# Marginalising vs Profiling of Nuisance Parameters (Physicist's View)

- **In the test statistic**
- **In simulation to get null distribution  of test statistic**

**...while maintaining frequentist coverage of the parameter of interest**

**Bob Cousins**

**Univ. of California, Los Angeles**

**Banff Workshop on Systematic Effects and Nuisance Parameters in Particle Physics Data Analyses**

**April 24, 2023 DRAFT 19 April**

## *Long* Tradition in HEP: Profile nuisance parameters and use asymptotic Wilks's Theorem to get approximate1D confidence intervals and (typically) 2D regions.
## *Name* "Profile likelihood" entered HEP in 2000, with Rolke, Lopez

**Classic FORTRAN* program: "Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations" by Fred James and Mats Roos, 1975.**

" MINUIT has been under continuous development at CERN since 1967. During this time the program has had enormous exposure to users at CERN and elsewhere. "

**"HESSE" errors from 2nd derivs at minimum;**

**"MINOS" errors from Δ(-lnL), as in Fig. 2**

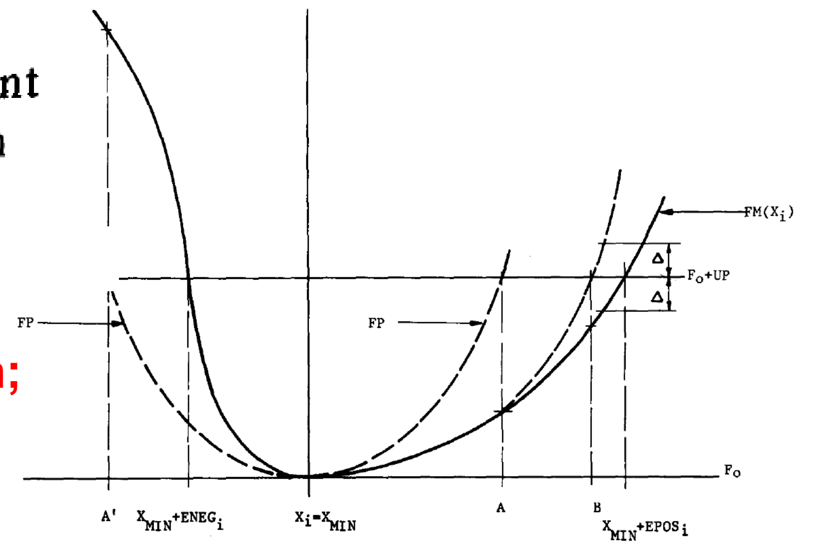



Fig. 2. Calculation of MINOS errors of parameter *i*. The (symmetric) dotted parabola FP is predicted from the covariance matrix, but the nonlinearity of the problem results in the solid curve FM which gives the asymmetric errors EPOS and ENEG (see text).

***Today it is in C++ and in ROOT**

# Approximate Confidence Regions Using Δ(-ln$\mathcal{L}$)
## (included in appendix to MINUIT users guide)

### "Interpretation of the Shape of the Likelihood Function around Its Minimum" by Fred James (1980)

It often happens that the solution of a minimum problem is itself straightforward, but the calculation or interpretation of the resulting parameter uncertainties, as determined by the shape of the function at the minimum, is considerably more complicated. The purpose of this note is to clarify the most commonly encountered difficulties in parameter error determination. These difficulties may arise in connection with any fitting program, but will be discussed here with the terminology of the program MINUIT for the convenience of MINUIT users.

The most common causes of misinterpretation may be grouped into three categories:

1. Proper normalization of the user-supplied chi-square or likelihood function, and appropriate ERROR DEF.
2. Non-linearities in the problem formulation, leading to different errors being calculated by different techniques, such as MIGRAD, HESSE and MINOS.
3. Multiparameter error definition and interpretation.

All these topics are discussed in some detail by Eadie et al., which may be consulted for further details.

Table 1
Table of UP for multiparameter confidence regions

| Number of parameters | Confidence level (probability contents desired inside hypercontour of "chi$^2$ = chi$^2$ min + UP") | | | | |
|---|---|---|---|---|---|
| | 50% | 70% | 90% | 95% | 99% |
| 1 | 0.46 | 1.07 | 2.70 | 3.84 | 6.63 |
| 2 | 1.39 | 2.41 | 4.61 | 5.99 | 9.21 |
| 3 | 2.37 | 3.67 | 6.25 | 7.82 | 11.36 |
| 4 | 3.36 | 4.88 | 7.78 | 9.49 | 13.28 |
| 5 | 4.35 | 6.06 | 9.24 | 11.07 | 15.09 |
| 6 | 5.35 | 7.23 | 10.65 | 12.59 | 16.81 |
| 7 | 6.35 | 8.38 | 12.02 | 14.07 | 18.49 |
| 8 | 7.34 | 9.52 | 13.36 | 15.51 | 20 09 |
| 9 | 8.34 | 10.66 | 14.68 | 16.92 | 21.67 |
| 10 | 9.34 | 11.78 | 15.99 | 18.31 | 23.21 |
| 11 | 10.34 | 12.88 | 17.29 | 19.68 | 24.71 |

If FCN is −log(likelihood) instead of chi-square, all values of UP should be divided by 2.

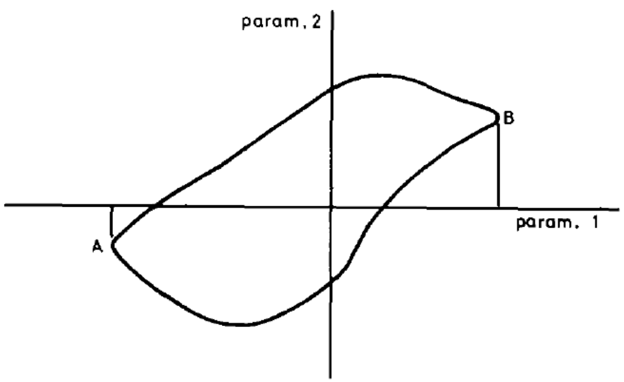



Fig. 1. MINOS errors for parameter 1.

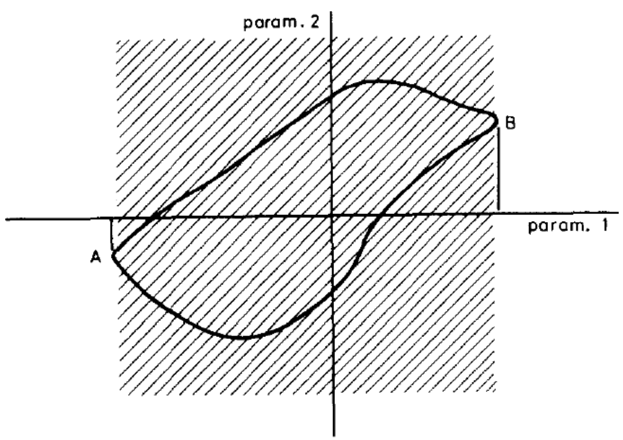



Fig. 2. MINOS error confidence region for parameter 1.

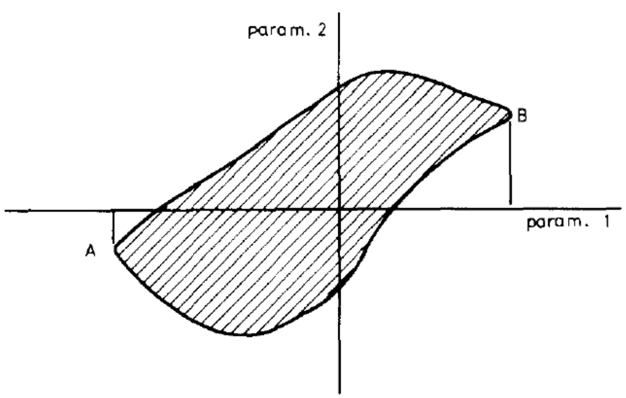



Fig. 4. Optimal confidence region for parameters 1 and 2.

**At some point, a few people started integrating out nuisance parameter(s), typically when Gaussian/Normal contribution to likelihood, while treating parameter of interest in frequentist manner.**
**E.g., (see Backup slides)**
**Robert Cousins and Virgil Highland, "Incorporating systematic uncertainties into an upper limit" (1992).**

**For a more enlightened discussion, see our paper on the on/off problem discussed later in this talk,**
**Cousins, Linnemann, Tucker (2008). "CLT"**

**Luc Demortier noted that, in these simple cases, what we did was the same math as George Box's prior predictive p-value!**

**Box calculated a tail probability after obtaining a Bayesian pdf, and we averaged a frequentist tail probability over a Bayesian pdf. (Simply reverse order of two integrals; see above paper.)**

# Profile likelihood ratio is most common test statistic in use today at LHC (in all of HEP?)

MINUIT MINOS history, still widely used.

1998 intervals by Feldman and Cousins were recognized as no-nuisance special case of inversion of “ exact” PLR test in “Kendall and Stuart”.

HEP-invented “modification” called $CL_S$ is ratio of two p-values that are both based on PLR test statistics (with some historical use of marginalization of nuisances).

CHAPTER 22

LIKELIHOOD RATIO TESTS AND TEST EFFICIENCY

**The LR statistic**

**22.1** The ML method discussed in Chapter 18 is a constructive method of obtaining estimators which, under certain conditions, have desirable properties. A method of test construction closely allied to it is the likelihood ratio (LR) method, proposed by Neyman and Pearson (1928). It has played a role in the theory of tests analogous to that of the ML method in the theory of estimation.

As before, we have the LF

$$L(x|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)$ is a vector of $r + s = k$ parameters ($r \geq 1, s \geq 0$) and $x$ may also be a vector. We wish to test the hypothesis

$$H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}, \tag{22.1}$$

which is composite unless $s = 0$, against

$$H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r0}.$$

We know that there is generally no UMP test in this situation, but that there may be a UMPU test – cf. **21.31**.

The LR method first requires us to find the ML estimators of $(\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)$, giving the unconditional maximum of the LF

$$L(x|\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s), \tag{22.2}$$

and also to find the ML estimators of $\boldsymbol{\theta}_s$, when $H_0$ holds,[1] giving the conditional maximum of the LF

$$L(x|\boldsymbol{\theta}_{r0}, \hat{\hat{\boldsymbol{\theta}}}_s). \tag{22.3}$$

$\hat{\hat{\boldsymbol{\theta}}}_s$ in (22.3) has been given a double circumflex to emphasize that it does not in general coincide with $\hat{\boldsymbol{\theta}}_s$ in (22.2). Now consider the likelihood ratio[2]

$$l = \frac{L(x|\boldsymbol{\theta}_{r0}, \hat{\hat{\boldsymbol{\theta}}}_s)}{L(x|\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s)}. \tag{22.4}$$

Since (22.4) is the ratio of a conditional maximum of the LF to its unconditional maximum, we clearly have

$$0 \leq l \leq 1. \tag{22.5}$$

Intuitively, $l$ is a reasonable test statistic for $H_0$: it is the maximum likelihood under $H_0$ as a fraction of its largest possible value, and large values of $l$ signify that $H_0$ is reasonably acceptable. The critical region for the test statistic is therefore

$$l \leq c_\alpha, \tag{22.6}$$

where $c_\alpha$ is determined from the distribution $g(l)$ of $l$ to give a size-$\alpha$ test, that is,

$$\int_0^{c_\alpha} g(l)\,dl = \alpha. \tag{22.7}$$

Neither maximum value of the LF is affected by a change of parameter from $\boldsymbol{\theta}$ to $\tau(\boldsymbol{\theta})$, the ML estimator of $\tau(\boldsymbol{\theta})$ being $\tau(\hat{\boldsymbol{\theta}})$ – cf. **18.3**. Thus the LR statistic is invariant under reparametrization.

**Key paper in 2010, first year of LHC data**

**Cowan, Cranmer, Gross, and Vitells (CCGV) “Asymptotic formulae for likelihood-based tests of new physics”.**

**Examined profile LR ratio and their preferred variants:**

***Widely used in HEP, with significant side effect:***

**Since asymptotic distributions were not known in HEP for other test statistics, the default became profile LR expressions in CCGV, *even for small sample size*, “for consistency”.**

# But…in stats literature, long history of criticism of simple profile LR

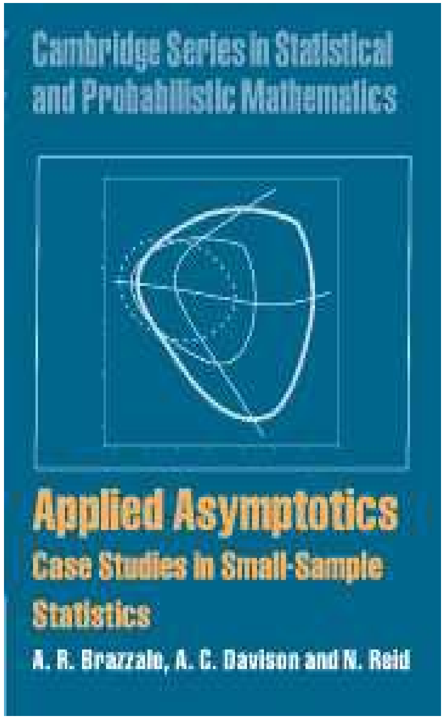
E.g., 2005 PhyStat at Oxford had various reviews of stats literature with higher-order approximations or other improvements: partial likelihood, adjusted likelihood, modified likelihood.

See in particular my talk, Nancy Reid's comments on it, and her workshop summary talk.

Talk by Anthony Davidson at PhyStat-nu 2019 earlier work in Banff challenge after Banff 2006

Talks by statistician Alessandra Brazzale and physicist Igor Volobouev at PhyStat-DM 2019

*Virtually none of these developments have been adopted in HEP, as far as I know.*



# And marginalizing nuisances is rare at LHC (I think)

IMO, three reasons:

1) The historical precedent of MINUIT MINOS,
2) Profiling is natural within context of LR test statistics
3) Since CCGV, asymptotic formulas are readily available for profile likelihood but not marginalization, and are the default in the now-dominant software tools.

Meanwhile our Bayesian friends have long advocated marginalizing nuisance parameters even if we stick to frequentist treatment of parameter of interest, e.g.,

James O. Berger, Brunero Liseo and Robert L. Wolpert, "Integrated Likelihood Methods for Eliminating Nuisance Parameters", (1999)

**From Berger et al rejoinder:**

**"Dr. Susko finishes by suggesting that it might be useful to compute both profile and integrated likelihood answers in application, as a type of sensitivity study. It is probably the case that, if the two answers agree, then one can feel relatively assured in the validity of the answer. It is less clear what to think in the case of disagreement, however. If the answers are relatively precise and quite different, we would simply suspect that it is a situation with a 'bad' profile likelihood."**

**I (BC) agree with Dr. Susko!**

**More studies needed in real cases to see if Berger et al are right.**

# Null distribution of the test statistic(s)

**Rather than pursuing higher-order corrections to the profile LR, the general practice in HEP has been to stick with the profile LR test statistic, and to use Monte Carlo simulation (known as "toy MC") to obtain the finite-sample-size distribution(s) of the profile LR under the null hypothesis, and under alternative(s) as desired. This is often compared to the relevant asymptotic formula.**

**So now one has the question: *How should nuisance parameters be treated in the simulation?***

**Keep in mind: we want correct frequentist coverage of the parameter of interest when nature is sampling from the *true* values of the nuisance parameters.**

**The usual procedure (evidently based on a desire to be "fully frequentist), is to "throw toys" *using the profiled values of the nuisance parameters*, i.e., their ML estimates conditional on whatever value of the parameter(s) of interest are being tested.**

**At some point, this was identified as the *parametric bootstrap*.**

**The hope is that the profiled values of the nuisance parameters are a good proxy for the unknown true values.**

**I would like to see more comparisons of these results with alternatives, particularly those from marginalizing the nuisance parameters with some judiciously chosen priors.**

# Example

I "cherry-picked" an example that

1) appears often in the statistical literature;
2) is an important prototype in HEP and astronomy; and
3) I happen to be a co-author on studies thereof:

*The ratio of Poisson means,* which maps onto:

a) the "on/off" problem in astronomy and
b) the "signal band plus sideband" problem in HEP.

For the algebra, see Cousins, Linnemann, Tucker, (2008). "CLT"

Here I will mostly just give concepts.

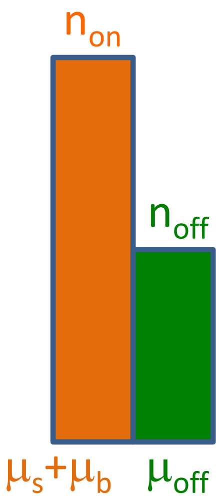
## Histogram with 2 bins, observed contents $n_{on}$ and $n_{off}$.

The "on" bin has (potentially) signal and bkg events with unknown Poisson means $\mu_s$ and $\mu_b$ , total Poisson mean $\mu_s + \mu_b$ .

The "off" bin has only bkg with unknown Poisson mean $\mu_{off}$ .

The ratio of bkg in the two bins, $\tau = \mu_{off} / \mu_b$ is ***known***.

Test $H_0$: $\mu_s = 0$. Rephrase $H_0$: $\mu_{off} / (\mu_s + \mu_b) = \tau$.
I.e., $H_0$: ratio of Poisson means of the two bins is $\tau$.
Choice of nuisance param; let it be $\mu_b$ .




$\tau = \mu_{off} / \mu_b$
known

Std solution: eliminate nuisance by "conditioning":
$n_{tot} = n_{on} + n_{off}$ has no information about ratio;
treat this ancillary statistic as fixed and consider binomial distribution of $n_{on}$ , $n_{off}$ .

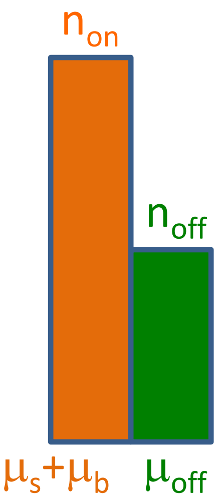
Again rephrase $H_0$: Binomial param $\rho$ is $\mu_b / (\mu_b + \tau\, \mu_b) = 1 / (1+\tau)$.

## Two ways to write the likelihood function $\mathcal{L}$ :

$$\mathcal{L}\,(n_{on}\,,\, n_{off}\,;\, \mu_s+\mu_b\,,\, \mu_{off})$$

$$= \text{Pois}(n_{on}\,;\, \mu_s+\mu_b\,)\ \text{Pois}(n_{off}\,;\, \mu_b)$$

$$= \text{Pois}(n_{tot}\,;\, \mu_s+\mu_b+\mu_{off})\ \text{Binom}(n_{on}\,;\, n_{tot}\,,\, \rho)$$




$\tau = \mu_{off} / \mu_b$

**Where recall $H_0$: $\mu_s = 0$, equiv to $H_0$: $\rho = 1 / (1+\tau)$.**
**So do binomial test of $\rho = 1 / (1+\tau)$, given data $n_{on}$ , $n_{tot}$ .**
**p-value obtained by inversion of binomial conf intervals.**
**Convert to Z-value with 1-tailed convention: $Z_{Bi}$.**

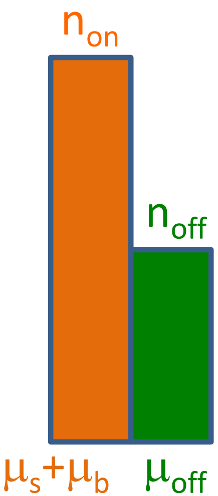
**CLT used standard Clopper-Pearson conf intervals.**
**Very conservative for small n's.**
**Later paper by Cousins, Hymes, Tucker *preferred Lancaster mid-p confidence intervals.***

**Compare other methods to these "exact" intervals.**

# Marginalization of nuisance $\mu_b$ (Bayes-Freq hybrid)

**First, consider case when $\mu_b$ is known *exactly*.**
**The p-value (call it $p_P$) for $H_0$: $\mu_s = 0$ is Poisson prob of obtaining $n_{on}$ or more events in the "on" bin with true mean $\mu_b$ .**

**Then introduce uncertainty in $\mu_b$ in Bayesian-like way, with uniform prior for $\mu_b$ (ugh) and likelihood $\mathcal{L}(\mu_b)$ from Poisson probability of observing $n_{off}$ in "off" bin, thus obtaining posterior prob $P(\mu_b|n_{off})$, a $\Gamma$ function.**




$\tau = \mu_{off} / \mu_b$

**Finally, take weighted average of $p_P$ over $P(\mu_b|n_{off})$:**

**$P_{\text{B-F hybrid}} = \int p_P \, P(\mu_b|n_{off}) \, d\mu_b$. and map to $Z_{\text{B-F hybrid}}$ .**

**Jim Linnemann discovered numerically, and then proved, that $Z_{\text{B-F hybrid}} = Z_{Bi}$ (!) (Called $Z_\Gamma$ in paper.)**
**[See Backup slides for note regarding simulation.]**

# Parametric Bootstrap treatment of nuisance $\mu_b$

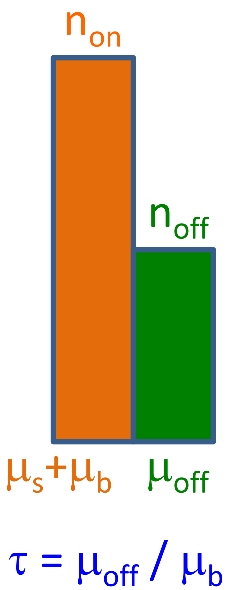
**For testing $H_0$: $\mu_s = 0$, we find the profile likelihood estimate of $\mu_b$. See Li and Ma, cited by CLT, for math.**

**Conceptually, we can use all the data by noting that $n_{tot} = n_{on} + n_{off}$ is sample from Poisson mean $\mu_b + \mu_{off} = \mu_b(1 + \tau)$.**

**So profiled MLE of $\mu_b$ for $\mu_s$=0 is $(n_{on} + n_{off}) / (1 + \tau)$.**

**Parametric bootstrap takes this value of $\mu_b$ as truth and proceeds to calculate p-value as in $p_P$ above.**

**Can be done by simulation or direct calculation, leading to $Z_{PB}$.**



# Numerical examples

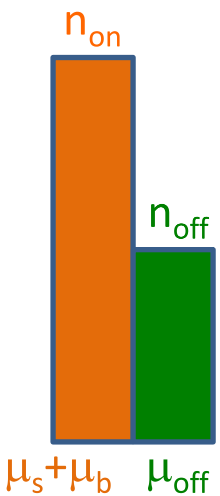
**Suppose $\tau$ = 1, $n_{on}$ = 10, $n_{off}$ = 2.**

**"Exact" $Z_{Bi}$ = $Z_{B\text{-}F\ hybrid}$ (aka $Z_{\Gamma}$) = 2.07**

**Using Lancaster mid-P binomial: $Z_{mid\text{-}P}$ = 2.28**

**Asymptotic result from Wilks's Thm is $Z_{PL}$ = 2.41**

**Parametric bootstrap:**
**For $\mu_s$ = 0, profiled MLE of $\mu_b$ = 6, and so $\mu_{off}$ = 6.**
**So generate toys with $\mu$ = 6 in each bin (!) $Z_{PB}$ = 2.32**




$\tau = \mu_{off} / \mu_b$

# Numerical examples

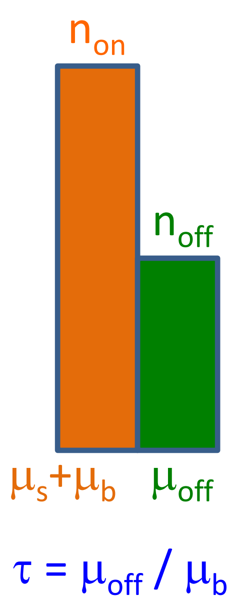
Suppose $\tau = 2$, $n_{on} = 10$, $n_{off} = 2$.

"Exact" $Z_{Bi} = Z_{\text{B-F hybrid}}$ (aka $Z_\Gamma$) = 3.27

Using Lancaster mid-P binomial: $Z_{\text{mid-P}} = 3.44$

Asymptotic result from Wilks's Thm is $Z_{PL} = 3.58$

Parametric bootstrap:
For $\mu_s = 0$, profiled MLE of $\mu_b = 4$, and so $\mu_{off} = 8$.
So generate toys with $\mu = 4, 8$ in the bins (!) $Z_{PB} = 3.46$




Superficially, B-F hybrid is "Exact", PB gives larger Z.
But recall: "Exact" overcovers due to discreteness – maybe PB is actually better. Need closer look.

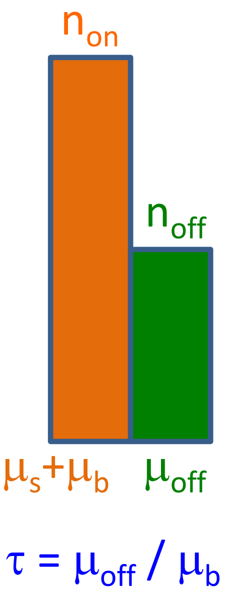
# Coverage Studies

**Following CLT, we pick a threshold $Z_{thresh}$ (e.g., 3 or 5) and consider a pair of true values of ($\mu_b$ , $\tau$).**

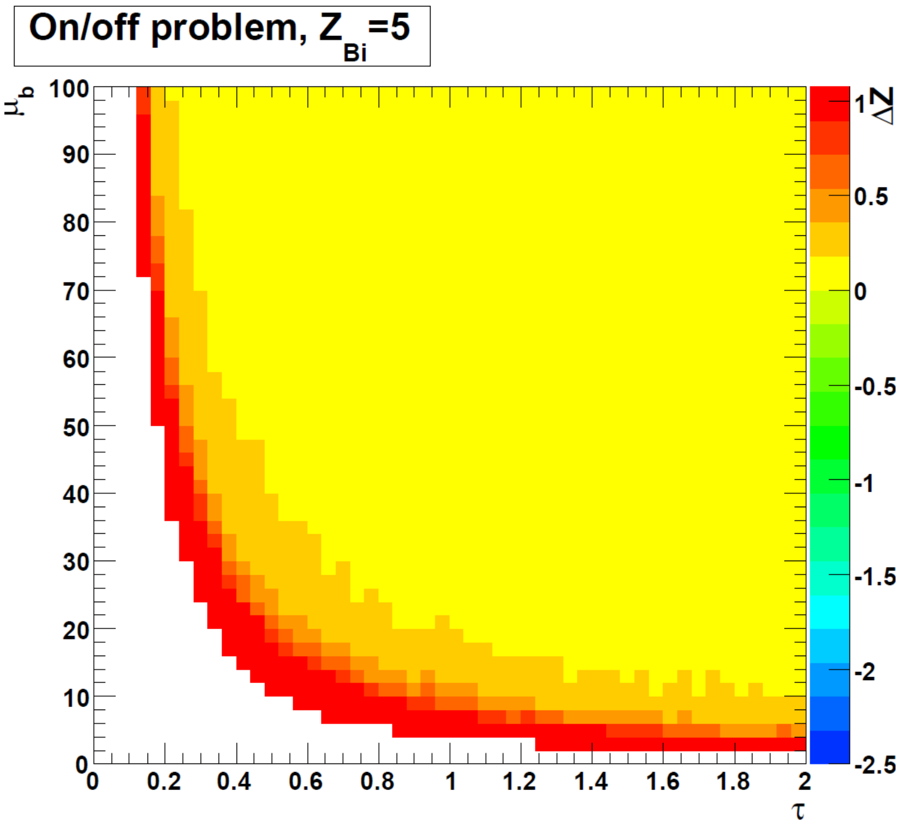
**For each pair, we consider all values of ($n_{on}$ , $n_{off}$ ), and calculate both the probability of obtaining that data and the claimed Z value for each recipe.**

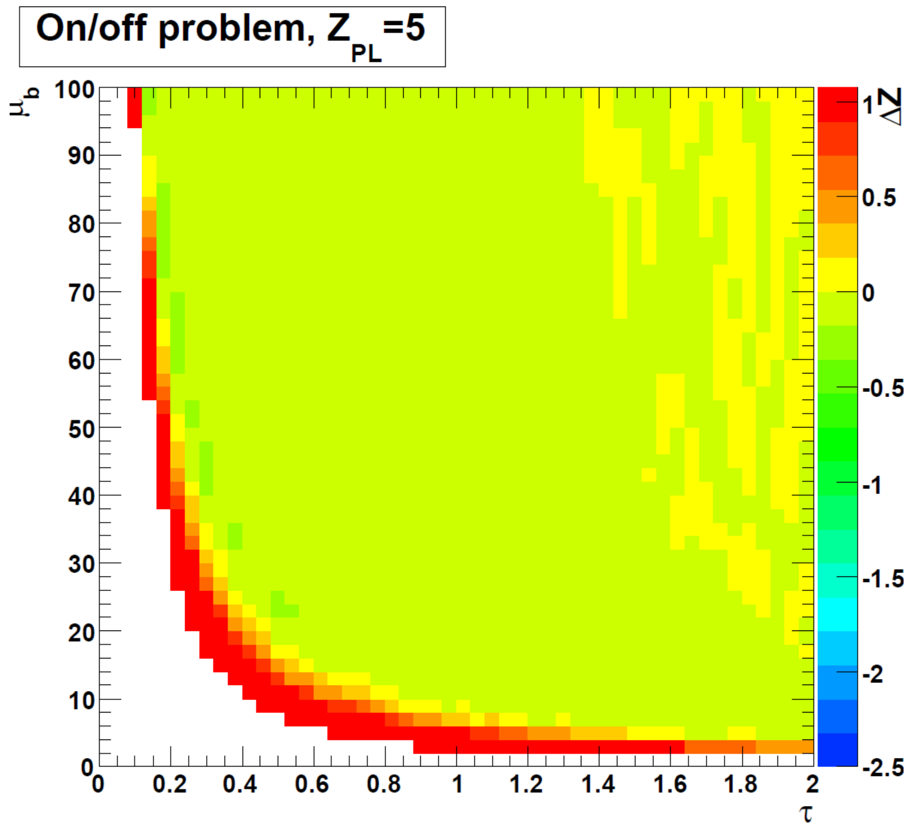
**We can then compute the probability that $Z \geq Z_{thresh}$ for each recipe. From this we map to a value $Z_{true}$ that represents the true Z , and compare to $Z_{thresh}$ .**

**(This could be done by simulation, but we do direct calculation.)**



# From the CLT (2008) paper




**For each fixed value of $\tau$ and $\mu_b$ , the color indicates $Z_{true} - Z_{claim}$ for the ensemble of expts quoting $Z_{claim} = 5$.**

**I calculated a couple points using parametric bootstrap $Z_{PB}$ :**

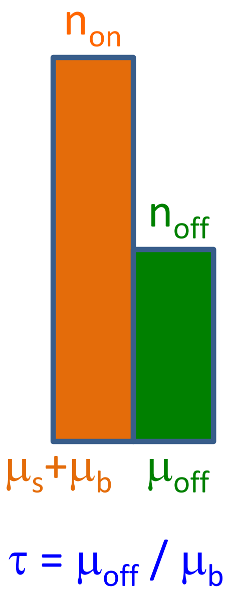
**For $\tau = 1$ and $\mu_b = 10$, $Z_{claim} = 3$: $Z_{true} = 3.0$ (!)**

**For $\tau = 1$ and $\mu_b = 20$, $Z_{claim} = 5$: $Z_{true} = 5.0$ (!)**

# Summary

How much is HEP losing by not examining more thoroughly higher-order asymptotic theory?

I cherry-picked my example problem to be one in which I knew that marginalizing the nuisance parameter (with judicious prior) yielded the standard frequentist Z, while doing "fully frequentist" parametric bootstrap as commonly done at the LHC would give higher $Z_{PB}$. But the discrete nature of the problem "saves" the coverage of $Z_{PB}$ !




For *Gaussian* uncertainty on mean $\mu_b$ , marginalizing the nuisance parameter can be badly anti-conservative (Kyle Cranmer at Oxford 2005, our paper).

I would urge more exploration of these issues in real LHC analyses!

# References (PhyStat at end)

Wolfgang A. Rolke and Angel M. López, Confidence intervals and upper bounds for small signals in the presence of background noise," Nucl. Instrum. Meth.A 458 (2001) 745
e-Print: hep-ph/0005187 [hep-ph]. See also other papers with Lopez et al.

Fred James and Mats Roos (CERN), "Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations", Computer Physics Commun. 10 (1975) 343

Fred James (CERN), "Interpretation of the Shape of the Likelihood Function around Its Minimum", Computer Physics Commun. 20 (1980) 29

Robert D. Cousins and Virgil L. Highland, "Incorporating systematic uncertainties into an upper limit", Nucl. Instr. Meth. A 320 (1992) 331.

Robert D. Cousins, James T. Linnemann, Jordan Tucker, "Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process", Nucl. Instrum. Meth. A 595 (2008) 480, arXiv:physics/0702156.

Robert D. Cousins, Kathryn E. Hymes, Jordan Tucker, "Frequentist Evaluation of Intervals Estimated for a Binomial Parameter and for the Ratio of Poisson Means", Nucl. Inst. Meth. A 612 (2010) 388, arXiv:0905.3831

G. Cowan, k. Cranmer, E. Gross, and I. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", Eur. Phys. J. C71 1554 (2011), arXiv:1007.1727. (See also arXiv:1210.6948)

# References (cont.)

**James O. Berger, Brunero Liseo, and Robert L. Wolpert, "Integrated Likelihood Methods for Eliminating Nuisance Parameters", Statistical Science 14 (1999) 1.**

**PhyStat 2005 at Oxford: Talks at https://confs.physics.ox.ac.uk/phystat05/programme.asp, Proceedings at https://confs.physics.ox.ac.uk/phystat05/proceedings/default.htm .**
**See, e.g., Bob Cousins, "Treatment of Nuisance Parameters in High Energy Physics, and Possible Justifications and Improvements in the Statistics Literature, and response by Nancy Reid; also Nancy Reid, "Summary of Some Statistical Issues".**

**PhyStat-nu 2019 at CERN: https://indico.cern.ch/event/735431/timetable/ , in particular Anthony Davidson, "Almost-perfect Signal Detection".**

**PhyStat-DM at Stockholm: https://indico.cern.ch/event/769726/timetable/#all.detailed , in particular, Alessandra R. Brazzale, "Likelihood Asymptotics and Beyond," and Igor Volobouev, "Improved Inference for the Signal Significance".**

# Thanks to all (see note), including my "sponsor", U.S. DOE Office of Science

# BACKUP

**Robert Cousins and Virgil Highland, “Incorporating systematic uncertainties into an upper limit” (1992).**

C-H looked at the case of a physics quantity

$\sigma = \mu_b / L$ ,

where $\mu_b$ = unknown Poisson mean, L is factor called luminosity.
One observes a sampled value n from the Poisson distribution.

From n, one obtains an 90% upper confidence limit on $\mu_b$.
If L known exactly, scale by 1/L to get upper conf limit on $\sigma$.

If instead one has unbiased estimate of L and assumes symmetric pdf for L with known variance, C-H integrate over L, what we now call marginalizing over the nuisance parameter L.

**Our motivation was to ameliorate an effect whereby confidence intervals derived from a discrete observable become shorter when a continuous nuisance parameter is added to the model.**

**In the initially submitted draft, we did not know that we were grafting a Bayesian pdf onto frequentist Poisson C.L.**

**(Fred James sorted us out before publication.)**

**As mentioned in main talk, for a more enlightened discussion, see our paper on the on/off problem discussed later in this talk, Cousins, Linnemann, Tucker (2008). "CLT"**

# Simulation of marginalization of nuisance $\mu_b$

**Sample $\mu_b$ from $P(\mu_b|n_{off})$, compute frequentist $p_P$ for each sampled value of $\mu_b$ (using same observed $n_{on}$).**

**Calculate arithmetic mean of these values of $p_P$ .**

**Convert to $Z_P$ as desired. (Note that $p_P$'s are averaged, not $Z_P$'s).**