Data Analysis with Theoretical Errors

Jérôme Charles

Centre de Physique Théorique (Marseille)

Fundamental Parameters from Lattice QCD, 2 September 2015

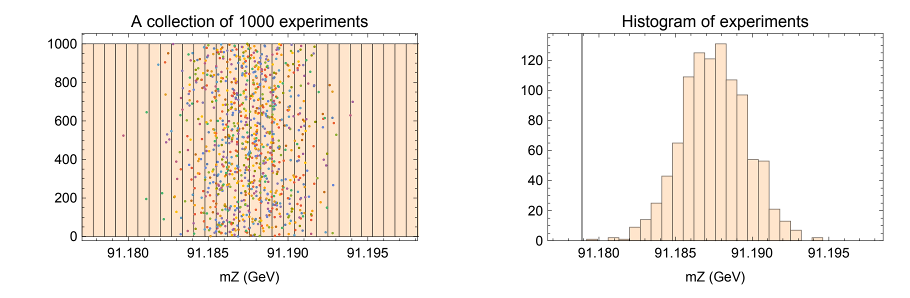
in collaboration with S. Descotes-Genon, V. Niess, L. Vale
and CKMfitter group

Warning: preliminary proposal !

Frequentist statistics in a nutshell

From measured (random) data, frequentist statistics answers the following question:
assuming some hypothesis $\mathcal{H}$ is true (the null hypothesis), are the observed data likely ?
Example: assuming the Standard Model is true, is my best fit value for m_{Z} likely ?
m_Z can be measured in e^+e^- collisions in the relevant invariant mass window. One can use the best fit value $\widehat{m_Z}$ of the resonance peak location as an *estimator* of the true value of m_Z. Estimators are functions of the data and thus are random variables. The estimator is said to be *consistent* if it converges to the true value when data statistics tends to infinity (*e.g.* maximum likelihood estimators are consistent). Another useful concept is the *bias*, which is defined as the difference between the average of the estimator among a large number of finite statistics experiments with the true value. Consistency implies that the bias vanishes asymptotically.

Assuming one can repeat many times the same experiment, one gets a collection of $\hat{m_Z}$ values. The histogram of this random sample brings information on the most likely value of m_Z and the average accuracy of the experiments.

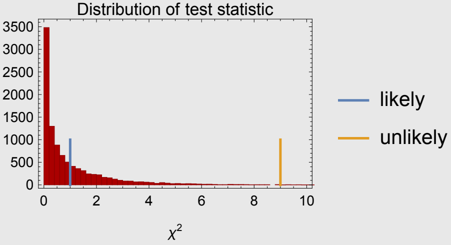


However in practice one only performs one (or a few) experiment(s). Thus one has to find a way to conclude whether the observation is likely from the information of a single experiment.

Repeated experiments and p-value

Whether given data are likely or not is usually quantified using a *test statistics* t, which is a function of data X such that *e.g.* low values supports the null hypothesis $\mathcal{H}$ whereas large values go against it.

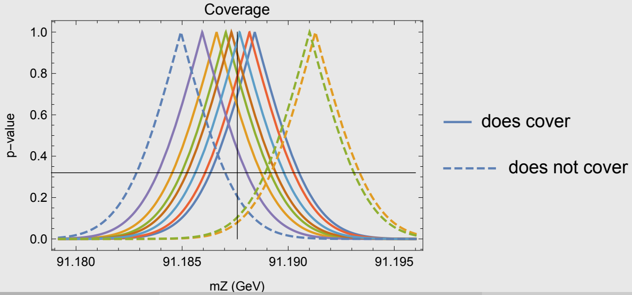
Then from the distribution of X one may compute the distribution of $t(X)$, as well as the probability $p(X_0)$ that the value $t(X)$ of a (often fictitious) repeated experiment is larger than the observed value $t(X_0)$: if $p(X_0)$ is large (small) it means that $t(X_0)$ is small (large) with respect to 'typical' values of $t(X)$, and thus that the observed data are in good (bad) agreement with the null hypothesis.



Confidence intervals and coverage

The hypothesis $\mathcal{H}$ is said to be *simple* if it completely specifies the distribution of the data X. In this case the p-value constructed from $t(X)$ is nothing else than the CDF of t, and thus the p-value is uniformly distributed with the observed value X_0.

In case of a numeric hypothesis $\mathcal{H} : X_{\mathrm{true}} = \mu$, the p-value curve allows the construction of *confidence intervals*: the interval of μ defined by $p \geq 1 - \mathrm{CL}$ contains X_{true} at the frequency CL, as follows from the uniformity of p.



Theoretical uncertainties

It often happens that an observable parameter is only related to a fundamental quantity through auxiliary (*nuisance*) parameters. Typical example: hadronic transitions depend on both quark fundamental couplings and hadronic matrix elements.

It would not be a problem if these hadronic matrix elements could be computed exactly. This is not the case in QCD ! Lattice QCD approach has the advantage that part of the computation uncertainty is of statistical (Monte-Carlo) origin; however others sources of uncertainties are not statistical: continuum extrapolation, finite volume, mass inter/extrapolations, partial quenching. . .

On the experimental side also there are model-dependent systematic uncertainties; however they are often controlled by auxiliary measurements, so that the usual consensus is to treat them on the same footing as the statistical contributions (usually modelled by Gaussian random variables).

The problem

How to interpret $\Delta(\mathrm{theo})$ in $X = X_0 \pm \sigma(\mathrm{exp}) \pm \Delta(\mathrm{theo})$?
as a pseudo-random error ? It might be justified in a fictitious world where one could do the same computation many times with a different technique such that it gives a different estimate around the true value; one would then end up with the widely used naive Gaussian approach, unless there is an argument to choose another pseudo-random distribution.
as a fixed bias ? One then defines

$$\delta = X_{\mathrm{true}} - \lim_{\sigma \to 0} X_0$$

where δ is a (variable) nuisance parameter related to the (fixed) theoretical uncertainty Δ. The above equation actually means that X_0 is not a consistent estimator, as the bias does not vanish asymptotically.

The nuisance δ-approach

Then from the frequentist point of view one tests the following null hypothesis:

$$\mathcal{H} : X_{\mathrm{true}} = \mu$$

through the construction of a p-value from the distribution of a given test statistic with

$$X_0 \sim \mathcal{N}(\mu + \delta, \sigma)$$

In this case $\mathcal{H}$ is *composite*, as one needs to know the value of δ in addition of μ to compute the distribution of X_0.

The quadratic statistic

Important point: the choice of the test statistic is free (as long as it models the null hypothesis one wants to test); it is perfectly legitimate to take the widely used quadratic form

$$\begin{aligned}\Delta\chi^2 &= \mathrm{Min}_\delta \left[\left(\frac{X_0-\mu-\delta}{\sigma}\right)^2 + \left(\frac{\delta}{\Delta}\right)^2\right] \\ &= \frac{(X_0-\mu)^2}{\sigma^2+\Delta^2}\end{aligned}$$

In the multidimensional case the quadratic form is the only one that keeps its form after minimization over some of the parameters

With $X_0 \sim \mathcal{N}(\mu + \delta, \sigma)$ the distribution of $\Delta\chi^2$ is a (rescaled) *non central χ^2 distribution*, with non centrality parameter $(\delta/\sigma)^2$.

The p-value is obtained from the cumulative distribution function, which is a Marcum *Q*-function, that reduces to the error function in one dimension.

$$p_\delta(\mu) = \frac{1}{2}\left[2 + \mathrm{Erf}\left(\frac{\delta - |\mu - X_0|}{\sqrt{2}\sigma}\right) - \mathrm{Erf}\left(\frac{\delta + |\mu - X_0|}{\sqrt{2}\sigma}\right)\right]$$

It depends explicitly on δ (but not Δ): one can take the supremum value for δ/Δ in some ensemble Ω, *e.g.*, $\Omega_1 = [-1, +1]$ (ambitious) or $\Omega_3 = [-3, +3]$ (reasonable).

Indeed this supremum p-value will allow to construct correct confidence intervals if and only if the (unknown) true value of δ/Δ belongs to the chosen Ω.

Conversely, if the true value of δ/Δ is outside the chosen Ω, the confidence intervals will suffer from undercoverage: one will exclude the null hypothesis ‘too quickly’.

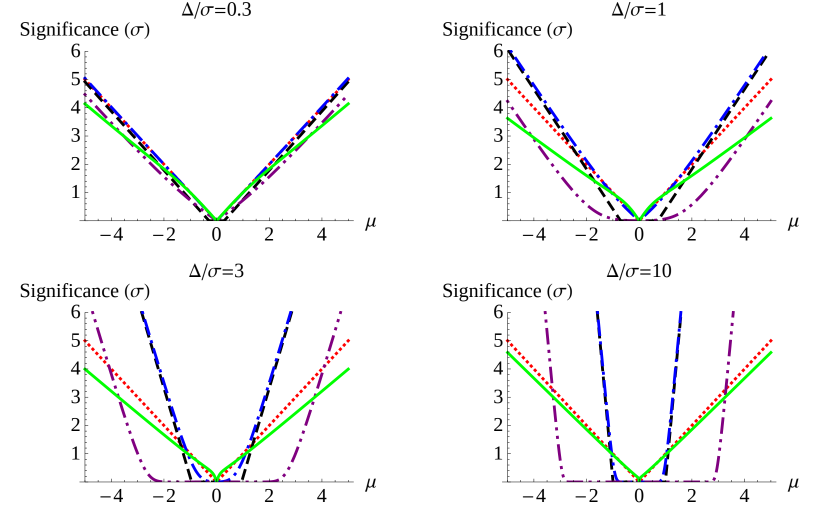
The external-δ approach

Another possibility is to forget, in a first step, that δ is unknown: thus one naturally tests the null hypothesis

$$\mathcal{H}' : X_{\mathrm{true}} = \mu + \delta$$

One gets a collection of p-values $p_\delta(\mu)$, and one has to define a procedure to combine them. An obvious possibility is to take the envelope over some ensemble Ω.

In 1D one recovers the CKMfitter Rfit Ansatz, with a plateau at $p = 1$ (also similar to the scan method).



red: naive Gaussian (nG), black: Ω_1-external, blue: Ω_1-nuisance, purple: Ω_3-nuisance

Choice of Ω

Problem with fixed Ω ensemble: the p-value (at large values) gets crazingly large when δ/Δ is varied in Ω_3 instead of Ω_1. Is the Ω_1 choice conservative ?

Key question: why bother to ensure good coverage for all $\delta/\Delta \in \Omega_3$ if one is only interested in a 1σ statement (metrology) ?

In contrast, is it safe to, *e.g*, exclude the Standard Model at 5σ is this statement assumes that all theoretical biases are within their 1Δ range ?

Possible solution: adapt Ω to the computed p-value; the smaller p, the larger Ω, and vice-versa.

This ‘feedback’ procedure does not blow out because the p-value is an increasing function of Ω.

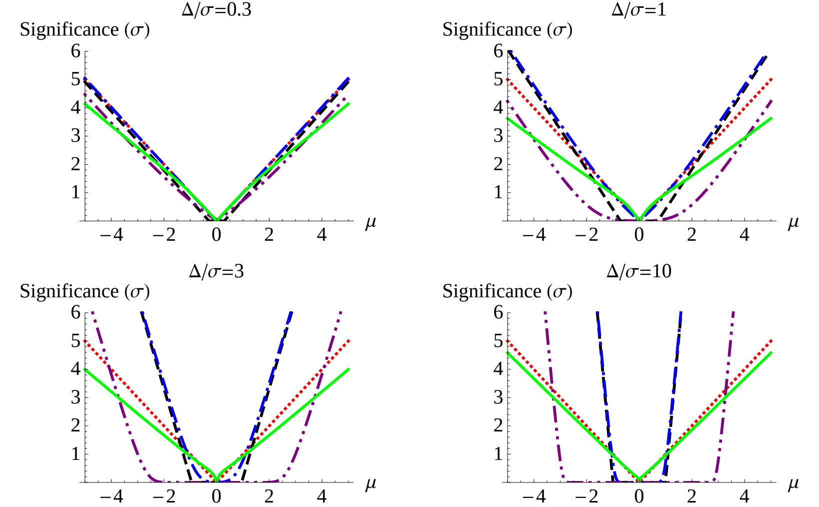
The adaptive Ω interval

The choice of how Ω depends on p is again rather free; however it looks very natural to choose the "would-be" $1-p$ confidence interval for δ/Δ, *i.e.* $\Omega(1-0.68) = [-1,+1]$ etc. In this case $\Omega(p)$ is independent of i when there are several δ_i.

Hence one has to maximize p_δ for δ/Δ varying in an interval that itself depends on p_δ; since p_δ is an increasing function of $|\delta|$, it amounts to solve the implicit equation

$$\begin{aligned} p &= \frac{1}{2}\left[2 + \mathrm{Erf}\left(\frac{\delta - |\mu - X_0|}{\sqrt{2}\sigma}\right) - \mathrm{Erf}\left(\frac{\delta + |\mu - X_0|}{\sqrt{2}\sigma}\right)\right] \\ (\delta/\Delta)^2 &= 2\mathrm{Erf}^{-1}(1-p) = n_\sigma(p)^2 \end{aligned}$$

It looks like an horrible equation, and indeed it is. However it is easily solvable numerically.



red: naive Gaussian (nG), black: Ω_1-external, blue: Ω_1-nuisance, purple: Ω_3-nuisance, green: adaptive Ω

Interpretation

The correct interpretation of this p-value is: p is a valid p-value if the true (unknown) value of δ/Δ belongs to the "would be" $1-p$ confidence interval around 0.

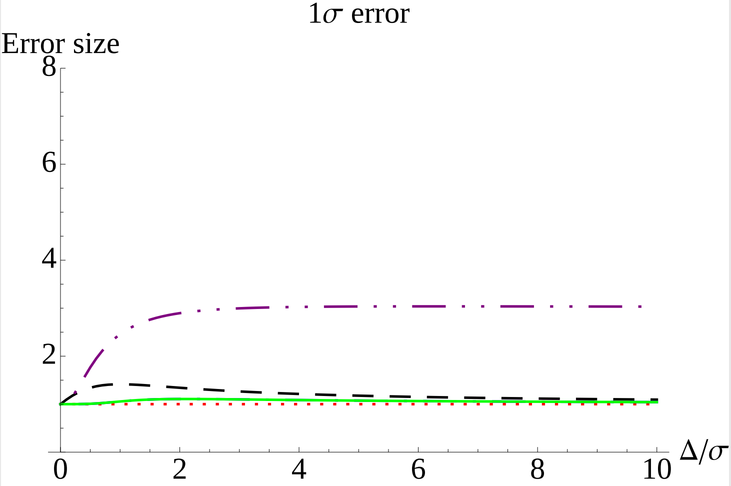
Alternative interpretation: if the true (unknown) value of δ/Δ belongs to the "would be" $1-\beta$ confidence interval around 0, then p is a valid p-value as long as it is sufficiently small $p \leq \beta$.

Thus this approach is agressive at large p-values (metrology), and conservative at small p-values (New Physics tests).

This is not a standard coverage criterion: one can use *adaptive* coverage, and *adaptively valid* p-value, to name this new concept.

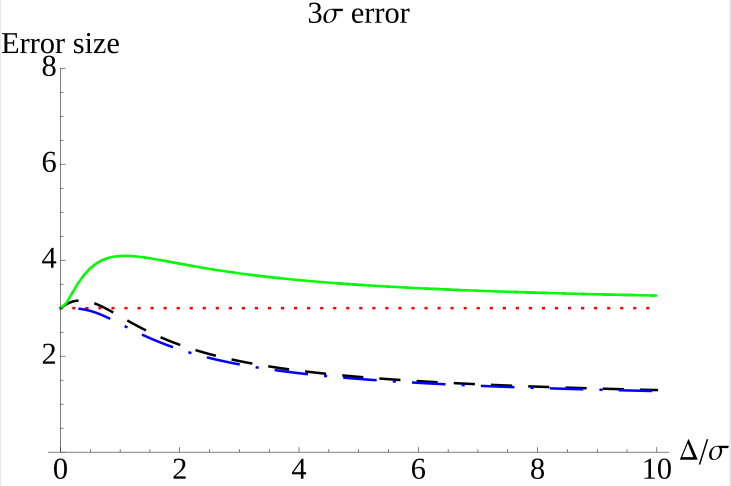
With this approach one can do a robust evidence (resp. discovery) statement, under the mild assumption that the true value of δ belongs to $[-3\Delta, +3\Delta]$ (resp. $[-5\Delta, +5\Delta]$) !

Size of confidence intervals



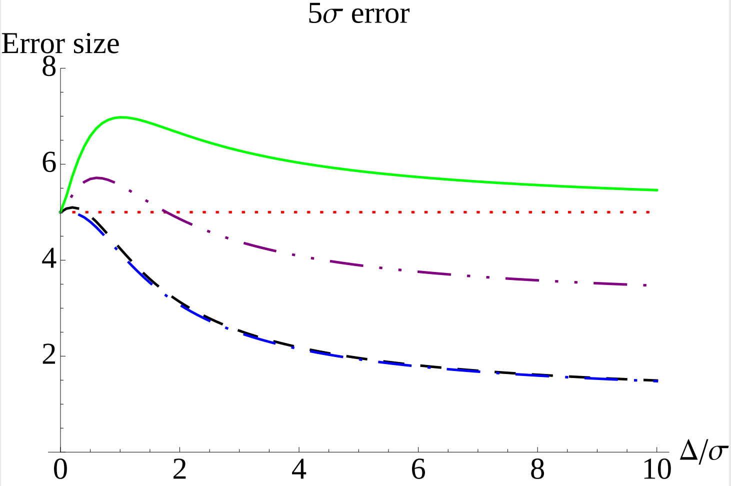
red: nG, black: Ω_1-external, blue: Ω_1-nuisance, purple: Ω_3-nuisance, green: adaptive Ω

Size of confidence intervals



red: nG, black: Ω_1-external, blue: Ω_1-nuisance, purple: Ω_3-nuisance, green: adaptive Ω

Size of confidence intervals



red: nG, black: Ω_1-external, blue: Ω_1-nuisance, purple: Ω_3-nuisance, green: adaptive Ω

Comparison with the naive Gaussian approach

In one dimension, the adaptive approach is numerically not very far from the nG method; maximum difference occurs for $\Delta/\sigma = 1$ (up to 50% larger error at a given CL).

The important point is that the adaptive approach allows a well-defined frequentist statement, while the nG does not.

The $g-2$ discrepancy

$$a_\mu^{SM} - a_\mu^{exp} = (288 \pm 63_{\mathrm{stat}} \pm 49_{\mathrm{theo}}) \times 10^{-11}$$

One finds the following pulls:

naive Gaussian	3.6σ
Ω_1-external	3.8σ
Ω_1-nuisance	4.0σ
adaptive Ω	2.7σ

Generally speaking, with $\Delta/\sigma = 1$, to see a evidence (resp. discovery) effet with adaptive Ω one needs a 4.1σ (resp. 7.0σ) effect with nG.

The multidimensional linear case

In a linear model the bias on a given parameter μ is a linear combination of all contributing biases $\delta_\mu = \sum_i w_i \delta_{X_i}$.

To compute the theoretical error on μ, Δ_μ, in a frequentist way, all we have to do is to choose a n-dimensional $\Omega^{(n)}$ space in which we maximize δ_μ.

The most natural generalization of the 1D interval is the nD hypercube; another possibility is the nD hyperball.

One can show that:

$$\delta_\mu = \sum_i w_i \delta_{X_i} \Rightarrow \Delta_\mu = \sum_i |w_i| \Delta_i$$

for nD hypercube, and

$$\delta_\mu = \sum_i w_i \delta_{X_i} \Rightarrow \Delta_\mu = \sqrt{\sum_i |w_i|^2 \Delta_i^2}$$

for nD hyperball.

Hypercube vs. hyperball

Thus the hypercube (resp. hyperball) corresponds to the linear (resp. quadratic) combination of individual uncertainties.

One may argue that the linear choice is too conservative, as it allows several δ_i's to lie is at their boundaries, whereas one may argue that the quadratic choice is not conservative.

Pure statistical arguments cannot solve this dilemma: this is an arbitrary (but well-defined) choice that must be made by the physicist.

Combination of n determinations of the same quantity

Important (but trivial) point: the linear addition scheme (hypercube) is the only one where the average of different determinations of the same quantity cannot lead to a weighted theoretical uncertainty that is smaller than the smallest uncertainty among all determinations.

Let's consider averaging X_1 with X_2, with $\sigma_1 = \sigma_2 = \Delta_1 = \Delta_2 = 1$, then the weighted bias is $\delta = (\delta_1 + \delta_2)/2$, which reaches $\Delta = 1$ only when both $\delta_1 = \delta_2 = 1$: cutting the $(+1, +1)$ corner of the square will necessarily lead to Δ smaller than 1

Price to pay ! either live with large errors coming from the linear addition of many uncertainties, or with the possibility that the averaged uncertainty among different determinations of the same quantity is more precise than each individual one.

Correlations

Different (experimental, lattice) determinations of the same quantity can be correlated. Known correlations between statistical uncertainties can be treated with the statistical covariance matrix in the usual way.

It often happens on the lattice that there are unknown (or not precisely known) correlations between theoretical uncertainties. In this case a conservative choice is the assumption that these correlations are $\sim 100\%$: in the bias approach it amounts to share a given δ bias parameters between different determinations.

Separation of statistical and theoretical contributions

In a linear model, where the data X is linearly dependent on the parameter of interest μ, the different nature of statistical and theoretical uncertainties allow to compute them separately, whatever the dimensionality of the problem.

Combination of marginally compatible measurements

When two determinations of the same quantity show marginal agreement, one may argue that at least one uncertainty is underestimated. For this reason the PDG traditionally uses a compatibility recipe that amounts to rescale the errors so that the combined χ^2 is 1. One may design a similar recipe if one thinks instead that the disagreement is due to theoretical uncertainties.

However, in any case, this kind of rescaling is ambiguous, especially from the point of view of global analyses. Indeed in a global fit one cannot perform such a rescaling without making the fit useless. The problem is then that there is no general argument that tells that different determinations of the same input are to be averaged before doing the global fit (with possibility of rescaling), or inside it (without possibility of rescaling).

Again pure statistical arguments cannot resolve these ambiguities.

The example of $B_K^{\bar{\mathrm{M}}\mathrm{S}}(2\mathrm{GeV})$

Inputs

Reference	Mean	Stat	Theo
ETMC10	0.532	± 0.019	±0.003 ± 0.007 ± 0.003 ± 0.008 ± 0.005
LVdW11	0.5572	± 0.0028	±0.0045 ± 0.0033 ± 0.0039 ± 0.0006 ± 0.0134
BMW11	0.5644	± 0.0059	±0.0022 ± 0.0008 ± 0.0006 ± 0.0006 ± 0.0002 ± 0.0056
RBC-UKQCD12	0.554	± 0.008	±0.007 ± 0.003 ± 0.012
SWME14	0.5388	± 0.0034	±0.0237 ± 0.0048 ± 0.0005 ± 0.0108 ± 0.0022 ± 0.0016 ± 0.0005

Combination

Method	Average	$1\ \sigma$ CI	$2\ \sigma$ CI	$3\ \sigma$ CI
nG	0.5577 ± 0.0063	0.5577 ± 0.0063	0.5577 ± 0.0126	0.5577 ± 0.0189
naive Rfit	0.5562 ± 0.0120 ± 0.0018	0.5562 ± 0.0138	0.5562 ± 0.0258	0.5562 ± 0.0379
educ Rfit	0.5562 ± 0.0020 ± 0.0100	0.5562 ± 0.0120	0.5562 ± 0.0139	0.5562 ± 0.0159
1-hypercube	0.5577 ± 0.0038 ± 0.0176	0.5577 ± 0.0193	0.5577 ± 0.0240	0.5577 ± 0.0281
adapt hyperball	0.5577 ± 0.0038 ± 0.0050	0.5577 ± 0.0068	0.5577 ± 0.0165	0.5577 ± 0.0257

Conclusion

The bias definition of theoretical uncertainties features very good frequentist properties.

It leads to a transparent splitting of the uncertainty in terms of the statistical and the theoretical contributions.

It makes explicit the unavoidable arbitrariness in combining theoretical uncertainties.

Still it remains well defined, both in 1D and nD.

Linear vs. quadratic combination is a choice to be made by the physicist, depending on his own prejudice.

The adaptive treatment of the bias ensemble is a new concept that allows more flexibility in the interpretation of p-values.