LECTURES ON

STATISTICS

@KyleCranmer
New York University
Department of Physics
Center for Data Science
Modeling:
The Scientific Narrative
CHOICE: DATA DRIVEN VS. SIMULATION

In the case of the CDF bump, the Z+jets control sample provides a data-driven estimate, but limited statistics. Using the simulation narrative over the data-driven is a choice. If you trust that narrative, it’s a good choice.

![Feynman diagrams for single top production.](image)

![Example of one of the numerous diagrams for the production of +jets.](image)

![Graph showing the dijet mass distributions.](image)
THE DATA-DRIVEN NARRATIVE

Regions in the data with negligible signal expected used as control samples
- simulated events are used to estimate extrapolation coefficients
- extrapolation coefficients may have theoretical and experimental uncertainties
WHAT DO WE MEAN BY UNCERTAINTY?

Let’s consider a simplified problem that has been studied quite a bit to gain some insight into our more realistic and difficult problems

- number counting with background uncertainty
  - in our main measurement we observe $n_{on}$ with $s + b$ expected

\[ \text{Pois}(n_{on} | s + b) \]

- and the background has some uncertainty
  - but what is “background uncertainty”? Where did it come from?
  - maybe we would say background is known to 10% or that it has some pdf $\pi(b)$
    - then we often do a smearing of the background:
    \[ P(n_{on} | s) = \int db \text{Pois}(n_{on} | s + b) \pi(b), \]
  - Where does $\pi(b)$ come from?
    - did you realize that this is a Bayesian procedure that depends on some prior assumption about what $b$ is?
THE “ON/OFF” PROBLEM

Now let’s say that the background was estimated from some control region or sideband measurement.

- We can treat these two measurements simultaneously:
  - main measurement: observe \( n_{on} \) with \( s+b \) expected
  - sideband measurement: observe \( n_{off} \) with \( \tau b \) expected

\[
P(n_{on}, n_{off} | s, b) = \text{Pois}(n_{on} | s + b) \text{Pois}(n_{off} | \tau b)
\]

- In this approach “background uncertainty” is a statistical error
- justification and accounting of background uncertainty is much more clear

How does this relate to the smearing approach?

\[
P(n_{on} | s) = \int db \text{Pois}(n_{on} | s + b) \pi(b),
\]

- while \( \pi(b) \) is based on data, it still depends on some original prior \( \eta(b) \)

\[
\pi(b) = P(b | n_{off}) = \frac{P(n_{off} | b)\eta(b)}{\int db P(n_{off} | b)\eta(b)}.
\]
A GENERAL PURPOSE STATISTICAL MODEL
I will represent PDFs graphically as below (directed acyclic graph)

- eg. a Gaussian $G(x|\mu, \sigma)$ is parametrized by $(\mu, \sigma)$
- every node is a real-valued function of the nodes below
RooFit: A data modeling toolkit

RooFit is a major tool developed at BaBar for data modeling. RooStats provides higher-level statistical tools based on these PDFs.

- Addition
- Composition (‘plug & play’)
- Multiplication
- Convolution

Possible in any PDF
No explicit support in PDF code needed

Wouter Verkerke, UCSB
**Marked Poisson Process**

**Channel:** a subset of the data defined by some selection requirements.
- eg. all events with 4 electrons with energy > 10 GeV
- $n$: number of events observed in the channel
- $\nu$: number of events expected in the channel

**Discriminating variable:** a property of those events that can be measured and which helps discriminate the signal from background
- eg. the invariant mass of two particles
- $f(x)$: the p.d.f. of the discriminating variable $x$

$$\mathcal{D} = \{x_1, \ldots, x_n\}$$

**Marked Poisson Process / Extended Likelihood:**

$$f(\mathcal{D}|\nu) = \text{Pois}(n|\nu) \prod_{e=1}^{n} f(x_e)$$
**Mixture Model**

**Sample:** a sample of simulated events corresponding to particular type interaction that populates the channel.

- statisticians call this a mixture model

\[
f(x) = \frac{1}{\nu_{\text{tot}}} \sum_{s \in \text{samples}} \nu_s f_s(x), \quad \nu_{\text{tot}} = \sum_{s \in \text{samples}} \nu_s
\]

---

**ATLAS**

H → eeνν (m_H = 400 GeV)

\[\int L \, dt = 35 \text{ pb}^{-1}\]

\[\sqrt{s} = 7 \text{ TeV}\]

![Graph showing data, Z+Jets, top, Diboson, W+jets, Multijet, and Signal (m_H=400 GeV)](image-url)
PARAMETRIZING THE MODEL \( \alpha = (\mu, \theta) \)

Parameters of interest \( (\mu) \): parameters of the theory that modify the rates and shapes of the distributions, eg.
- the mass of a hypothesized particle
- the “signal strength” \( \mu=0 \) no signal, \( \mu=1 \) predicted signal rate

Nuisance parameters \( (\theta \text{ or } \alpha_p) \): associated to uncertainty in:
- response of the detector (calibration)
- phenomenological model of interaction in non-perturbative regime

Lead to a parametrized model:

\[
\nu \rightarrow \nu(\alpha), \quad f(x) \rightarrow f(x|\alpha)
\]

\[
f(D|\alpha) = \text{Pois}(n|\nu(\alpha)) \prod_{e=1}^{n} f(x_e|\alpha)
\]
Tabulate effect of individual variations of sources of systematic uncertainty

- typically one at a time evaluated at nominal and “± 1 σ”
- use some form of interpolation to parametrize $p^{th}$ variation in terms of nuisance parameter $\alpha_p$

$f(D|\alpha) = \text{Pois}(n|\nu(\alpha)) \prod_{e=1}^{n} f(x_e|\alpha)$
Visualizing the model for one channel
After parametrizing each component of the mixture model, the pdf for a single channel might look like this.
**Simultaneous Multi-Channel Model**

**Simultaneous Multi-Channel Model:** Several disjoint regions of the data are modeled simultaneously. Identification of common parameters across many channels requires coordination between groups such that meaning of the parameters are really the same.

\[
f_{\text{sim}}(D_{\text{sim}}|\alpha) = \prod_{c \in \text{channels}} \left[ \text{Pois}(n_c|\nu_c(\alpha)) \prod_{e=1}^{n_c} f_c(x_{ce}|\alpha) \right]
\]

where \( D_{\text{sim}} = \{D_1, \ldots, D_{c_{\text{max}}} \} \)

**Control Regions:** Some channels are not populated by signal processes, but are used to constrain the nuisance parameters

- attempt to describe systematics in a statistical language
- Prototypical Example: “on/off” problem with unknown \( \nu_b \)

\[
f(n, m|\mu, \nu_b) = \text{Pois}(n|\mu + \nu_b) \cdot \text{Pois}(m|\tau \nu_b)
\]

\( \text{signal region} \quad \text{control region} \)}
Often detailed statistical model for auxiliary measurements that measure certain nuisance parameters are not available.

- one typically has MLE for $\alpha_p$, denoted $a_p$ and standard error

**Constraint Terms:** are idealized pdfs for the MLE.

$$f_p(a_p | \alpha_p) \quad \text{for} \quad p \in \mathbb{S}$$

- common choices are Gaussian, Poisson, and log-normal
- New: careful to write constraint term a frequentist way
- Previously: $\pi(\alpha_p | a_p) = f_p(a_p | \alpha_p) \eta(\alpha_p)$ with uniform $\eta$

**Simultaneous Multi-Channel Model with constraints:**

$$f_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \alpha) = \prod_{c \in \text{channels}} \left[ \text{Pois}(n_c | \nu_c(\alpha)) \prod_{e=1}^{n_c} f_c(x_{ce} | \alpha) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p | \alpha_p)$$

where

$$\mathcal{D}_{\text{sim}} = \{ \mathcal{D}_1, \ldots, \mathcal{D}_{c_{\text{max}}} \}, \quad \mathcal{G} = \{ a_p \} \quad \text{for} \quad p \in \mathbb{S}$$
2.1 Index Convention

2.2 Auxiliary measurements

We will use the following mnemonic index conventions:

- \( e \in \) events
- \( b \in \) bins
- \( c \in \) channels
- \( s \in \) samples
- \( p \in \) parameters
EXAMPLE OF DIGITAL PUBLISHING

RooFit’s Workspace now provides the ability to save in a ROOT file the full likelihood model, any priors you might want, and the minimal data necessary to reproduce likelihood function.

Need this for combinations, as p-value is not sufficient information for a proper combination.
**HistFactory**

32 page documentation of HistFactory tool + manual

- currently a “living document”

http://cds.cern.ch/record/1456844
**State of the art:** At the time of the discovery, the combined Higgs search included 100 disjoint channels and >500 nuisance parameters

- Models for individual channels come from about 11 sub-groups performing dedicated searches for specific Higgs decay modes
- In addition low-level performance groups provide tools for evaluating systematic effects and corresponding constraint terms

<table>
<thead>
<tr>
<th>Higgs Decay</th>
<th>Subsequent Decay</th>
<th>Additional Sub-Channels</th>
<th>$m_H$ Range</th>
<th>L [fb$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H \rightarrow \gamma\gamma$</td>
<td>–</td>
<td>9 sub-channels ($p_T, \eta, \gamma$ conversion)</td>
<td>110-150</td>
<td>4.9</td>
</tr>
<tr>
<td>$H \rightarrow ZZ$</td>
<td>$\ell\ell\ell'$</td>
<td>${4e, 2e2\mu, 2\mu2e, 4\mu}$</td>
<td>110-600</td>
<td>4.8</td>
</tr>
<tr>
<td>$H \rightarrow ZZ$</td>
<td>$\ell\ell\nu\nu$</td>
<td>${ee, \mu\mu} \otimes {\text{low pile-up, high pile-up}}$</td>
<td>200-280-600</td>
<td>4.7</td>
</tr>
<tr>
<td>$H \rightarrow ZZ$</td>
<td>$\ell\ell qq$</td>
<td>${b\text{-tagged, untagged}}$</td>
<td>200-300-600</td>
<td>4.7</td>
</tr>
<tr>
<td>$H \rightarrow WW$</td>
<td>$\ell\nu\ell\nu$</td>
<td>${ee, e\mu, \mu\mu} \otimes {0\text{-jet, 1-jet, VBF}}$</td>
<td>110-300-600</td>
<td>4.7</td>
</tr>
<tr>
<td>$H \rightarrow WW$</td>
<td>$\ell\nu qq$</td>
<td>${e, \mu} \otimes {0\text{-jet, 1-jet}}$</td>
<td>300-600</td>
<td>4.7</td>
</tr>
<tr>
<td>$H \rightarrow \tau^+\tau^-$</td>
<td>$\ell\ell 4\nu$</td>
<td>${e\mu} \otimes {0\text{-jet}} \oplus {1\text{-jet, VBF, VH}}$</td>
<td>110-150</td>
<td>4.7</td>
</tr>
<tr>
<td>$H \rightarrow \tau^+\tau^-$</td>
<td>$\ell\tau_{\text{had}} 3\nu$</td>
<td>${e, \mu} \otimes {0\text{-jet}} \otimes {E_{T}^{\text{miss}} \geq 20 \text{ GeV}}$</td>
<td>110-150</td>
<td>4.7</td>
</tr>
<tr>
<td>$H \rightarrow \tau^+\tau^-$</td>
<td>$\ell\tau_{\text{had}} 2\nu$</td>
<td>${1\text{-jet}}$</td>
<td>110-150</td>
<td>4.7</td>
</tr>
</tbody>
</table>

| VH $\rightarrow b\bar{b}$ | $Z \rightarrow \nu\bar{\nu}$ | $E_{T}^{\text{miss}} \in \{120 - 160, 160 - 200, \geq 200 \text{ GeV}\}$ | 110-130 | 4.6 |
| VH $\rightarrow b\bar{b}$ | $W \rightarrow \ell\nu$ | $p_T^W \in \{< 50, 50 - 100, 100 - 200, \geq 200 \text{ GeV}\}$ | 110-130 | 4.7 |
| VH $\rightarrow b\bar{b}$ | $Z \rightarrow \ell\ell$ | $p_T^Z \in \{< 50, 50 - 100, 100 - 200, \geq 200 \text{ GeV}\}$ | 110-130 | 4.7 |
VISUALIZING THE COMBINED MODEL

**State of the art:** At the time of the discovery, the combined Higgs search included 100 disjoint channels and >500 nuisance parameters.

**RooFit / RooStats:** is the modeling language (C++) which provides technologies for collaborative modeling:

- provides technology to publish likelihood functions digitally
- and more, it’s the full model so we can also generate pseudo-data

\[
f_{\text{tot}}(D_{\text{sim}}, G | \alpha) = \prod_{c \in \text{channels}} \left[ \text{Pois}(n_c | \nu_c(\alpha)) \prod_{e=1}^{n_c} f_c(x_{ce} | \alpha) \right] \cdot \prod_{p \in \mathcal{S}} f_p(a_p | \alpha_p)
\]
**Evolution of Model Complexity**

- **Number of Datasets Combined**
- **Number of Model Components**
- **Number of Parameters in Likelihood**

![Graphs showing the evolution of model complexity](image-url)
HYPOTHESIS TESTING
HYPOTHESIS TESTING

One of the most common uses of statistics in particle physics is Hypothesis Testing (e.g. for discovery of a new particle)

› assume one has pdf for data under two hypotheses:
  • Null-Hypothesis, \( H_0 \): eg. background-only
  • Alternate-Hypothesis \( H_1 \): eg. signal-plus-background

› one makes a measurement and then needs to decide whether to reject or accept \( H_0 \)
Before we can make much progress with statistics, we need to decide what it is that we want to do.

- first let us define a few terms:
  - Rate of Type I error \( \alpha \)
  - Rate of Type II \( \beta \)
  - Power = \( 1 - \beta \)

Treat the two hypotheses asymmetrically
- the Null is special.
  - Fix rate of Type I error, call it “the size of the test”

Now one can state “a well-defined goal”
- Maximize power for a fixed rate of Type I error
The idea of a “5σ” discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
- eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy

[G. Cowan]
THE NEYMAN-PEARSON LEMMA

In 1928-1938 Neyman & Pearson developed a theory in which one must consider competing Hypotheses:
- the Null Hypothesis $H_0$ (background only)
- the Alternate Hypothesis $H_1$ (signal-plus-background)

Given some probability that we wrongly reject the Null Hypothesis

$$\alpha = P(x \notin W|H_0)$$

(Convention: if data falls in W then we accept $H_0$)

Find the region $W$ such that we minimize the probability of wrongly accepting the $H_0$ (when $H_1$ is true)

$$\beta = P(x \in W|H_1)$$
The Neyman-Pearson Lemma

The region $W$ that minimizes the probability of wrongly accepting $H_0$ is just a contour of the Likelihood Ratio

$$\frac{P(x|H_1)}{P(x|H_0)} > k\alpha$$

Any other region of the same size will have less power

The likelihood ratio is an example of a Test Statistic, eg. a real-valued function that summarizes the data in a way relevant to the hypotheses that are being tested.
Consider the contour of the likelihood ratio that has size a given size (eg. probability under $H_0$ is $1-\alpha$)
Now consider a variation on the contour that has the same size
Now consider a variation on the contour that has the same size (e.g., same probability under $H_0$)
A SHORT PROOF OF NEYMAN-PEARSON

Because the new area is outside the contour of the likelihood ratio, we have an inequality

\[
\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha
\]

\[
P(H_1) < P(H_0)k_\alpha
\]
A short proof of Neyman-Pearson

\[
\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha
\]

\[
P(\bigcap | H_0) < P(\bigcup | H_0)k_\alpha
\]

\[
\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha
\]

\[
P(\bigcup | H_1) > P(\bigcup | H_0)k_\alpha
\]

And for the region we lost, we also have an inequality

Together they give...
A short proof of Neyman-Pearson

The new region has less power.

\[
\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha \quad \Rightarrow \quad P(H_1|x) < P(H_0|x) k_\alpha
\]

\[
\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha \quad \Rightarrow \quad P(H_1|x) > P(H_0|x) k_\alpha
\]

\[
P(\cup | H_1) < P(\cup | H_0) k_\alpha
\]

\[
P(\cup | H_1) > P(\cup | H_0) k_\alpha
\]
2 DISCRIMINATING VARIABLES

Often one uses the output of a neural network or multivariate algorithm in place of a true likelihood ratio.

› That’s fine, but what do you do with it?
› If you have a fixed cut for all events, this is what you are doing:

\[
L_{tot} = L_1 \cdot L_2
\]

\[
q_{12} = \ln L_{12} = \ln L_1 + \ln L_2 = q_1 + q_2
\]
Experiments vs. Events

Ideally, you want to cut on the likelihood ratio for your experiment

- equivalent to a sum of log likelihood ratios

Easy to see that includes experiments where one event had a high likelihood and the other one was relatively small.
AN OPTIMAL WAY TO COMBINE

Special case of our general probability model
(no nuisance parameters)

\[
Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_{i}^{N_{\text{chan}}} \text{Pois}(n_i|s_i + b_i) \prod_{j}^{n_i} \frac{s_if_s(x_{ij})+b_if_b(x_{ij})}{s_i+b_i}}{\prod_{i}^{N_{\text{chan}}} \text{Pois}(n_i|b_i) \prod_{j}^{n_i} f_b(x_{ij})}
\]

\[
\ln Q = -s_{\text{tot}} + \sum_{i} \sum_{j} \ln \left(1 + \frac{s_if_s(x_{ij})}{b_if_b(x_{ij})}\right)
\]

Instead of simply counting events, the optimal test statistic is equivalent to adding events weighted by

\[
\ln(1+\text{signal/background ratio})
\]

The test statistic is a map \( T: \text{data} \rightarrow \mathbb{R} \)

By repeating the experiment many times, you obtain a distribution for \( T \)
P-VALUES

Instead of choosing to accept/reject $H_0$ one can compute the p-value

$$p = \int_{T_0}^{\infty} f(T|H_0)$$

If the model for the data depends on parameters $\alpha$ the p-value also depends on $\alpha$.

$$p(\alpha) = \int_{T_0}^{\infty} f(T|\alpha) dT = \int f(D|\alpha) \theta(T(D) - T_0) dD = P(T \geq T_0|\alpha)$$
P-VALUES

When the model has nuisance parameters, only reject the null if $p(\alpha)$ sufficiently small for all values of the nuisance parameters.

If the model for the data depends on parameters $\alpha$, the p-value also depends on $\alpha$. 

$$p(\alpha) = \int_{T_0}^{\infty} f(T|\alpha) dT = \int f(D|\alpha) \theta(T(D) - T_0) dD = P(T \geq T_0|\alpha)$$
THE PROFILE LIKELIHOOD RATIO

Consider our general model with a single parameter of interest $\mu$

- let $\mu=0$ be no signal, $\mu=1$ nominal signal

In the LEP approach the likelihood ratio is equivalent to:

$$ Q_{\text{LEP}} = \frac{L(\mu = 1, \theta)}{L(\mu = 0, \theta)} = \frac{f(D|\mu = 1, \theta)}{f(D|\mu = 0, \theta)} $$

- but this variable is sensitive to uncertainty on $\theta$ and makes no use of auxiliary measurements $a$

Alternatively, one can define profile likelihood ratio

$$ \lambda(\mu) = \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(D, G|\mu, \hat{\theta}(\mu; D, G))}{f(D, G|\hat{\mu}, \hat{\theta})} $$

- where $\hat{\theta}(\mu; D, G)$ is best fit with $\mu$ fixed (the constrained maximum likelihood estimator, depends on data)
- and $\hat{\theta}$ and $\hat{\mu}$ are best fit with both left floating (unconstrained)
- Tevatron used $Q_{\text{Tev}} = \lambda(\mu=1)/\lambda(\mu=0)$ as generalization of $Q_{\text{LEP}}$
AN EXAMPLE

Essentially, you need to fit your model to the data twice: once with everything floating, and once with signal fixed to 0

\[
\lambda(\mu = 0) = \frac{L(\mu = 0, \hat{\theta}(\mu = 0))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(D, G|\mu = 0, \hat{\theta}(\mu = 0; D, G))}{f(D, G|\hat{\mu}, \hat{\theta})}
\]

Figure 14 shows the fit to the signal candidates for a \(m_H=120\) GeV Higg with (a,c) and without (b,d) the signal contribution. It can be seen that the background shape and normalizations are trying to accommodate the excess near \(m_{\tau\tau} = 120\) GeV, but the control samples are constraining the variation.

Table 13 shows the significance calculated from the profile likelihood ratio for the \(ll\)-channel, the \(lh\)-channel, and the combined fit for various Higgs boson masses.
After a close look at the profile likelihood ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(D, G|\mu, \hat{\theta}; D, G)}{f(D, G|\hat{\mu}, \hat{\theta})}$$

one can see the function is independent of true values of $\theta$

› though its distribution might depend indirectly

Wilks’s theorem states that under certain conditions the distribution of

$-2 \ln \lambda(\mu=\mu_0)$ given that the true value of $\mu$ is $\mu_0$ converges to a chi-square distribution

› more on this later, but the important points are:

› “asymptotic distribution” is known and it is independent of $\theta$!

- more complicated if parameters have boundaries (eg. $\mu \geq 0$)

Thus, we can calculate the p-value for the background-only hypothesis without having to generate Toy Monte Carlo!
TOY MONTE CARLO

Explicitly build distribution by generating “toys” / pseudo experiments assuming a specific value of $\mu$ and $\nu$.

- randomize both main measurements $\mathcal{D}=$\{X\} and auxiliary measurements $\mathcal{G}=$\{a\}
- fit the model twice for the numerator and denominator of profile likelihood ratio
- evaluate $-2\ln \lambda(\mu)$ and add to histogram

Choice of $\mu$ is straightforward: typically $\mu=0$ and $\mu=1$, but choice of $\theta$ is less clear

- more on this later

This can be very time consuming. Plots below use millions of “toy” pseudo-experiments
“THE ASIMOV PAPER”

Recently we showed how to generalize this asymptotic approach

- generalize Wilks’s theorem when boundaries are present
- use Wald’s result for distribution for alternate hypothesis $f(-2 \log \lambda(\mu) | \mu')$

Asymptotic formulae for likelihood-based tests of new physics
Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells


http://arxiv.org/abs/1007.1727v2
This is a significant development as building this distribution from Monte Carlo approaches can take 100,000 CPU hours for Higgs search!

G. Cowan, KC, E. Gross, O. Vitells
[arXiv:1007.1727]
So far this looks a bit like magic. How can you claim that you incorporated your systematic just by fitting the best value of your uncertain parameters and making a ratio?

It won’t unless the parametrization is sufficiently flexible.

So check by varying the settings of your simulation, and see if the profile likelihood ratio is still distributed as a chi-square.

Here it is pretty stable, but it’s not perfect (and this is a log plot, so it hides some pretty big discrepancies).

For the distribution to be independent of the nuisance parameters your parametrization must be sufficiently flexible.
A VERY IMPORTANT POINT

If we keep pushing this point to the extreme, the physics problem goes beyond what we can handle practically.

The p-values are usually predicated on the assumption that the true distribution is in the family of distributions being considered:

- eg. we have sufficiently flexible models of signal & background to incorporate all systematic effects
- but we don’t believe we simulate everything perfectly
- ..and when we parametrize our models usually we have further approximated our simulation.

- nature -> simulation -> parametrization

At some point these approaches are limited by honest systematics uncertainties (not statistical ones). Statistics can only help us so much after this point. Now we must be physicists!
**LOOK-ELSEWHERE EFFECT**

Approximation best above $3\sigma$

Typically our signal model has some parameter (e.g. $m_H$), which does not affect the null (background only).

This modifies the distribution of the likelihood ratio test statistic we call this the “look-elsewhere effect”

Recently Gross & Vitells found the results of Rice, Davies, and Leadbetter for a fast asymptotic approximation for the global $p$-value

\[
{p_0}^{\text{global}} \approx {p_0}^{\text{local}} + \left\langle N(q_{ref}) \right\rangle e^{-\frac{1}{2}(q_{test} - q_{ref})}
\]


R. B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, Biometrika 64 (1977); Biometrika 74 (1987).
Even if we fix the location of the signal some systematic effects are equivalent to small uncertainty in the location (e.g. energy calibration).

**Figure 1.** An example pseudo-experiment with background only. The trial factor estimated from toy Monte Carlo simulations is q0=6.6, corresponding to 3.5σ significance.

**Figure 3.** A comparison of the profile likelihood ratio at 126 GeV with the bound calculated from 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.

**Figure 4.** The resulting distributions and trial factors for the look-elsewhere effect. The solid line shows the best signal fit, while the dotted line shows the background fit. (bottom) The trial factor for the look-elsewhere effect is q0=6.3, corresponding to 3.6σ significance.

**Figure 5.** The profile likelihood ratio for the look-elsewhere effect at 126 GeV compared to the bound calculated from 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.
A More Subtle Effect

Even if we fix the location of the signal, some systematic effects are equivalent to small uncertainty in the location (e.g. energy calibration).

These parameters are slowing convergence to the asymptotic distribution and variance may not reduce with more data.

O. Vitells found exact solution by Leadbetter for the case of only one such nuisance parameter.

\[ E[N_u] = \sigma_2 \int \phi(u(M)) \left[ \phi \left( \frac{u'(M)}{\sigma_2} \right) + \frac{u'(M)}{\sigma_2} \left\{ \Phi \left( \frac{u'(M)}{\sigma_2} \right) - \frac{1}{2} \right\} \right] dM \]

(H.R. Leadbetter, 1965)

note: used in Higgs discovery papers