Statistical issues in modern flavour physics experiments

PHYSTAT Flavour Workshop 2020

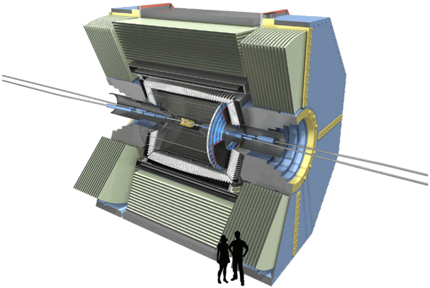
H. Dembinski[1] M. Kenzie[2]

[1]TU Dortmund, Germany; [2]Warwick University, UK

October 19, 2020

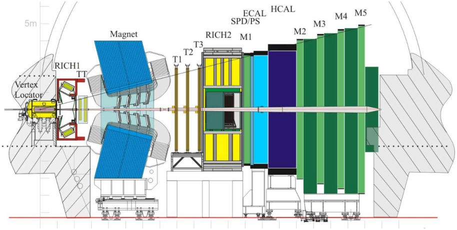
Modern heavy flavour physics experiments

Belle-II (2017-)



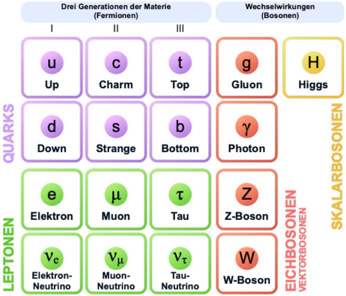
- e^+e^- collisions (clean)
- $\sqrt{s} = 10.58\,\mathrm{GeV}$ ($\Upsilon(4S)$), decays to $B\bar{B}$ in 96 %
- Fully hermetic acceptance
- Specialised for study of B-hadrons
- Target luminosity $50\,000\,\mathrm{fb}^{-1}$

LHCb (2010-)



- pp, p-Pb, Pb-Pb collisions (messier)
- $\sqrt{s} = 0.9 - 14\,\mathrm{TeV}$
- Fixed-target: $\{p,\mathrm{Pb}\}$-$\{\mathrm{He},\mathrm{Ne},\mathrm{Ar}\}$ @ 69-110 GeV
- Single-arm spectrometer
- General purpose forward spectrometer
- Target luminosity $300\,\mathrm{fb}^{-1}$ (Run 1-6)

Heavy flavour physics



CP violation $\leftrightarrow$ matter/antimatter asymmetry

- K^0: Cronin & Fitch 1964
- B^0: BABAR 2001
- D^0: LHCb 2019

- Search for new physics in B decays
  - Time-integrated CP violation[a]
  - Time-dependent CP violation[a]
  - Search for rare decays[b]
- Direct search for light new particles[b]
- Search for flavour-changing neutral currents beyond SM
- Hadron spectroscopy and precise measurement of Standard Model
  - Tetraquarks, Pentaquarks, excited mesons and baryons
- Ion collisions (LHC): Study of Quark Gluon Plasma and collective effects
- Spin/resonance analysis with Dalitz method[a,c]

[a] sPlot technique, [b] limit setting, [c] multi-dimensional data

High-level goals 1

Identify and promote optimal methods and best practices

Image credit: Matt Flores, Unsplash

- *Gold standard*: Unbiased estimators with minimum variance
- Maximum likelihood estimation
- Blind searches
- Report sufficient information
  - Full covariance matrices of statistical and systematic errors
  - Likelihood functions for limits (needs software: HistFactory, pyhf, . . .)
  - Symmetric intervals preferred that behave like std.deviations
- Use ensemble methods to check estimators

See Nicholas Wardle's talk (combination of results) on Tuesday, 14:00 CEST

High-level goals 2

Ensure coherent meaning of uncertainty intervals and limits

- Particle physics has Frequentist tradition
- Confidence intervals need to have specified coverage probability
  - Exception for limit setting where overcovering is accepted
  - Users of Bayesian methods need to demonstrate coverage
- OK: high-statistics unweighted case
- Ongoing research: low statistics, weighted data
- Rules needed for consistent treatment of systematic uncertainties
- Consistency with wider community when reporting limits (CL_s)

See Giovanni Punzi's talk (interval estimation) on Monday, 16:30 CEST

Typical challenges for measurements in flavour physics

- Sophisticated non-linear models with many nuisance parameters
  - Profile likelihood method replaces Neyman construction
  - Uncertainty of control variables (calibration factors, efficiencies, ...) propagated into final result either by
    - Likelihood profiling
    - Marginal likelihood (seems to be rare?)
    - Error propagation (first order or simulation based)
  - Exact coverage probability not guaranteed, has to be checked
  - Common practical challenge is to fully automate these fits
- Unbinned fits are popular, especially for multi-dimensional data
  - Computationally expensive when samples are large
- sWeights as a statistical tool are popular and correspondingly analysis of weighted data (more on that later)

See Peter Stangl's talk (global fits) on Wednesday, 16:30 CEST

Intervals from HESSE or MINOS method?

F. James, "Statistical Methods in Experimental Physics" (2nd) edition, World scientific, p. 240

Table 9.2. Functions for interval estimation in the general case.

Method	$Q^2(\theta)$	Asymptotic Expansion	Mean
(1a) Information computed analytically	Q^2_{1a}	$\frac{X^2}{1} + \frac{2X^2Y}{I^2\sqrt{N}} + \frac{KX^3}{I^3\sqrt{N}}$	$1 + \frac{1}{N}(2a+b)$
(1b) Information estimated from data	Q^2_{1b}	$\frac{X^2}{I} + \frac{X^2Y}{I^2\sqrt{N}} + \frac{KX^3}{I^3\sqrt{N}}$	$1 + \frac{1}{N}(a+b)$
(1c) Information estimated from data, at $\theta = \hat{\theta}$	Q^2_{1c}	$\frac{X^2}{I} + \frac{X^2Y}{I^2\sqrt{N}}$	$1 + \frac{1}{N}a$
(2a) Information computed analytically	Q^2_{2a}	$\frac{X^2}{I}$	1
(2b) Information estimated from data	Q^2_{2b}	$\frac{X^2}{I} + \frac{X^2Y}{I^2\sqrt{N}}$	$1 + \frac{1}{N}a$
(2c) Information estimated from data, at $\hat{\theta}$	Q^2_{2c}	$\frac{X^2}{I} + \frac{X^2Y}{I^2\sqrt{N}} + \frac{KX^3}{I^3\sqrt{N}}$	$1 + \frac{1}{N}(a+b)$
(3) Likelihood ratio	Q^2_3	$\frac{X^2}{I} + \frac{X^2Y}{I^2\sqrt{N}} + \frac{KX^3}{3I^3\sqrt{N}}$	$1 + \frac{1}{N}\left(a + \frac{b}{3}\right)$

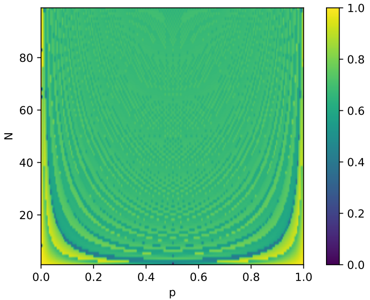
- HESSE method
  - Based on asymptotic normality of estimator
  - Symmetric intervals
  - Full covariance matrix
  - Asymptotic bias wrt to χ^2: $1 + \frac{1}{N} a$
- MINOS method
  - Based on on asymptotic χ^2 distribution of likelihood ratio
  - Asymmetric intervals
  - Asymptotic bias wrt to χ^2: $1 + \frac{1}{N}\left(a + \frac{b}{3}\right)$
- MINOS intervals
  - Asymmetric intervals seem to offer more detail
  - Coverage probability in small samples should be worse in general to due larger bias
  - Difficult to combine with other results (likelihood cannot be recovered from 3 points)

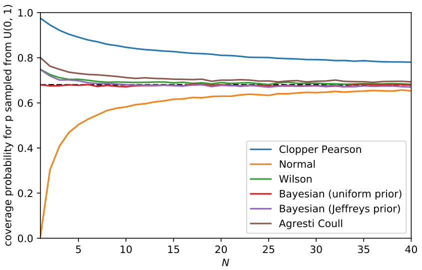
Intervals used for distributions with discrete samples

We prefer "right coverage probability on average" over "always conservative" to be consistent.

Poisson distribution: sqrt(N) estimate preferred over exact Neyman construction

Binomial distribution: Wilson interval preferred over Clopper-Pearson





Remarks on ensemble methods

- Study distribution of estimator on generated pseudo-datasets
  - Parametric bootstrap: Fit model to data, draw samples from model
  - Non-parametric bootstrap: sample uniformly from original data set
- Generic method applicable to estimators of arbitrary complexity
  - Construct confidence intervals with good coverage (bca method), estimate bias, estimate coverage probability of interval estimator
- Challenges
  - Samples must be independent and identically distributed (i.i.d.)
    - Arbitrary sample of particle tracks may not be i.i.d.
    - Events within time blocks of constant detector performance may be i.i.d.
  - Ensemble methods can be prohibitively time-consuming

See Brad Efron's talk on Tuesday, 18:00 h CEST

Efron & Tibshirani, "An Introduction to the Bootstrap", CRC Press 1994

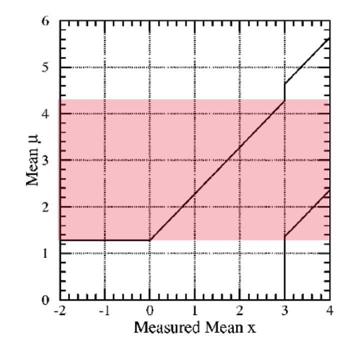
Limit setting

- Typical scenario
  - Observed $n = n_s + n_b$, want to estimate signal expectation μ_s
  - Background expectation μ_b not exactly known, estimate $\hat{\mu}_b$ has statistical uncertainty (e.g. background estimated from off-signal region)
  - $\hat{\mu}_s$ and $\hat{\mu}_b$ are usually estimated from fits with various nuisance parameters (calibration factors, efficiencies, ...)
  - Want to report **central interval** when evidence for signal is strong and **upper limit** otherwise (with well-defined coverage probability)
- Undesired
  - Methods that yield empty or unphysical intervals (e.g. $\mu_s \in [-3, 1]$)
  - Methods that undercover through flip-flopping
  - Experiment with higher expected background should not give better limit when $n = 0$ is observed

Feldman-Cousins approach

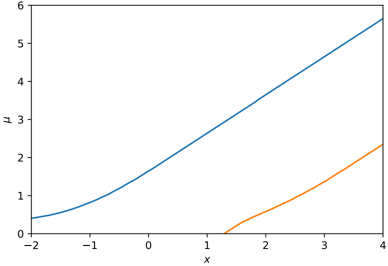
Feldman & Cousins, Phys. Rev. D 57 (1998)

- FC approach: Refinement of classic Neyman construction with guaranteed coverage properties



Red area has less than 90 % coverage probability (image from FC paper, red overlay added)

- Educational example from FC paper: Gaussian for x with $\sigma = 1$, $\mu \geq 0$
- Bad algorithm to report result at 90 % CL
  - If result less than 3σ, report upper limit
  - If result greater than 3σ, report central confidence interval
  - If $x < 0$, report upper limit for $x = 0$
- Intervals constructed in this way contain μ in only 85 % of cases if $\mu = 2$



Confidence belt constructed with FC method for normal with $\sigma = 1, \mu > 0$

- FC method
  - Neyman construction: Constructed belt horizontally, read off vertically
  - For each μ: start with empty interval and iteratively grow in direction of higher likelihood ratio $R = L(x|\mu)/L(x|\hat{\mu})$ with $\hat{\mu} \geq 0$
- No flip-flopping due to transition from upper limit to central interval
- No empty intervals

See Robert Cousins' talk on Monday, 15:45 CEST

CL_s approach

- Criticism of FC method
  - May give better limit for experiment with higher expected background
- CL$_s$ generalised originally Bayesian limit for counting experiments
  - Classic derivation offered by Zech, Nucl. Instrum. Meth. A 277 (1989) 608, but not frequentist in Neyman sense, see comment by Highland, NIM A 398 (1997) 429 and reply by Zech, NIM A 398 (1997) 431
- Counts replaced with likelihood ratio test statistic $t = -2\ln[L_{s+b}/L_b]$
  - L_{s+b} likelihood of signal and background fit
  - L_b likelihood of background-only fit
- Set limit on s: Solve $\text{CL}_s(s) = 1 - \text{CL}$ for s_{max}

$$\text{CL}_s(s) = \frac{P(t \leq t_{\text{obs}}; s)}{P(t \leq t_{\text{obs}}; 0)} = \frac{\text{CL}_{s+b}}{\text{CL}_b}$$

Read, J. Phys. G 28 (2002) 2693-2704

- Arbitrary nuisance parameters can be included
  - Maximise likelihoods L_{s+b} and L_b over nuisance parameters
- No solution for flip-flopping
- Practical issues
  - t distribution often computed from simulations to get $P(t \leq t_{\text{obs}})$
  - Each computation of t requires maximum-likelihood fit
  - Simulation of $P(t)$ requires many generated data samples for several values of parameter μ
- Options to reduce computational burden
  - Binned fits instead of unbinned fits
  - Use of asymptotic formulas (next slide)

Specialised likelihood ratio test statistics

- Cowan, Cranmer, Gross, Vitells, Eur.Phys.J.C 71 (2011) 1554 studied test statistics for fits to histograms
  - Ansatz $E[n_i] = \mu\, s_i(\vec{\theta}_s) + b_i(\vec{\theta}_b)$ for bin i with nuisance parameters $\vec{\theta} = \{\vec{\theta}_s, \vec{\theta}_b\}$
  - General statistic $t_\mu = -2\ln[L(\mu; \hat{\hat{\vec{\theta}}}(\mu))/L(\hat{\mu}; \hat{\vec{\theta}})]$
  - $\tilde{t}_\mu$ for measurement of non-negative signal
  - $\tilde{q}_0$ for discovery of non-negative signal
  - q_μ for upper limits
  - $\tilde{q}_\mu$ for upper limits on non-negative signal
- Systematic uncertainties handled as nuisance parameters
- Asymptotic formulas for their pdfs are given based on classic results from Wald and Wilks and so-called Asimov data sets
  - Useful for sensitivity studies to compute expected median limit
- Can be combined with CL_s limit setting or Feldman-Cousins approach

Open issues: Limit setting

- Flip-flopping remains an issue
  - Only avoided by Feldman-Cousins method
  - But Feldman-Cousins method incompatible with CL_s and any other non-Neyman construction like Bayesian limits
- Simulating distribution of likelihood ratio test statistic
  - Should nuisance parameters be varied within uncertainties or fixed to data estimates?
  - Should data-constrained nuisance parameters be treated differently from nuisance parameters that represent systematic uncertainties?

Handling and reporting systematic uncertainties

- Systematic uncertainties can be Frequentist or Bayesian
  - Frequentist example: calibration parameter from control measurement
  - Bayesian example: choice of a particular background model
- Expressed in σ, but usually no well-defined confidence level for intervals
  - Chebyshev's inequality applies: $1 - 1/k^2$ of results must be within $k\sigma$
- Guiding principle: consistency of statistical and systematic uncertainties
- Do not estimate systematic uncertainties overly conservative
- Distinguish checks from systematic variations
- Only failed checks should add to total systematic uncertainty

See Roger Barlow's talk on Monday, 14:45 h CEST

Barlow (2002), "Systematic errors: Facts and fictions", hep-ex/0207026
Barlow (2019), "Practical Statistics for Particle Physics", arXiv:1905.12362v1

Rules for discrete systematic variations

"These are just ballpark estimates. Do not push them too hard." (RB)

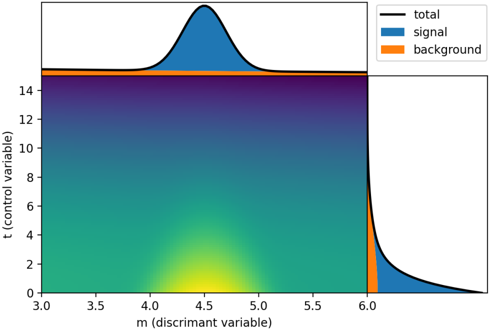
- Systematic uncertainty should behave like standard deviation
  - People will use it in least-squares fits and gaussian pdfs
- Distinguish "reasonable" and "extreme" variations
- Reasonable variation
  - Variance is $\frac{1}{N-1}\sum_i (R_i - \bar{R})^2$
  - Distribution-free
- Extreme variations
  - Extreme ends of assumed uniform distribution
  - Variance is $(R_{\max} - R_{\min})^2/12$

Two results R_1, R_2	reasonable variation	extreme variation
None preferred	$\bar{R} \pm \lvert R_1 - R_2 \rvert$	$\bar{R} \pm \lvert R_1 - R_2 \rvert/\sqrt{12}$
R_1 preferred	$R_1 \pm \lvert R_1 - R_2 \rvert$	$R_1 \pm \lvert R_1 - R_2 \rvert/\sqrt{6}$

Open issues: systematic uncertainties

- How to include discrete variations in likelihood profiling?
  - Example: Changing background or signal model
  - Discrete variations cannot be handled by nuisance parameter
  - Suggested solution discrete profiling: Dauncey, Kenzie, Wardle, Davies, JINST 10 (2015) P04015
- Likelihood profiling or marginalisation?
  - Profiling (Frequentist): Applicable to uncertainties that originate from measurements in control samples (detector calibration, beam luminosity, etc.), see Cowan, Cranmer, Gross, Vitells, Eur.Phys.J.C 71 (2011) 1554
  - Marginalisation (Bayesian): Some systematic uncertainties are Bayesian in nature, see Cousins, Highland, Nucl.Instrum.Meth.A 320 (1992) 331 for application to limit setting

sPlot method (aka sWeights)



- Signal and background events with variables m and t (t may be multi-dimensional)
- Signal and background each independent in m and t

$$f(m,t) = z\, g_s(m)\, h_s(t) + (1-z)\, g_b(m)\, h_b(t)$$

- sPlot technique: compute weights $w_s(m)$ to estimate parameters of $h_s(t)$ without modelling $h_b(t)$
  - Parametric models needed only for $g_s(m)$, $g_b(m)$, $h_s(t)$
  - Asymptotically unbiased and minimum variance for weights
- Very popular in flavour physics experiments

Pivk & Le Diberder, Nucl.Instrum.Meth.A 555 (2005) 356-369

sWeight trick

- Integrate

$$f(m,t) = z\, g_s(m)\, h_s(t) + (1-z)\, g_b(m)\, h_b(t)$$

over t to get

$$g(m) = z\, g_s(m) + (1-z)\, g_b(m)$$

- Fit this to get $\hat{z}, \hat{g}_s(m), \hat{g}_b(m)$
- sWeights with projection property $\int \mathrm{d}m\, w_s(m)\, f(m,t) = z\, h_s(t)$

$$w_s(m) = \frac{W_{bb}\, g_s(m) - W_{sb}\, g_b(m)}{(W_{ss} W_{bb} - W_{sb}^2)\, g(m)}$$

with $W_{xy} = \int \frac{g_x(m) g_y(m)}{g(m)} \mathrm{d}m$

- Estimates for W_{xy} can be computed from $\hat{z}, \hat{g}_s(m), \hat{g}_b(m)$

See Michael Schmelling's talk on Wednesday, 14:00 CEST

sWeights: (Semi-)open issues

- Classic sPlot technique only applicable if signal and background both factorise in m, t; independence needs to be demonstrated in practice
  - Insufficient: test for zero correlation of m, t
  - Proper: test for zero Kendall rank coefficient (credit to Sara Algeri)
- Combining sWeights with detection efficiencies
  - Detector efficiency may vary over m, t
  - Efficiency weights cannot be trivially combined with sWeights
- Non-factorising background in m, t
  - Factorisation usually good for signal but not necessarily for background
  - How to handle (mildly) non-factorising background?

HD, M. Kenzie, C. Langenbruch, M. Schmelling, paper in preparation with extensions to sPlot method to handle detector efficiencies and non-factorising background

Fits of (s)weighted data

- Binned fit
  - Per bin: Estimates of expectation $\sum_i w_i$ and variance $\sum_i w_i^2$
  - Use least-squares fit or maximum-likelihood fit with scaled Poisson distribution (better) Bohm & Zech, Nucl.Instrum.Meth.A 748 (2014) 1-6
  - Asymptotically unbiased
  - Biased when samples per bin become small (no info from empty bins)

- Unbinned fit
  - Maximise "weighted log-likelihood" $\ln \mathcal{L}_w(\theta) = \sum_i w_i \ln f(x_i; \theta)$
    - Not really a likelihood = product of probabilities, modified properties
    - Still proper estimator with proven asymptotic properties
  - Asymptotically unbiased
  - Modified covariance matrix $V_\theta = H^{-1} D H^{-1}$ Langenbruch, arXiv:1911.01303

$$H = \frac{\partial^2 \ln \mathcal{L}_w}{\partial^2 \theta}\Big|_{\theta=\hat{\theta}} \qquad D = \sum_i w_i^2 \frac{\partial \ln f(x_i)}{\partial \theta}\Big|_{\theta=\hat{\theta}} \frac{\partial \ln f(x_i)}{\partial \theta}\Big|_{\theta=\hat{\theta}}$$

See Christoph Langenbruch's talk on Wednesday, 14:45 CEST

(s)weighted fits: Open issues

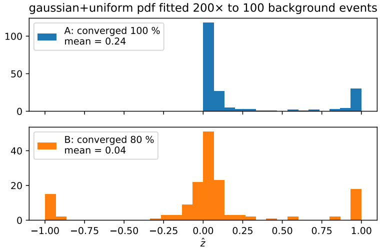
- Are weighted fits less accurate than full parametric fits?
  - Skipping background model $h_b(t)$ suggests loss of information
  - At least in some toys accuracy reduction is found to be negligible
- Modified covariance matrix V_θ assumes known w_i, but w_i are estimated from data and therefore deviate from asymptotic weights
  - Additional contribution to V_θ or contribution zero?
  - To be addressed in upcoming paper
- How to obtain confidence intervals with MINOS method?

$$\Delta \ln \mathcal{L}_w = ?$$

- How to combine with weighted with normal likelihood, e.g. to add gaussian nuisance parameter ϕ

$$f_\mathrm{corr} \ln \mathcal{L}_w - \frac{(\phi - \phi_0)^2}{2\sigma_\phi^2} \quad \text{with} \quad f_\mathrm{corr} = ?$$

Open issues: signal+background fits with vanishing signal



- Setting: maximum likelihood fit with

$$f(z, \theta_s, \theta_b) = z\, f_s(\theta_s) + (1 - z)\, f_b(\theta_b)$$

- Many background-only fits needed for e.g. CL$_s$

- Option A: use boundary condition $0 \leq z \leq 1$
  - Biased estimate $\hat{z}$ for $z \to 0$
- Option B: allow $z < 0$
  - Bias of $\hat{z}$ avoided, but ordinary fits become unstable
  - $z\, f_s(x_i, \theta_s) + (1 - z)\, f_b(x_i, \theta_b) > 0$ must be valid for all x_i
    - Condition not supported by MINUIT (bound on z depends on θ_s, θ_b)
  - Can this be fixed in MINUIT?
  - Different minimizer? Different analytical approach?

Please contact me (HD) if you interested in solving this.

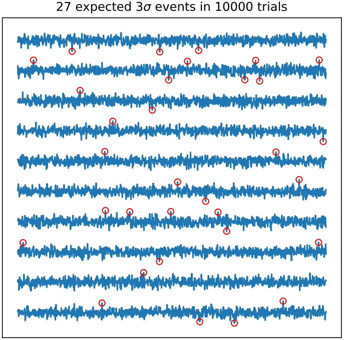
Open issues: GoF for unbinned fits

- Our common χ^2 GoF statistic requires binned data
- Unbinned fits
  - Cannot use likelihood value as GoF statistic, see Heinrich, PHYSTAT 2003, arXiv:physics/0310167
  - GoF statistic directly from fitted model and unbinned data?
- In binning of high-dimensional data: difficult to maintain enough counts per bin so that χ^2 statistic follows asymptotic distribution

See Sara Algeri's talk on Tuesday, 14:45 CEST

See Francois Le Diberder's talk on Wednesday, 15:45 CEST

Open issues: Look-elsewhere effect



Look-elsewhere effect

- Expected number of rare deviations from H_0 proportional to number of trials
  - Win German lottery $P = 7 \times 10^{-9}$
  - $N_{\text{trial}}/\text{yr} \approx 4 \times 10^8$ (7M regular players)
  - $P \times N_{\text{trial}}/\text{yr} \approx 3\,\text{wins/yr}$ (152 lottery millionaires in 2018)
- Important for model-independent searches
- Dilution factor computed by repeating experiment on H_0 simulations many times

- What if $N > 1$ unexpected peaks were found? How to compute dilution factor for this?
- Dilution factor for non-compact spaces: where to stop looking?

See André David's talk on Tuesday, 15:45 CEST

Concluding words

- Statistics is a science
  - Where methods with proven optimal properties are known, we use them
  - Conventions are needed when there is no clear optimal choice
  - Consistency/comparability important guide in making choices
    - Comparability to previous results
    - Comparability to fellow CERN experiments
- Many thanks for comments and discussion on this talk to:

  Roger Barlow, Olaf Behnke, Christoph Langenbruch, Louis Lyons, Michael Schmelling, and Diego Tonelli
- PHYSTAT has successful history in bringing experts together and to advance state-of-the-art

I am looking forward to a fruitful workshop!