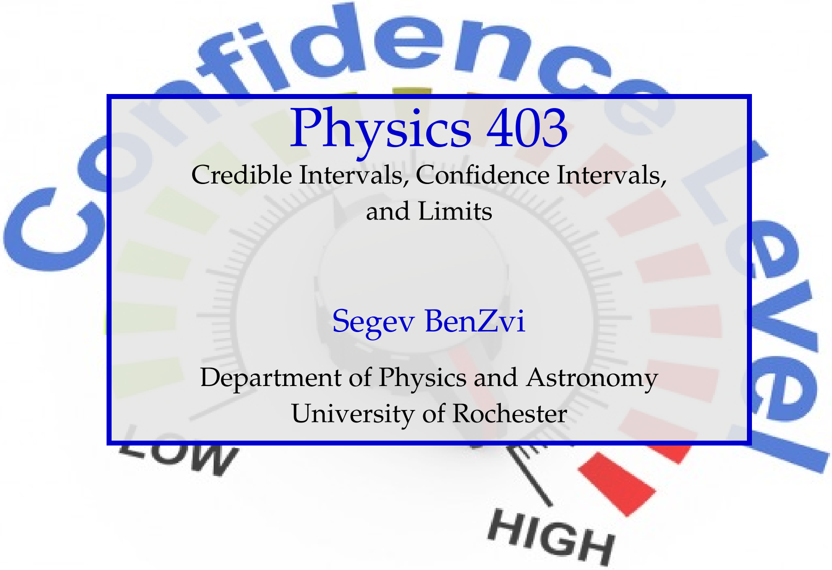


Physics 403

Credible Intervals, Confidence Intervals, and Limits

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Reading

- Cowan, Ch. 9
- Feldman and Cousins paper (PRD **57**:3873, 1998)

Table of Contents

Parameter Intervals

- We often want to make a statement about some parameter μ whose true value μ_t (in the frequentist sense) is unknown.
- We measure an observable x whose PDF depends on μ. I.e., we have $p(x|\mu) = \mathcal{L}(x|\mu)$
- From Bayes' Theorem, we want to calculate

$$p(\mu_t|x) = \frac{\mathcal{L}(x|\mu_t)\, p(\mu_t)}{p(x)}$$

- A Bayesian interval $[\mu_1, \mu_2]$ corresponding to a confidence level α is constructed by requiring

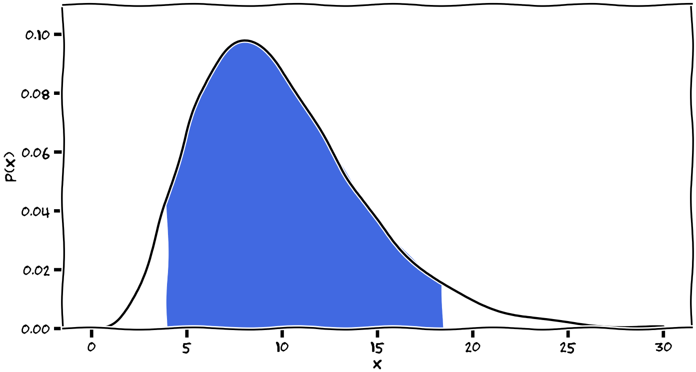
$$\int_{\mu_1}^{\mu_2} p(\mu_t|x)\, d\mu_t = \alpha$$

This is called the credible interval of μ_t

Bayesian Credible Intervals

Central Interval

Given the posterior PDF, it is easy to quote a range for a paramter:



Central 90% of the PDF gives a credible region $x \in [4.0, 18.4]$.

Credible Intervals in the Wild

Selected for a Viewpoint in *Physics*

PRL **116,** 061102 (2016) PHYSICAL REVIEW LETTERS week ending 12 FEBRUARY 2016

Observation of Gravitational Waves from a Binary Black Hole Merger

B. P. Abbott *et al.**

(LIGO Scientific Collaboration and Virgo Collaboration)

(Received 21 January 2016; published 11 February 2016)

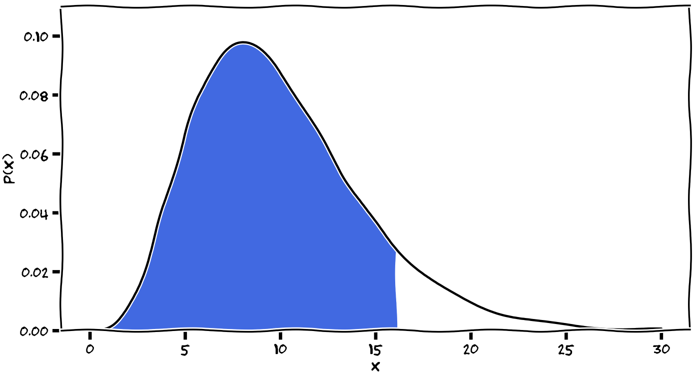
On September 14, 2015 at 09:50:45 UTC the two detectors of the Laser Interferometer Gravitational-Wave Observatory simultaneously observed a transient gravitational-wave signal. The signal sweeps upwards in frequency from 35 to 250 Hz with a peak gravitational-wave strain of 1.0×10^{-21}. It matches the waveform predicted by general relativity for the inspiral and merger of a pair of black holes and the ringdown of the resulting single black hole. The signal was observed with a matched-filter signal-to-noise ratio of 24 and a false alarm rate estimated to be less than 1 event per 203 000 years, equivalent to a significance greater than 5.1σ. The source lies at a luminosity distance of 410^{+160}_{-180} Mpc corresponding to a redshift $z = 0.09^{+0.03}_{-0.04}$. In the source frame, the initial black hole masses are $36^{+5}_{-4} M_{\odot}$ and $29^{+4}_{-4} M_{\odot}$, and the final black hole mass is $62^{+4}_{-4} M_{\odot}$, with $3.0^{+0.5}_{-0.5} M_{\odot} c^2$ radiated in gravitational waves. All uncertainties define 90% credible intervals. These observations demonstrate the existence of binary stellar-mass black hole systems. This is the first direct detection of gravitational waves and the first observation of a binary black hole merger.

DOI: 10.1103/PhysRevLett.116.061102

Bayesian Credible Intervals

Upper Limit

If you wanted to quote an upper limit instead you would just integrate to find the 90$^{\text{th}}$ percentile:



Here $x \in [0, 16.0]$, or, "the upper limit of x at 90% C.L. is 16.0"

Advantages and Disadvantages of the Bayesian Approach

- With the Bayesian approach you can account for prior knowledge when calculating the credible region, which can be very useful for quoting limits near a physical boundary
- Example from Cowan [1]: you measure $m^2 = E^2 - p^2$. Because of measurement uncertainties the maximum likelihood estimator $\hat{m}^2 < 0$
- A Bayesian would be able to use a prior that vanishes for $m < 0$, so you don't have to publish an unphysical value. This option is not available to a frequentist
- However, a 90% Bayesian credible interval may not mean that 90% you will measure a value in a certain range, because the PDF does not have to refer to long-run frequencies
- The frequentist range (confidence interval) is sometimes what you want, but interpreting it is tricky

Table of Contents

Classical Confidence Intervals

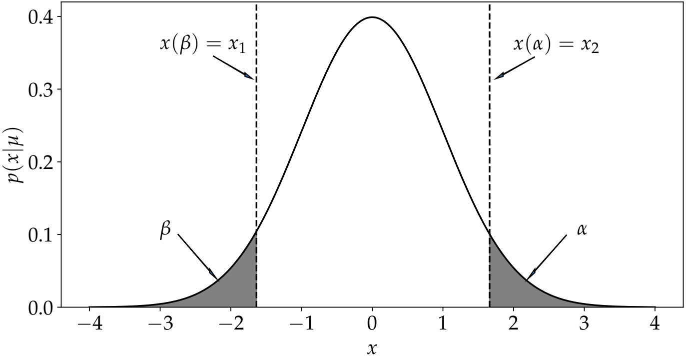
- Since frequentists do not work with posterior PDFs, classical intervals are not statements about μ_t given an observable x
- For a frequentist, the range $[\mu_1, \mu_2]$, called the **confidence interval**, is a member of a set such that

$$p(\mu \in [\mu_1, \mu_2]) = \alpha$$

- The values μ_1 and μ_2 are functions of x, and refer to the varying intervals from an ensemble of experiments with fixed μ
- Frequentist: $[\mu_1, \mu_2]$ contains the fixed, unknown μ_t in a fraction α of hypothetical experiments
- Bayesian: the degree of belief that μ_t is in $[\mu_1, \mu_2]$ is α
- These views *can* correspond, but they don't have to

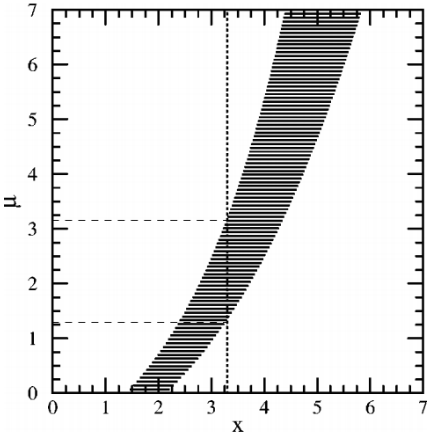
Constructing a Neyman-Pearson Interval

To construct a confidence interval, we begin with $p(x|\mu)$, the PDF of the observable given a fixed value of the parameter μ:



The observable x has probability $1 - \alpha - \beta$ to fall in the unshaded region. We define a central confidence interval by setting $\alpha = \beta$

Constructing a Confidence Belt



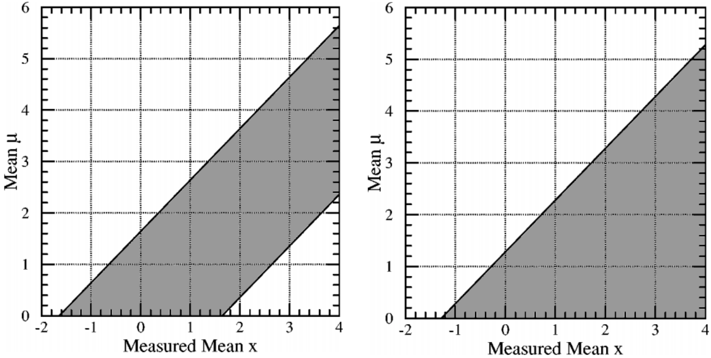
- Since μ varies, we now repeat this procedure for different values of μ
- Construct a confidence belt by calculating x_1 and x_2 such that

$$p(x < x_1|\mu) = p(x > x_2|\mu) = (1-\alpha)/2$$

  for each value of μ
- If we observe x_0, the confidence interval $[\mu_1, \mu_2]$ is the union of all values of μ defined by the vertical slice through the belt

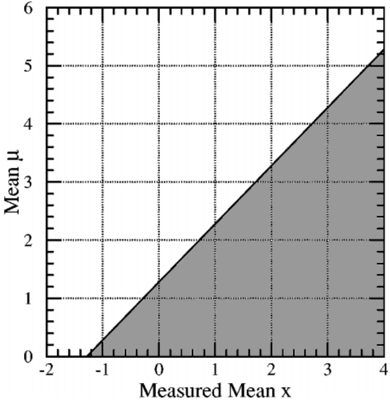
Central and Upper Intervals

To construct an upper interval, calculate $p(x < x_1|\mu) = 1 - \alpha$ for all μ.



Left: confidence belt for 90% C.L. central intervals for the mean of a Gaussian with a boundary at 0. Right: confidence belt for 90% C.L. upper limits for the mean μ (from [2])

Using an Upper Limit



This plot was constructed using the PDF

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2}\right].$$

Only positive values of μ are physically allowed, so the plot cuts off at $\mu < 0$.

This is perfectly valid, but what happens when the measurement is $x = -1.8$?

Draw a vertical line at $x = -1.8$; the confidence interval is an empty set. How can we interpret this?

Interpreting the Confidence Interval

- **Problem**: we set up the problem to ignore μ in the non-physical region $\mu < 0$, but observed $x = -1.8$ and found that $[\mu_1, \mu_2] = \emptyset$
- **Temptation**: we might want to conclude that all values of μ are ruled out by the measurement. Is this correct?

Interpreting the Confidence Interval

- **Problem**: we set up the problem to ignore μ in the non-physical region $\mu < 0$, but observed $x = -1.8$ and found that $[\mu_1, \mu_2] = \emptyset$
- **Temptation**: we might want to conclude that all values of μ are ruled out by the measurement. Is this correct?
- Nope! Remember what the 90% confidence interval tells you: given an ensemble of identical experiments, you should expect to construct an interval that contains the true value of μ 90% of the time
- If $[\mu_1, \mu_2] = \emptyset$ then conclude that you have conducted one of the 10% of experiments that fail to contain the true value
- It's a common mistake to conclude that the confidence interval tells you about the true value μ_t. It doesn't. A frequentist test won't tell you about μ_t, just the long-run outcome of many identical experiments

Frequentist Coverage

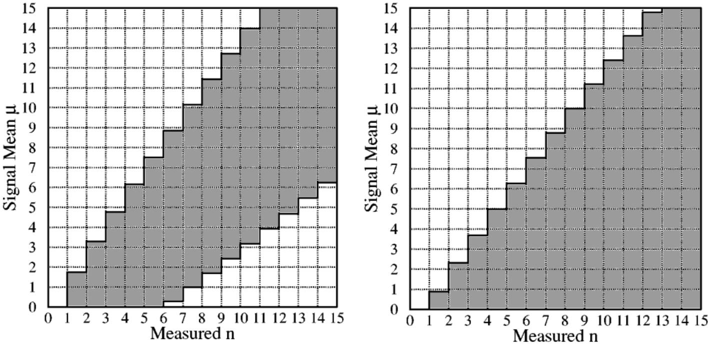
- If $p(\mu \in [\mu_1, \mu_2]) = \alpha$ is satisfied, one says that the intervals cover μ at the stated confidence (or equivalently, that the intervals have the correct "coverage")
- **Undercoverage**: there exists a μ such that $p(\mu \in [\mu_1, \mu_2]) < \alpha$
- **Overcoverage**: there exists a μ such that $p(\mu \in [\mu_1, \mu_2]) > \alpha$
- Undercoverage is a serious problem because it can lead to a Type I error, i.e., failure to accept a true null hypothesis (false discovery)
- If a set of intervals overcovers for some values of μ but never undercovers it is called "conservative"
- Conservative intervals are not considered as big of a problem, but they result in Type II errors, i.e., failure to reject a false null hypothesis (loss of power)

Poisson Process with Background

Built-in Overcoverage

Suppose you are counting discrete events $x \to n$ where n consists of signal events with unknown mean μ and background events with known mean b:

$$p(n|\mu) = (\mu + b)^n \exp\left[-(\mu + b)\right] / n!$$



Poisson Process with Background

Built-in Overcoverage

- Because in the Poisson process n is an integer, we will sometimes find that

$$p(\mu \in [\mu_1, \mu_2]) \neq \alpha$$

  simply because the discrete intervals cannot cover μ

- Convention: for the Poisson process, instead choose to satisfy

$$p(\mu \in [\mu_1, \mu_2]) \geq \alpha$$

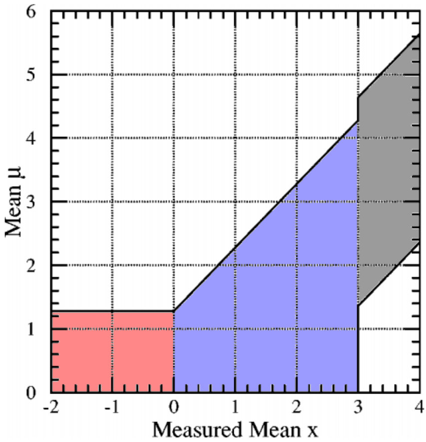
  in all edge cases. I.e., systematically overcover when necessary

- This is pretty sub-par. We want a 90% interval to fail to contain the true value 10% of the time. Overcoverage means this happens less than that
- Unfortunately, the overcoverage in the Poisson case is not done by choice. It's a consequence of the discreteness of the counts

Other Limits on Producing Confidence Intervals

- There is a serious limitation with classical confidence intervals: for coverage to be meaningful, you must decide ahead of time what kind of interval to calculate
- If it is *determined before conducting an experiment* that an upper limit is appropriate, then the triangular confidence belt shown earlier is perfectly fine
- If it is *determined before conducting an experiment* that a central limit is appropriate, then the central confidence belt shown earlier is perfectly fine
- But, if the experimenter decides to publish an upper or central interval *based on the results of the experiment* – a completely reasonable thing to do, by the way – then things go bad very quickly

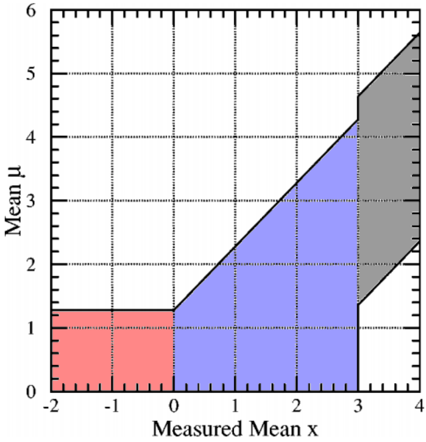
Flip Flopping



Suppose a physicist measures a quantity x and decides to publish results about μ as follows:

- If $x < 0$, publish an upper limit on μ to be "conservative" (red)
- If the measurement of x is $< 3\sigma$, calculate an upper limit on μ (blue)
- If the measurement of x is $\geq 3\sigma$, calculate a central confidence interval (gray)
- We say the physicist "flip-flops" between publishing central intervals and upper limits

The Problem with Flip Flopping



- Flip-flopping shows up as kinks in the confidence belt. It's a problem because μ is undercovered if $x < 3\sigma$
- For $\mu = 2$, the interval $[x_1 = 2 - 1.28, x_2 = 2 + 1.64]$ contains only 85% of the probability defined by

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2}\right]$$

- Hence, most of the intervals on this plot don't cover μ and are not conservative

Table of Contents

Alternative Methods of Constructing a Confidence Interval

There is quite a bit of freedom in how you can construct a confidence interval, so there are several approaches for how to draw the 90% interval over x for a fixed μ:

- Upper or lower limits: add all x greater than or less than a given value
- Central intervals: draw a central region with equal probability of x falling above or below the region
- Ranking: starting from x which maximizes $p(x|\mu)$, keep adding values of x, ranked by $p(x|\mu)$, until the interval contains 90% of the probability

This last ordering scheme is closely related to the so-called Feldman-Cousins method, which can be used to get around the flip-flopping problem [2]

The Feldman-Cousins Method

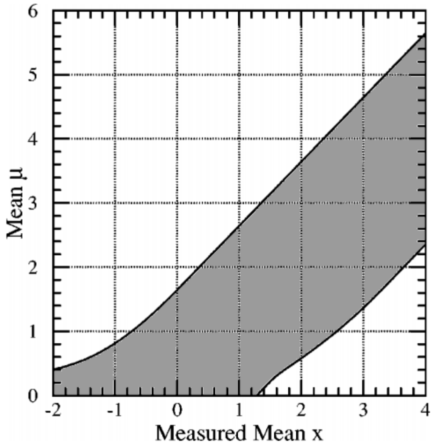
- For each x, let $\hat{\mu}$ be the physically allowed value of the mean μ which maximizes $p(x|\mu)$. I.e., $\hat{\mu}$ is the MLE
- Then calculate the likelihood ratio

$$R = \frac{p(x|\mu)}{p(x|\hat{\mu})}$$

- For μ fixed, add values of x to the interval from higher to lower R until the desired probability content is realized, e.g., 90%
- Gaussian example: $\hat{\mu} = x$ if $x \geq 0$ and 0 if $x < 0$. So

$$R = \frac{p(x|\mu)}{p(x|\hat{\mu})} = \begin{cases} \exp\left[-\frac{(x-\mu)^2}{2}\right] \Big/ 1 & x \geq 0 \\ \exp\left[-\frac{(x-\mu)^2}{2}\right] \Big/ \exp\left[-\frac{x^2}{2}\right] & x < 0 \end{cases}$$

A Feldman-Cousins Interval



- Left: Feldman-Cousins confidence interval for a Gaussian μ with a boundary at 0
- The ratio ensures that the confidence interval is never an empty set
- There are no more kinks in the confidence belt; it transitions smoothly between upper limits and central intervals given a measurement x
- Procudure: take data and calculate the FC interval. If x is small, then the method automatically returns $\mu_1 = 0$

Using Feldman-Cousins Intervals in Practice

Common application: quote a limit on the size of a signal given a known background:

$$p(\mu|b, n_0) = (\mu + b)^{n_0} \exp\left[-(\mu + b)\right] / n_0!$$

$n_0 \backslash b$	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	5.0
0	0.00, 2.44	0.00, 1.94	0.00, 1.61	0.00, 1.33	0.00, 1.26	0.00, 1.18	0.00, 1.08	0.00, 1.06	0.00, 1.01	*0.00, 0.98*
1	0.11, 4.36	0.00, 3.86	0.00, 3.36	0.00, 2.91	0.00, 2.53	0.00, 2.19	0.00, 1.88	0.00, 1.59	0.00, 1.39	0.00, 1.22
2	0.53, 5.91	0.03, 5.41	0.00, 4.91	0.00, 4.41	0.00, 3.91	0.00, 3.45	0.00, 3.04	0.00, 2.67	0.00, 2.33	0.00, 1.73
3	1.10, 7.42	0.60, 6.92	0.10, 6.42	0.00, 5.92	0.00, 5.42	0.00, 4.92	0.00, 4.42	0.00, 3.95	0.00, 3.53	0.00, 2.78
4	1.47, 8.60	1.17, 8.10	0.74, 7.60	0.24, 7.10	0.00, 6.60	0.00, 6.10	0.00, 5.60	0.00, 5.10	0.00, 4.60	0.00, 3.60
5	1.84, 9.99	1.53, 9.49	1.25, 8.99	0.93, 8.49	0.43, 7.99	0.00, 7.49	0.00, 6.99	0.00, 6.49	0.00, 5.99	0.00, 4.99
6	2.21,11.47	1.90,10.97	1.61,10.47	1.33, 9.97	1.08, 9.47	0.65, 8.97	0.15, 8.47	0.00, 7.97	0.00, 7.47	0.00, 6.47
7	3.56,12.53	3.06,12.03	2.56,11.53	2.09,11.03	1.59,10.53	1.18,10.03	0.89, 9.53	0.39, 9.03	0.00, 8.53	0.00, 7.53
8	3.96,13.99	3.46,13.49	2.96,12.99	2.51,12.49	2.14,11.99	1.81,11.49	1.51,10.99	1.06,10.49	0.66, 9.99	0.00, 8.99
9	4.36,15.30	3.86,14.80	3.36,14.30	2.91,13.80	2.53,13.30	2.19,12.80	1.88,12.30	1.59,11.80	1.33,11.30	0.43,10.30

The lookup table for 90% C.L. is reprinted from the Feldman-Cousins paper [2]. For $b = 4$, we have to observe at least $n_0 = 8$ events before $\mu_1 \neq 0$. We'd then say that we exclude $\mu = 0$ at the 90% C.L.

Remaining Conceptual Problems

Problems remain with the interpretation of data when the number of events are fewer than the expected background

Example

Experiment 1: $b = 0,\ \ n_0 = 0 \implies \mu \in [0.00, 2.44]$ at 90% C.L.
Experiment 2: $b = 15, n_0 = 0 \implies \mu \in [0.00, 0.92]$ at 90% C.L.

What's going on here? Experiment 1 worked hard to remove their background. Experiment 2 did not and expected a much higher background.

Neither experiment observed any events, but Experiment 2, by common sense the "worse" of the two experiments, has a smaller Feldman-Cousins confidence interval than Experiment 1. Seems pretty unfair, no?

Don't Confuse Intervals with Posterior Probabilities

- The origin of the paradox is that it's easy to think of the smaller confidence interval as a tighter constraint on μ_t
- But that is thinking about the interval as if it is equivalent to

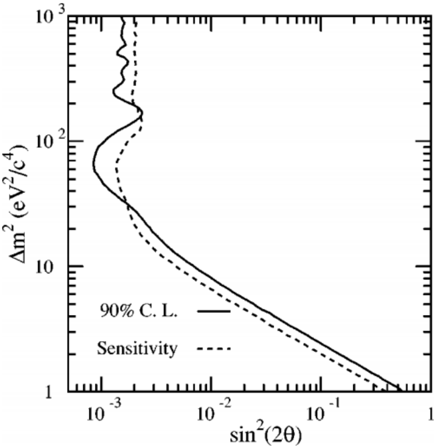
$$p(\mu_t|x_0) = \text{posterior probability of } \mu_t$$

- Remember, we are calculating $p(x_0|\mu)$. I.e., μ is fixed. If we need to make an inference about μ_t, we should be using a Bayesian framework (according to Feldman and Cousins themselves [2])
- So why even bother constructing a frequentist interval? Perhaps the question answers itself...
- If you're interested in the long-run behavior of many identical experiments, frequentist intervals are useful. But it's really easy to make basic conceptual mistakes when using them

Reporting Sensitivity

If $n_0 < b$, report the "sensitivity," the average upper limit obtained with an ensemble of background-only experiments, as well as calculated limits [2]

b	68.27% C.L.	90% C.L.	95% C.L.	99% C.L.
0.0	1.29	2.44	3.09	4.74
0.5	1.52	2.86	3.59	5.28
1.0	1.82	3.28	4.05	5.79
1.5	2.07	3.62	4.43	6.27
2.0	2.29	3.94	4.76	6.69
2.5	2.45	4.20	5.08	7.11
3.0	2.62	4.42	5.36	7.49
3.5	2.78	4.63	5.62	7.87
4.0	2.91	4.83	5.86	8.18
5.0	3.18	5.18	6.32	8.76
6.0	3.43	5.53	6.75	9.35
7.0	3.63	5.90	7.14	9.82
8.0	3.86	6.18	7.49	10.27
9.0	4.03	6.49	7.81	10.69
10.0	4.20	6.76	8.13	11.09
11.0	4.42	7.02	8.45	11.46
12.0	4.56	7.28	8.72	11.83
13.0	4.71	7.51	9.01	12.22
14.0	4.87	7.75	9.27	12.56
15.0	5.03	7.99	9.54	12.90



Reporting Sensitivity

Back to the example:

Example

Experiment 1: $b = 0, \;\; n_0 = 0 \implies \mu \in [0.00, 2.44]$ at 90% C.L. The sensitivity is 2.44 at 90% C.L.

Experiment 2: $b = 15, n_0 = 0 \implies \mu \in [0.00, 0.92]$ at 90% C.L. The sensitivity is 4.83 at 90% C.L.

The upper limit from Experiment 2 (0.92) is much smaller than its sensitivity (4.83), implying that the experiment benefitted from a huge and rather unlikely downward fluctuation in n_0.

Fluctuations happen, even into non-physical regions (remember the m^2 example). Frequentists have to publish these fluctuations no matter what since the results from many experiments is of interest. Failure to do so will bias meta-analyses of the literature

Summary

Constructing a frequentist confidence interval means that you identify some confidence level α and then build a set $[\mu_1, \mu_2]$ that has probability α of containing μ_t. Unfortunately:

- Sometimes the confidence interval is an empty set
- Intervals have kinks if you flip-flop between upper limits and central measurements
- You can't simply cut data in unphysical regions

If your data imply an unphysical result, too bad; you ran one of the $1 - \alpha$ fraction of experiments with an interval that doesn't contain μ_t.

The Feldman-Cousins method exploits the fact that you can construct a Neyman interval in several ways. Ranking x by its likelihood ratio allows you to fix some of the pathologies in interval construction

References I

[1] Glen Cowan. *Statistical Data Analysis*. New York: Oxford University Press, 1998.

[2] Gary J. Feldman and Robert D. Cousins. "A Unified approach to the classical statistical analysis of small signals". In: *Phys.Rev.* D57 (1998), pp. 3873–3889. arXiv: `physics/9711021 [physics.data-an]`.