High-order Asymptotic Expansions for Likelihood-based Statistics With Application to Testing for Signal Presence in Particle Physics Experiments

Basitha Hewa[*] & Igor Volobouev[†] & Alex Trindade[*]

(†) *Dept. of Physics & Astronomy, Texas Tech University*

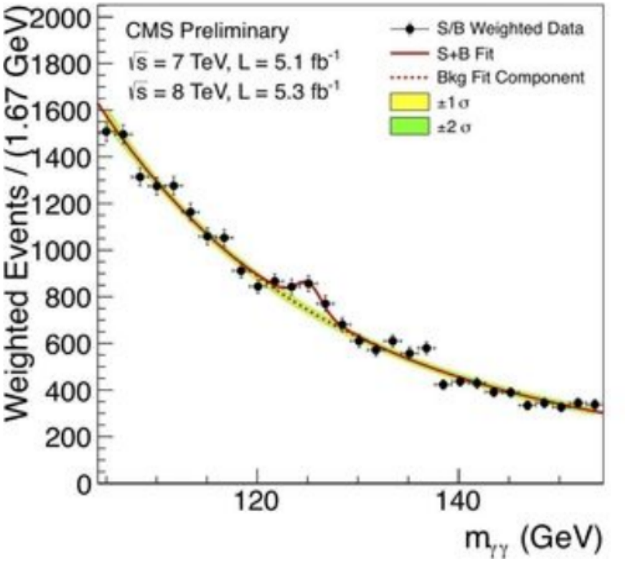
(∗) *Dept. of Mathematics & Statistics, Texas Tech University*

September 2020

Outline (new stuff in "blue" highlight)

- **Classical Setup:** Test a one-sided hypothesis for a single parameter via Likelihood Ratio, Score, and Wald tests.
- Goal: Accurate p-values.
- **Classical Solution:** Use asymptotics or simulation!
- **Non-Classical Setup:** Required Type I error rate is $\alpha \sim 10^{-7}$.
- Solution 1 (asymptotic): Derive high-order Edgeworth approximations to p-values.
- Solution 2 (asymptotic): Derive high-order saddlepoint approximations to p-values (with new twists).
- Compare accuracies on simulated data (bump-hunting expmts).
- Practical implementation: going beyond the toy problem. . .

Motivation: Discovery of Higgs Boson (The God Particle, Nobel Prize 2013)



- Model 1: Overall density is a mixture of signal $s(x)$ and background $b(x)$ densities:
$$p(x|\alpha) = \alpha s(x) + (1-\alpha)b(x) \tag{1}$$
- Signal fraction α based on IID sample $x_1, \ldots, x_n$ is estimated by maximizing the log-likelihood
$$\ell(\alpha) = \sum_{i=1}^{n} \log p(x_i|\alpha). \tag{2}$$
- Leading to the MLE
$$\hat{\alpha} := \arg\max_{\alpha \in \mathbb{R}} \ell(\alpha) \tag{3}$$
- Goal: produce accurate tests of $\mathcal{H}_0 : \alpha = 0$ vs. $\mathcal{H}_1 : \alpha > 0$.
- Only unknown parameter is $\alpha \in \mathbb{R}$.

- Model 2: Sample size is not *a priori* known, so treat data $x_1, \ldots, x_N$ as arising from a Poisson process with intensity function:

$$\Lambda(x|\lambda) = \lambda s(x) + \mu b(x) \tag{4}$$

- Signal fraction λ is estimated by maximizing the log-likelihood

$$\ell(\lambda) = -(\lambda + \mu) + \sum_{i=1}^{N} \log \Lambda(x_i|\lambda) \tag{5}$$

- Leading to the MLE

$$\hat{\lambda} := \arg\max_{\lambda \in \mathbb{R}} \ell(\lambda) \tag{6}$$

- Goal: produce accurate tests of $\mathcal{H}_0 : \lambda = 0$ vs. $\mathcal{H}_1 : \lambda > 0$.
- Only unknown parameter is $\lambda \in \mathbb{R}$.

- Background is either standard Uniform or Exponential on $[0,1]$:

$$b(x) = \begin{cases} 1, & \text{if } x \in [0,1] \\ 0, & \text{if } x \notin [0,1] \end{cases}, \quad b(x) = \begin{cases} e^{-x}/(1-e^{-1}), & \text{if } x \in [0,1] \\ 0, & \text{if } x \notin [0,1] \end{cases}$$

- Signal is truncated Gaussian on $[0,1]$:

$$s(x) = \begin{cases} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Big/ \int_0^1 e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy, & \text{if } x \in [0,1] \\ 0, & \text{if } x \notin [0,1] \end{cases}$$

- Whenever specific settings of the signal are needed, we use

$$\mu = 0.5, \qquad \text{and} \qquad \sigma = 0.1.$$

- Toy problem! Everything known except mix proportion (α or λ)...

For $\boldsymbol{\alpha}$ and $\hat{\boldsymbol{\alpha}}$ (similar statements hold for λ and $\hat{\lambda}$ with $n \mapsto \mu$):

- $\ell_i(\alpha) = \partial^i \ell / \partial \alpha^i$, the i-th derivative of $\ell(\alpha)$
- $J(\alpha) = -\ell_2(\alpha)$
- *Expected information* number: $I(\alpha) = \mathbb{E}[J(\alpha)]$
- *Observed information* number: $J(\hat{\alpha}) = -\ell_2(\hat{\alpha})$

Assume usual regularity conditions for consistency and asymptotic normality of $\hat{\alpha}$ are satisfied:

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\alpha)^{-1}), \qquad \mathcal{I}(\alpha) = \lim_{n \to \infty} \frac{1}{n} I(\alpha)$$

$$\Longrightarrow \hat{\alpha} \mathrel{\dot\sim} \mathcal{N}(\alpha, \sigma^2_{\hat{\alpha}}(\alpha)), \qquad \sigma^2_{\hat{\alpha}}(\alpha) = I(\alpha)^{-1}$$

(By not restricting $\alpha \in [0, 1]$ we avoid "exotic" asymptotics at the boundaries...)

In lack of a UMP test, we have the following:

Table: Promising statistics for tests on α.

Method	Statistic	Value
Likelihood Ratio	T_{LR}	$2[\ell(\hat{\alpha}) - \ell(0)]$
Wald (Expected)	T_{W}	$\hat{\alpha}^2 I(0)$
Wald (Observed)	T_{W2}	$\hat{\alpha}^2 J(\hat{\alpha})$
Score	T_{S}	$\ell_1(0)^2 / I(0)$
Wald-type 3	T_{W3}	$\hat{\alpha}^2 / \sigma_3^2$
Wald-type 4	T_{W4}	$\hat{\alpha}^2 / \sigma_4^2$

The Wald-type 3 & 4 statistics are variants of T_{W2} (used by physicists) that use shortcuts for computing $J(\alpha)$ so as to avoid differentiating $\ell(\alpha)$.

- For one-sided testing use *signed* version of any of the statistics (say T) in the Table:
$$R = \text{sgn}(\hat{\alpha})\sqrt{T}.$$
- Under $\mathcal{H}_0$, to first order $R \sim Z$, where $Z \sim \mathcal{N}(0, 1)$, whence
$$p\text{-value} \ = P(Z > r), \qquad r = \text{sgn}(\hat{\alpha})\sqrt{t}$$
- In general, $R_n \sim Z$ to k-th order, means that
$$\text{approx error} = R_n - Z = O_p(n^{-k/2})$$

$$\Rightarrow P(R_n \leq r) = \Phi(r) + \frac{a_{1,n}}{n^{1/2}} + \frac{a_{2,n}}{n^1} + \frac{a_{3,n}}{n^{3/2}} + \cdots + \frac{a_{k-1,n}}{n^{(k-1)/2}} + O(n^{-k/2})$$

- Taylor expansions of $\ell(\alpha)$ near true value of α
- Joint cumulants for the derivatives of $\ell(\alpha)$ under $\mathcal{H}_0$; in our case can express everything as a function of:

$$V_k = \mathbb{E}\ell_k(0), \qquad k = 1, 2, \ldots$$

- Edgeworth-type series: approx pdf for $X \approx \mathcal{N}(\kappa_1, \kappa_2)$ via the Gram-Charlier series

$$f(x) = \frac{\phi(z)}{\sqrt{\kappa_2}} \left[1 + \sum_{k=3}^{\infty} \beta_k' H_k(z) \right], \qquad z = \frac{x - \kappa_1}{\sqrt{\kappa_2}}$$

- $H_j(z)$ are the Hermite polynomials.
- Coefficients β_j' are chosen to match cumulants κ_j of X (by inversion of its CGF $K(s) = \log \mathbb{E} \exp\{sX\}$).

- CDF of X obtained by integrating $f(x)$, grouping together terms in powers of $n^{-1/2}$, resulting in the Edgeworth expansion.
- For a "typical" likelihood-based statistic we obtain

$$F(x) = \Phi(z) - \phi(z)\left[\sum_{k=2}^{11} \beta_k H_k(z) + \mathcal{O}(n^{-5/2})\right], \qquad z = \frac{x - \kappa_1}{\sqrt{\kappa_2}},$$

Table: Value of coefficient of $\beta_k \kappa_2^{(k+1)/2}$ in Edgeworth expansion for CDF of R.

Statistic R	Value of k: 2	3	4	5	6	7	8	9	10	11
$R \neq R_{\mathrm{LR}}$	$\frac{\kappa_3}{6}$	$\frac{\kappa_4}{24}$	$\frac{\kappa_5}{120}$	$\frac{10\kappa_3^2+\kappa_6}{720}$	$\frac{\kappa_3\kappa_4}{144}$	$\frac{8\kappa_3\kappa_5+5\kappa_4^2}{5760}$	$\frac{\kappa_3^2}{1296}$	$\frac{\kappa_3^2\kappa_4}{1728}$	0	$\frac{\kappa_3^4}{31104}$
$R = R_{\mathrm{LR}}$	$\frac{\kappa_3}{6}$	$\frac{\kappa_4}{24}$	0	0	0	0	0	0	0	0

- The challenge now is to express (approximate) the κ_j (which are unknown) in terms of the V_k (which can be computed)!!!
- Has to be done case-by-case for each statistic R.
- Start from suitable Taylor expansions in probability for $\hat{\alpha}$, the maximizer of $\ell(\alpha)$, and use some tricks...
- Required A LOT OF BOOKKEEPING (20th century).
- In 21st century this can be replaced with careful programming of a symbolic algebra system (Maple/Mathematica).
- Above challenge has been worked out to 3rd order for classical statistics (LR, Wald, Score), by assuming $X \approx \mathcal{N}(0,1)$, so we:
  - assumed $X \approx \mathcal{N}(\kappa_1, \kappa_2)$ (gives greater accuracy), and
  - worked out 5th order expansions for all statistics in Table 1.

- Represent the log-likelihood derivatives by

$$\left.\frac{d^k \ell(\alpha)}{d\alpha^k}\right|_{\alpha=0} = nV_k + \sqrt{n} Z_k, \qquad \text{recall} \quad V_k := \mathbb{E}\ell_k(0)$$

  and Z_k is an $\mathcal{O}_p(1)$ random variable with zero mean.
- Construct high order Taylor expansion for $\ell(\alpha)$ at $\alpha = 0$, and solve for $\hat{\alpha}$ in terms of V_k and Z_k:

$$\sqrt{n}\hat{\alpha} = \sum_{k=0}^{k_{\max}} a_k(Z, V) n^{-k/2} + \mathcal{O}_p(n^{-(k_{\max}+1)/2})$$

  - The multivariate polynomials $a_k(Z, V)$ are functions of $(Z_1, Z_2, \ldots)$ and $(V_1, V_2, \ldots)$.
  - These polynomials are complicated but only need to be derived once (e.g., using symbolic computing).
  - They do not depend on the model or statistic!

- Very simple, and holds for all orders of accuracy!!!
- All other statistics are more complicated...
- Makes it possible to analytically derive all cumulants.
- Cumulants depend only on following (dimensionless & location-scale invariant) expressions:

$$\gamma = \frac{V_3}{2(-V_2)^{3/2}}, \quad \rho = -\frac{V_4}{6V_2^2}, \quad \xi = \frac{V_5}{24(-V_2)^{5/2}}, \quad \zeta = \frac{V_6}{120V_2^3}$$

Table: Approximations to the first 6 cumulants of R_S for the two models under consideration. The error in these approximations is $\mathcal{O}(n^{-5/2})$ (Mixture model) or $\mathcal{O}(\mu^{-5/2})$ (Poisson model).

Model	Cumulant $\hat{\kappa}_1$	$\hat{\kappa}_2$	$\hat{\kappa}_3$	$\hat{\kappa}_4$	$\hat{\kappa}_5$	$\hat{\kappa}_6$
Mixture	0	1	$\gamma/\sqrt{n}$	$(\rho-3)/n$	$(\xi-10\gamma)/n^{3/2}$	$(30+\zeta-10\gamma^2-15\rho)/n^2$
Poisson	0	1	$\gamma/\sqrt{\mu}$	ρ/μ	$\xi/\mu^{3/2}$	ζ/μ^2

- When n is small (not necessarily the case in these experiments).
- When Type I error rate (q_o) is very small..., how small?
- In "signal-hunting" particle physics experiments the gold standard is 5σ:

$$q_0 = P(Z > 5) = 2.87 \times 10^{-7}$$

- This puts us way out in the tail of the $\mathcal{N}(0,1)$...
- (And is the reason why simulation is undesirable; to get 100 values exceeding q_0 requires $\sim 10^9$ runs!)

- 5th order Edgeworth-approx: $F_R(r) - F_R^{\text{edge}}(r) = \mathcal{O}(n^{-5/2})$.
- Consider normal approx error

$$\Delta R(r) = r - \tilde{r}, \qquad \tilde{r} = \Phi^{-1}(F_R^{\text{edge}}(r))$$

- **Implies:**

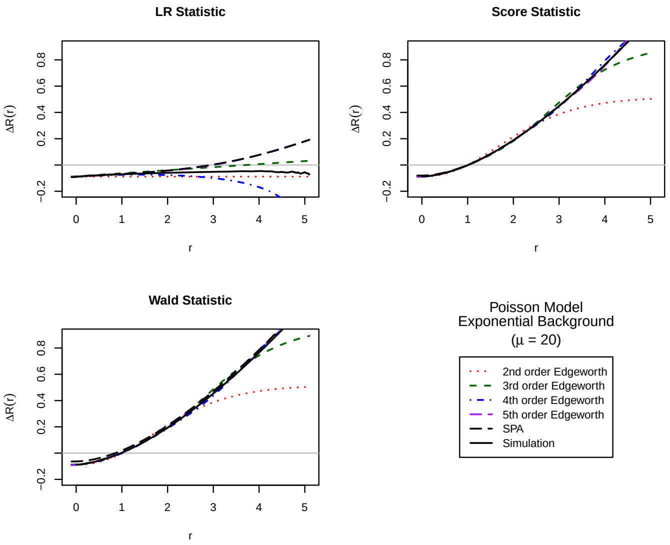
$$\tilde{r} = r \text{ to an accuracy of } O(n^{-5/2}) \text{ under } \mathcal{H}_0$$

  - If R is exactly $\mathcal{N}(0,1)$:

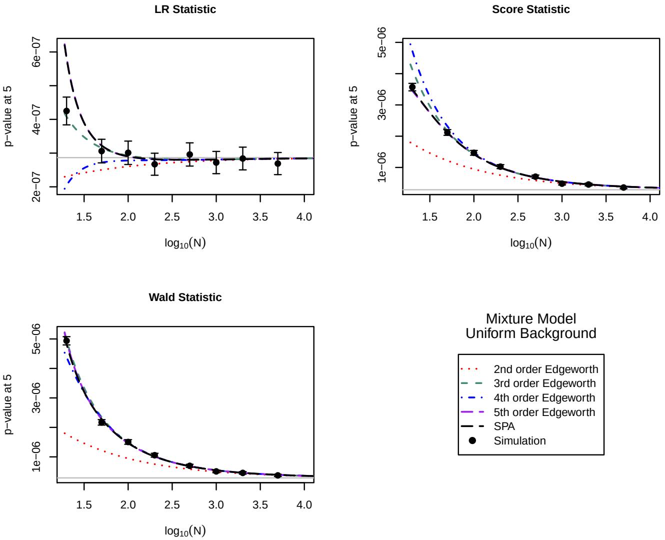
$$\Delta R(r) = 0$$

  - If R differs greatly from $\mathcal{N}(0,1)$:

$$\text{large values of } \Delta R(r)$$


LR Statistic
Score Statistic
Wald Statistic
ΔR(r)
r
−0.2
0.2
0.4
0.6
0.8
0
1
2
3
4
5
Poisson Model
Exponential Background
(μ = 20)
2nd order Edgeworth
3rd order Edgeworth
4th order Edgeworth
5th order Edgeworth
SPA
Simulation

Example: P-values at $r = 5$ for Mixture Model With Unif. Background (under $\mathcal{H}_0$)



- SPA is an efficient "automatic" procedure to perform the inversion:

$$K(s) = \sum_{j=1}^{\infty} \frac{s^j}{j!}\kappa_j \quad \mapsto \quad F(x) = P(X \leq x)$$

- $(k+1)$-th order SPA for the CDF of $\bar{X}_n$ (Daniels, 1987):

$$\hat{F}_{n,k}(x) = \Phi(\hat{w}\sqrt{n}) - \phi(\hat{w}\sqrt{n})\left[\frac{c_0}{n^{1/2}} + \frac{c_1}{n^{3/2}} + \cdots + \frac{c_k}{n^{k+1/2}}\right]$$

- The (asymptotic) truncation error of $\hat{F}_{n,k}(x)$ is:

$$\frac{\hat{F}_{n,k}(x)}{F(x)} = 1 + \mathcal{O}(n^{-k-3/2}) \qquad \Longleftrightarrow \qquad F(x) - \hat{F}_{n,k}(x) = \mathcal{O}(n^{-k-3/2})$$

- Since we have $\{\hat{\kappa}_1, \ldots, \hat{\kappa}_6\}$ for R (Table 3), SPA with $n = 1$ is an alternative to the Edgeworth approximation of p-values.
- Starting with $\hat{K}_m(s) = \sum_{j=1}^{m} \hat{\kappa}_j s^j / j!$, we note from Figs 1 & 2 that 5th order Edge and SPA give same $F(r)$... why?

Theorem

Let $\hat{G}_{1,k}(x)$ be **estimated** *$\hat{F}_{1,k}(x)$ by using $\hat{K}_m(s) = \sum_{j=1}^{m} \hat{\kappa}_j s^j / j!$ with $\hat{\kappa}_j = \kappa_j + \mathcal{O}_p(n^{-\alpha})$. Then:*

$$\frac{\hat{G}_{1,k}(x)}{F(x)} = 1 + \mathcal{O}_p(n^{-\min\{\alpha,(m-1)/2,k+3/2\}})$$

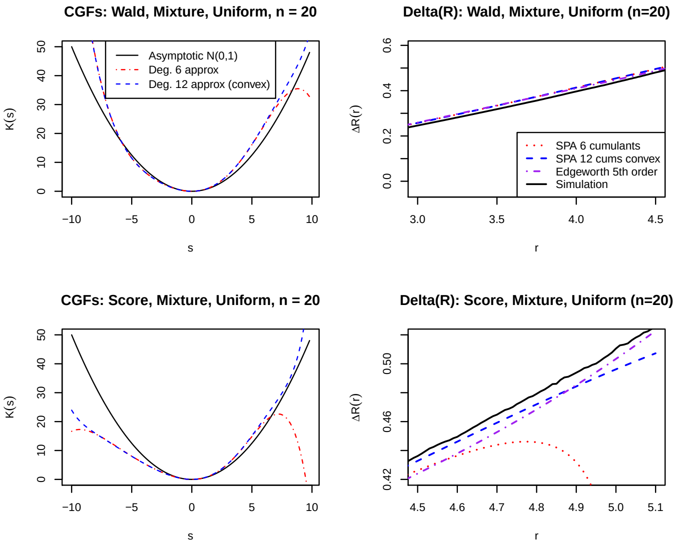
- In our case, $m = 6$ and $\alpha = 5/2$, so if we take $k = 1$, we get:

$$\frac{\hat{G}_{1,1}(x)}{F(x)} = 1 + \mathcal{O}_p(n^{-5/2})$$

- Thus with $\hat{K}_6(s)$, both Edge and SPA give same 5th order estimated $F(r)$... provided CGF is convex!
- Edge: doesn't care about convexity, but requires **new** painstaking analytical computations as m changes...
- SPA: remains essentially the **same** as m changes, but CGF must be convex...
- Idea: convexify CGF by doubling number of cumulants:

$$\begin{aligned} \{\hat{\kappa}_1, \ldots, \hat{\kappa}_6\} &= \text{approx cumulants on hand (rest are } \mathcal{O}(n^{-5/2})) \\ \{\kappa_7, \ldots, \kappa_{12}\} &= \text{solve for these by minimizing} \\ & \qquad \sum_{j=7}^{12} \kappa_j^2 \\ & \qquad \text{subject to convexity (quadratic programming)} \end{aligned}$$

Example: P-values at $r = 5$ for Mixture Model With Unif. Background (under $\mathcal{H}_0$)



- SPA vs. Edge: if both use same $\hat{K}_6(s)$, and it's convex, then

$$F_R^{\mathrm{edge}}(r) \approx F_R^{\mathrm{spa}}(r).$$

- If CGF not convex, then SPA can easily be "fixed", whereas Edge may give results of dubious quality...
- SPA CDF: guaranteed to be positive; Edge can be negative...
- Score statistic has a very simple asymptotic expansion, which makes it (relatively) easy to derive any number of (estimated) cumulants!
- Application of SPA to these instances is immediate, whereas Edge requires substantial analytical effort!!!
- (Question: Is it possible to combine these good properties of Score statistic with efficiency of LR statistic into a new statistic?)

Going Beyond Toy Problem: Nuisance Parameters in Signal & Background

- Doable: $s(x)$ & $b(x) \mapsto b(x|\phi)$.
  - extend everything we have done to the nuisance parameter setting (multivariate Edge/SPA).
- Problem: $s(x) \mapsto s(x|\boldsymbol{\theta})$ means $\boldsymbol{\theta}$ is **not identifiable** under $\mathcal{H}_0$:
  - classical inference for treating nuisance parameters then breaks down...
  - Davies (Biometrika, 1987): appropriate p-value is an **excursion probability**

$$\text{p-value} \ = P(\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} R(\boldsymbol{\theta}) > c)$$

  - Theory of Random Fields (TRF): emerged as only **analytical** solution so far (large-scale searches in neuroimaging, astrophysics, etc.)
    - $R(\boldsymbol{\theta})$ is viewed as Gaussian random field over manifold $\boldsymbol{\Theta} \subset \mathbb{R}^d$
    - ϕ has been profiled out of $R(\boldsymbol{\theta}, \phi) : \phi \mapsto \hat{\phi}$
    - provides closed-form approximaton when c is large...

- Excursion set of field above level c:

$$\mathcal{A}_c = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : R(\boldsymbol{\theta}) > c\}$$

- **Euler characteristic of excursion set:**

$$\phi(\mathcal{A}_c) = \text{geometric property of field}$$

- Fundamental result in TRF:

$$\mathbb{E}[\phi(\mathcal{A}_c)] = \sum_{i=0}^{d} a_i f_i(c)$$

  - a_i: positive constants (to be determined by Monte Carlo)
  - $f_i(\cdot)$: known "universal" functions

- For large c: (Taylor *et al.*, *Annals of Probability*, 2005)

$$\text{p-value } = P(\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} R(\boldsymbol{\theta}) > c) \approx \mathbb{E}[\phi(\mathcal{A}_c)] \equiv p_{global}$$

(Volobouev, I. & Trindade, A., JINST, 2018)

Suppose $\boldsymbol{\theta} \neq 0$ & $\boldsymbol{\phi} \neq 0$

- Solution 1 (straightforward): treat **all** parameters via TRF in conjuction with Edge/SPA $O(n^{-5/2})$ normalized versions of LR statistic
$$r \mapsto \tilde{r} = \Phi^{-1}(\hat{F}_{R_{\mathrm{LR}}}(r))$$
- Solution 2 (exotic): adjust global significance of test statistic, leading to (conservative) estimate of p_{global} in context of TRF...

$$p_{global} = P(R_{\mathrm{LR}}(\hat{\theta}) > r(\hat{\theta}))$$

  - $r(\theta)$ is observed (local) value of $R_{\mathrm{LR}}(\theta)$ computed from sample,
  - $\hat{\theta} = \arg\max_{\theta \in \Theta} r(\theta)$.

- Normal approx error for each observed (local) $r \equiv r(\theta)$ as before:

$$\Delta R(r(\theta)) = r(\theta) - \tilde{r}(\theta)$$

- Locate:

$$\theta^* = \arg\max_{\theta \in \Theta} \Delta R(r(\theta))$$

  - Search can use same grid as TRF search for $\hat{\theta} = \arg\max \; r(\theta)$.

- Calculate **global significance** of signal p_{global} via TRF, and express it in terms of the **global** r:

$$r_{global} = \Phi^{-1}(1 - p_{global})$$

- Adjust global r:

$$r_{global}^{adj} = r_{global} - \Delta R(r(\theta^*))$$

- Global (adjusted) p-value is then:

$$p_{global}^{adj} = 1 - \Phi(r_{global}^{adj})$$

References

- Algeri, S., van Dyk, D., Conrad, J., & Anderson, B. (2016), "On methods for correcting for the look-elsewhere effect in searches for new physics", *Journal of Instrumentation*, 11 P12010.
- Butler, R. (2007), *Saddlepoint Approximations With Applications*, Cambridge University Press.
- Easton, G. & Ronchetti, E. (1986), "General saddlepoint approximations with applications to L statistics", *Journal of the American Statistical Association*, 81, 420–430.
- Gross, E. & Vitells, O. (2010), "Trial factors for the look elsewhere effect in high energy physics", *The European Physical Journal C*, 70, 525–530.
- Ohman-Strickland, P. & Casella, G. (2002), "Approximate and estimated saddlepoint approximations", *Canadian Journal of Statistics*, 30, 97–108.
- Severini, T. (2000), *Likelihood Methods in Statistics*, Oxford University Press.
- Volobouev, I. & Trindade, A. (2018), "Improved Inference for the Signal Significance", *Journal of Instrumentation*, 13 P12011.

THE END!