Harmonizing discovery thresholds and reporting two-sided confidence intervals: a modified Feldman & Cousins method

K. D. Morá

Oskar Klein Centre, Department of Physics, Stockholm University
Albanova University Center, SE-10691 Stockholm, Sweden

E-mail: knut.mora@fysik.su.se

Abstract: When searching for new physics effects, collaborations will often wish to publish upper limits and intervals with a lower confidence level than the threshold they would set to claim an excess or a discovery. However, confidence intervals are typically constructed to provide constant coverage, or probability to contain the true value, with possible overcoverage if the random parameter is discrete. In particular, that means that the confidence interval will contain the 0-signal case with the same frequency as the confidence level. This paper details a modification to the Feldman-Cousins method to allow a different, higher excess reporting significance than the interval confidence level.

Keywords: Analysis and statistical methods, Dark Matter detectors (WIMPs, axions, etc.)
1 Introduction

Many physics experiments, in particular searches for new physics, look for very low event rates where the asymptotic methods of constructing frequentist confidence intervals do not work. Confidence intervals are required have coverage $1 - \alpha$-confidence level intervals should contain the true value a fraction $1 - \alpha$ of repeated experiments. The actual coverage of a statistical method may vary with the true signal properties— for example, an asymptotic 0.68 confidence-level interval for a counting experiment observing $n; [n - \sqrt{n}, n + \sqrt{n}]$ events measuring the expectation value $\mu$ will cover asymptotically as $\mu \to \infty$, but may cover as little as 0.55 and as much as 1 depending on $\mu$.

A method that provides confidence intervals that provides coverage is known since 1937 as the Neyman construction [1]. The Neyman construction initially consists of constructing for each possible true value of parameter of interest $\theta$ a confidence belt:

$$1 - \alpha = \int_a^b f(x|\theta)dx$$  \hspace{1cm} (1.1)

where $f(x)$ is the probability density function for the observed parameter $x$, and $[a, b]$ denote the limits of the confidence belt. The confidence interval on $\theta$ can then be found by constructing $a(\theta)$ and $b(\theta)$, which will express the upper and lower range in which $x$ would fall $1 - \alpha$ of the time if the true parameter of interest is $\theta$. Inverting these functions yields the Neyman construction limits for an observation $x$:

$$[a^{-1}(x), b^{-1}(x)]$$  \hspace{1cm} (1.2)

The condition for the confidence belt provided in equation 1.1 is not unique, the limits of the confidence belt have to be set by a boundary condition. This condition has traditionally consisted in either the desire to set upper or lower limits (for example in absence of a signal) or in reporting (symmetric or asymmetric) two-sided intervals in case of a measurement of a physical parameter (for example a signal rate). In 1998, Feldman & Cousins [2] noted that the fact that this choice has
to be made prior to the experimental outcome can lead to situations where the intervals do not cover. For example, in a search for a new particle, experimenters will be inclined to present upper limits below a threshold (defined to correspond to a required p-value in some suitably chosen test statistic for testing the no signal hypothesis), whereas two-sided confidence intervals (i.e. error bars on a parameter estimate) will be presented above it. As the question whether the test statistic lies above threshold can only be answered after the measurement has been made, the boundary condition on equation 1.1 in this cases changes depending on the outcome of the experiment, implying undercoverage, the probability for the interval to contain the true value will be smaller than desired. The suggested remedy (hereafter referred to as the "FC method") is to construct confidence intervals by a single Neyman construction that switches from one-sided to two-sided intervals based on the experimental outcome, providing correct coverage by construction.

Conventionally, upper limits are reported with confidence levels of less than 95%, and two-sided intervals are presented only in the case of discovery or at least some reasonably significant indication, which would have a (one-sided) p-value, usually much smaller than the p-value of 5% or more implied by the confidence interval. While statistically presenting a two-sided interval and not claiming a discovery does not pose a problem (the fact that the confidence interval excludes the non existence of a signal at some confidence level should not be confused with a discovery claim), in practice experimenters are reluctant to present a two sided limit even if the FC method provides it. A common remedy is to report only the upper edge of the interval provided. This leads to a signal-dependent over-coverage, or, equivalently, some confidence intervals or upper limits could be more constraining without violating coverage. In this paper, we suggest a modification of the FC method that will provide two sided intervals only at a desired discovery threshold, while still providing a unified confidence interval calculation method and improving the coverage. The paper is organized as follows: in section 2 we review briefly the FC method, and the procedure for assessing the existence of an excess or a discovery. In section 3 we introduce the modified version of the FC method. In sections 4 and 5 we present results for simple cases of Gaussian and Poisson probability density functions, and conclude in section 6.

2 Feldman Cousins construction

For an experiment where one measures some data $\bar{x}$ with a probability distribution $f(\bar{x}|\theta)$ that depends on a parameter $\theta$, the likelihood is given as $\mathcal{L}(\theta) = f(\bar{x}|\theta)$. The method proposed by Feldman and Cousins uses the log-ratio $R$ between the likelihood given the $\theta$ that minimizes the likelihood, $\hat{\theta}$:

$$R(\theta) = 2 \cdot \log \left[ \mathcal{L}(\hat{\theta}) / \mathcal{L}(\theta) \right]$$

(2.1)

to decide which $\theta$ to include. Either constructing the confidence belt from Equation 1.1 with the constraint that the $x$ with the lowest $R$ are included first, or constructing the confidence belt directly in the $R$ parameter:

$$1 - \alpha = \int_0^{R_{\max}} f(R|\theta) dR$$

(2.2)

for each value of $\theta$ will yield the FC construction. The confidence interval, whether one- or two-sided will be the region where $R(\theta) < R_{\max}(\theta)$. Note that the threshold likelihood ratio $R_{\max}(\theta)$
also depends on the parameter of interest. The method has coverage for each $\theta$ by construction. Therefore, if an experiment is looking to constrain a parameter $\theta$ that has a null-hypothesis and lower bound, $\theta_0$, as, for example the production cross-section of an unknown particle, the method must give confidence intervals that do not contain $\theta_0$ in $\alpha$ of the cases. If $\theta_0$ is a lower bound, that means that in those cases, the method must yield a two-sided interval.

The p-value of a result with respect to the null-hypothesis $\theta = \theta_0$ is the percentile of $R(\theta_{\text{result}})$ under the null-hypothesis:

$$ p_{\text{result}}(R_{\text{result}}) = \int_{R_{\text{result}}}^{\infty} f(R|\theta_0) dR $$

(2.3)

This may also be inverted to yield discovery thresholds; $p_{\text{result}}^{-1}(\alpha)$ in equation 2.3 is the discovery threshold for an $\alpha$ excess. Note that this equation shows that at the null-hypothesis $\theta = \theta_0$, the FC threshold for inclusion in the confidence interval, $R_{\text{max}}(\theta_0)$ implies a p-value of $\alpha$, and that a confidence interval that does not include $\theta_0$ implies a p-value below $\alpha$. Typical confidence intervals for upper limits, and thus the FC construction are $\alpha = 0.1, 0.05, 0.01$. Using the FC method consistently will report two-sided intervals at those same thresholds. However, a conventional discovery threshold in particle physics is $5\sigma$, or $p = 3 \cdot 10^{-7}$, and experiments may not wish to publish measurements of excesses lower than, say, $3\sigma$, or $p = 1.3 \cdot 10^{-3}$. A pragmatic solution to this is to only report the upper edge of the confidence interval as an upper limit until the discovery significance has exceeded the required discovery threshold. This will only lead to over-coverage, as one extends a confidence interval constructed to cover with an $1-\alpha$ frequency. The over-coverage will, however, lead to some loss of sensitivity.

3 Modified Feldman-Cousins

The aim of modifying the FC method in this work is to avoid the over-coverage induced by experiments reporting their results only above a higher threshold than their interval confidence level. The proposed modification to the FC method uses an idea very similar as that which was used by the ATLAS Higgs search[3]; where the ordering ratio $R$ is multiplied by the sign of $\hat{\theta} - \theta$:

$$ R'(\theta) = \text{sgn}(\hat{\theta} - \theta) \cdot R(\theta) $$

(3.1)

This separates the cases where the signal prefers a lower and higher signal than the tested hypothesis. Close to a boundary, say a requirement that $\theta_0 \leq \theta$, $R(\theta_0)$ will be only positive, and for slightly larger $\theta$, the distribution of $R$ will still be asymmetric between upwards and downwards fluctuations. Denoting the confidence level of the interval $1-\alpha$, and the significance above which two-sided intervals are reported $\gamma$, the upper edge of the confidence interval will be defined as:

$$ R'(\theta)^+ = w(\theta) \cdot R_{\text{max},y}(\theta_0) + (1-w(\theta)) \cdot R_{\text{max},\alpha}(\theta) $$

(3.2)

where $w(\theta)$ is a weighting function that monotonically decreases from 1 at $\theta = \theta_0$ to 0 as $\theta$ increases. The threshold functions, $R_{\text{max},y}$ and $R_{\text{max},\alpha}(\theta)$ are the discovery threshold at $\theta = 0$, and the threshold function defined by equation 2.3, respectively. The lower edge of the interval, $R'(\theta)^-$ is then defined so that for any signal, $1-\alpha$ is contained between $R'(\theta)^-$ and $R'(\theta)^+$:

$$ 1-\alpha = \int_{R'(\theta)^-}^{R'(\theta)^+} f(R'|\theta) dR' $$

(3.3)
At $\theta = \theta_0$, the above equation would indicate a coverage of $1 - \alpha$. However, at the border of the domain for $\alpha$, the distribution of $R(\theta_0)$ will be peaked towards 0, and by defining the confidence interval with strict inequalities, the coverage at the boundary can be arranged to be $\gamma$. Below, two examples, using toy Monte Carlo computations, are shown, illustrating the method and showing the coverage of the method.

4 Gaussian Example

As an example, we consider an experiment measuring $x$, distributed as a Gaussian with standard deviation 1 around an expectation value $\mu = s + b$, where $b = 0.5$, and $0 \leq s$. A toy Monte-Carlo computation is used to construct the threshold function $R(s)$, which is the upper blue line in figure 1. The dashed black line shows the log-likelihood ratio. The FC confidence interval satisfying equation 2.3 is the region of $s$ where the likelihood ratio is below the upper blue threshold curve.

The weighting function $w(s)$ is a freedom of the method, but a natural choice would be the curve that most steeply joins the FC interval from the discovery threshold. In the case of this gaussian likelihood, the shape of the log-likelihood ratio is fixed modulo a shift in $s$ due to the observed $x$. Therefore, the log-likelihood ratio curve that exactly yields the desired discovery threshold is the steepest curve that can interpolate between the two without requiring split confidence intervals where the log-likelihood ratio would exit and then re-enter the acceptance region below the threshold. This threshold curve is shown in orange. For a likelihood where the width of the likelihood is not fixed, for example due to nuisance parameters, it may be necessary to use toy Monte-Carlo computations to construct an envelope of log-likelihood ratio curves to ensure that the weighting function is sufficiently shallow.

The upper and lower limits as a function of the observed $x$ is shown in figure 2. Note that unlike the standard FC interval, the modified FC interval may return upper limits approaching 0 for downwards fluctuations, similar to a one-sided Neyman construction. To avoid publishing empty intervals, experiments may choose to apply a power-constraint [4] or similar methods to keep upper limits over a threshold. Visually, this would enter figure 2 as a horizontal limit below which no upper limit would be set.

The coverage of the FC and modified FC methods are shown in fig 3. In addition the green curve illustrates the over-coverage caused by an experiment choosing to only report the lower limit if the discovery significance exceeds $3\sigma$. Note that for the modified FC method, the coverage is maintained at the nominal 0.9 for all signals except at 0 signal, where the coverage steeply rises up to 0.9986, the coverage of a $3\sigma$ confidence interval.

For this simple case, it is also possible to compute the FC and modified FC thresholds numerically, starting from noting that the log-likelihood ratio for a downwards and upwards fluctuation
The log-likelihood ratio, \( \hat{s} = 2 \), is multiplied by the sign of \( \hat{s} - s \). The range of \( s \) where the log-likelihood ratio curve lies between the upper and lower blue curves constitute the FC interval. The upper orange curve shows the modified FC threshold, interpolating between the \( 3\sigma \) and \( 90\% \) thresholds, with the lower orange curve following from demanding \( 90\% \) coverage.

For an upwards fluctuation, \( 0 < x - b \):

\[
R_{up}(s) = 2 \cdot \log \left[ \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{(x_{up} - s - b)^2}{2\sigma^2}} \right] - 2 \cdot \log \left[ \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{(x_{up} - s - b)^2}{2\sigma^2}} \right]
\]

\[
R_{up}(s) = \frac{1}{\sigma^2} \left[ (x_{up} - s - b)^2 - (x_{up} - (x_{up} - b) - b)^2 \right]
\]

\[
R_{up}(s) = -5 - \frac{1}{\sigma^2} \left[ (x_{up} - s - b)^2 \right]
\]
Where \( x_{up} \) is the observed \( x \) that corresponds to an upwards fluctuation giving \( R_{up}(s) \). Similarly, for a downwards fluctuation, the relation between \( R_{dn}(s) \) and \( x_{dn} \) is:

\[
R_{dn}(s) = \frac{1}{\sigma^2} \left[ (x_{dn} - s - b)^2 - (x_{dn} - \max(x_{dn} - b, 0) - b)^2 \right] \tag{4.4}
\]

Setting the thresholds equal to each other, and defining \( z \equiv x - s - b \), gives a relation between the upper and lower edge of the confidence interval:

\[
z_{up} = \sqrt{(z_{dn} + s - \max(z_{dn} + s, 0))^2 - z_{dn}^2} \tag{4.5}
\]

Using the Neyman construction condition from equation 1.1, and that \( z \) is distributed according to a normal gaussian distribution, with cumulative distribution function \( \Theta \), gives an equation that may be solved numerically for \( z_{dn} \) for a certain confidence level \( 1 - \alpha \):

\[
\alpha = \Theta(z_{dn}) - 1 + \Theta\left(\sqrt{(z_{dn} + s - \max(z_{dn} + s, 0))^2 - z_{dn}^2}\right) \tag{4.6}
\]

Solving this equation for \( z_{dn,\alpha}(s) \), computing \( z_{up,\alpha}(s) \) and inverting them gives a Neyman construction as shown in figure 2. Alternatively, the \( R(s) \frac{1}{\alpha^2}(z_{up,\alpha}(s))^2 \) threshold may be computed as shown in figure 1. Note that the threshold in this case only depends on the signal expectation value, and is independent of \( b \).

---

**Figure 3**: Coverage of confidence intervals for the gaussian variable as function of true signal \( s \). The blue line shows the FC construction coverage, which also covers the no-signal case in 0.9 of the cases. The green curve shows the FC coverage if an experiment decides to only report an upper limit for discovery significances under 3\( \sigma \). The modified FC curves in orange displays nominal coverage above \( s = 0 \), and covers \( s = 0 \) with a confidence interval corresponding to a 3\( \sigma \) confidence interval. The gray band shows the binomial 1\( \sigma \) band around the correct 0.9 coverage due to toy Monte Carlo statistics.
Table 1: Table of confidence intervals using the modified FC method for a range of observed \( x - b \). The \( \alpha = 0.1 \) column corresponds to the modified FC construction shown in Figure 2. For comparison, the last column shows the standard 0.68 FC interval.

| \( x - b \) | \( \alpha=0.01 \) | \( \alpha=0.05 \) | \( \alpha=0.10 \) | \( \alpha=0.32 \) | FC, \( \alpha=0.32 \) |
|---|---|---|---|---|---|
| -2.00 | (0.0, 1.10) | (0.0, 3.80e-06) | (0.0, 1.60e-06) | (0.0, 2.96e-07) | (0.0, 0.08) |
| -1.00 | (0.0, 1.68) | (0.0, 0.80) | (0.0, 0.30) | (0.0, 7.73e-07) | (0.0, 0.27) |
| 0.00 | (0.0, 2.58) | (0.0, 1.96) | (0.0, 1.64) | (0.0, 0.49) | (0.0, 1.00) |
| 1.00 | (0.0, 3.58) | (0.0, 2.96) | (0.0, 2.64) | (0.0, 2.00) | (0.24, 2.00) |
| 2.00 | (0.0, 4.58) | (0.0, 3.96) | (0.0, 3.64) | (0.0, 3.00) | (1.00, 3.00) |
| 3.00 | (0.60, 5.58) | (1.10, 4.96) | (1.30, 4.64) | (2.00, 4.00) | (2.00, 4.00) |
| 4.00 | (1.54, 6.58) | (2.04, 5.96) | (2.36, 5.64) | (3.00, 5.00) | (3.00, 5.00) |

5 Poisson Example

In the case of a discrete variable, equation 1.1 must be replaced by an inequality:

\[
1 - \alpha \geq \sum_a^{b} f(x|\theta)
\]  

(5.1)

which can give over-coverage for some true parameters. Figure 4 shows the coverage plot of the FC interval as well as modified FC intervals as a function of signal for a random variable \( x \) that is distributed according to the Poisson distribution with an expectation value \( s + b \), where \( s \) is the signal and \( b = 0.5 \) a known background rate. As the confidence intervals and discovery significance depend only on the single random variable \( x \), the steepest interpolation between the discovery and FC thresholds will correspond to a simple step in \( x \). The discreteness of the random variable means that unlike in the continuous Gaussian case, \( R^s(s)^- \) remains at the FC boundaries, leading to equal confidence intervals and coverage as when using the FC construction. In cases where the FC interval provides over-coverage due to discreteness, the overcoverage of the pragmatic solution of only reporting the upper edge of confidence intervals below the discovery threshold may be irreducible.

6 Summary

This paper has demonstrated a method for constructing a modified FC method where the discovery significance is different from the confidence level of the upper limits and intervals. For an example Gaussian case, the coverage at 0 signal corresponds to the discovery significance, and moves to the required confidence level for all signals larger than 0. This allows experiments to avoid the over-coverage that results from expanding the standard FC intervals, and simplifies reporting or discussion of coverage properties.

Experiments may still wish to limit upper limits to positive value using, for example a power-constraint [4]. The more discrete coverage using this method will result in discrete coverage regimes that only depends on the true signal, and so may still be preferred on didactical grounds.
Figure 4: Coverage of confidence intervals for a Poisson variable with expectation value $s + 0.5$ as function of true signal $s$, computed using a toy Monte Carlo computation. The gray band shows the binomial $1\sigma$ band around the correct 0.9 coverage due to toy Monte Carlo statistics. The blue line shows the FC construction coverage, which also covers the no-signal case over 0.9 of the cases. The green curve shows the FC coverage if an experiment decides to only report an upper limit for discovery significances under $3\sigma$, which for this example is equal to the modified FC curve in orange, displaying over-coverage above $s = 0$, and covering $s = 0$ as a $3\sigma$ confidence interval.

Acknowledgements

The author would like to thank Jan Conrad and Jelle Aalbers for fruitful discussions and suggestions. This research was supported by a grant of the Knut and Alice Wallenberg Foundation, PI: J. Conrad

References

[1] J. Neyman. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Phil. Trans. Roy. Soc. Lond., A236(767):333–380, 1937. doi: 10.1098/rsta.1937.0005.
[2] Gary J. Feldman and Robert D. Cousins. A Unified approach to the classical statistical analysis of small signals. Phys. Rev., D57:3873–3889, 1998. doi: 10.1103/PhysRevD.57.3873.
[3] Georges Aad et al. Combined search for the Standard Model Higgs boson in $pp$ collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. Phys. Rev., D86:032003, 2012. doi: 10.1103/PhysRevD.86.032003.
[4] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Power-Constrained Limits. pre-print, 2011. [physics.data-an/1105.3166].