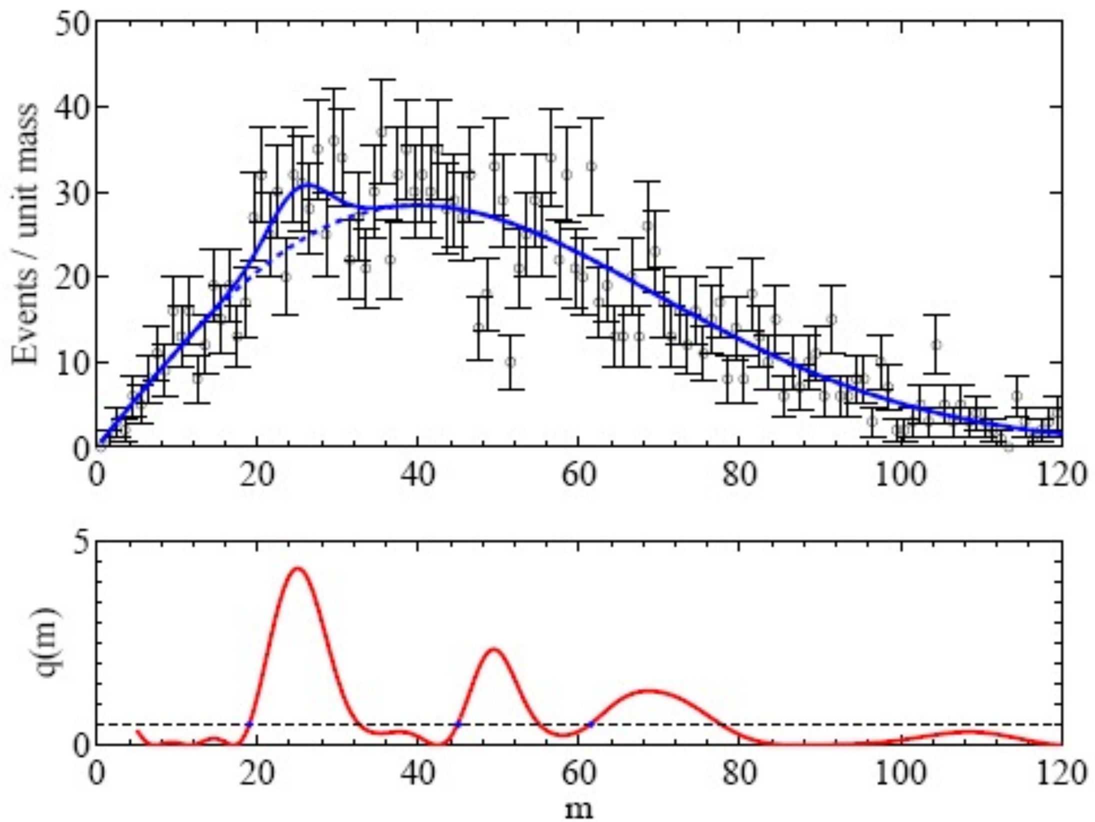
Upcrossings of $-2\ln L$

The Gross-Vitells formula for the trials factor requires $\langle N(c) \rangle$, the mean number "upcrossings" of $t_{\rm fix} = -2\ln\lambda$ in the fit range based on a threshold $c = t_{\rm fix} = Z_{\rm fix}^2$.

$\langle N(c) \rangle$ can be estimated from MC (or the real data) using a much lower threshold c_0:

$$\langle N(c) \rangle \approx \langle N(c_0) \rangle e^{-(c-c_0)/2}$$

In this way $\langle N(c) \rangle$ can be estimated without need of large MC samples, even if the the threshold c is quite high.



MC study by Gross and Vitells validating approximation for finding mean number of upcrossings ($c_0 = 0.5$)

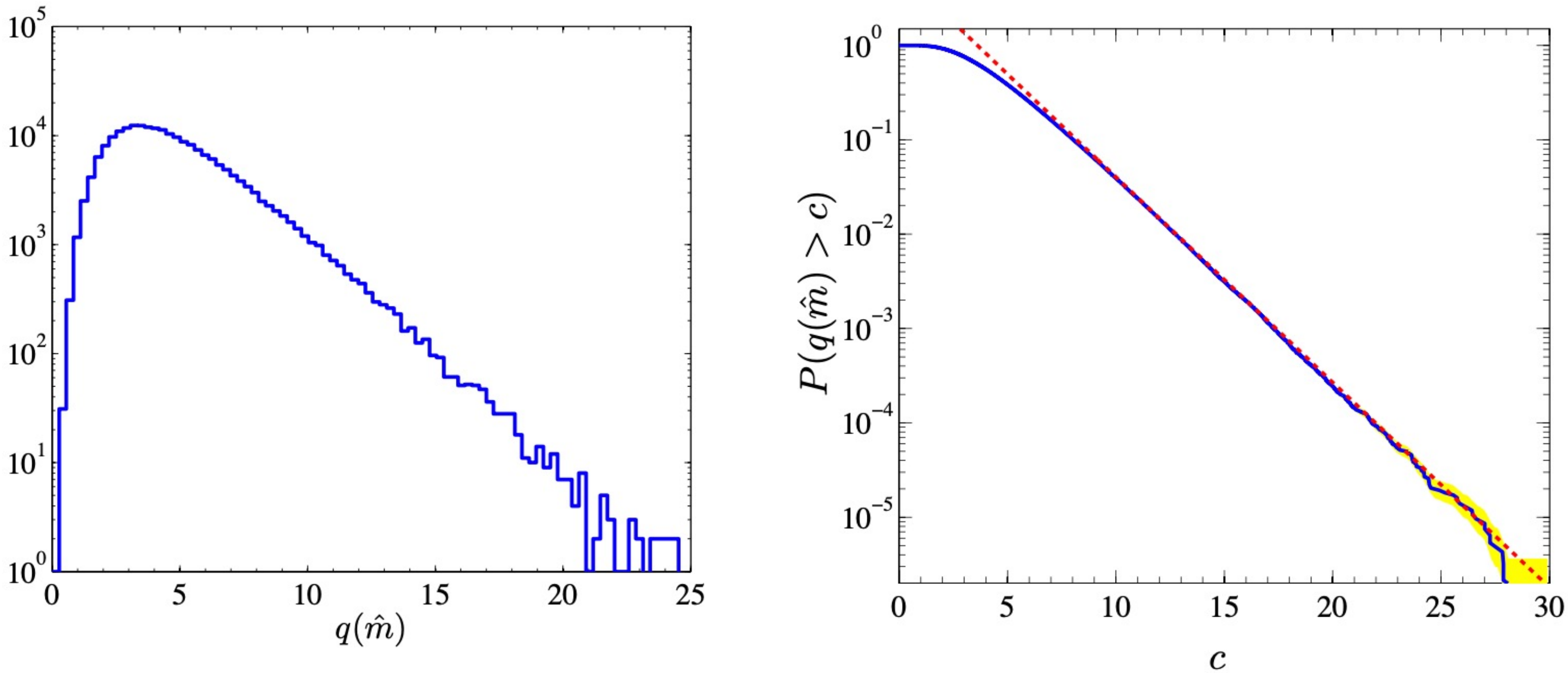


Fig. 2. (top) Distribution of $q(\hat{m})$. (bottom) Tail probability of $q(\hat{m})$. The solid line shows the result of the Monte Carlo simulation, the dotted red line is the predicted bound (eq. 3) with the estimated $\langle N(c_0)\rangle$ (see text). The yellow band represents the statistical uncertainty due to the limited sample size.

Trails factor for example of Gross and Vitells

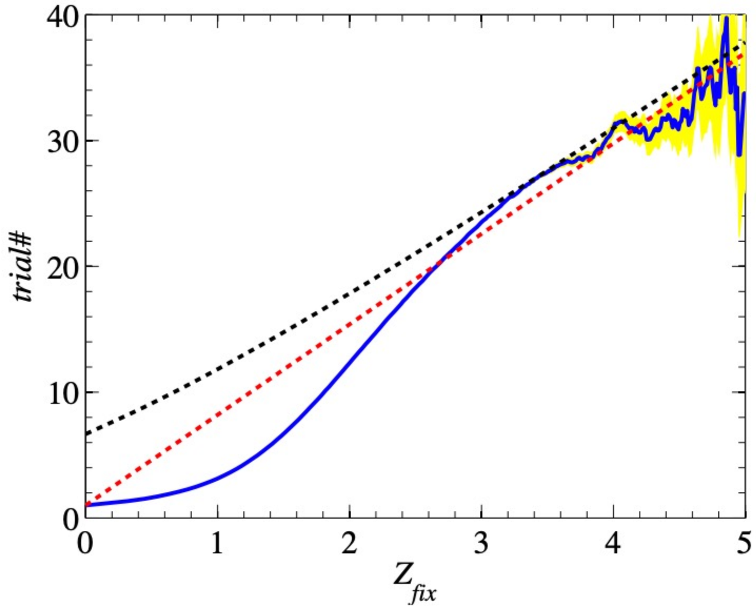


Fig. 3. The trial factor estimated from toy Monte Carlo simulations (solid line), with the upper bound of eq.(3) (dotted black line) and the asymptotic approximation of eq.(12) (dotted red line). The yellow band represents the statistical uncertainty due to the limited sample size.

$$\text{trials factor} = \frac{p_{\text{global}}}{p_{\text{local}}}$$

$$\approx \frac{p_{\text{local}} + \langle N \rangle}{p_{\text{local}}}$$

$$Z_{\text{fix}} = \Phi^{-1}(1 - p_{\text{local}})$$

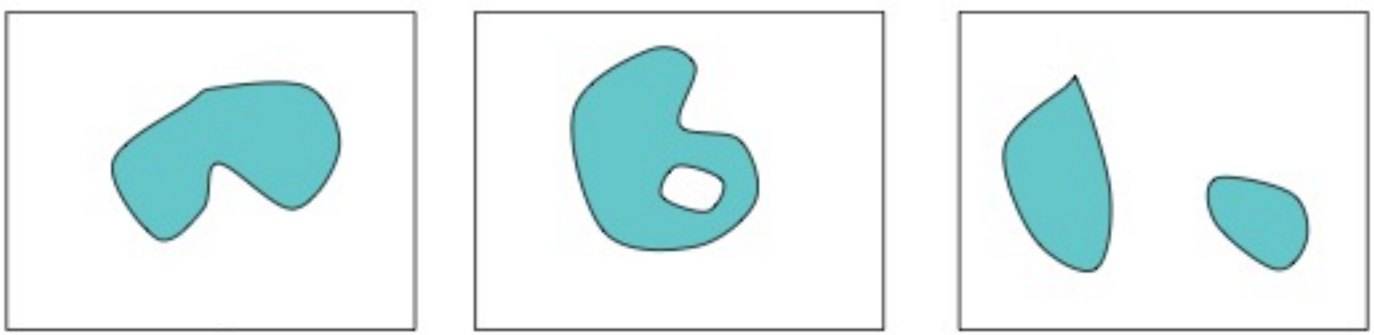
Approximate correction is good for $Z > 3$, i.e., relevant for claiming signal at 3-sigma or more.

Multidimensional look-elsewhere effect

Generalization to multiple dimensions: number of upcrossings replaced by expectation of Euler characteristic:

$$\mathrm{E}[\varphi(A_u)] = \sum_{d=0}^{n} \mathcal{N}_d \rho_d(u)$$

- Number of disconnected components minus number of \`holes'



φ=1 φ=0 φ=2

Applications: astrophysics (coordinates on sky), search for resonance of unknown mass and width, ...

Bayesian approach to LEE

See, e.g., James Berger, Bayesian approach to discovery, PHYSTAT11 contribution, https://indico.cern.ch/event/107747/

In Bayesian statistics, probability is associated with hypotheses.

A Bayesian tool for discovery of a new signal is the Bayes factor:

$$B_{10} = \frac{\int P(\mathbf{x}|H_1(\theta))\, \pi(\theta)\, d\theta}{P(\mathbf{x}|H_0)}$$

= posterior odds if prior odds one.

The large parameter space of the alternative H_1 is automatically taken into account by integrating over the internal parameter.

The prior pdf $\pi(\theta)$ encodes what region of the parameter space is deemed relevant (i.e., "where else you need to look").

Summary on Look-Elsewhere Effect

The Look-Elsewhere Effect is when we test a single model (e.g., SM) with multiple observations, i.e., in multiple places.

This is distinct from the case of exclusion limits. There we test different signal hypotheses (typically once) and say whether each is excluded (result is a confidence interval).

With exclusion there is, however, the also problematic issue of testing many signal models (or parameter values) and thus excluding some for which one has little or no sensitivity.

Approximate correction for LEE should be sufficient, and one should also report the uncorrected significance.

Extra slides

Thoughts on the LEE

[13]The extent to which other people's searches should be included in an allowance for the "look elsewhere" effect depends subtly on the implied question being addressed. Thus are we considering the chance of obtaining a statistical fluctuation in any of the analyses we have performed; or by anyone analysing data in our experiment; or by any Particle Physicist this year? Anyone observing a possible Higgs signal at the LHC would be very unhappy about having to reduce the significance of their result because of the statistical fluctuations that could occur in speculative searches performed elsewhere.

Louis Lyons, Open statistical issues in particle physics, Annals of Applied Statistics 2008, Vol. 2, No. 3, 887-915

"There's no sense in being precise when you don't even know what you're talking about." — John von Neumann

Some papers I didn't manage to get through

S. Algeri, D.A. van Dyk, J. Conrad, B. Anderson, *On methods for correcting the look-elsewhere effect in searches for new physics*, Journal of Instrumentation 11 P12010, 2016, arXiv:1602.03765.

Multidimensional method; also nice description of the formalism.

Adrian E. Bayer, Uros Seljak, *The look-elsewhere effect from a unified Bayesian and frequentist perspective,* JCAP 10 (2020) 009, arXiv:2007.13821

...a continuous generalization of the Bonferroni and Sidak corrections by applying the Laplace approximation to evaluate the Bayes factor, and in turn relating the trials factor to the prior-to-posterior volume ratio. We use this to define a test statistic whose frequentist properties have a simple interpretation in terms of the global p-value,...