Statistics 3/2

19th JINR-ISU Baikal Summer School on Physics of Elementary Particles and Astrophysics

Denis Derkach

Contents

Frequentist Confidence Intervals

Frequentist Confidence intervals

Definition

The confidence interval is an interval constructed using a random sample from a distribution with an unknown parameter, such that it contains the given parameter with a given probability. I.e

$$\mathbb{P}(L \leqslant \theta \leqslant U) = p.$$

Note that for the Bayesian approach:

$$\mathbb{P}(L \leqslant \theta \leqslant U|X)$$

Coverage of confidence intervals

From frequentist point of view.

Definition

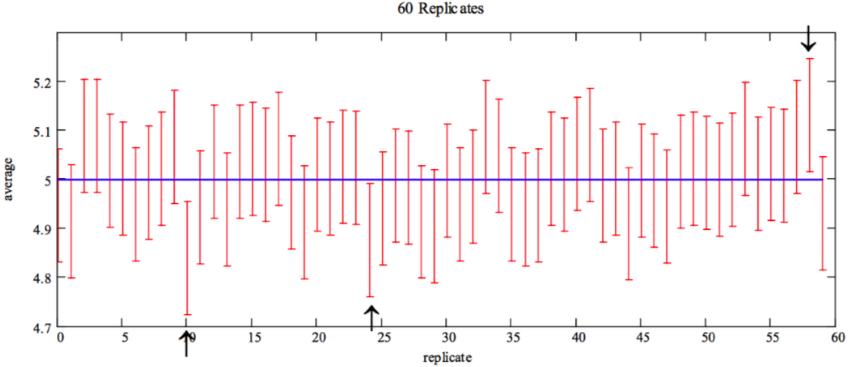
The coverage probability for an interval estimate is the proportion of instances in which the sample statistic obtained from infinite independent and identical replications of the experiment is contained.

NB: The existence of term coverage for Bayesian varies from book to book.

Coverage of confidence intervals

Example

60 experiments from $\mathcal{N}(5; 0.1)$. The 95% Confidence interval is given by $\mu = \text{avg} \pm 0.116$



We have $3/60 \approx 0.05$ intervals that contain true value.

Observed coverage

In practice, methods that have only asymptotic coverage are mainly used. They are characterised by the observed coverage, p_{method}. If $p \geq p_{\mathrm{method}}$ this is called undercoverage, if $p \leq p_{\mathrm{method}}$, this is overcoverage.

NB: overcoverage is less of a problem (but this reduces the quality of the experiment).

Normal theory

Let us take $X \sim N(\mu; \sigma^2)$. For known μ and σ^2:

$$\beta = \mathbb{P}(a \leq X \leq b) = \int_a^b N(\mu, \sigma^2) dX'.$$

If μ is unknown, we can no longer numerically calculate this integral; instead, we can estimate the probability $[\mu + c, \mu + d]$:

$$\beta = \mathbb{P}(\mu + c < X < \mu + d) = \int_{\mu+c}^{\mu+d} N(\mu, \sigma^2) dX' =$$

$$= \int_{c/\sigma}^{d/\sigma} \frac{1}{\sqrt{2\pi}} \exp[\frac{1}{2} Y^2] dY.$$

which means that $\beta = \mathbb{P}(X - d \leq \mu \leq X - c)$

Normal theory for interval estimation

The normal theory worked since:

- we were able to obtain a function that depends on $(X - \mu)^2$;
- the function is integrable for any limit.

These properties are fulfilled asymptotically for likelihood functions.
NB: We need more events for this.

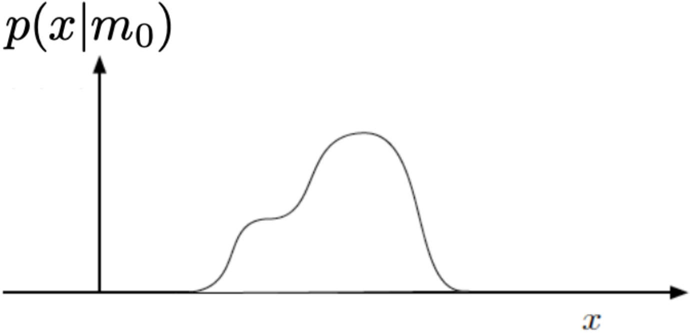
NB2: All said is easily extrapolated for multidimensional models.

Neyman construction

The Neyman construction for constructing frequentist confidence intervals involves the following steps:

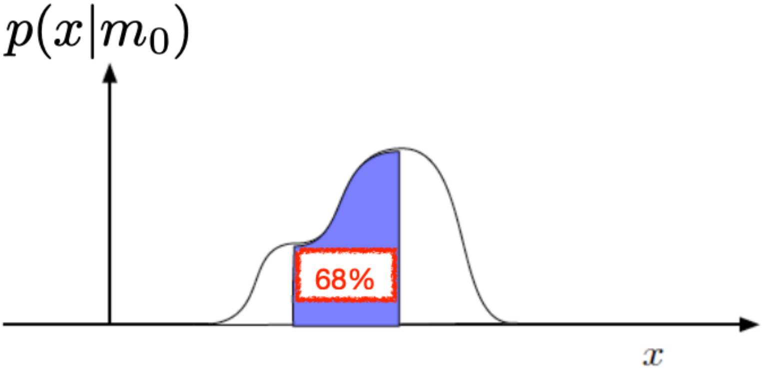
- Given a true value of the parameter m, determine a PDF $p(x|m)$ for the outcome of the experiment.
- Using some procedure, define an interval in x that has a specified probability (say, 68%) of occurring.
- Do this for all possible true values of m, and build a confidence belt of these intervals.
- Compute the confidence belt given the value of x observed.

Neyman construction



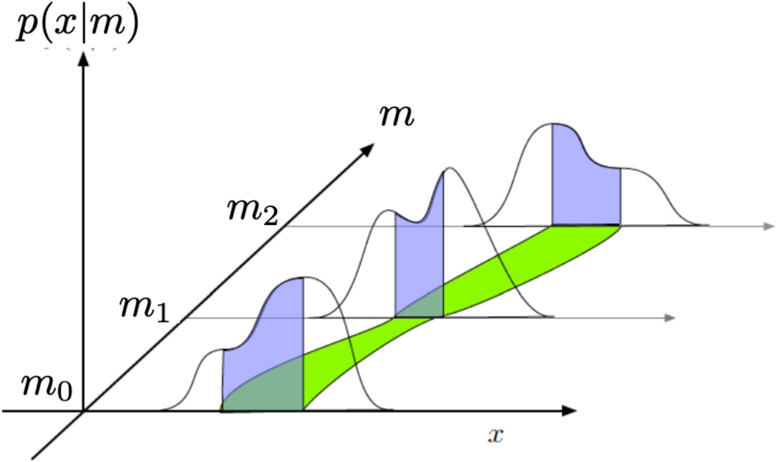
Given a true value of the parameter m, determine a PDF $p(x|m)$ for the outcome of the experiment.

Neyman construction



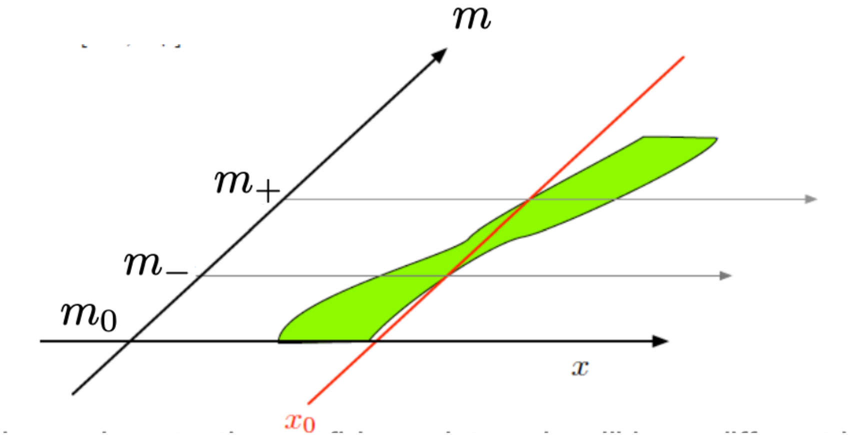
Using some procedure, define an interval in x that has a specified probability (say, 68%) of occurring.

Neyman construction



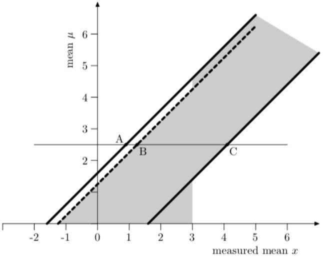
Do this for all possible true values of m, and build a confidence belt of these intervals.

Neyman construction



Compute the confidence belt given the value of x_0 observed.
Finally, the parameter m lies in the interval $[m_-; m_+]$.

Neyman construction: problems

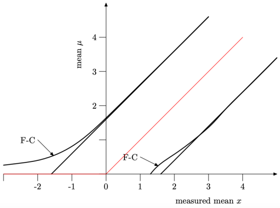


Close to the boundary problems:

- empty intervals;
- "flip-flop" in the regions close to but not touching physics boundaries.

These problems are solved using additional constructions, for example, a unified approach (Feldman-Cousins, see below) proposes to supplement forbidden regions by analyzing the relative likelihood.

Two words on Feldman-Cousins



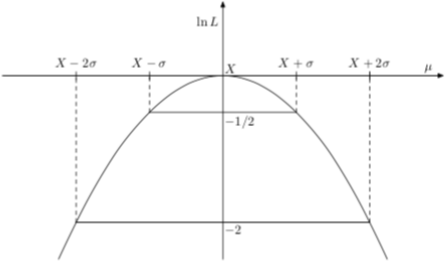
- introduce ordering principle:

$$\frac{\mathcal{L}(x;m)}{\mathcal{L}(x;m_{best})}$$

- mostly solves the problems mentioned before
- have difficulties addressing case of many nuisance parameters.

Likelihood based confidence intervals

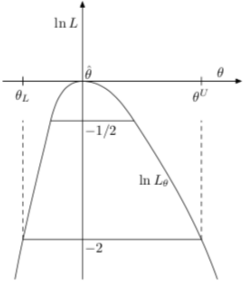
Motivation



Log-likelihood function for Gaussian X, distributed $N(\mu, \sigma^2)$.

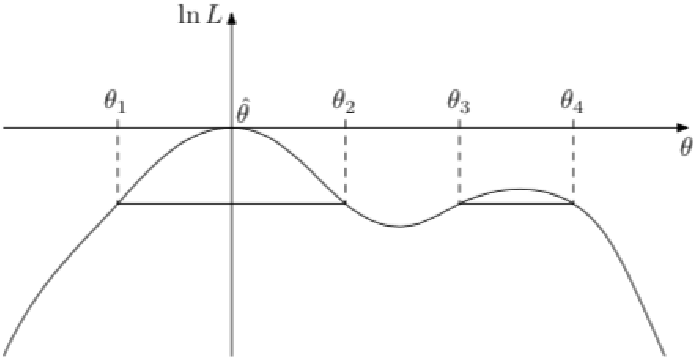
In previous slides, we have seen that the normal theory allows one to get an honest confidence interval for quantities distributed normally. This result can be read differently: if the log-likelihood is parabolic, then we can honestly calculate the confidence intervals.

Likelihood independent on the parameterization



If the likelihood function is nonparabolic, we can (almost) always bring it to a parabolic form by some transformation $g(\theta)$. At the same time, the function itself does not depend on the parameterization, therefore we can evaluate θ_L and θ_H in terms of $\ln L = \ln L_{\max} - 1/2$ (for the 68 % interval).

Nontrivial Cases



"Pathological" log-likelihood function.

In case of a multimodal likelihood function there is a chance to find a second peak (and use it in disjoint CL).

Multidimensional case

The biggest problems begin in the multidimensional case.

- Use normal theory (if likelihood is Gaussian).
- Easy way to use the likelihood profile function:

$$g(x_k) = \max_{x_i, i \neq k} \ln L(X).$$

  This method will make it possible to analyze simple non-Gaussian likelihoods.
- Use plugin method (create a set of toys in each point of parameter of interest and check the likelihood value of you fit for the toy). (recommended, but very CPU consuming).

Systematic uncertainties

In general, each source of systematic error is characterized by its own random variable (or rather, almost every). Suppose we know the density of this random variable:

- Bayesian way: no problem, just marginalize credibility;
- frequentist way: the task becomes very multidimensional;
- mixed way: let's pretend that we are Bayesians, marginalize, and then use as a frequentist inference.
- make a combination of the profiling and classic ways.

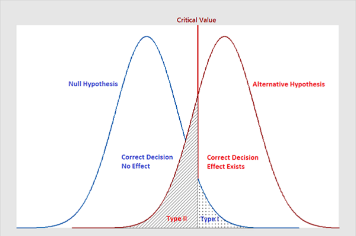
Hypotheses testing

Hypotheses

Statistical tests are often formulated using a

- Null hypothesis (eg, Standard Model (SM) background only)
- Alternative hypothesis (eg, SM background + new physics)

Hypothesis being some statement about parameter. To run hypothesis test we construct some summary statistics for both hypotheses and select a critical value.



Type I and II error

		Null Hypothesis	
		TRUE	FALSE
Test result	accept	OK True Negative	type 2 error ($\mathbb{P} = \beta$) False Negative
	reject	type 1 error($\mathbb{P} = \alpha$) False Positive	OK True Positive

We want the test to provide low α and β simultaneously. For this, we are looking for the most powerful test.

Neyman-Pearson test for two simple hypotheses

Лемма (Neyman-Pearson)

$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta_1$

Neyman-Pearson test statistics:

$$T = \frac{\mathfrak{L}(\theta_1)}{\mathfrak{L}(\theta_0)} = \frac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)}. \tag{1}$$

Suppose that H_0 is rejected for $T \geq k$. Choose k such that $\mathbb{P}_{\theta_0}(T \geq k) = \alpha$.

Then, the Neumann-Pearson criterion (based on statistics (1)) will be the most powerful test $W(\theta_1)$ among all the criteria of size α.

Problems

- statistics must be fully known for any x;
- we can evaluate only simple hypotheses.

Instead, we can try to approximate likelihood locally. Or take a less powerful test (t-test, χ^2 test).

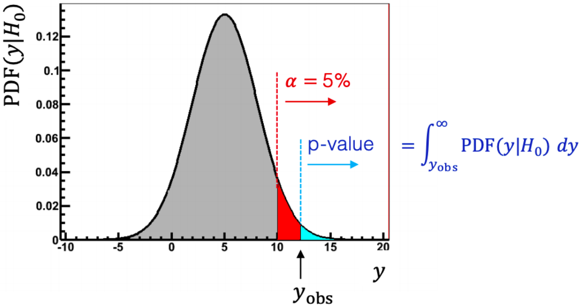
p-value

In frequentist statistics one cannot make a probabilistic statement about the true value of a parameter given the data. Instead:

- One defines acceptance / rejection regions of a test statistic (α).
- The measurement (data) is one specific outcome of an ensemble of possible data.
- One accepts or rejects H_0 confidence level given by α.
- It is also possible to state how probable a particular or worse outcome (test statistic measurement) is for a given hypothesis (eg.H_0 p-value).

One then shows the data and quotes the H_0 outcome given the required confidence level and the hypothesis p-value.

p-value



NB: α must be predefined!

p value

- A p-value can tell you that a difference is statistically significant, but it tells you nothing about the size or magnitude of the difference.
- A low p-value can give us a statistical evidence to support rejecting the null hypothesis, but it does not prove that the alternative hypothesis is true.
- p-value doesn't tell you anything directly about what you're observing, it tells you about your odds of observing it.

Bayes approach

We need to find:

$$\mathbb{P}(hyp|data) = \frac{\mathbb{P}(data|hyp)\mathbb{P}(hyp)}{\mathbb{P}(data)}$$

Normalization can be found by integrating over all possible parameter values, which is rather difficult for some types of hypotheses. We can study the bf Bayev factor:

$$R = \frac{\mathbb{P}(H_0|data)\mathbb{P}(H_1)}{\mathbb{P}(H_1|data)\mathbb{P}(H_9)}$$

The resulting ratio can be considered as a chance of success at the rate of H_0 versus H_1. The ratio will still depend on a priori knowledge.

Lindley paradox

Testing a point null hypothesis against a non-point alternative. For example, coin tosses:

- $H_0 : p = 0.5.$
- $H_1 : p! = 0.5.$

In an experiment by Jahn, Dunne and Nelson (1987), it says that at 104490000 attempts, 52263471 eagles and 52226529 tails were received. What does this mean in terms of statistics?

Lindley paradox

- Frequentist approach:

$$z(x) = \sqrt{\frac{N}{\theta_0(1-\theta_0)}}\left(\frac{1}{N}\sum x_i - \theta_0\right),$$

  i.e p-value: $p \ll 0.01$, H_0 is not supported.

- Bayes factor:

$$R = \frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \approx 19.$$

  H0 Should be accepted!

Solution: the answers are different!

- frequentist approach says that the null hypothesis poorly explains the data;
- Bayesian approach says that the null hypothesis describes the data better than all alternative ones.