Multiple hypothesis testing and testing one hypothesis multiple times: a unified (re)view

Sara Algeri[1,2]
Highlights of joint works with:
Prof. David van Dyk[1] and Prof. Jan Conrad[2]

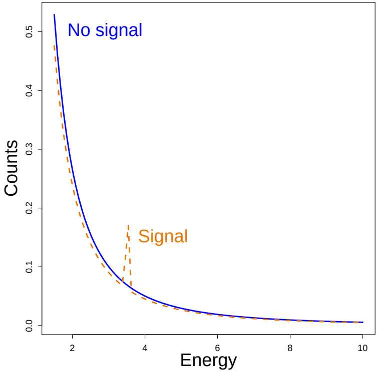
[1]Department of Mathematics, Imperial College London, London, UK
[2]Department of Physics, Stockholm University, Stockholm, Sweden.

PhyStat-ν · June 1, 2016

General framework

Goal of statistical signal detection in physics

We would like to distinguish signals of physics phenomena from the random fluctuations of the data.



- **E.g.,** Neutrinos, Higgs boson, quark.
- We want to detect a bump (the signal of the particle) on top of a background flux.

How does statistics tackle this problem?

- Approach 1:
  Multiple hypotesis testing $\Rightarrow$ Bonferroni's correction.

- Approach 2:
  Simulations $\Rightarrow$ Monte Carlo, Bootstrap.

- Approach 3: **Testing one hypothesis multiple times**.

  Statistics:

  Hypothesis testing when a nuisance parameter is present only under the alternative $\Rightarrow$ Davies (1977, 1987).

  Physics:

  "Look-elsewhere effect" $\Rightarrow$ Gross and Vitells (2010).

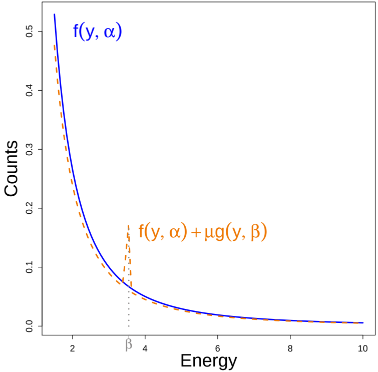
Questions I would like to address with this talk

1. What does it mean exactly to "test one hypothesis multiple times", and in what sense is it equivalent to the "look-elsewhere effect"?
2. Can we tackle both nested and non-nested models with this approach?
3. What is the difference between testing one hypothesis multiple times and multiple hypothesis testing?
4. When do multiple hypothesis testing and testing one hypothesis multiple times coincide in some sense?
5. What else can we do, and what is the potential of working in this direction?

Outline

1. What does it mean exactly to "test one hypothesis multiple times", and in what sense is it equivalent to the "look-elsewhere effect"?
2. Can we tackle both nested and non-nested models with this approach?
3. What is the difference between testing one hypothesis multiple times and multiple hypothesis testing?
4. When do multiple hypothesis testing and testing one hypothesis multiple times coincide in some sense?
5. What else can we do and what is the potential of working in this direction?

A statistical framework for a physics problem



The model of interest is proportional to

$$\underbrace{f(y,\alpha)}_{\text{background}} + \underbrace{\mu}_{\substack{\text{signal}\\\text{strength}}} \underbrace{g(y, \overbrace{\beta}^{\substack{\text{signal}\\\text{location}}})}_{\text{bump}} \qquad (1)$$

and we test

$$H_0 : \mu = 0 \quad \text{vs.} \quad \mu > 0. \qquad (2)$$

Problems

μ is on the boundary of its parameter space + β is not defined under H_0.

Solutions

$\underbrace{\text{Chernoff, 1954 + Davies, 1987}}_{\text{Theoretical solutions}}, \underbrace{\text{Gross and Vitells, 2010}}_{\text{Practical solution}}.$

Testing on the boundary of the parameter space

- **Model:**

$$\propto f(y,\alpha) + \mu g(y,\beta) \qquad \mu \geq 0 \tag{3}$$

For now, let β be fixed, the model in (3) is identifiable.

- **Test**

$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu > 0$$

- **Test statistics***:

$$LRT = -2\log[\underbrace{L(0,\hat{\alpha}_0,\text{-})}_{\text{Likelihood under } H_0} - \underbrace{L(\hat{\mu},\hat{\alpha},\beta)}_{\text{Likelihood under } H_1}] \tag{4}$$

* for the specific case of β be fixed.

- η is on the boundary $\Rightarrow$ WE CAN USE Chernoff, 1954 i.e.:

$$\boxed{LRT \xrightarrow[n\to\infty]{d} \tfrac{1}{2}\chi^2_1 + \tfrac{1}{2}\delta(0) \quad \text{under } H_0} \tag{5}$$

Testing one hypothesis multiple times (1)

- If β fixed, under H_0 the LRT is asymptotically $\frac{1}{2}\chi^2_1 + \frac{1}{2}\delta(0)$.
- If we let β vary$\Rightarrow$ Under H_0, $\{LRT(\beta), \beta \in \boldsymbol{B}\}$ is asymptotically a $\frac{1}{2}\chi^2_1 + \frac{1}{2}\delta(0)$ random process indexed by β.
- In practice:
  - Define a grid $\boldsymbol{B}_R$ of R β_r values over the energy spectrum $\boldsymbol{B}$.
  - $\forall \beta_r \in \boldsymbol{B}_R$ calculate $LRT(\beta_r)$.

Many "sub"-alternatives...

It is like if we had many alternative hypothesis $H_{11}, \dots, H_{1r}, \dots, H_{1R}$, one for each value $\beta_r \in \boldsymbol{B}_R$, and for each of them we have one value $LRT(\beta_r)$.

...but yet just one test statistic...

We finally combine the R $LRT(\beta_r)$ values in a unique test statistics

$$\max_{\beta_r \in \boldsymbol{B}_R} LRT(\beta_r)$$

Testing one hypothesis multiple times (2)

... and one global p-value...

The **p-value** of our test $H_0 : \eta = 0$ versus $H_a : \eta > 0$ is in the form

$$\boxed{P(\sup_{\beta \in \boldsymbol{B}} LRT(\beta) > c)} \tag{6}$$

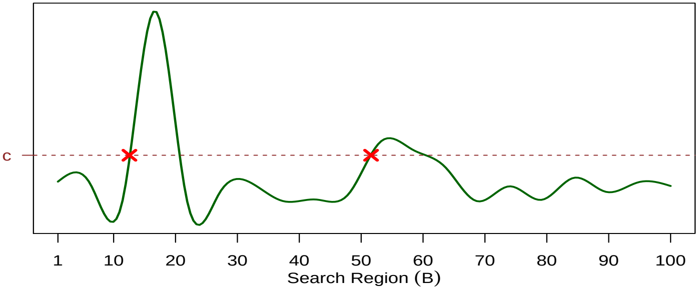
with $c = \max_{\beta_r \in \boldsymbol{B}_R} LRT(\beta)$.

...which we must calculate/approximate somehow!

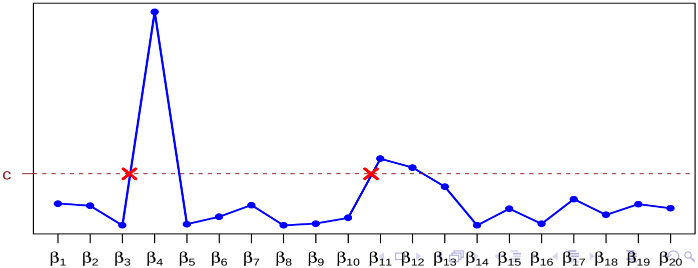
To do so, we first need to introduce the concept of **upcrossings** of the LRT-process $\{LRT(\beta), \beta \in \boldsymbol{B}\}$.

What do we mean by "upcrossings"?

True LRT-process under H_0



Discretized version we deal with in practice



Approximation of $P\left(\sup_{\beta\in\boldsymbol{B}} LRT(\beta) > c\right)$

- From **Davies, 1987** we have that as $c \to +\infty$

$$P(\sup LRT(\beta) > c) \approx \frac{P(\chi^2_1 > c)}{2} + \underbrace{\frac{e^{\frac{c}{2}}}{\sqrt{2\pi}} \int_L^U \kappa(\beta) d\beta}_{\text{Expected \# of upcrossings over } c \text{ of the LRT process under } H_0} \quad (7)$$

- if $c \not\to +\infty \Rightarrow$ we have an upper bound for $P(\sup LRT(\beta) > c)$.
- $\kappa(\beta)$ is complicated $\Rightarrow$ (7) proposed in **Gross and Vitells, 2010**

$$P(\sup LRT(\beta) > c) \approx \frac{P(\chi^2_1 > c)}{2} + \underbrace{e^{-\frac{c-c_0}{2}} E[N(c_0)|H_0]}_{=E[N(c)|H_0]} \quad \text{Expected \# of upcrossings over } c_0 \text{ of the LRT process under } H_0 \quad (8)$$

- where $c_0 << c$ and $E[N(c_0)|H_0]$ is estimated using (few) Bootstrap simulations.

For more details and an alternative approach to the problem, check out:

- S. Algeri, D.A. van Dyk, J. Conrad and B. Anderson *Looking for a Needle in a Haystack? Look Elsewhere! A statistical comparison of approximate global p-values.* Submitted, 2016.

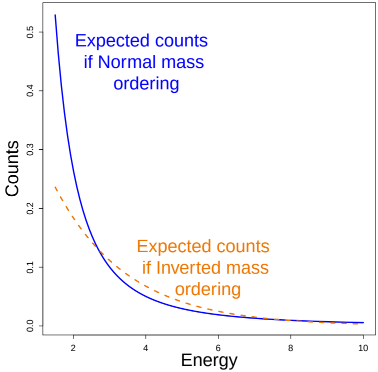
Outline

1. What does it mean exactly to "test one hypothesis multiple times", and in what sense is it equivalent to the "look-elsewhere effect"?
2. Can we tackle both nested and non-nested models with this approach?
3. What is the difference between testing one hypothesis multiple times and multiple hypothesis testing?
4. When do multiple hypothesis testing and testing one hypothesis multiple times coincide in some sense?
5. What else can we do and what is the potential of working in this direction?

Non-nested models comparison in physics

Goal

Establish the neutrino mass hierarchy.



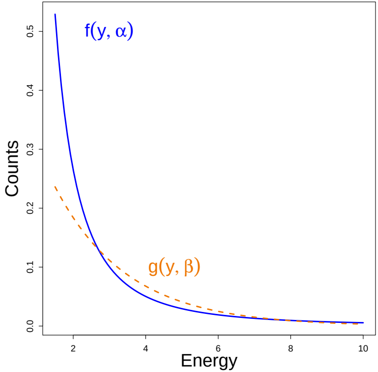
Let m_2 and m_3 be the masses of neutrinos type 2 and 3 respectively. We would like to know if

$$H_0 : \underbrace{m_2 \text{ is lighter than } m_3}_{\textbf{Normal ordering (NO)}}$$

versus

$$H_1 : \underbrace{m_3 \text{ is lighter than } m_2}_{\textbf{Inverted ordering (IO)}}$$

The statistical problem



- Let $x_1, \ldots, x_B$ be the (binned) observed counts of neutrino-induced reactions;
- $y_1, \ldots, y_B$ are the expected counts in each of the B energy bins;
- $f(y, \alpha)$ is the model for the expected counts if NO;
- $g(y, \beta)$ is the model for the expected counts if IO;
- $f \not\equiv g$ for any α and β.

Which one between f and g fits the data better?

$$H_0 : \underbrace{f(y, \alpha)}_{\textbf{NO}} \qquad \text{vs.} \qquad H_1 : \underbrace{g(y, \beta)}_{\textbf{IO}}$$

Problem

f and g are non-nested.

Solutions

$\underbrace{\text{Cox, 1961-1962, Atkinson, 1970; etc.}}_{\text{Theoretical solutions}}$, $\underbrace{\text{Bootstrap, } \textbf{next slides.}}_{\text{Practical solutions}}$.

What if there were no nuisance parameters α and β?

- If instead of $f(y,\alpha)$, $g(y,\beta)$, the models to be compared were simply $f(y)$ and $g(y)$, then the Likelihood Ratio is simply the sum of n independent terms.

$$LR = \sum_{i=1}^{B} \log \frac{f(y_i)}{g(y_i)} \tag{9}$$

- Thus, the Central Limit Theorem guarantees that for large B (and a large enough number of counts in each bin).

$$\boxed{LR \approx N(\mu_{LR}, \sigma^2_{LR})} \tag{10}$$

where μ_{LR} and σ^2_{LR} are the mean and the variance of LR respectively, and can be typically be replaced by their sample counterparts.

But what can we do when nuisance parameters α and β are present?

Blennow M. et al., JHEP, 2014 - Method

Consider

$$\chi^2_{NO} = \min_{\alpha} \sum_{i=1}^{B} \left[\frac{x_i - f(y_i,\alpha)}{\sigma_i}\right]^2 \quad \text{and} \quad \chi^2_{IO} = \min_{\beta} \sum_{i=1}^{B} \left[\frac{x_i - g(y_i,\beta)}{\sigma_i}\right]^2$$

for simplicity we assume $\sigma^2_{NO_i} = \sigma^2_{IO_i} = \sigma^2_i$.

The test statistics is

$$\boxed{T = \chi^2_{NO} - \chi^2_{IO} \sim N(T_0; 4T_0) \quad \text{(under } H_0\text{, i.e., NO)}} \tag{11}$$

with $T_0 = \min_\beta \sum_{i=1}^{B} \left[\frac{g(y_i,\beta)-f(y_i,\alpha_0)}{\sigma_i}\right]^2$, and α_0 s.t. $\alpha_{\min} \to \alpha_0$.

This sounds like a very useful result....

Blennow M. et al., JHEP, 2014 - Assumptions

...BUT

To make it work we need:

- The $x_1, \ldots, x_B$ are independent and approximately Gaussian. (If not independent, the variance is no longer as simple as $4T_0$.)
- We need assumptions similar to those required by Wilks' theorem.

PLUS one (or more) of the following:

1. There are no free parameters α and β.
2. OR The hyperplanes of the two hypotheses at their minima are parallel, i.e.,

$$\left.\frac{\partial f(y_i, \alpha)}{\partial \alpha}\right|_{\alpha=\alpha_{\min}} = \left.\frac{\partial g(y_i, \beta)}{\partial \beta}\right|_{\beta_{\min}}.$$

   This is very restrictive!
   E.g.: Let $f(\alpha) = \alpha$, $\alpha \in [0;10]$, $g(\beta) = \beta^2$, $\beta \in [15;20]$ and $\beta_{\min} = 1$

$$1 = \left.\frac{\partial f(\alpha)}{\partial \alpha}\right|_{\forall \alpha \in [0;10]} \neq \left.\frac{\partial g(\beta)}{\partial \beta}\right|_{\beta=1} = 2 \quad \Rightarrow \textbf{NOT OK.}$$

3. OR if T_0 much greater than the number of parameters in the hypothesis.

An new formulation of the problem

Consider a comprehensive model which includes $f(y, \alpha)$ and $g(y, \beta)$ as special cases:

$$\boxed{(1-\eta)f(y,\alpha)+\eta g(y,\beta)} \tag{12}$$

Thus, considering the model in (12) we test

$$H_0 : \eta = 0 \quad \text{versus} \quad H_1 : \eta > 0$$

To exclude intermediate values of η we can interchange the roles of the hypotheses and test

$$H_0 : \eta = 1 \quad \text{versus} \quad H_1 : \eta < 1.$$

From a new formulation to a well known problem

Model:

$$(1-\underbrace{\eta}_{\text{Tested on the boundary}})f(y,\alpha)+\eta g(y,\underbrace{\beta}_{\text{Not defined under } H_0}) \quad \text{with} \quad 0 \le \eta \le 1 \tag{13}$$

Test:

$$H_0 : \eta = 0 \quad \text{versus} \quad H_1 : \eta > 0$$

similar argument for $H_0 : \eta = 1$ versus $H_1 : \eta < 1$

Note!

These are precisely the same issues we encounter when detecting new particles, i.e., when testing one hypothesis multiple times

$\Longrightarrow$ **we already have a solution!**

For more details, check out:

- S. Algeri, J. Conrad and D.A. van Dyk. *A method for comparing non-nested models with application to astrophysical searches for new physics.* MNRAS: Letters, 2016.
- S. Algeri, R package 'NONnest', 2015.

Outline

1. What does it mean exactly to "test one hypothesis multiple times", and in what sense is it equivalent to the "look-elsewhere effect"?
2. Can we tackle both nested and non-nested models with this approach?
3. What is the difference between testing one hypothesis multiple times and multiple hypothesis testing?
4. When do multiple hypothesis testing and testing one hypothesis multiple times coincide in some sense?
5. What else can we do and what is the potential of working in this direction?

Multiple hypothesis testing - Framework

Also in this case:

- We define a grid $\boldsymbol{B}_R$ of R β_r values over the energy spectrum $\boldsymbol{B}$.
- $\forall \beta_r \in \boldsymbol{B}_R$ calculate $LRT(\beta_r)$.

However, now we have:

Many ~~sub-~~alternatives...

We have many alternative hypothesis
$H_{11}, \ldots, H_{1r}, \ldots, H_{1R}$, one for each value $\beta_r \in \boldsymbol{B}_R$.

...many test statistics...

$\forall \beta_r \in \boldsymbol{B}_R$ we have one test statistics $LRT(\beta_r)$, and such that $LRT(\beta_r) \sim \frac{1}{2}\chi^2_1 + \frac{1}{2}\delta(0)$ asymptotically.

...many p-values!

$\forall \beta_r \in \boldsymbol{B}_R$ we have $p_r = \frac{P(\chi^2_1 > LRT(\beta_r))}{2}$.

Local p-values and type I error

- We have an ensemble of R local p-values $p_1, \dots, p_r, \dots, p_R$.
- Their minimum, p_{L}, is compared with the target probability of type I error α_{L}.

Global and local probability of false detection

α_{L} = specific probability of false detection for each of the R tests

$\neq$

α_{G} = probability of having at least one false detection over the whole ensemble of R tests.

$\Rightarrow$ **we must correct p_{L} accordingly:**

- **If the R tests were independent**

$$\alpha_{\mathrm{G}} = 1 - (1 - \alpha_{\mathrm{L}})^R \quad \Rightarrow \quad p_{\mathrm{G}} = 1 - (1 - p_{\mathrm{L}})^R \tag{14}$$

- **If the R tests were dependent** (which is generally the case)

$$\alpha_{\mathrm{G}} \leq R\alpha_{\mathrm{L}} \quad \Rightarrow \quad p_{\mathrm{B}F} = Rp_{\mathrm{L}} \quad \textbf{Bonferroni's correction} \tag{15}$$

Outline

1. What does it mean exactly to "test one hypothesis multiple times", and in what sense is it equivalent to the "look-elsewhere effect"?
2. Can we tackle both nested and non-nested models with this approach?
3. What is the difference between testing one hypothesis multiple times and multiple hypothesis testing?
4. When do multiple hypothesis testing and testing one hypothesis multiple times coincide in some sense?
5. What else can we do and what is the potential of working in this direction?

Two approaches, same inference

- Let $l(\eta|\alpha, \beta_r, y)$ be the log-likelihood of the model of reference.
- The respective score function evaluated at H_0 is $S(\beta_r) = \frac{\partial l(\eta|\alpha,\beta_r,y)}{\partial \eta}\Big|_{\eta=0}$
- Take the normalized version of it, i.e.,

$$S^{\star}(\beta_r) = \frac{S(\beta_r)}{\sqrt{cov(S(\beta_r),S(\beta_r))}} \tag{16}$$

A sufficient condition on $S^{\star}(\beta_r)$ (Berman's condition)

If the covariance function of $S^{\star}(\beta_r)$ satisfies

$$\sup_{|\beta_r-\beta_t|>\tau} |cov(S^{\star}(\beta_r), S^{\star}(\beta_t))| \log(\tau) \to 0 \quad \text{as} \quad \tau \to +\infty, \tag{17}$$

THEN TESTING ONE HYPOTHESIS MULTIPLE TIMES AND MULTIPLE HYPOTHESES TESTING ARE (APPROXIMATELY) EQUIVALENT!

This result will be discussed in:

S. Algeri, D.A. van Dyk and J. Conrad. *Testing one hypothesis multiple times.* In preparation, 2016.

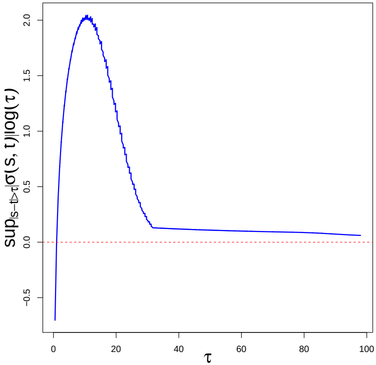
(Hopefully, available on ArXiv by the end of the summer.)

Example

Consider a power-law distributed background with index τ and a Gaussian signal with dispersion proportional to the signal location.
The full model is

$$\frac{1}{k_\tau y^{\tau+1}} + \frac{\mu}{k_{M_\chi}} \exp\left\{ -\frac{(y - M_\chi)^2}{0.02 M_\chi^2} \right\} \tag{18}$$

with k_τ and k_{M_χ} normalizing constants, $y \in [1; 100]$, $\tau = 1.4$ and $M_\chi \in [1; 100]$.



Realistic data analysis

We simulated observation of monochromatic feature by the Fermi Large Area Telescope (LAT).

- 2391 events from an astrophysical background corresponding to isotropic emission following a spectral power-law with index 2.4, i.e., $\tau = 1.4$.
- 64 events from a Gaussian signal with mass of 35 GeV.
- 80 energy bins, spaced equally from 10-350 GeV.

Method	Signal Location	Signal Strength	Sig.
Unadjusted local	35.82	0.042	5.920σ
Bonferroni adj. local	35.82	0.042	5.152σ
Gross & Vitells	35.82	0.042	5.192σ

Outline

1. What does it mean exactly to "test one hypothesis multiple times", and in what sense is it equivalent to the "look-elsewhere effect"?
2. Can we tackle both nested and non-nested models with this approach?
3. What is the difference between testing one hypothesis multiple times and multiple hypothesis testing?
4. When do multiple hypothesis testing and testing one hypothesis multiple times coincide in some sense?
5. What else can we do and what is the potential of working in this direction?

What can we do more?

- Berman's condition can be used as diagnostic tool to assess the validity of the Davies (1987) and Gross and Vitells (2010) approximations for the global p-value $P(\sup LRT(\beta) > c)$.
  - $\Rightarrow$ We would like to provide a formal explanation of cases where the global p-value approximations do and do not work.
  - $\Rightarrow$ We would like to provide precise indications on how to spot these cases.
- Further:
  - $\Rightarrow$ We would like to exploit the information on the dependence structure of the underlying LRT processes to improve, if possible, the global p-value approximations discussed in this talk.
  - $\Rightarrow$ We would like to provide guidelines for the choice of the resolution R.

Thank you for listening!

References

- S. Algeri et al. *Looking for a Needle in a Haystack? Look Elsewhere! A statistical comparison of approximate global p-values.* Submitted, 2016.
- S. Algeri, J. Conrad and D.A. van Dyk. *A method for comparing non-nested models with application to astrophysical searches for new physics.* In: MNRAS: Letters, 2016.
- S. Algeri, *R package 'NONnest'* (2015). Download: `http://wwwf.imperial.ac.uk/~sa2514/Research.html`.
- S. Algeri, D.A. van Dyk and J. Conrad. *Testing one hypothesis multiple times.* In preparation, 2016.
- A. C. Atkinson. *"A Method For Discriminating Between Models"*. In: Journal of the Royal Statistical Society. Series B (Methodological) 32.3 (1970).
- M. Blennow et al. *Quantifying the sensitivity of oscillation experiments to the neutrino mass ordering*. In: Journal of High Energy Physics 03 (2014).
- H. Chernoff. *"On the Distribution of the Likelihood Ratio"*. In: The Annals of Mathematical Statistics 25.3 (1954).
- D. R. Cox. *"Tests of Separate Families of Hypotheses"*. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. (1961).
- D. R. Cox. *"Further Results on Tests of Separate Families of Hypotheses"*. In: Journal of the Royal Statistical Society. Series B (Methodological) 24.2 (1962).
- R. B. Davies. *"Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternatives"*. In: Biometrika 74.1 (1987).
- R. B. Davies. *"Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternatives"*. In: Biometrika 64.2 (1977).
- B. Efron. *Large-Scale Inference.* Cambridge Books Online. Cambridge University Press, 2010.
- E. Gross and O. Vitells. *"Trial factors for the look elsewhere effect in high energy physics"*. In: The European Physical Journal C 70.1-2 (2010), pp. 525-530.
- M.R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes.* Springer Series in Statistics. Springer-Verlag New York Inc., 1983.