Higher-order asymptotic corrections and their application to the Gamma Variance Model

Alessandra Brazzale[1], Enzo Canonero[2], Glen Cowan[2]

[1] Department of Statistical Sciences University of Padova, Italy
[2] Physics Department, Royal Holloway, University of London, U.K.

April 20, 2023

DRAFT 0.99

Abstract

We present improved methods for calculating confidence intervals and p-values in specific class of statistical model that can incorporate uncertainties in parameters that themselves represent uncertainties (informally, "errors on errors") called the Gamma Variance Model (GVM). This model contains fixed parameters, generically called ε, that represent the relative uncertainties in estimates of standard deviations of Gaussian distributed measurements. If the ε parameters are small, one can construct confidence intervals and p-values using standard asymptotic methods. This is formally similar to the familiar situation of a large data sample, in which estimators for all adjustable parameters have Gaussian distributions. Here we address the important case where the ε parameters are not small and as a consequence the asymptotic distributions do not represent a good approximation. The improvements are based on the technology of higher-order asymptotics (p^* approximation and Bartlett correction).

Keywords: Gamma Variance Model, higher-order asymptotics, p^* approximation, Bartlett correction

1 Introduction

In experimental sciences such as Particle Physics one collects data, here denoted as $\boldsymbol{y}$, and seeks to make inferences about a hypothesis H that defines the probability distribution for the data, $P(\boldsymbol{y}|H)$. Often $P(\boldsymbol{y}|H)$ is indexed by a set of *parameters of interest* $\boldsymbol{\mu}$ and by a set of *nuisance parameters* $\boldsymbol{\theta}$, thus $P(\boldsymbol{y}|H) = P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta})$. The parameters of interest are the main objective of the analysis, whereas nuisance parameters are often introduced to account for systematic uncertainties in the model.

We focus here on frequentist tests of the hypothesized parameters that are based on the likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\boldsymbol{y}|H) = P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta})$. These tests lead to confidence intervals or regions for the parameters of interest as well as p-values that quantify goodness of fit. To find these results, one requires the sampling distribution of test statistics that are obtained from the likelihood function and are described in greater detail below. For appropriately defined test statistics, the corresponding distributions can often be found using asymptotic results based on theorems due to Wilks [1] and Wald [2] (see, e.g., [3]). The asymptotic distributions are valid in specific limits, which usually correspond to having a large data sample, whose size we will denote generically as n.

In this paper we are interested specifically in the case where n is not sufficiently large for the asymptotic distributions of the relevant test statistics to represent a good approximation. In such problems one could use Monte Carlo methods to obtain the distributions, but this involves additional time-consuming computation. Instead, one can modify the test statistic using the higher-order asymptotic methods, specifically, the p^* approximation of Barndorff-Nielsen [5] and the Bartlett correction [9], as described, e.g., in [10, 11]. With these methods, the distribution of the modified statistic becomes closer to the asymptotic form, allowing one to find confidence intervals and p-values without use of Monte Carlo.

In this paper we consider applications of higher-order asymptotic methods to the Gamma Variance Model (GVM), which was proposed in Ref. [12]. In the GVM, measured values are modeled as following Gaussian distributions with a mean that depends on the parameters of the problem, and with variances σ^2 whose values are themselves not certain. The variances as well are thus taken as adjustable parameters, and the values one would assign to them are treated as measurements that follow a Gamma distribution with parameters α and β (see Sec. 4 below). These parameters are chosen so that the Gamma distribution's relative width reflects the desired uncertainty on σ^2. This is quantified using the quantity $\varepsilon = 1/2\sqrt{\alpha}$, which to first approximation is the relative uncertainty on the estimate of the standard deviation σ, informally referred to as the "error on the error".[1]

In the Gamma Variance Model there is a correspondence between the error-on-error parameters ε and an effective sample size n of

$$n = 1 + \frac{1}{2\varepsilon^2} \; . \tag{1}$$

That is, the large-sample limit corresponds to the case where $\varepsilon \to 0$ and thus the values of σ are accurately estimated. For many analyses, however, the assigned values of standard deviations for individual measurements may easily be uncertain at the level of several tens of percent or more. In this case the effective sample size is low and thus the asymptotic distributions of likelihood-based test statistics are not necessarily valid. The goal of this paper is to apply higher-order asymptotics to this model and thus achieve more accurate confidence levels and p-values.

In Sec. 2 we briefly review the basic techniques for finding confidence intervals and p-values in a general likelihood-based analysis. Section 3 describes how in general these techniques can be improved using the methods of higher-order asymptotics and in Sec. 4 we recall the important properties of the Gamma Variance Model. In Sec. 5 we apply higher-order asymptotic corrections to a simple example of the GVM based on a single measurement, and in Secs. 6 and 7 to averages of measured values. A summary and conclusions are given in Sec. 8.

2 Parameter inference using the profile likelihood ratio

In this section we review the basic technology used to find confidence intervals and p-values from test statistics derived from the likelihood ratio by using the first-order asymptotic distributions

[1] In earlier references, e.g., [12, 14], the parameter ε was denoted as r.

based on Wilks' theorem. Further details on these methods as applied in Particle Physics analyses can be found, e.g., in Ref.[3].

In statistical data analysis, the central object needed to carry out inference related to the parameters of interest $\boldsymbol{\mu}$ using measured data $\boldsymbol{y}$ is the likelihood function: $L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta})$. Here nuisance parameters $\boldsymbol{\theta}$ are introduced to account for systematic uncertainties. Their presence enlarges the model's parameter space and allows it to better approximate the truth, even though this reduces the sensitivity to the parameters of interest.

In frequentist statistics, a test of hypothesized parameter values can be carried out by defining a test statistic based on the (profile) likelihood ratio

$$w_{\boldsymbol{\mu}} = -2 \log \frac{L(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})} \,. \tag{2}$$

Here $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\theta}}$ are the Maximum Likelihood Estimators (MLEs) for the parameters of interest and the nuisance parameters, respectively, and $\hat{\hat{\boldsymbol{\theta}}}$ are the profiled (or constrained) estimators of the nuisance parameters, given by the values of $\boldsymbol{\theta}$ that maximize the likelihood for a fixed value of $\boldsymbol{\mu}$. The likelihood ratio is used to test the compatibility of a value of $\boldsymbol{\mu}$ with the experimental data, with greater $w_{\boldsymbol{\mu}}$ corresponding to increasing incompatibility between the hypothesized $\boldsymbol{\mu}$ and the data.

To mitigate the negative impact of the nuisance parameters, one often performs independent control measurements, here denoted as $\boldsymbol{u} = (u_1, \ldots, u_N)$, that provide information on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$. Here we will suppose, as is often the case, that these are direct estimates of the nuisance parameters $\boldsymbol{\theta}$ and are treated as independently Gaussian distributed random variables with standard deviations $\boldsymbol{\sigma_u} = (\sigma_{u_1}, ..., \sigma_{u_N})$. The resulting likelihood is, therefore,

$$L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \times P(\boldsymbol{u}|\boldsymbol{\theta}) = P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \times \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{u_i}^2}} \, \exp\left[-\frac{(u_i - \theta_i)^2}{2\sigma_{u_i}^2} \right] , \tag{3}$$

or equivalently, the log-likelihood is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\theta}) = \log L(\boldsymbol{\mu}, \boldsymbol{\theta}) = \log P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) - \sum_{i=1}^{N} \frac{(u_i - \theta_i)^2}{2\sigma_{u_i^2}} + C \,, \tag{4}$$

where C is a constant that can be dropped since it does not depend on the parameters. The likelihood ratio defined by Eq. (2) can be used to derive a confidence region for the parameters of interest $\boldsymbol{\mu}$ (or a confidence interval if there is just one parameter of interest).

The p-value for a hypothesized value $\boldsymbol{\mu}$ is found from the statistic $w_{\boldsymbol{\mu}}$ as

$$p_{\boldsymbol{\mu}} = \int_{w_{\boldsymbol{\mu},\mathrm{obs}}}^{\infty} f(w_{\boldsymbol{\mu}}|\boldsymbol{\mu}, \boldsymbol{\theta}) dw_{\boldsymbol{\mu}} = 1 - F[w_{\boldsymbol{\mu},\mathrm{obs}}] \,, \tag{5}$$

where $f(w_{\boldsymbol{\mu}}|\boldsymbol{\mu}, \boldsymbol{\theta})$ is the probability density function of $w_{\boldsymbol{\mu}}$ under the hypothesis that $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ are the true parameters of the model, $w_{\boldsymbol{\mu},\mathrm{obs}}$ is the observed value of the likelihood ratio, and F is the cumulative distribution of $w_{\boldsymbol{\mu}}$. The boundary of the confidence region for $\boldsymbol{\mu}$, with confidence level $1 - \alpha$, is found from the p-value of Eq. (5) by solving $p_\mu = \alpha$. This gives a region in parameter space that satisfies Prob($\boldsymbol{\mu} \in$ confidence regions) $\geq 1 - \alpha$.

In many realistic applications, the computation of the probability density function $f(w_{\boldsymbol{\mu}}|\boldsymbol{\mu}, \boldsymbol{\theta})$ is a major challenge since it is usually not known in closed form. Monte Carlo simulations are often used to compute it, but this can be very time-consuming for complex models with many measurements. Moreover, to ensure that $p_{\boldsymbol{\mu}}$ is greater than α for all $\boldsymbol{\theta}$ values, one needs to know $f(w_{\boldsymbol{\mu}})$ for every point in the $\boldsymbol{\theta}$ space, which further complicates the problem.

However, it is possible to avoid the numerical computation of $f(w_{\boldsymbol{\mu}})$ in the *asymptotic limit*, which is defined as the limit where all the MLEs of the model are Gaussian distributed. This limit is typically reached when the experimental sample size n approaches infinity, i.e., in the so-called *large sample limit*. In this limit, the MLEs have a Gaussian distribution with an error term of order $\mathcal{O}(n^{-1/2})$. Moreover, according to Wilks' theorem [1], $w_{\boldsymbol{\mu}}$ follows a chi-square distribution with M degrees of freedom, where M is the dimension of the parameters of interest space. Wilks' theorem thus states that the asymptotic distribution of $w_{\boldsymbol{\mu}}$ is independent of nuisance parameters, and is given by

$$f(w_{\boldsymbol{\mu}})|\boldsymbol{\mu}, \boldsymbol{\theta}) = \chi^2_M + \mathcal{O}(n^{-1}) \,. \tag{6}$$

This allows us to compute the confidence regions by substituting $f(w_{\boldsymbol{\mu}})$ with χ^2_M. Note that the Gaussianity of the MLEs can also be controlled by other parameters of the likelihood, as shown in Sec. 4.

In the asymptotic limit it is also possible to show that the likelihood ratio can be approximated using a quadratic expansion,

$$w_{\boldsymbol{\mu}} = (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})V^{-1}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})\,, \tag{7}$$

where $V_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$ is found using the "observed information matrix" $j_{ij}(\boldsymbol{\mu})$,

$$V_{ij} = -j_{ij}(\hat{\boldsymbol{\mu}}) = -\frac{\partial^2 \ell}{\partial \mu_i \partial \mu_j}\Big|_{\hat{\boldsymbol{\mu}}}\,. \tag{8}$$

This equation says that in the asymptotic limit, the confidence region in the parameters of interest space is a hyper-ellipsoid centered in $\hat{\boldsymbol{\mu}}$. Equivalently the profile likelihood can be approximated as

$$\ell(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}}) = \ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})V^{-1}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})\,, \tag{9}$$

Deviations from the quadratic approximations of the likelihood root and the profile likelihood are expected when the conditions of the asymptotic limit ar not satisfied.

3 Higher-order asymptotic corrections

As shown in the previous section, the asymptotic distributions greatly simplify the procedure for finding p-values and confidence regions. These distributions do not represent valid approximations, however, if the MLEs of the model parameters are not Gaussian distributed, which usually happens when the experimental sample size is small. In cases where the asymptotic limit cannot be used, one may resort to MC simulations as an alternative. However, this approach is computationally expensive, therefore alternative analytical approaches are preferred.

Broadly speaking, there are two possible strategies to obtain test statistics with known distributions: one can either better approximate the distribution of the likelihood ratio, or one can adjust the test statistic itself such that its distribution is better approximated by its asymptotic formula, even for small sample sizes. An example of the former is given in Ref. [4]; in this paper, we focus on the latter approach. Specifically, this section explores two potential solutions, the p^* approximation [5, 6, 7, 8] and the Bartlett correction [9, 13].

In addition to the likelihood ratio test,

$$w_{\boldsymbol{\mu}} = -2\log\frac{L(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})} = 2\left[\ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})\right]\,, \tag{10}$$

the other relevant first-order statistic we want to study is the profile likelihood root test (here simply the "likelihood root"). This is defined as the square root of the likelihood ratio multiplied by the sign of $\hat{\mu} - \mu$:

$$r_\mu = \mathrm{sign}(\hat{\mu} - \mu)\sqrt{2\left[\ell(\hat{\mu}, \hat{\boldsymbol{\theta}}) - \ell(\mu, \hat{\hat{\boldsymbol{\theta}}})\right]}\,. \tag{11}$$

In contrast to the likelihood ratio, the likelihood root can be defined only when there is a single parameter of interest. Therefore, this statistic can be relevant in applications such as averages of scalar point estimates. The asymptotic distribution of the likelihood root is a standard Gaussian, with an error term of $\mathcal{O}(n^{-1/2})$:

$$f(r_\mu|\mu, \boldsymbol{\theta}) = \mathcal{N}(0, 1) + \mathcal{O}(n^{-1/2})\,. \tag{12}$$

Here n generally represents the sample size of the experiment, but it can also be another parameter of the likelihood that controls the convergence of the likelihood to the asymptotic limit. It should be noted that the error term in Eq. (12) is larger than the error on the asymptotic distribution of the likelihood ratio, as shown in Eq. (6). However, the likelihood ratio only allows for two-sided tests, while the likelihood root allows for one-sided tests as well.

The objective is to derive corrections for w_μ and r_μ such that the distributions of the refined statistics, denoted as w^*_μ and r^*_μ, are more accurately approximated by the asymptotic distributions

provided earlier, with error terms of order $\mathcal{O}(n^{-3/2})$ or smaller. Furthermore, these enhanced statistics will be designed to preserve the same local power as the original test statistics at the pertinent perturbative order, specifically $\mathrm{P}(S^* > x^*) = \mathrm{P}(S > x) + \mathcal{O}(n^{-3/2})$, where S represents one of the first-order statistics and S^* corresponds to its higher-order improved counterpart.

3.1 The p^* approximation

The asymptotic distributions of the likelihood root and the likelihood ratio are derived from the assumption that the MLEs are Gaussian in the asymptotic limit. However, this assumption is only valid up to error terms of order $\mathcal{O}(n^{-1/2})$. A major development in likelihood-based inference is to improve the distributions of the MLEs. For models with a single parameter μ, the Barndorff-Nielsen p^* approximation [5, 6, 7, 8] is the basic higher-order approximation to the distribution of $\hat{\mu}$. The p^* density function of $\hat{\mu}$ is given by

$$f(\hat{\mu}) = p^*(\hat{\mu}) = c\,|j(\hat{\mu})|^{1/2}\, e^{-w_\mu/2}\,. \tag{13}$$

Here, w_μ is the likelihood ratio defined in Eq. (2), c is a normalization constant equal to $\sqrt{2\pi} + \mathcal{O}(n^{-1})$, and j is the observed information matrix defined as $-\frac{\partial^2 \ell}{\partial \mu^2}$. The error term on the p^* approximation is of order $\mathcal{O}(n^{-3/2})$, which is a significant improvement compared to $\mathcal{O}(n^{-1/2})$ from the first-order Gaussian approximation. If the p^* approximation is expanded at order $\mathcal{O}(n^{-1/2})$, the Gaussian density of $\hat{\mu}$ is recovered. At order $\mathcal{O}(n^{-1/2})$, the likelihood ratio w_μ can be approximated using Eq. (7) as

$$w_\mu = (\hat{\mu} - \mu)^2 |j| + \mathcal{O}(n^{-1/2})\,, \tag{14}$$

where j is a constant since the likelihood is a quadratic function of μ at this order in n. Therefore the p^* approximation reduces to a Gaussian:

$$f(\hat{\mu}) = \frac{1}{\sqrt{2\pi}} |j|^{1/2} e^{-\frac{(\hat{\mu}-\mu)^2}{2|j|^{-1}}} + \mathcal{O}(n^{-1/2})\,. \tag{15}$$

The p^* approximation provides a way to improve the cumulative distribution of the likelihood root r_μ by reducing its error term. Specifically, integrating the p^* approximation, as shown in Ref. [8], yields the following expression for the cumulative distribution of r_μ,

$$F(r_\mu) = \Phi(r_\mu) + (r_\mu^{-1} - q_\mu^{-1})\phi(r_\mu) + \mathcal{O}(n^{-3/2})\,, \tag{16}$$

where ϕ is the standard normal density function and Φ is its cumulative distribution. The statistic q_μ is defined as

$$q_\mu = \left(\frac{\partial \ell}{\partial \hat{\mu}}\Big|_{\hat{\mu}} - \frac{\partial \ell}{\partial \hat{\mu}}\Big|_{\mu} \right) j(\hat{\mu})^{1/2} \tag{17}$$

and applies to models with one parameter of interest and no nuisance parameters.

In the asymptotic limit, the statistic q_μ approaches r_μ, resulting in the cumulative distribution of r_μ approaching the cumulative distribution of a standard normal. This can be intuitively demonstrated by utilizing the asymptotic limit expression of the log-likelihood, as provided in Eq. (9), provided the model contains only one parameter μ.

An alternative expression to (16) can be obtained by modifying the statistic r_μ itself, rather than its cumulative distribution. In particular, as shown in [7], the modified statistic

$$r_\mu^* = r_\mu + \frac{1}{r_\mu} \log \frac{q_\mu}{r_\mu} \tag{18}$$

follows a standard normal distribution with an error term of order $n^{-3/2}$:

$$f(r_\mu^*) = \mathcal{N}(0,1) + \mathcal{O}(n^{-3/2})\,. \tag{19}$$

That is, the error on the r_μ^* asymptotic distribution falls off three powers of $n^{-1/2}$ faster than that of the likelihood root (see Eq. (11)). In addition, by squaring r_μ^* one obtains the statistic r_μ^{*2}, which is asymptotically distributed as chi-squared with one degree of freedom. This can be interpreted as a higher-order correction to the likelihood ratio statistic w_μ.

An intuitive interpretation of the r^* statistics can be obtained, despite its non-trivial derivation (see, for example, Brazzale et al. (2007) [10]). The following equation i (see e.g. [10]):

$$r_\mu^* = \frac{r_\mu - \mathrm{E}[r_\mu]}{\mathrm{V}[r_\mu]^{1/2}} + \mathcal{O}(n^{-3/2})\,, \tag{20}$$

this tells that at the n^{-1} perturbative order r_μ^* in the standardized version of r_μ.

For statistical models that include nuisance parameters, the structure of the statistic r_μ^* remains the same as before, but the definition of q_μ changes. Consider a model with full parameter space $\boldsymbol{\psi} = \{\mu, \boldsymbol{\theta}\}$, where μ is the parameter of interest and $\boldsymbol{\theta}$ represents the nuisance parameters. In such a model, the statistic q_μ can be defined using either of two equivalent expressions:

$$q_{\mu,1} = \frac{\det\left[\ell_{\hat{\boldsymbol{\psi}}}(\hat{\mu},\hat{\boldsymbol{\theta}}) - \ell_{\hat{\boldsymbol{\psi}}}(\mu,\hat{\hat{\boldsymbol{\theta}}}) \quad \ell_{\boldsymbol{\theta}\hat{\boldsymbol{\psi}}}(\mu,\hat{\hat{\boldsymbol{\theta}}})\right]}{\det\left[\ell_{\boldsymbol{\psi}\hat{\boldsymbol{\psi}}}(\hat{\mu},\hat{\boldsymbol{\theta}})\right]} \left(\frac{\det[j_{\boldsymbol{\psi}\boldsymbol{\psi}}(\hat{\mu},\hat{\boldsymbol{\theta}})]}{\det[j_{\boldsymbol{\theta}\boldsymbol{\theta}}(\mu,\hat{\hat{\boldsymbol{\theta}}})]}\right)^{1/2}, \tag{21}$$

or

$$q_{\mu,2} = \frac{\det\left[\boldsymbol{\phi}(\hat{\mu},\hat{\boldsymbol{\theta}}) - \boldsymbol{\phi}(\mu,\hat{\hat{\boldsymbol{\theta}}}) \quad \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mu,\hat{\hat{\boldsymbol{\theta}}})\right]}{\det\left[\boldsymbol{\phi}_{\boldsymbol{\psi}}(\hat{\mu},\hat{\boldsymbol{\theta}})\right]} \left(\frac{\det[j_{\boldsymbol{\psi}\boldsymbol{\psi}}(\hat{\mu},\hat{\boldsymbol{\theta}})]}{\det[j_{\boldsymbol{\theta}\boldsymbol{\theta}}(\mu,\hat{\hat{\boldsymbol{\theta}}})]}\right)^{1/2}. \tag{22}$$

The numerator and denominator of both $q_{\mu,1}$ and $q_{\mu,2}$ are determinants of $d \times d$ matrices, where d is the dimension of the full parameter space. In both the expressions j is the information matrix, defined as

$$j_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}) = -\frac{\partial^2 \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}. \tag{23}$$

Moreover, the subscripts on ℓ and ϕ indicate derivatives, e.g., $\ell_{\boldsymbol{\psi}}$ and $\boldsymbol{\phi}_{\boldsymbol{\psi}}$ indicate the partial derivatives of ℓ and $\boldsymbol{\phi}$ with respect to the parameters $\boldsymbol{\psi}$.

It is important to note that the definition of $q_{\mu,1}$ involves derivatives with respect to the MLEs of the model. This means that $q_{\mu,1}$ can only be used if the likelihood's dependence on the MLEs can be explicitly expressed. If this is not possible, $q_{\mu,2}$ can be used instead. However, this requires defining a new set of parameters $\boldsymbol{\phi}$, called the *canonical parameters*, defined as

$$\boldsymbol{\phi}^T(\boldsymbol{\psi}, \boldsymbol{y}_{\mathbf{obs}}) = \sum_{i=1}^{N} \frac{\partial l}{\partial y_i}\Big|_{\boldsymbol{y}_{\mathbf{obs}}} \times V. \tag{24}$$

Here $\boldsymbol{y}_{\mathbf{obs}}$ are the observed data and V is $N \times d$ matrix defined as:

$$V = -\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}^T}\right)^{-1} \left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\psi}^T}\right)\Big|_{\hat{\boldsymbol{\psi}}_{\text{obs}}}. \tag{25}$$

In the last expression $\boldsymbol{z} = \{z_1(y_1), ..., z_N(y_N)\}$ is a vector of *pivotal quantities*. Pivotal quantities are statistics that have a fixed distribution under the model, i.e., they are not dependent on the parameters of the model. Such a vector always exists in the form of cumulative distributions $F(y_i)$, which are always uniformly distributed in $[0, 1]$. But alternative choices are often available, e.g., for a Gaussian distributed random variable y with mean μ and standard deviation σ one can define the pivotal statistic $z = (y - \mu)/\sigma$, whose distribution is a standard normal for any chosen μ and σ. For a comprehensive understanding of the construction of the r_μ^* see e.g. [10].

In Secs. 5, 6 and 7 below, the statistic r_μ^* will be used to derive improved confidence intervals for the mean μ of a set of measured values in the context of the Gamma Variance Model.

3.2 The Bartlett correction

A different approach to higher-order asymptotics due to Bartlett [9] involves a scaling of the likelihood ratio statistic, rather than a correction to the distributions of the MLEs. Bartlett's argument is as follows: For a model with M parameters of interest $\boldsymbol{\mu}$, the likelihood ratio $w_{\boldsymbol{\mu}}$ follows a chi-square distribution for M degrees of freedom in the asymptotic limit, and for a sample size n suppose

$$E[w_{\boldsymbol{\mu}}] = M + b + \mathcal{O}(n^{-2}). \tag{26}$$

Therefore the modified statistic

$$w_{\boldsymbol{\mu}}^* = w_{\boldsymbol{\mu}} \frac{M}{E[w_\mu]} = \frac{w_{\boldsymbol{\mu}}}{1 + b/M} \tag{27}$$

is expected to follow a distribution closer to the asymptotic χ^2_M. The factor $1 + b/M$ is known as the Bartlett correction.

Lawley [13] developed a general method to compute the expectation value up to order $\mathcal{O}(n^{-2})$ proving that all the cumulants of $w^*_{\boldsymbol{\mu}}$ match with the cumulants of a χ^2_M distribution up to order $\mathcal{O}(n^{-2})$. Specifically, Lawley's formula is based on a quartic expansion of both the likelihood ratio and the score equation, $\frac{\partial \ell}{\partial \boldsymbol{\mu}}(\hat{\boldsymbol{\mu}}) = 0$, in powers of $\hat{\mu}_i - \mu_i$ (see, e.g., [11]). Here, i is an index running over the parameter space of the model. The two expansions can be combined to obtain an approximation of the expectation value

$$E[w_{\boldsymbol{\mu}}] = 2E[l(\hat{\boldsymbol{\mu}}) - l(\boldsymbol{\mu})] = M + \epsilon_M + \mathcal{O}(n^{-2})\,, \tag{28}$$

where ϵ_M is the Lawley correction factor to the asymptotic expectation value M. The correction term ϵ_M has a complicated structure involving derivatives of the likelihood up to the fourth order and their expectation values. Nevertheless, for many applications, it is possible to compute it analytically. Specifically, the definition of ϵ_M is given by

$$\epsilon_q = \sum_{rstu} \lambda_{rstu} - \sum_{rstuvw} \lambda_{rstuvw}\,, \tag{29}$$

where the indexes r,s,t,u,v, and w label all the parameters of the model. The two terms inside the sum are defined as

$$\begin{aligned}
\lambda_{rstu} &= k^{rs}k^{tu}\left(\frac{1}{4}k_{rstu} - k^{(u)}_{rst} + k^{(tu)}_{rs}\right)\,, \\
\lambda_{rstuvw} &= k^{rs}k^{tu}k^{vw}\left(\frac{1}{6}k_{rtv}k_{suw} + \frac{1}{4}k_{rtu}k_{svw} - k_{rtv}k^{(u)}_{sw} - k_{rtu}k^{(v)}_{sw} + k^{(v)}_{rt}k^{(v)}_{sw} + k^{(u)}_{rt}k^{(v)}_{sw}\right)\,.
\end{aligned} \tag{30}$$

The terms inside the above definitions can be computed as

$$k_{rs} = E\left[\frac{\partial^2 l}{\partial\mu_r\partial\mu_s}\right]\,, \quad k_{rst} = E\left[\frac{\partial^3 l}{\partial\mu_r\partial\mu_s\partial\mu_t}\right]\,, \quad k_{rstu} = E\left[\frac{\partial^4 l}{\partial\mu_r\partial\mu_s\partial\mu_t\partial\mu_u}\right]\,, \tag{31}$$

$$k^{(t)}_{rs} = \frac{\partial k_{rs}}{\partial\mu_t}\,, \qquad k^{(tu)}_{rs} = \frac{\partial^2 k_{rs}}{\partial\mu_t\partial\mu_u}\,, \qquad k^{(u)}_{rst} = \frac{\partial k_{rs}}{\partial\mu_u}\,, \tag{32}$$

where the matrices with upper indices are the inverses of the corresponding matrices with lower indices. The general expression for the Bartlett correction is quite involved, but its computation is not conceptually complicated, since it only involves computing derivatives and expectation values of them.

When the parameters of the likelihood can be split into two subsets, one consisting of parameters of interest $\boldsymbol{\mu} = \{\mu_1, ..., \mu_d\}$, and the other of nuisance parameters $\boldsymbol{\theta} = \{\theta_1, ..., \theta_{M-d}\}$, one is typically interested in testing specific points in $\boldsymbol{\mu}$ space. In such scenarios, the Lawley formula can be used to compute the expected value as

$$\begin{aligned}
E[w_{\boldsymbol{\mu}}] &= 2E[l(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}) - l(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})] \\
&= 2E[l(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}) - l(\boldsymbol{\mu}, \boldsymbol{\theta})] - 2E[l(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}}) - l(\boldsymbol{\mu}, \boldsymbol{\theta})] \\
&= d + \epsilon_M - \epsilon_{M-d} + \mathcal{O}(n^{-2})\,.
\end{aligned} \tag{33}$$

The notation ϵ_{M-d} indicates that the summation in Eq.(29) is only performed over indices labeling the nuisance parameters. However, a more efficient way to compute Eq. (33), is to directly calculate the difference $\epsilon_M - \epsilon_{M-d}$ by summing the terms in Eq. (30) over all permutations of the indices that contain at least one parameter of interest. It is worth noting that, for composite hypotheses, the expectation value of the likelihood ratio is dependent on the nuisance parameters, as its distribution still has a dependence on them. Therefore, to evaluate the expectation value, and thus Eq.(33), one should use $\boldsymbol{\theta} = \hat{\hat{\boldsymbol{\theta}}}$, the MLEs of the nuisance parameters.

The Lawley formula is a valuable tool in situations where it is not feasible to analytically compute the exact expectation value of $w_{\boldsymbol{\mu}}$, a common scenario in realistic applications. An alternative approach is to numerically estimate the expectation value of $w_{\boldsymbol{\mu}}$ using MC methods. Specifically, E[$w_{\boldsymbol{\mu}}$] can be estimated by generating data and setting the parameters of interest $\boldsymbol{\mu}$ to the value in the parameter space being tested, and the nuisances $\boldsymbol{\theta}$ to their profile values. When the exact expectation value of $w_{\boldsymbol{\mu}}$ is computed, the error term on the asymptotic distribution of $w^*_{\boldsymbol{\mu}}$ is expected to be smaller than $\mathcal{O}(n^{-2})$.

4 Overview of the Gamma Variance Model

Having outlined in the preceding section the general formalism for higher-order asymptotic corrections, we now demonstrate their use with the Gamma Variance Model (GVM). This model, introduced in Ref. [12], extends the likelihood of Eq. (3) by regarding the variances $\sigma_{u_i}^2$ as adjustable rather than known parameters. The values that one would have assigned to them before are now treated as independent gamma distributed estimates v_i, i.e.,

$$v_i \sim \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} v_i^{\alpha_i - 1} e^{-\beta_i v_i} \,. \tag{34}$$

Here the parameters of the gamma distribution α_i and β_i are defined such that the expected value is $E[v_i] = \alpha_i/\beta_i$ and the variance is $\sigma_{v_i}^2 = \alpha_i/\beta_i^2$. These are chosen such v_i is an unbiased estimator for $\sigma_{u_i}^2$, (i.e., $E[v_i] = \sigma_{u_i}^2$) and the width of the gamma distribution is adjusted to reflect the appropriate level of uncertainty by defining

$$\varepsilon_i \equiv \frac{1}{2}\frac{\sigma_{v_i}}{E[v_i]} = \frac{1}{2}\frac{\sigma_{v_i}}{\sigma_{u_i}^2} \,, \tag{35}$$

which using error propagation becomes

$$\varepsilon_i \simeq \frac{s_i}{E[s_i]} \,, \tag{36}$$

where $s_i = \sqrt{v_i}$. Therefore, the new quantity ε_i is the relative uncertainty on the assigned systematic error, which we refer to informally as the *relative error-on-error* parameter.

Including the v_i as measurements into the likelihood gives

$$L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \times \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{u_i}^2}} e^{-\frac{(u_i - \theta_i)^2}{2\sigma_{u_i}^2}} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} v_i^{\alpha_i - 1} e^{-\beta_i v_i} \,. \tag{37}$$

Although treating the $\sigma_{u_i}^2$ as adjustable in the GVM in effect doubles the number of nuisance parameters in comparison to the model where the $\sigma_{u_i}^2$ are known, one can profile over them in closed form. After some manipulation (see Ref. [12]), the profile log-likelihood is found to be

$$\ell(\boldsymbol{\mu}, \boldsymbol{\theta}, \hat{\hat{\boldsymbol{\sigma}}}_{\boldsymbol{u}}^{\boldsymbol{2}}) = \ell_p(\boldsymbol{\mu}, \boldsymbol{\theta}) = \log P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) - \frac{1}{2}\sum_{i=1}^{N} \left(1 + \frac{1}{2\varepsilon_i^2}\right) \log\left[1 + 2\varepsilon_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right] . \tag{38}$$

Comparing with the log-likelihood based on fixed $\sigma_{u_i}^2$ from Eq. (4), one sees that the usual quadratic term, $(u_i - \theta_i)^2/\sigma_{u_i}^2$, is replaced by

$$\left(1 + \frac{1}{2\varepsilon_i^2}\right) \log\left[1 + 2\varepsilon_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right] . \tag{39}$$

As discussed in Refs. [12], the Gamma Variance model leads to interesting and useful consequences for inference about the parameters of interest $\boldsymbol{\mu}$. In particular, the size of the confidence region for $\boldsymbol{\mu}$ becomes coupled to the goodness of fit, with increasing incompatibility between the input data leading to larger regions. Furthermore, the point estimate for $\boldsymbol{\mu}$ shows a decreased sensitivity to outliers in the data. It is therefore of particular interest to apply the GVM in cases where the input values are in tension either amongst themselves or with the predictions of a hypothesis of interest. For example, the tension between measured and predicted values of the anomalous muon magnetic moment was explored in Ref. [14]. The GVM represents a purely frequentist approach to this type of problem. Bayesian methods have been found to yield qualitatively similar results, e.g., in Refs. [15, 16, 17, 18].

A practical difficulty with the Gamma Variance Model arises in connection with the use of asymptotic formulae to obtain p-values and confidence regions when the ε parameters exceed a value of around 0.2. As discussed in Ref. [12], there is a correspondence between the parameters ε_i and an effective sample size, n_i, which can be found by considering a sample of n_i independent observations of u_i and using their sample variance as an estimate of $\sigma_{u_i}^2$. This estimator is found to be gamma distributed with an error-on-error parameter ε_i related to the sample size by

$$n_i = 1 + \frac{1}{2\varepsilon_i^2} \,. \tag{40}$$

Thus when ε_i becomes too large, then n_{eff} drops to become of order unity and the large-sample limit required for use of asymptotic distributions is no longer satisfied. Values of ε are expected to be roughly 0.2 to 0.5 or even larger in many applications, which could make it far more difficult to compute p-values and confidence regions.

The breakdown of the asymptotic formulae for large ε_i can be understood intuitively by expanding the logarithmic term (39) in powers of ε_i:

$$\left(1 + \frac{1}{2\varepsilon_i^2}\right) \log\left[1 + 2\varepsilon_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right] = \left(1 + 2\varepsilon_i^2\right) \frac{(u_i - \theta_i)^2}{v_i} - 2\varepsilon_i^2 \frac{(u_i - \theta_i)^4}{v_i^2} \,. \tag{41}$$

Thus as ε_i approaches zero, the logarithmic constraint reduces to the standard quadratic one, and the fact that the Gaussian-distributed u_i enter in this fashion leads to the asymptotic distributions for the statistics w_μ and ε_μ discussed above. However, for large ε_i, the Gamma Variance Model deviates from the quadratic approximation by an error term of order $\mathcal{O}(\varepsilon_i^2)$, as shown in Eq. (41).

Consequently, when ε_i is not equal to zero, the asymptotic formulae used to obtain p-values and confidence regions are not guaranteed to represent valid approximations. Furthermore, the interval of convergence of the logarithm in Eq. (39) is

$$2\varepsilon_i^2 \frac{(u_i - \theta_i)^2}{v_i} < 1 \,, \tag{42}$$

thus problems in the convergence of the asymptotic formulae may arise if the above condition is not satisfied.

In principle, this difficulty can be overcome by using Monte Carlo calculations, but this can entail substantial additional work and computing time. It is therefore of great use to have a method of finding p-values and confidence regions without MC, and thus the primary goal of this paper is to investigate the use of higher-order asymptotics with the GVM to obtain results that remain accurate even for large ε_i.

5 Single-measurement model

In order to investigate the asymptotic properties of a statistical model with uncertain error parameters, it is convenient to use the simplified model introduced in Ref. [12]. Here for completeness we reproduce several results shown in that paper using the Bartlett correction and extend them in Sec. 5.1 using the formulae from Lawley and the p^* approximation.

The single-measurement model describes a single Gaussian distributed measurement y with mean μ and standard deviation σ. We take μ to be the parameter of interest and σ^2 to be a nuisance parameter constrained by an independent gamma-distributed estimate v. Therefore, the resulting likelihood is

$$L(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} \frac{\beta}{\Gamma(\alpha)} v^{\alpha-1} e^{-\beta v} \,, \tag{43}$$

where $\alpha = 1/4\varepsilon^2$ and $\beta = 1/4\varepsilon^2\sigma^2$, and ε is the relative error on the standard deviation σ. Additionally, the log-likelihood of the model is given by

$$\ell(\mu, \sigma^2) = -\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} - \left(\frac{1}{2} + \frac{1}{4\varepsilon^2}\right) \log\sigma^2 - \frac{v}{4\varepsilon^2\sigma^2} \,. \tag{44}$$

The goal is to compute the likelihood ratio w_μ (see Eq. (2)) to study its asymptotic properties and to apply to it the higher-order corrections defined in Sec.3. This requires the estimators

$$\hat{\mu} = y \,, \tag{45}$$

$$\widehat{\sigma^2} = \frac{v}{1 + 2\varepsilon^2} \,, \tag{46}$$

$$\widehat{\widehat{\sigma^2}} = \frac{v + 2\varepsilon^2(y-\mu)^2}{1 + 2\varepsilon^2} \,. \tag{47}$$

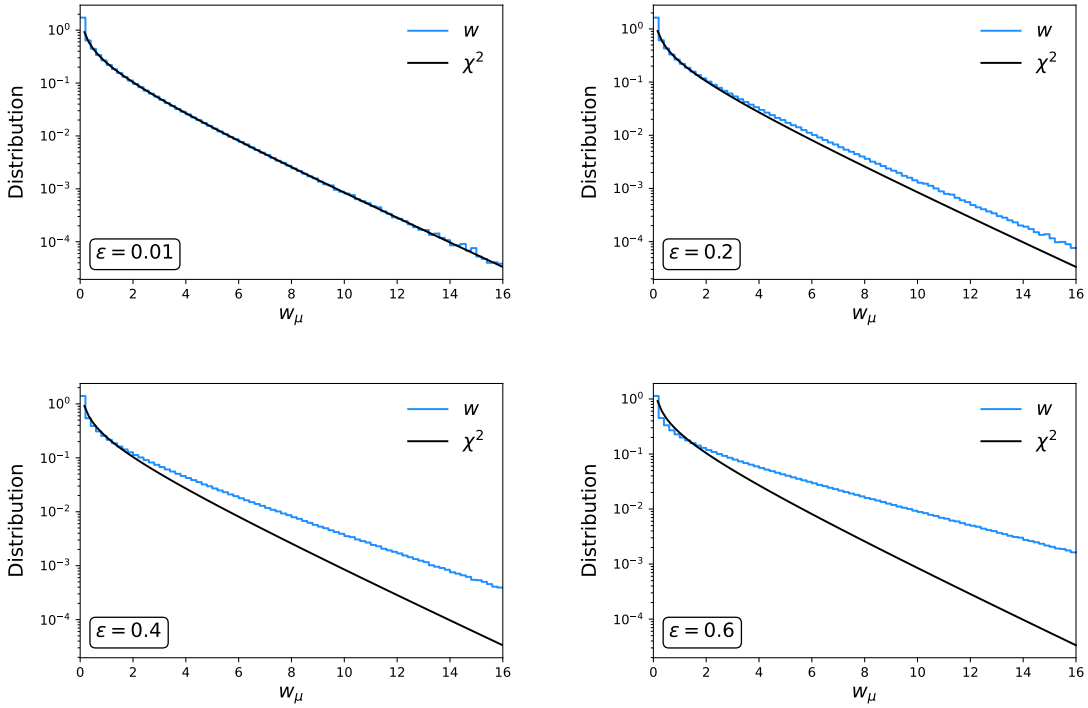

Figure 1. Distributions of w_μ (blue) for different values of the parameter ε compared with the asymptotic χ^2 distribution (black).

With the help of the above expressions it is easy to derive the likelihood ratio w_μ,

$$w_\mu = \left(1 + \frac{1}{2\varepsilon^2}\right) \log\left[1 + 2\varepsilon^2 \frac{(y-\mu)^2}{v}\right], \tag{48}$$

which, in the limit $\varepsilon \to 0$, becomes

$$w_\mu = \frac{(y-\mu)^2}{v} + \mathcal{O}(\varepsilon^2)\,. \tag{49}$$

In this limit, the likelihood ratio can be approximated by a quadratic expression, as expected in the asymptotic limit. Consequently, the parameter ε is related to an effective sample size, as it measures the extent to which the model deviates from the asymptotic limit. In particular, it is expected that the distribution of w_μ deviates from its asymptotic χ^2_1 distribution by an error term of order $\mathcal{O}(\varepsilon^2)$. Figure 1 displays the distributions of data generated according to Eq. (43) setting $\mu = 0$, $\sigma = 1$ and $\varepsilon = 0.01$, 0.2, 0.4, 0.6. As expected, the distribution deviates from the asymptotic χ^2_1 distribution as the ε parameter increases. The simple dependence of the single measurement model on the parameter ε makes it an ideal candidate for studying the effectiveness of higher-order asymptotic methods in improving asymptotic formulae.

5.1 Higher-order asymptotics for the single-measurement model

As one can see in Fig. 1, the likelihood ratio exhibits noticeable deviations from its asymptotic χ^2_1 distribution even for moderate values of ε. Hence, it is important to investigate whether higher-order statistics, namely r^*_μ and w^*_μ, can be better approximated by their asymptotic distributions, particularly for larger values of ε. The r^*_μ statistic is given by

$$r^*_\mu = r_\mu + \frac{1}{r_\mu} \log \frac{q_\mu}{r_\mu}\,, \tag{50}$$

where r_μ is the likelihood root defined as

$$r_\mu = \mathrm{sign}(\mu - \hat{\mu})\, w_\mu^{1/2}\,. \tag{51}$$

The asymptotic distribution of r^* is a standard normal and it has an associated error term of $\mathcal{O}(n^{-3/2})$. For the single-measurement model, ε is the effective sample size parameter of the model ($n = 1 + 1/2\varepsilon^2$), hence the error term is of order $\mathcal{O}(\varepsilon^3)$. In order to compute r^*_μ it is necessary to compute the statistic q_μ. The likelihood of the single-measurement model can be explicitly written in terms of the MLEs defined in Eq.(45) and Eq.(46):

$$\ell(\hat{\mu}, \hat{\sigma^2}) = -\frac{1}{2}\frac{(\hat{\mu} - \mu)^2}{\sigma^2} - \left(\frac{1}{2} + \frac{1}{4r^2}\right)\log\sigma^2 - \frac{\hat{\sigma^2}(1 + 2r^2)}{4r^2\sigma^2}\,. \tag{52}$$

Therefore, it is possible to use Eq.(21) to compute the statistic q_μ:

$$q_\mu = \frac{\sqrt{(1 + 2\varepsilon^2)v}}{v + 2\varepsilon^2(y - \mu)^2}(y - \mu)\,. \tag{53}$$

Since the asymptotic distribution of r^*_μ is a standard normal, the asymptotic distribution of r^{*2}_μ is a χ^2_1 distribution, and therefore it can be seen as a higher-order correction to the likelihood ratio. The second higher-order statistic we want to study is the Bartlett-corrected likelihood ratio,

$$w^*_\mu = w_\mu \frac{M}{E[w_\mu]} \equiv \frac{w_\mu}{1 + b/M}\,, \tag{54}$$

where $E[w_\mu] = M + b$ is what one must find to obtain the Bartlett correction, i.e., $E[w_\mu]/M = 1 + b/M$. The Bartlett-corrected likelihood ratio w^*_μ is expected to be χ^2_1 distributed in the asymptotic limit. The expectation value $\mathrm{E}[w_\mu]$ can be estimated using the Lawley formula defined by Eq. (33):

$$\mathrm{E}[w_\mu] = 1 + 3\varepsilon^2 + \mathcal{O}(\varepsilon^4)\,. \tag{55}$$

The asymptotic distribution of w^*_μ will have an error term of $\mathcal{O}(n^{-2})$, or equivalently $\mathcal{O}(\varepsilon^4)$ for the single-measurement model. All of the higher-order statistics described above, namely r^{*2}_μ and w^*_μ, follow a χ^2 distribution in the asymptotic limit.

In Fig. 2, we show the distributions of these two statistics for data generated according to Eq. (43) with $\mu = 0$, $\sigma = 1$, and ε values of 0.01, 0.2, 0.4, and 0.6. The distributions of two statistics are much better approximated by a χ^2_1 distribution compared to the original likelihood ratio w_μ, indicating that higher-order statistics provide significant improvements in this application.

5.2 Confidence intervals for the single-measurement model

The likelihood ratio is a commonly used tool for deriving confidence regions, typically obtained by finding the p-value of $\boldsymbol{\mu}$ and then solving the equation $p_{\boldsymbol{\mu}} = \alpha$, where $1 - \alpha$ represents the desired confidence level. In the case of the single-measurement model, which involves only one parameter of interest μ, our goal is to construct a confidence interval for it as described in Sec. 2. To obtain the p-value, the distribution of w_μ must be determined. In the asymptotic limit, w_μ follows a χ^2_1 distribution, but Fig. 1 shows that for large values of ε, this approximation is not very accurate. To address this, we can use higher-order statistics such as r^*_μ and w^*_μ instead. This involves computing the p-value using w^* or r^{*2}, i.e.,

$$p_\mu = \int_{w^*_{\mathrm{obs}}}^{\infty} \chi^2_1(w^*)\,dw = 1 - F_{\chi^2_1}[w^*_{\mathrm{obs}}]\,, \tag{56}$$

or

$$p_\mu = \int_{r^{*2}_{\mathrm{obs}}}^{\infty} \chi^2_1(r^{*2})\,dr^{*2} = 1 - F_{\chi^2_1}[r^{*2}_{\mathrm{obs}}]\,. \tag{57}$$

To illustrate this we find the confidence interval for μ as a function of the parameter ε setting the observed values of y and v to 0 and 1, respectively. Figure 3 presents a comparison of the confidence intervals obtained using the likelihood ratio w_μ and the higher-order statistics r^*_μ and w^*_μ. Additionally, the confidence interval is also computed by calculating the p-value exactly, as described in [12]. The plot in Fig. 3 shows that the use of higher-order statistics significantly improves the accuracy of the confidence interval.

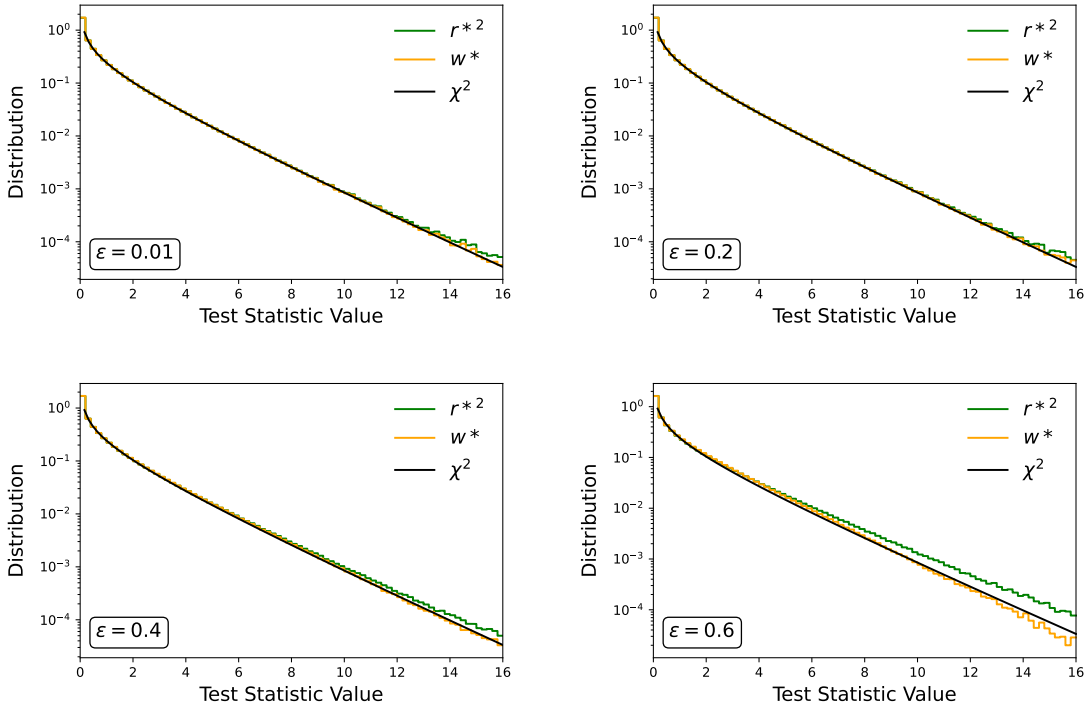

Figure 2. Distributions of r^{*2} (green) and w^* computed with the Lawley formula (orange) for different values of the parameter ε compared with the χ^2 asymptotic distribution (black).

6 Simple-average model

The single-measurement model can be extended to an average of N measurements $\boldsymbol{y} = \{y_1, ..., y_N\}$, which are assumed to follow a Gaussian distribution with mean μ and standard deviations $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_N\}$. In addition, the standard deviations are assumed to be uncertain with associated best estimates $\boldsymbol{v} = \{v_1, ..., v_N\}$ and relative errors $\boldsymbol{\varepsilon} = \{\varepsilon_1, ..., \varepsilon_N\}$. The likelihood of the model is thus

$$L(\mu, \boldsymbol{\sigma^2}) = \prod_i^N \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(y_i-\mu)^2/2\sigma_i^2} \prod_i^N \frac{\beta_i}{\Gamma(\alpha_i)} v^{\alpha_i - 1} e^{-\beta_i v_i} , \tag{58}$$

where $\alpha_i = 1/4\varepsilon_i^2$ and $\beta_i = 1/4\varepsilon_i^2\sigma_i^2$. Equivalently, the log-likelihood of the model is

$$\ell(\mu, \boldsymbol{\sigma^2}) = \sum_{i=1}^N \left[-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma_i^2} - \left(\frac{1}{2} + \frac{1}{4\varepsilon_i^2}\right) \log \sigma_i^2 - \frac{v_i}{4\varepsilon_i^2\sigma_i^2} \right] . \tag{59}$$

In contrast to the full Gamma Variance Model described in Sec. 4, it does not include nuisance parameters θ_i or their estimates, but rather treats the variances σ_i^2 of the primary measurements y_i as uncertain. It can be easily generalized to a curve-fitting problem where the expectation value of each measurement y_i can be defined as a function of the parameters of interest $\boldsymbol{\mu}$ and a control measurement x_i, i.e., $\mathrm{E}[y_i] = f(x_i; \boldsymbol{\mu})$.

The log-likelihood of Eq. (59) profiled over the $\boldsymbol{\sigma^2}$ is given by

$$\ell_p(\mu) = -\sum_i^N \frac{1}{2}\left(1 + \frac{1}{2\varepsilon_i^2}\right) \log\left[1 + 2\varepsilon_i^2 \frac{(y_i-\mu)^2}{v_i}\right] , \tag{60}$$

which has been computed using the profile value of σ_i^2:

$$\widehat{\widehat{\sigma_i^2}} = \frac{v_i + 2\varepsilon_i^2(y_i-\mu)^2}{1 + 2\varepsilon_i^2} . \tag{61}$$

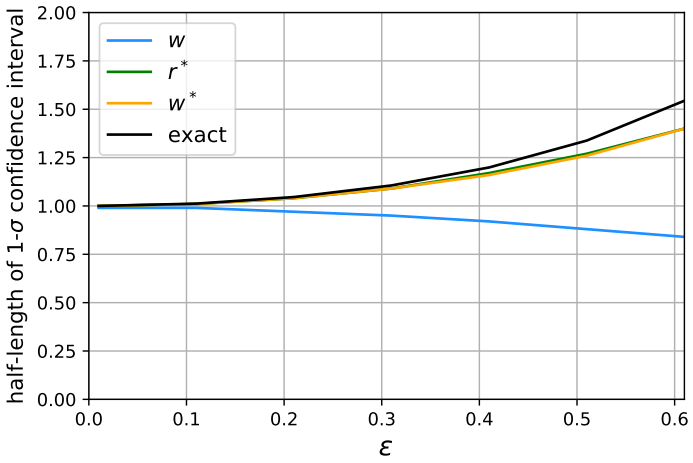

Figure 3. Half-length of 1-σ confidence intervals for μ (Eq. (43)) as a function of ε, computed using w_μ (blue), r^*_μ (green) and w^*_μ calculated using the Lawley formula (orange). The black curve represents the exact half-length of the confidence interval.

As in the example of the single-measurement model from Sec. 5, we compute the likelihood ratio w_μ, and the statistics r^*_μ and w^*_μ. To do this, one must first calculate the MLE $\hat{\mu}$ by solving

$$\frac{\partial \ell_p}{\partial \mu} = 0\,. \tag{62}$$

This must be found numerically, as it cannot be solved in closed form for arbitrary N. As discussed in Sec. 4, the distribution of the likelihood ratio w_μ is expected to deviate from its asymptotic χ^2_1 form by an error term of order $\mathcal{O}(\varepsilon_i^2)$ for each measurement y_i. Therefore, it is important to investigate whether the higher-order statistics r^*_μ and w^*_μ can improve the precision of the inference on μ.

To compute r^*_μ, it is necessary to calculate the statistic q_μ using Eq.(22), as the log-likelihood of Eq.(59) cannot be expressed explicitly as a function of the MLEs. Additionally, to compute q_μ, as discussed in Sec. 3.1, a vector of pivotal quantities $\boldsymbol{z} = \{z_{y_1}, ..., z_{y_N}, z_{v_1}, ..., z_{v_N}\}$ must be defined. For the simple-average model, the best choices for these are

$$\begin{aligned} z_{y_i} &= \frac{(y_i - \mu)^2}{\sigma_i^2}\,, \\ z_{v_i} &= \frac{v_i}{\sigma_i^2}\,. \end{aligned} \tag{63}$$

These pivotal quantities can be used to compute q_μ exploiting Eqs. (24) and (25). However, it is not possible to derive an analytical expression for q_μ for arbitrary N, since it requires computing the determinants of generic $(N+1)\times(N+1)$ matrices. Nevertheless, once N is specified, q_μ can be found in closed form.

The Bartlett corrected likelihood ratio w^*_μ can be estimated using the Lawley formula, as discussed earlier. The result can be expanded at order ε_i^2 as

$$\mathrm{E}[w_\mu] = 1 + \frac{4}{\sum_{i=1}^N 1/v_i}\sum_{i=1}^N \frac{\varepsilon_i^2}{v_i} - \frac{1}{(\sum_{i=1}^N 1/v_i)^2}\sum_{i=1}^N \frac{\varepsilon_i^2}{v_i^2} + \sum_{i=1}^N \mathcal{O}(\varepsilon_i^4)\,, \tag{64}$$

which in the limit $\varepsilon_i \to 0$ gives 1 as expected. Using Eq. (64) one can thus find the corrected statistic $w^*_\mu = w_\mu M/E[w_\mu]$.

6.1 Confidence intervals for the parameter of interest

In Particle Physics, the likelihood ratio and its higher-order corrections are used to estimate confidence regions for the parameters of interest. Hence, these statistics can be tested by examining

how well they predict the size of confidence regions when their density functions are approximated using their asymptotic distributions.

In this section, we aim to compute a 68.3% confidence interval for the parameter μ of Eq. (59). Specifically, we consider the simple case of averaging two measurements, y_1 and y_2, with observed values of $+\delta$ and $-\delta$, respectively, where δ takes on values of 0.5 and 1.5. The estimate of the standard deviations v_1 and v_2 is set to 1. Both measurements are assigned equal error on error parameters, $\varepsilon_1 = \varepsilon_2 = \varepsilon$, and we present results as a function of ε. We use the likelihood ratio w_μ and the higher-order statistics r^*_μ and w^*_μ to compute the 68.3% confidence intervals for μ.

In addition, the confidence interval is found by estimating the p-value of the likelihood ratio using MC. This is done by generating the exact distribution of the data for a fixed value of μ while setting the nuisance parameters σ_i^2 to their profiled values. This technique is commonly known as the *profile construction* [19] or *hybrid resampling* [20, 21] method. The half lengths of the confidence intervals are plotted in Fig. 4, and the results are compared with the numerical predictions computed using the profile construction technique. In both examples, the Bartlett-corrected likelihood ratio w^*_μ provides reliable estimates of the size of the confidence interval, which closely match those obtained using the profile construction technique.

The r^*_μ statistic performs well when the averaged data are internally compatible (left panel of Fig.(4)). However, for larger values of the parameter ε, r^*_μ breaks down as the tension in the observed data grows (right panel of Fig.(4)). A conservative approach to determine the applicability of r^*_μ in improving the likelihood ratio predictions is to verify whether the arguments of the logarithmic terms of Eq. (60) are within their radius of convergence. Specifically, for the endpoints of the confidence interval, one should check whether the inequality

$$2\varepsilon_i^2 \frac{(y_i - \mu)^2}{v_i} < 1 \tag{65}$$

is satisfied for every measurement y_i. In the example above, this condition implies that one should not trust the accuracy of the result from r^*_μ if $\varepsilon \geq 0.5$ for $\delta = 0.5$ and $\varepsilon \geq 0.3$ for $\delta = 1.5$.

6.2 Goodness-of-fit

The likelihood ratio is commonly used to construct confidence regions for the parameters of interest. It does not, however, provide a measure of how well the selected model describes the observed data. To address this issue, a new statistic is typically defined, known as the goodness-of-fit statistic. For the simple-average model we use

$$q = -2 \log \frac{L(\hat{\mu}, \hat{\boldsymbol{\sigma}}^2)}{L_{\mathrm{s}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\sigma}}^2)}\,, \tag{66}$$

where L_{s} represents the likelihood of the *saturated model*. The saturated model is obtained by replacing the expectation values $\mathrm{E}[y_i] = \mu$ with a set of independent parameters $\boldsymbol{\phi} = \{\phi_1, \,, ..., \,, \phi_n\}$, such that $\mathrm{E}[y_i] = \phi_i$. Since the log-likelihood of the saturated model $L_{\mathrm{s}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\sigma}}^2)$ is equal to zero, the goodness-of-fit statistic reduces to

$$q = -2 \log L(\hat{\mu}, \hat{\boldsymbol{\sigma}}^2) = \sum_i^N \left(1 + \frac{1}{2\varepsilon_i^2}\right) \log \left[1 + 2\varepsilon_i^2 \frac{(y_i - \hat{\mu})^2}{v_i}\right]\,. \tag{67}$$

If the above expression is expanded in powers of ε_i^2, in the limit $\varepsilon_i^2 \to 0$ one finds

$$q = -\sum_i^N \frac{(y_i - \hat{\mu})^2}{v_i} + \mathcal{O}(\varepsilon_i^2)\,. \tag{68}$$

In this limit, q reduces to a sum of squares of Gaussian-distributed quantities, and thus its distribution follows a χ^2_{N-1} distribution with $N-1$ degrees of freedom. This is because the expectation values are constrained by the single fitted parameter $\hat{\mu}$. However, for large values of the ε_i parameters, deviations from the χ^2_{N-1} asymptotic distribution are expected.

To correct the goodness-of-fit statistic using higher-order asymptotics, q needs to be defined as a likelihood ratio. This can be done by defining the saturated model such that the simple-average model is nested within it. A possible choice is to define the saturated model as

$$\ell_{\mathrm{s}}(\boldsymbol{\alpha}, \mu, \boldsymbol{\sigma}^2) = \log L_{\mathrm{s}}(\boldsymbol{\alpha}, \mu, \boldsymbol{\sigma}^2) = \sum_i^N -\frac{1}{2}\frac{(y_i - \alpha_i - \mu)^2}{\sigma_i^2} - \left(\frac{1}{2} + \frac{1}{4\varepsilon_i^2}\right) \log \sigma_i^2 - \frac{v_i}{4\varepsilon_i^2 \sigma_i^2}\,, \tag{69}$$

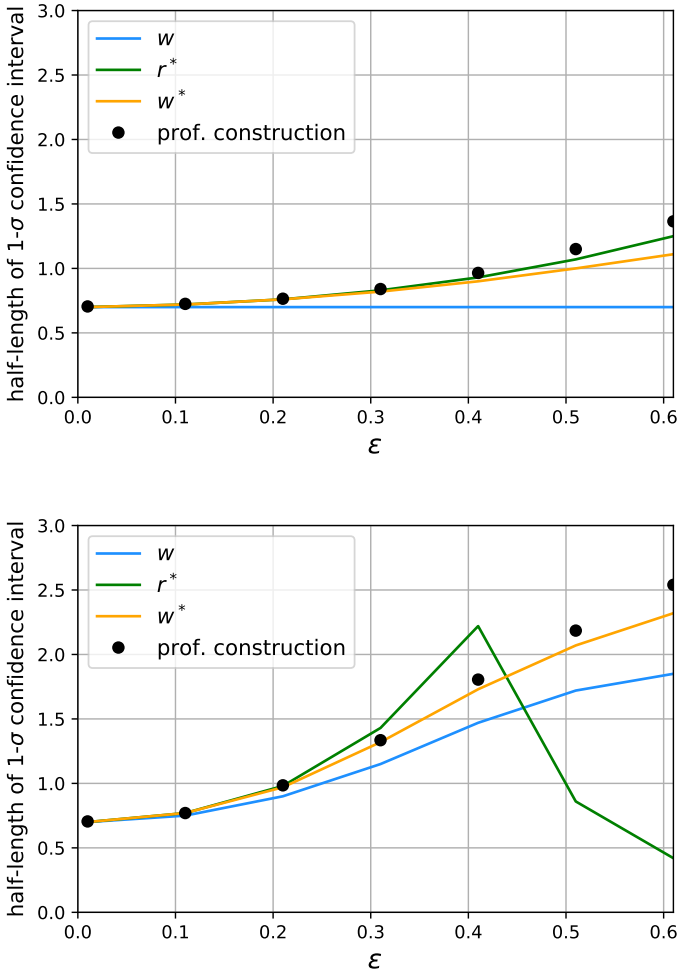

Figure 4. Half-length of 1-σ confidence intervals for parameter μ (Eq. (59)) as a function of ε for $\delta = 0.5$ (top) and $\delta = 1.5$ (bottom), computed using w_μ (blue), r^*_μ (green) and w^*_μ calculated using the Lawley formula (orange). The black dots represent our most precise estimate of the interval, computed using the profile construction.

where we fix $\alpha_N = -\sum_{i=1}^{N-1} \alpha_i$ so that $\sum_{i=1}^{N} \alpha_i = 0$. Given this definition, the simple-average model is recovered by fixing all the α_i to zero, hence $\ell = \ell_\mathrm{s}(\boldsymbol{\alpha} = \mathbf{0}, \mu, \boldsymbol{\sigma^2})$. Therefore, the goodness-of-fit can be written as a likelihood ratio of the saturated model,

$$q = -2 \log \frac{L_\mathrm{s}(\boldsymbol{\alpha} = \mathbf{0}, \hat{\hat{\mu}}, \hat{\hat{\boldsymbol{\sigma}}}^2)}{L_\mathrm{s}(\hat{\boldsymbol{\alpha}}, \hat{\mu}, \hat{\boldsymbol{\sigma}}^2)}\,, \tag{70}$$

and its Bartlett correction can be computed using the Lawley formula. This is done by treating the α_i as parameters of interest and μ and the σ_i^2 as nuisance parameters. The p^* approximation instead is not, in general, useful to correct the goodness-of-fit, as it can only be applied to models with one parameter of interest, e.g., averages of two measurements.

To measure how well the model describes the observed data, one can compute the p-value of the goodness-of-fit,

$$p = \int_{q_\mathrm{obs}}^{\infty} f(q)\, dq = 1 - F[q_\mathrm{obs}]\,, \tag{71}$$

In general, small values of the p-value are associated with a bad agreement between the model and the data. In Particle Physics, p-values are typically converted to a related quantity Z called the *significance*, defined as

$$Z = \Phi^{-1}(1 - p)\,, \tag{72}$$

where Φ^{-1} is the inverse cumulative distribution of a standard normal. The significance tells us how many standard deviations away from the mean the observed p-value is.

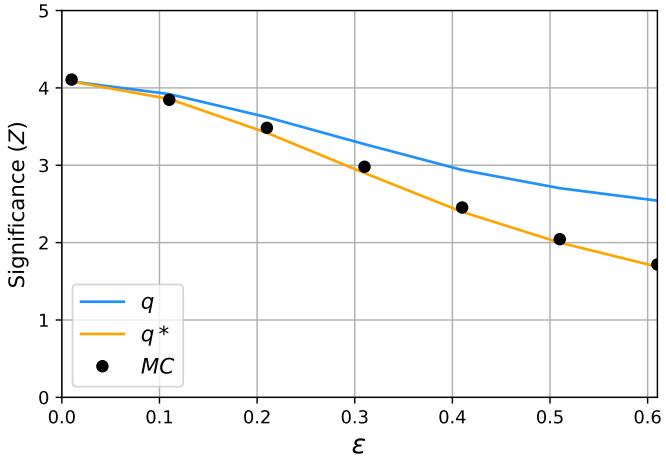

Figure 5. Significance Z as a function of ε, computed using the goodness-of-fit (blue) and its Bartlett corrected counterpart calculated using the Lawley formula (orange). The black dots represent the significance computed by generating the distribution of q with MC.

To illustrate this method, we find the p-value and corresponding significance Z for an average of two incompatible measurements, namely $y_1 = -3$ and $y_2 = 3$, with estimated variances of $v_1 = v_2 = 1$. We set $\varepsilon_1 = \varepsilon_2 = \varepsilon$ and allow ε to vary. The significance is estimated using both the goodness-of-fit statistic q and the Bartlett-corrected version q^*. Figure 5 shows the significance as a function of the parameter ε. The use of the Bartlett correction results in a significant improvement in estimating the significance, eliminating the need for MC simulation.

7 Averages using the full Gamma Variance Model

An important application of the Gamma Variance Model is the average of N measurements, each if which are reported with known statistical but uncertain systematic errors. In this scenario, N measurements are assumed to be independent and Gaussian distributed with means $\mathrm{E}[y_i] = \mu + \theta_i$ and known variances (the "statistical errors") $\mathrm{V}[y_i] = \sigma_{y_i}^2$. Here, the nuisance parameters θ_i represent potential biases to the means of the y_i. As described in Sec. 4, their values are estimated with independent Gaussian distributed control measurements u_i, whose variances $\sigma_{u_i}^2$ (the "systematic errors") are treated as adjustable parameters. The $\sigma_{u_i}^2$ are estimated by measurements v_i, whose gamma distributions are characterized by the error-on-error parameters ε_i. The log-likelihood of the model becomes

$$\ell(\mu, \boldsymbol{\theta}, \boldsymbol{\sigma_u^2}) = -\frac{1}{2}\sum_{i=1}^{N}\left[\frac{(y_i - \mu - \theta_i)^2}{\sigma_{y_i}^2} - \frac{1}{2}\frac{(u_i - \theta_i)^2}{\sigma_{u_i}^2} - \left(\frac{1}{2} + \frac{1}{4\varepsilon_i^2}\right)\log\sigma_{u_i}^2 - \frac{v_i}{4\varepsilon_i^2\sigma_{u_i}^2}\right], \tag{73}$$

and the profiled log-likelihood l_p can be computed using

$$\widehat{\widehat{\sigma_{u_i}^2}} = \frac{v_i + 2\varepsilon_i^2(u_i - \theta_i)^2}{1 + 2\varepsilon_i^2}, \tag{74}$$

leading to

$$l_p(\mu, \boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{N}\left[\frac{(y_i - \mu - \theta_i)^2}{\sigma_{y_i}^2} - \frac{1}{2}\left(1 + \frac{1}{2\varepsilon_i^2}\right)\log\left(1 + 2\varepsilon_i^2\frac{(u_i - \theta_i)^2}{v_i^2}\right)\right]. \tag{75}$$

The MLEs $\hat{\mu}$ and $\hat{\theta}_i$ can be found numerically or by solving a system of cubic equations (see Ref. [12]).

As before, we compute the likelihood ratio w_μ and the higher-order statistics r_μ^* and w_μ^*. To compute r_μ^*, one needs to calculate q_μ as defined in Eq. (22). To do this, one requires a vector of

pivotal quantities $\boldsymbol{z} = \{z_{y_1}, ..., z_{y_N}, z_{u_1}, ..., z_{u_N}, z_{v_1}, ..., z_{v_N}\}$, which can be defined as

$$\begin{aligned} z_{y_i} &= \frac{(y_i - \mu - \theta_i)^2}{\sigma_{y_i}^2}\,, \\ z_{u_i} &= \frac{(u_i - \theta_i)^2}{\sigma_{u_i}^2}\,, \\ z_{v_i} &= \frac{v_i}{\sigma_{u_i}^2}\,. \end{aligned} \tag{76}$$

The Bartlett corrected likelihood ratio $w_\mu^* = w_\mu M/E[w_\mu]$ can be estimated numerically using the Lawley formula defined by Eq. (33), which predicts the expectation value of w_μ to be

$$\mathrm{E}[w_\mu] = 1 + \sum_i^N \mathcal{O}(\varepsilon_i^4)\,. \tag{77}$$

Therefore, the Bartlett correction is zero up to $\mathcal{O}(\varepsilon_i^4)$, indicating that any deviations of the likelihood ratio density function from its asymptotic distribution are expected to be of the same order, $\mathcal{O}(\varepsilon_i^4)$.

To further improve the accuracy of the Lawley formula, one can compute the Bartlett correction numerically. Specifically, one can estimate the expectation value of w_μ by generating data with all the model parameters set to their maximum likelihood estimates and approximating it as a constant that is independent of the model parameters:

$$\mathrm{E}[w_\mu] \simeq \mathrm{E}[w_{\hat{\mu}}]\,. \tag{78}$$

This approximation has been found to yield highly accurate results. Moreover, this approximation is important because it significantly speeds up the computation of confidence intervals. Rather than generating a new set of data to estimate the Bartlett correction for every tested value of μ, data only needs to be generated once.

In certain scenarios, an analyst may wish to conduct inference on one or more nuisance parameters θ_i for example to generate a ranking plot of the systematics or obtain the correlation matrix of the nuisances. In such cases, the nuisance parameters must be treated as parameters of interest. According to the Lawley formula, the expected value of the likelihood root will be

$$\mathrm{E}[w_{\mu,\boldsymbol{\theta}}] = 1 + M + \sum_i^M 4k^{\theta_i\theta_i}\frac{\varepsilon_i^2}{v_i} - \sum_i^M \left(k^{\theta_i\theta_i}\right)^2 \frac{\varepsilon_i^2}{v_i^2} + \sum_i^N \mathcal{O}(\varepsilon_i^4)\,. \tag{79}$$

Here, M represents the number of nuisance parameters that have been promoted to parameters of interest. The term $k^{\theta_i\theta_i}$ refers to the θ_i component of the expectation value of the inverse Hessian matrix of the likelihood, which is defined by the first term of equation Eq. (31) and computed for $\sigma_{u_i}^2 = v_i$.

7.1 Confidence regions

As in the previous examples, we compute confidence intervals for the parameter of interest μ using the likelihood ratio and higher-order statistics, assuming their density functions are given by the asymptotic distributions. Specifically, consider an example similar to what was used in Sec. 6, namely, the mean of two measurements, $y_1 = -\delta$ and $y_2 = +\delta$, here with associated statistical errors σ_1 and σ_2 both equal to $1/\sqrt{2}$ and $\delta = 0.5/1.5$. Additionally, we assume that the control measurements u_1 and u_2 have observed values of 0, and the estimates of the systematic errors to be $1/\sqrt{2}$, or equivalently, the estimates of the variances v_1 and v_2 to be 1/2. Both measurements are assumed to have equal error on error parameters, $\varepsilon_1 = \varepsilon_2 = \varepsilon$, look at the results for different ε.

Figure 6 displays the confidence interval estimated for the parameter μ using the likelihood ratio w_μ, as well as the higher-order statistics w_μ^* and r_μ^*. The resulting confidence intervals are compared to the confidence interval predicted using the profile construction method, which is taken as the best available estimate of such intervals. Among the three statistics, w_μ^* provides the most accurate estimates, almost perfectly overlapping with the numerical predictions obtained using the profile construction method. Moreover, w_μ^* is significantly faster to compute as it only requires data generation for $\mu = \hat{\mu}$.

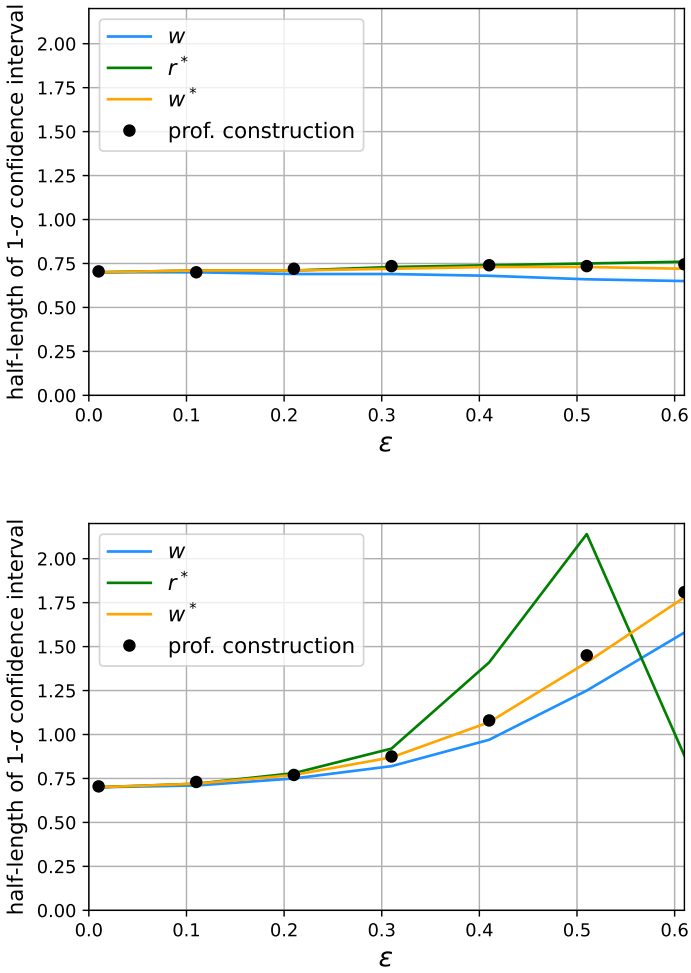

Figure 6. Half-length of 1-σ confidence intervals for parameter μ (Eq. (73)) as a function of ε for $\delta = 0.5$ (top) and $\delta = 1.5$ (bottom), computed using w_μ (blue), r^*_μ (green) and w^*_μ calculated with MC (orange). The black dots represent our most precise estimate of the interval, computed using the profile construction.

In contrast, the profile construction method entails generating a new set of data for every tested value of μ. The r^*_μ statistic, on the other hand, provides accurate predictions for internally consistent data (see the left plot of Fig. 6). However, for growing discrepancies between the measurements (right plot of Fig. 6), it is reliable only for small values of ε. To determine if r^*_μ can be used, one can verify whether the logarithms appearing in the profile log-likelihood satisfy the perturbative condition given by Eq. (42). For $\delta = 1.5$, this condition limits the applicability of r^*_μ to $\varepsilon \simeq 0.3$, whereas, for $\delta = 0.5$, the threshold is higher, above 0.6.

Figure 7 shows the 2D confidence region in the (μ, θ_1) plane, which was obtained using the same measured data as in the previous example, whereas the error on error parameters ε_1 and ε_2 were fixed to 0.5. The confidence regions were computed using the likelihood ratio $w_{\mu,\boldsymbol{\theta}}$ and the Bartlett-corrected likelihood ratio $w^*_{\mu,\boldsymbol{\theta}}$ computed via Eq. (79), and were then compared with the confidence regions estimated using the profile construction technique. The results indicate that the Bartlett correction improves the accuracy of predictions significantly compared to the uncorrected likelihood ratio.

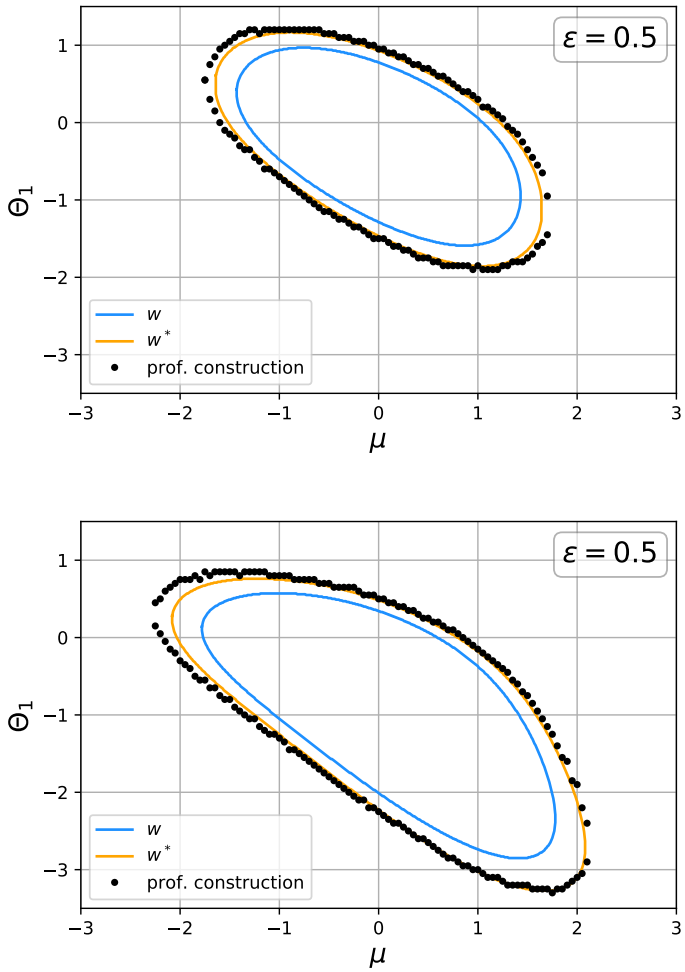

Figure 7. 68.3% confidence regions in (μ, θ_1) plane for $\delta = 0.5$ (top) and $\delta = 1.5$ (bottom), computed using the likelihood ratio (blue) and the Bartlett correction calculated with the Lawley formula (orange). The black dots represent the most precise estimate of the interval, computed using the profile construction. The error on error parameters ε_1 and ε_2 are fixed to 0.5.

7.2 Goodness of fit

The goodness-of-fit statistic for the Gamma Variance Model can be defined using the same approach as used for the simple-average model in Sec. 6.2, leading to

$$q = -2 \log L(\hat{\mu}, \hat{\boldsymbol{\theta}}, \widehat{\widehat{\boldsymbol{\sigma}_{\boldsymbol{u}}^{\boldsymbol{2}}}}) = \sum_{i=1}^{N} \left[\frac{(y_i - \hat{\mu} + \hat{\theta}_i)^2}{\sigma_{y_i}^2} - \left(1 + \frac{1}{2\varepsilon_i^2}\right) \log \left(1 + 2\varepsilon_i^2 \frac{(u_i - \hat{\theta}_i)^2}{v_i^2}\right) \right] . \tag{80}$$

In this case, however, constructing a saturated model is not useful as the Lawley correction is 0 at order ε_i^2. Nonetheless, the Bartlett correction can still be computed using MC simulations. This method allows for a significant computational improvement over generating the exact distribution of q using pseudo-experiments to estimate its p-value. The latter approach would require an order of $\mathcal{O}(10^5)$ simulations to accurately capture a 4σ effect, while the expectation value of q can be estimated with good precision using only $\mathcal{O}(10^3)$ pseudo-experiments.

To illustrate these techniques we compute the significance of the p-value for an average of two incompatible measurements using Eq. (73). The observed values of y_1 and y_2 are assumed to be -3 and 3 whereas the control measurements u_1 and u_2 are set to 0. The statistical uncertainties, σ_1 and σ_2, are set to $1/\sqrt{2}$ as the estimates of the systematic errors (which is equivalent to set $v_1 = v_2 = 1/2$). Figure 8 compares the significance computed using the goodness-of-fit q and the

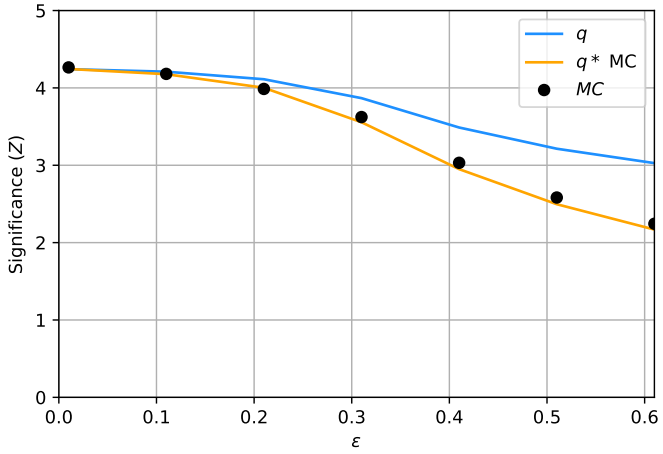

Figure 8. Significance as a function of ε, computed using the goodness-of-fit (blue) and its Bartlett corrected counterpart calculated using MC (orange). The black dots represent the significance computed by generating the distribution of q with MC.

Bartlett-corrected q^*, with the significance computed generating the distribution of q^* numerically, all of them as a function of ε. Consistent with our earlier findings, the Bartlett correction yields highly accurate predictions.

8 Conclusions

In conclusion, we have demonstrated the efficacy of higher-order asymptotics in the computation of confidence intervals and p-values within the framework of the Gamma Variance Model. The GVM is a specialized statistical model designed to address uncertainties in parameters that themselves represent uncertainties. The methods studied in this paper hold particular relevance when the fixed parameters, ε, indicative of the relative uncertainties in estimates of standard deviations for Gaussian-distributed measurements, are not negligible. In such scenarios, standard asymptotic methods prove inadequate, and the asymptotic distributions fall short in providing accurate approximations.

Our investigation specifically focused on the Barndorff-Nielsen p^* approximation and the Bartlett correction, both of which are higher-order asymptotic techniques that offer adjustments to the first-order (profile) likelihood ratio and likelihood root test statistics. These adjustments enable the test statistics to be more accurately approximated by their asymptotic distributions, even when the ε parameter is large.

Both the Barndorff-Nielsen p^* approximation and the Bartlett correction demonstrated their value as tools to enhance the accuracy and reliability of confidence interval and p-value calculations using Gamma Variance Models. However, it should be noted that the p^* approximation exhibited instabilities in the presence of internally incompatible data for large values of ε. Additionally, while the r^* approximation can be computed analytically for all the examples examined in this paper, the expressions become complex for models associated with realistic applications, such as the simple-average and full GVM models.

Conversely, the Bartlett correction, calculated using the Lawley formula, offers a more elegant expression for the expectation value of the likelihood ratio, which is employed to compute the Bartlett correction factor for the likelihood ratio. Specifically, refer to Eq.(64) for the simple-average model (see Sec.6) and Eqs.(77) and (79) for combinations utilizing the full GVM (see Sec. 7.1). Nevertheless, the r^* approximation is suitable for inference on the parameter of interest in combinations executed with the full GVM, provided the condition in Eq.(42) is met for every nuisance parameter. This is because the Lawley formula does not introduce corrections at the order of ε^2 for the likelihood ratio, allowing the r^* approximation to offer an analytical expression, while the Bartlett correction can only be computed using MC.

The Bartlett correction also proved to be an effective technique for improving the goodness-of-fit statistic. This was true for cases where the statistic could be computed analytically using

the Lawley formula, such as the simple-average model, as well as for cases where it was estimated using Monte Carlo methods, as in the full Gamma Variance Model. The application of the Bartlett correction in the latter scenario significantly reduced the number of pseudo-experiments required for accurately estimating the significance of rare effects.

These findings highlight the potential of higher-order asymptotics to refine inference on the parameters of interest in a various contexts, not only the GVM. Higher-order asymptotics are valuable tools when the MLEs of statistical models do not follow Gaussian distributions or, equivalently, when log-likelihoods are not well approximated by quadratic expressions. For Gamma Variance Models this occurs when ε is large; however, in general, such deviations are typically associated with small experimental-sample sizes. In Particle Physics it is not uncommon to search for new signal processes by counting collision events with very specific characteristics, such that the expected number of background events may be order unity. With sample sizes of this order it is expected that asymptotic distributions will not be accurate and high-order asymptotic formulae should prove valuable (see, e.g., Refs. [10, 22]).

The introduction of higher-order asymptotic corrections removes a potential stumbling block for use of the Gamma Variance Model. As many estimate of systematic uncertainties may themselves be uncertain at the level of 20% to 50% or more, one would not expect asymptotic confidence intervals or p-values to be accurate. By using higher-order corrections, accurate results can be achieved without with minimal or no Monte Carlo simulation, greatly simplifying use of the model.

Acknowledgements

EC and GC are grateful to Bogdan Malaescu and other colleagues in the ATLAS Collaboration who have provided valuable feedback on this work and to the U.K. Science and Technology Facilities Council for its support.

References

[1] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

[2] A. Wald, *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*, Transactions of the American Mathematical Society, Vol. **54**, No. 3 (Nov., 1943), pp. 426-482.

[3] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C 71 (2011) 1554.

[4] Li-Gang Xia, *Improved Asymptotic Formulae for Statistical Interpretation Based on Likelihood Ratio Tests*, arXiv:2101.06944 [physics.data-an] (2021).

[5] O. Barndorff-Nielsen, *Conditionality Resolutions*, Biometrika , 67 (1980) 2, pp. 293-310.

[6] O. Barndorff-Nielsen, *On a formula for the distribution of the maximum likelihood estimator*, Biometrika, 70 (2) 1983, pp. 343–365, doi.org/10.1093/biomet/70.2.343.

[7] O. Barndorff-Nielsen, *Inference on full or partial parameters based on the standardized signed log likelihood ratio*, Biometrika, Volume 73, Issue 2, August 1986, pp. 307–322.

[8] Barndorff-Nielsen, O.E. (1990), *Approximate Interval Probabilities.*, Journal of the Royal Statistical Society: Series B (Methodological), 52: 485-496.

[9] M.S. Bartlett, *Properties of sufficiency and statistical tests*, Royal Society of London Proceedings Series A 160, (1937) 268-282.

[10] A.R. Brazzale, A.C. Davison and N. Reid, *Applied Asymptotics: Case Studies in Small-Sample Statistics*, Cambridge University Press (2007).

[11] Gauss M. Cordeiro and Francisco Cribari-Neto, *An Introduction to Bartlett Correction and Bias Reduction*, Springer Verlag, 2014.

[12] G. Cowan, *Statistical Models with Uncertain Error Parameters*, Eur. Phys. J. C (2019) 79:133; arXiv:1809.05778.

[13] D.N. Lawley, *A general method for approximating to the distribution of likelihood ratio criteria*, Biometrika, Vol. 43, Issue 3-4, (1956) 295-303.

[14] G. Cowan, *Effect of Systematic Uncertainty Estimation on the Muon $g-2$ Anomaly*, EPJ Web of Conferences 258, 09002 (2022); arXiv:2107.02652.

[15] Wolfgang von der Linden, Volker Dose and Udo von Toussaint, *Bayesian Probability Theory: Applications in the Physical Sciences*, Cambridge University Press, 2014.

[16] G. D'Agostini, *Sceptical combination of experimental results: General considerations and application to epsilon-prime/epsilon*, arXiv:hep-ex/9910036 (1999).

[17] G. Cowan, *Bayesian Statistical Methods for Parton Analyses*, in *Proceedings of the 14th International Workshop on Deep Inelastic Scattering (DIS2006)*, M. Fuzz, K. Nagano, and K. Tokushuku (eds.), Tsukuba, 2006.

[18] Jens Erler and Rodolfo Ferro-Hernández, *Alternative to the application of PDG scale factors*, Eur. Phys. J. C 80 (2020) 6, 541, arXiv:2004.01219.

[19] K. Cranmer, *Statistical challenges for searches for new physics at the LHC*, in Proceedings of PHYSTAT05, L. Lyons and M.K. Unel (eds.), Imperial College Press, pp. 112-123 (2005)

[20] C. Chuang and T.L. Lai, *Hybrid resampling methods for confidence intervals*, Statistica Sinica 10 (2000) 1-50.

[21] Bodhisattva Sen, Matthew Walker and Michael Woodroofe, *On the Unified Method with Nuisance Parameters*, Statistica Sinica 19 (2009) 301-314.

[22] A.C. Davison and N. Sartori, *The Banff Challenge: Statistical Detection of a Noisy Signal*, Statistical Science, Vol. 23, No. 3 (2008), pp. 354-364.