Statistics 2/2

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

Denis Derkach

Contents

PDF in Multidimensional Case

Multidimensional distributions

We often encounter situations where we have to analyze several random variables at once. In this case, we need to analyze a more complex entity, the multidimensional PDF $\mathbb{P}(\xi_1 \leq x_1, \ldots, \xi_n \leq x_n)$ for a random vector $\xi = (\xi_1, \ldots, \xi_n)$.

Independence of random variables

Definition

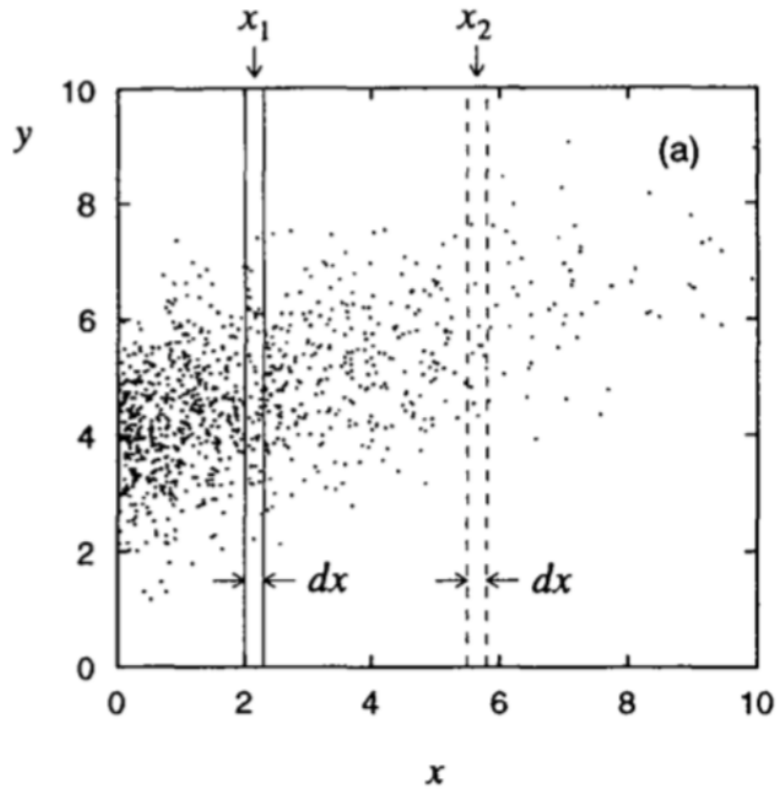
Let random variables X and Y have a joint density $p(x, y)$. X and Y will be called independent if

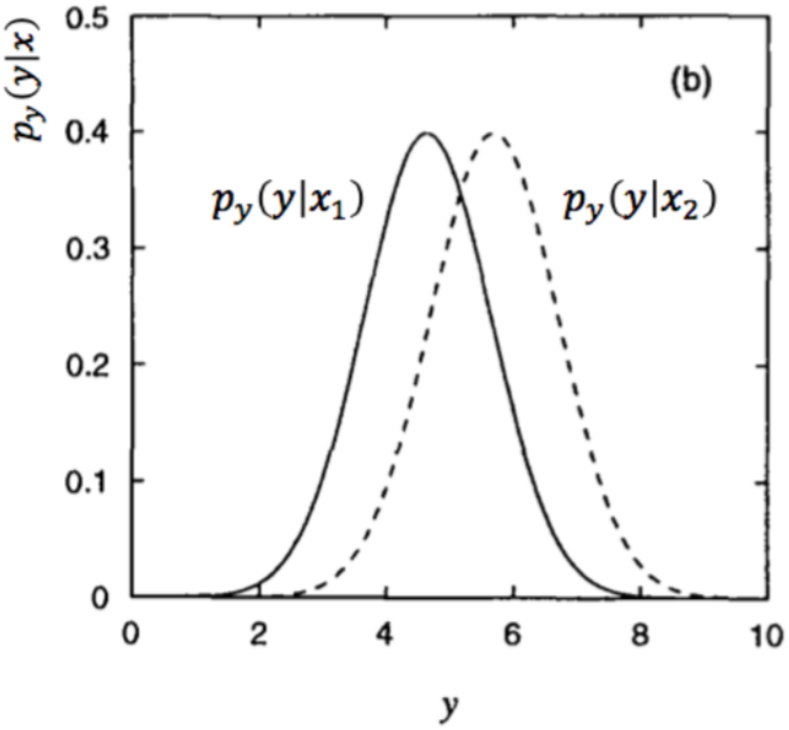
$$p(x, y) = p(x) \cdot p(y).$$

Back to 1D

We like 1D data. Remember:

- Frequentist: the more frequent - the more probable.
- Bayesian: the more I believe - the more probable.





Covariance

Let X, Y be two random variables.

Definition

The covariance of X and Y will be defined:

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$$

Covariance coefficients we can written as matrix.

Covariance properties

- X и Y independent, than $\mathrm{cov}(X, Y) = 0$ (not vice versa).
- $\mathrm{cov}(X, X) = \mathbb{V}arX$.
- $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$.
- $\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.
- $\mathrm{cov}(aX, bY) = ab \cdot \mathrm{cov}(X, Y)$.
- $\mathrm{cov}(X + a, Y + b) = \mathrm{cov}(X, Y)$.
- $\mathrm{cov}^2(X, Y) \leq \mathbb{V}arX\mathbb{V}arY$.

Pearson Correlation Coefficient

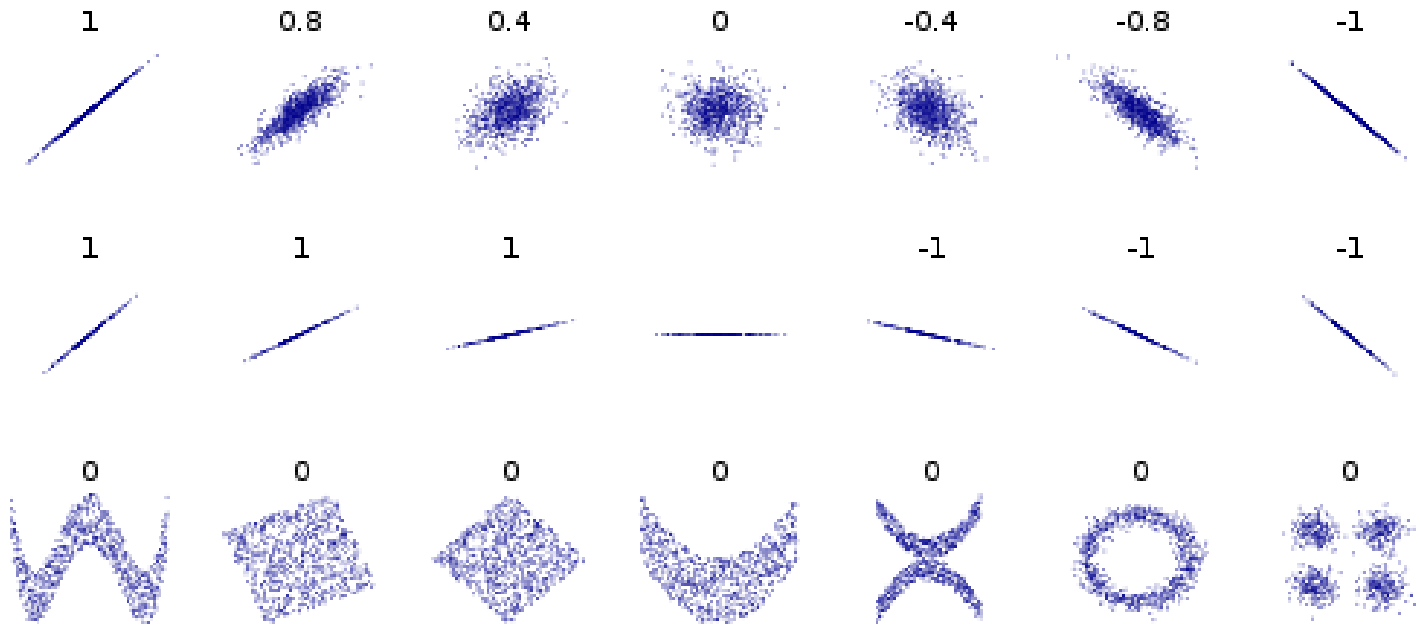
Definition

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\mathbb{V}arX\mathbb{V}arY}}$$

If X, Y are independent: $\rho = 0$, ie, they are uncorrelated.

Pearson coefficient values

The Pearson coefficient however can be misleading. For cases, where nonlinearities are expected it is better to use mutual information.



Estimate of the Pearson Correlation Coefficient

Evident estimate:

$$\widehat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

This estimate is biased for low number of events in sample. There are

other possibilities:

`https://arxiv.org/pdf/1707.09037.pdf`

Likelihood

Likelihood

Notice, that when we wrote PDF, we did not assume anything about parameters. What if we have the data already:

$$\mathbb{P}(data|parameters)\Big|_{data\ obs.} = \mathcal{L}(parameters)$$

$\mathcal{L}$ is called the Likelihood Function.
NB: it's not a probability,

How many events in the sample?

We consider each measurement in the sample as an independent, identically distributed random variable, which means the PDF is a product of individual PDFs. The same applies to likelihood:

$$\mathcal{L}_{\backslash}(\theta) = \prod_{i=1}^{n} \mathcal{L}_{\rangle}(\theta)$$

Transformation Behaviour

What happens, if we transform data points or parameters?

- likelihood function remains invariant:

$$\mathcal{L}(\theta) = \mathcal{L}(\tau(\theta))$$

.

- for PDF invariant is the integrated probability between corresponding points. So, tranformation takes into account Jacobian $X \to Y(X)$:

$$PDF(X) = J(X,Y)PDF(Y)$$

, where $J = \frac{\partial X}{\partial Y}$.

Point Estimation

Problem Statement

Parametric estimation: We need to estimate the value $T(\theta)$, where T — some function of the model parameter θ.

$$\begin{aligned} T: \quad & \Theta \to \mathcal{Y}, \\ & \theta \mapsto T(\theta). \end{aligned}$$

This means that we need to have an **estimate** $\hat{T}$ using data sample X:

$$\hat{T} : \mathcal{X} \to \hat{\mathcal{Y}}.$$

NB: $\mathcal{Y}$ and $\hat{\mathcal{Y}}$ are not always the same.
NB2: Estimates can be determined and randomized.

Moments of distribution

Let $\theta = (\theta_1, \ldots, \theta_k)$ — be our parameters. We can than obtain for $1 \leq j \leq k$ the j-th moment:

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E}(X^j) = \int x^j dF_\theta(x),$$

the j-th sample moment can be obtained using formula:

$$\widehat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

Method of Moments

Definition

$\widehat{\theta}_n$ — parameter estimate of $\theta = (\theta_1, \dots, \theta_k)$ if

$$\alpha_1(\widehat{\theta}_n) = \widehat{\alpha}_1,$$
$$\alpha_2(\widehat{\theta}_n) = \widehat{\alpha}_2,$$
$$\dots$$
$$\alpha_k(\widehat{\theta}_n) = \widehat{\alpha}_k.$$

Example

Let $X_1, \dots, X_n \sim Bernoulli(p)$, then

- $\alpha_1 = \mathbb{E}(X) = p$,
- $\widehat{\alpha}_1 = n^{-1} \sum_{i=1}^n X_i$,
- then $\widehat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Example

Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, than

$$\alpha_1 = \mathbb{E}(X_1) = \mu,$$

$$\alpha_2 = \mathbb{E}(X_1^2) = \mathbb{V}ar(X_1) + (\mathbb{E}(X_1))^2 = \sigma^2 + \mu^2,$$

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\widehat{\sigma}^2 + \widehat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Solving the system of equations:

$\widehat{\mu} = \overline{X}_n$ и $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.

Теорема

Let $\widehat{\theta}_n$ — estimate of parameter θ that uses MoM, than (under certain conditions):

1. $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ when $n \to \infty$;
2. Estimate is asymptotically normal, i.e.

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, \Sigma),$$

where $\Sigma = g\mathbb{E}(XX^T)g^T$, $X = (X^1, X^2, \ldots, X^k)^T$, $g = (g_1, \ldots, g_k)$ и $g_j = \partial \alpha_j^{-1}(\theta)/\partial\theta$.

NB: The second bullet can be used to find confidence intervals.

Problem 1.

Comments about Method of Moments

- sub-optimal;
- easy to use;
- often use to have a preestimate of parameters for more precise methods.

Maximum Likelihood Estimator

Definition

Maximum Likelihood Estimator (MLE) is defined as the estimate $\widehat{\theta}_n$ of parameter θ, which maximizes likelihood: $\mathcal{L}_n(\theta)$ (with n being the number of events in a sample).

Practicalities

Example

Let us have an exponential distribution (say, we want to study a lifetime of a particle):

$$f(t,\tau) \sim \frac{1}{\tau} e^{-t/\tau}.$$

And a sample of i. i. d.measurements $t_1, \ldots, t_n \sim f$.
The likelihood function in this case is: $\mathcal{L}_n(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t/\tau}$. Taking a logarithm: $\ell_n(\tau) = \sum_{i=1} n f(t;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$.
Find maximum of $\mathcal{L}$ (taking into account that $\log$ is monotonic):

$$\frac{\partial L_n(\tau)}{\partial \tau} = 0 \rightsquigarrow \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \rightsquigarrow \widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i.$$

Some MLE properties

1. MLE is consistent: $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.
2. MLE does not depend on the parameterisation: $\widehat{\theta}_n$ — MLE for θ, than $g(\widehat{\theta}_n)$ — MLE for $g(\theta)$;
3. MLE is asymptotically normal: $(\widehat{\theta} - \theta_*)/\widehat{se} \rightsquigarrow \mathcal{N}(0,1)$;
4. MLE is asymptotically optimal.

Example of MLE:

Find $\widehat{\mu}$ and $\widehat{\sigma}$ for Normal function with number of events in sample n:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Rewrite as log-likelihood:

$$\ell_n(\mu,\sigma) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma - \frac{(x-\mu)^2}{2\sigma^2} \right)$$

Take derivatives:

$$\frac{\partial \ell_n}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \quad \frac{\partial \ell_n}{\partial \sigma} = \sum_{i=1}^{n} \left(\frac{(x_i-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right)$$

Example of MLE:

Thus:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and:

$$\widehat{\sigma} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

MLE estimate is not biased σ!

Signal and Background MLE

- usually samples contain signal and background;
- introduce mixed likelihood for every event:

$$\mathcal{L}_i(\mu) = \mu S \cdot p_s(x_i) + B \cdot p_b(x_i)$$

, $p_{s,b}$ are pdf's for signal and background, S and B are relevant number of events.

- add normalisation term $exp[-\nu)\frac{\nu^n}{n!}$, where $\widehat{\nu}$ is an estimate of n to account for the possibility of varying events in the sample.

We get:

$$\mathcal{L}(\mu) = exp[-\nu)\frac{\nu^n}{n!}\prod_{i=1}^{n}(\mu S \cdot p_s(x_i) + B \cdot p_b(x_i))$$

NB: More complications: nuisance parameters, many background contributions, normalisation that depends on parameters etc.

Problem 2.

How to test your maximum likelihood fit

- Try on several simulated samples (if available).
- Generate samples using model and fit them back (make it at least 100 times).
- Perform a goodness-of-fit test (for example, check χ^2 of the final parametrisation vs. data).

Maximum aposteriori estimate, MAP

Formally, MLE estimates the parameter values, for which our data is the most probable.

$$f(X;\theta) \sim f(X|\theta).$$

In fact, we normally ask, which are the values of the parameters that are most probable:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)},$$

for f, g и h — relevant pdfs.

Maximum aposteriori estimate

Definition

Maximum aposteriori estimate (MAP) is an estimate $\widehat{\theta}_n$ for parameter θ, that maximuses $f(\theta|X)$.

Connection to MLE

MAP and MLE are evidently connected:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)} = \frac{\prod_{i=1}^{n} f(X_i;\theta)g(\theta)}{h(X)} \sim \text{const} \prod_{i=1}^{n} f(X_i;\theta)g(\theta)$$

Taking log:

$$\log f(\theta|X) = \log(g(\theta)) + \sum_{i=1}^{n} \log(f(X_i|\theta)).$$

Basically, both methods give the same estimate modulo logarithm of apriori knowledge $\log(g(\theta))$.

Conjugate priors

Which $g(\theta)$ to use?

- any;
- normally, it's easier if the prior is coming from the same functional form as a posterior (a.k.a conjugate distributions).

The values of the parameters of conjugate distributions make sense of previous measurements.

See Wikipedia list

Thoughts on MAP

- allows you to take into account previous knowledge;
- gives a point estimate (thus not strictly Bayesian);
- depends on parameterization;
- for relatively large n coincides with MLE (and also in the case of $g(X) = \ rmconst$!).

Interval estimation

Interval estimation: motivation

- Usually we try to measure a parameter on a final sample.
- It would be interesting to understand our confidence in the result.

That is why we need interval estimation.

Bayes vs. Frequentist

As always, we have different approaches to solving the problem, depending on the interpretation of probabilities.

Bayesian credibility intervals

Credibility interval

Definition

A Bayesian p confidence interval is the $[L, U]$ interval to which the value of the θ parameter belongs with the posterior probability p

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

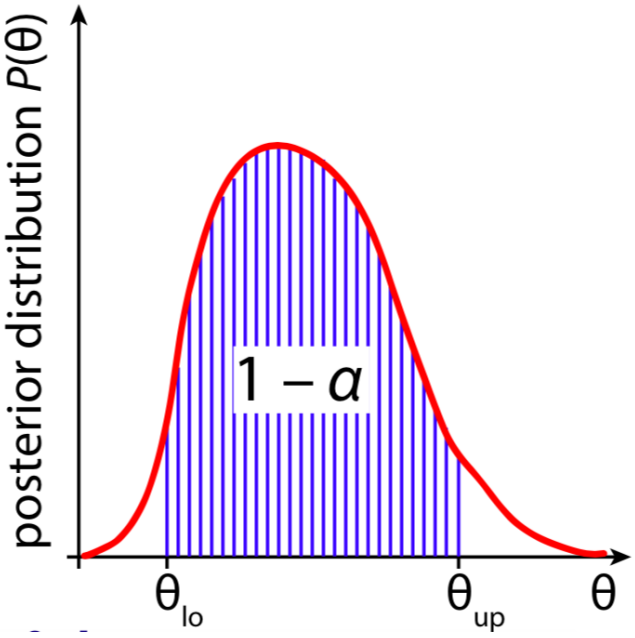
NB: normally written as Cr.L. (but often C.L. is used).

Bayesian credibility interval

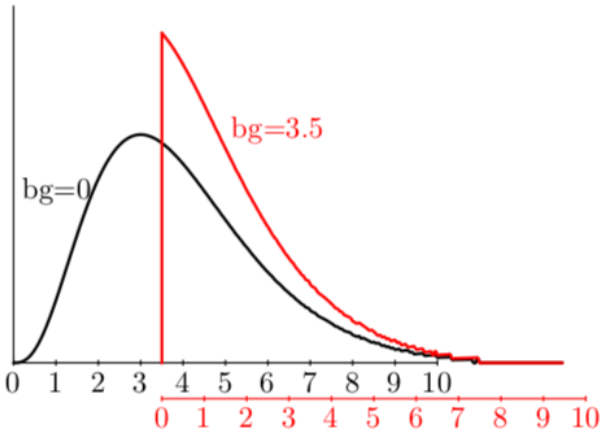
$$1-\alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|X)d\theta$$

How to choose $\theta_{\rm lo}$ и $\theta_{\rm hi}$:

- HPD (highest probability density) — only the highest probabilities to be chosen.
- Central interval — start integration from the peak.
- One sided interval — integrate from infinity.



Close to the boundary behavior



Bayesian 90% Upper Limits (Uniform Prior)

observed =	0	1	2	3
background = 0.0	2.30	3.89	5.32	6.68
0.5	2.30	3.50	4.83	6.17
1.0	2.30	3.26	4.44	5.71
2.0	2.30	3.00	3.87	4.92
3.0	2.30	2.83	3.52	4.37

The behavior near the boundaries as obtained in the Bayesian approach is very simple - we use an apriori distribution with information about the physical boundary.

The results obtained are very logical if we use a flat prior distribution with a clear left border.

Nuisance parameters

Definition

A nuisance parameter is any unknown parameter of the probability distribution in a statistical problem related to the study of other parameters of this distribution.

In the Bayesian approach, the inclusion of interfering parameters also occurs in a simple way (if we know its distribution $P(b)$, simply take integral:

$$\mathbb{P}(\theta|\text{data}) = \int_b \frac{\mathbb{P}(\text{data}|\theta, b)\mathbb{P}(\theta)}{\mathbb{P}(\text{data})}\mathbb{P}(b)db.$$

Measurement Combination

Another good feature of the Bayesian approach is a simple way combine several measurements:

$$\mathbb{P}(\theta|\mathrm{data}) = \frac{\mathbb{P}_1(\mathrm{data}|\theta)\ldots\mathbb{P}_N(\mathrm{data}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\mathrm{data})}.$$

It is enough to use prior distribution only once.
NB: it is sometimes useful to count the work in several passes.

Problem: prior probability

n the process of determining the boundaries of the experiment, we cared only about the left boundary. And what happens to the right? Should she be at infinity? in this case:

$$\int_b^a \text{Uniform}(x)dx = 0, \forall a, b$$

That is, we must also limit the right side, and we don't have information about it (or almost no information).

In addition, the use of a flat prior distribution is quite arbitrary.

Jeffreys prior probability

For each family of curves, the Jeffreys prior probability can be calculated from this condition. For example, for the Poisson distribution, Jeffreys proposed to make it invariant to scale. $\sim 1/\mu$.

Bayesian 90% Upper Limits ($1/\mu$ Jeffreys Prior)

observed =	0	1	2	3
background = 0.0	0.00	2.30	3.89	5.32
0.5	0.00	0.00	0.00	0.00
1.0	0.00	0.00	0.00	0.00
2.0	0.00	0.00	0.00	0.00
3.0	0.00	0.00	0.00	0.00

Jeffreys prior probability

Now, typically it is preferred to use distributions that minimize Fisher information for a given parameter. For the Poisson distribution:

$$P(\mu) = \frac{1}{\sqrt{\mu}}.$$

This does not produce correct intervals for noisy background. In order to correct this, we need to use:

$$\mathbb{P}(\mu) = \frac{1}{\sqrt{\mu + b}}.$$

Which means that our prior probability depends on our knowledge of background :-(

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

The method that allows you to construct an interval $(\theta_a; \theta_b)$ such that $\mathbb{P}(\theta_a \leq \theta_0 \leq \theta_b) = \beta$, where θ_0 is the real value of the parameter, has the property cover.

Frequentist intervals will fluctuate with new samples. Therefore, coverage is defined as the proportion of intervals that contains the current value of θ_0.

NB: the existence of coverage for the Bayesian approach is questionable.

Coverage of confidence intervals

In practice, methods that have only asymptotic coverage are mainly used. If $\mathbb{P} \leq \beta$ this is called undercoverage, if $\mathbb{P} \geq \beta$, this is overcoverage.

vspace 10pt

NB: overcoverage is less of a problem (but from the point of view of the experimenter, 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

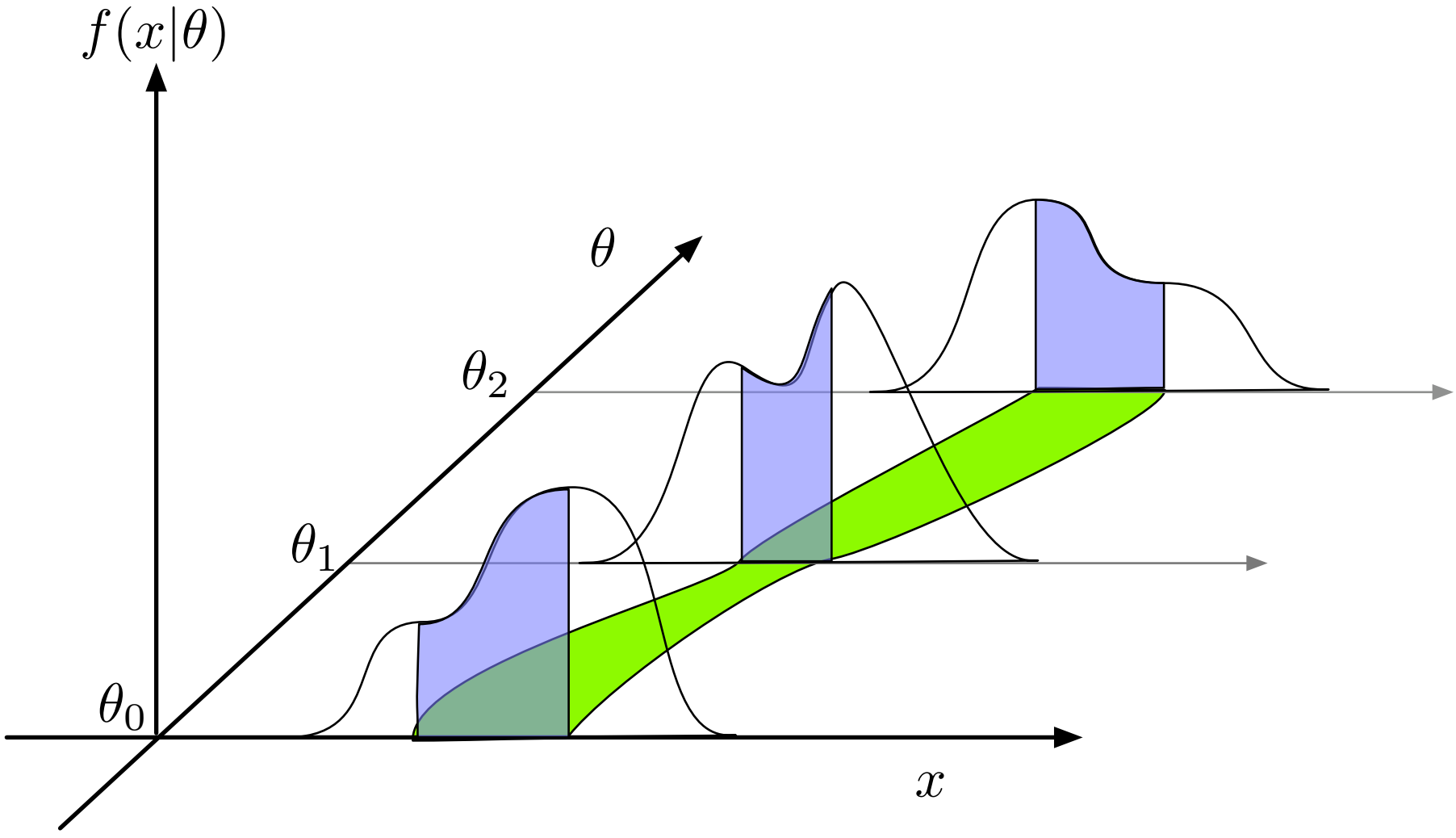
for multidimensional models.

Neyman construction

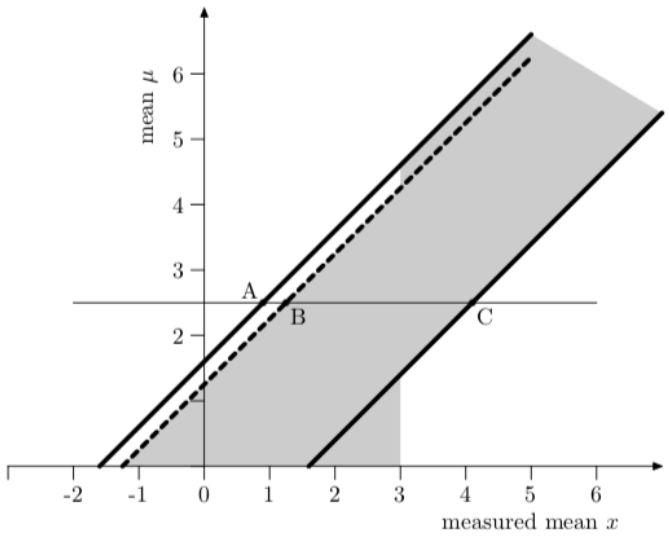
The Neyman construction for constructing frequentist confidence intervals involves the following steps:

- Given a true value of the parameter θ, determine a p.d.f. $f(x;\theta)$ for the outcome of the experiment. Often x is an estimator for the θ.
- Using some procedure, define an interval in x that has a specified probability (say, 90%) of occurring
- Do this for all possible true values of θ, and build a confidence belt of these intervals
- Compute the confidence belt given the value of x.

Neyman construction



Neyman construction: problems



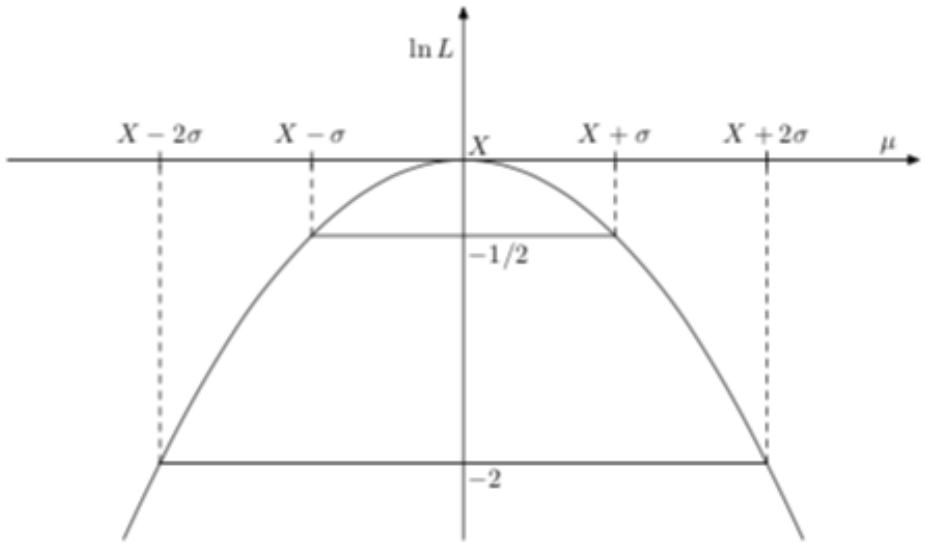
Close to the boundary problems:

- 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.

Likelihood based confidence intervals

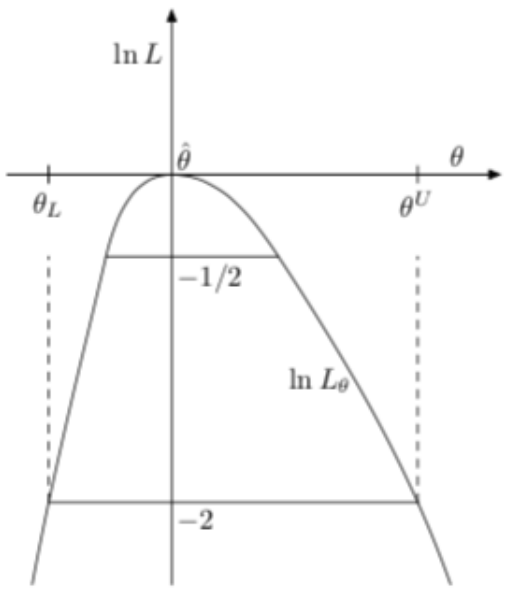
Motivation



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

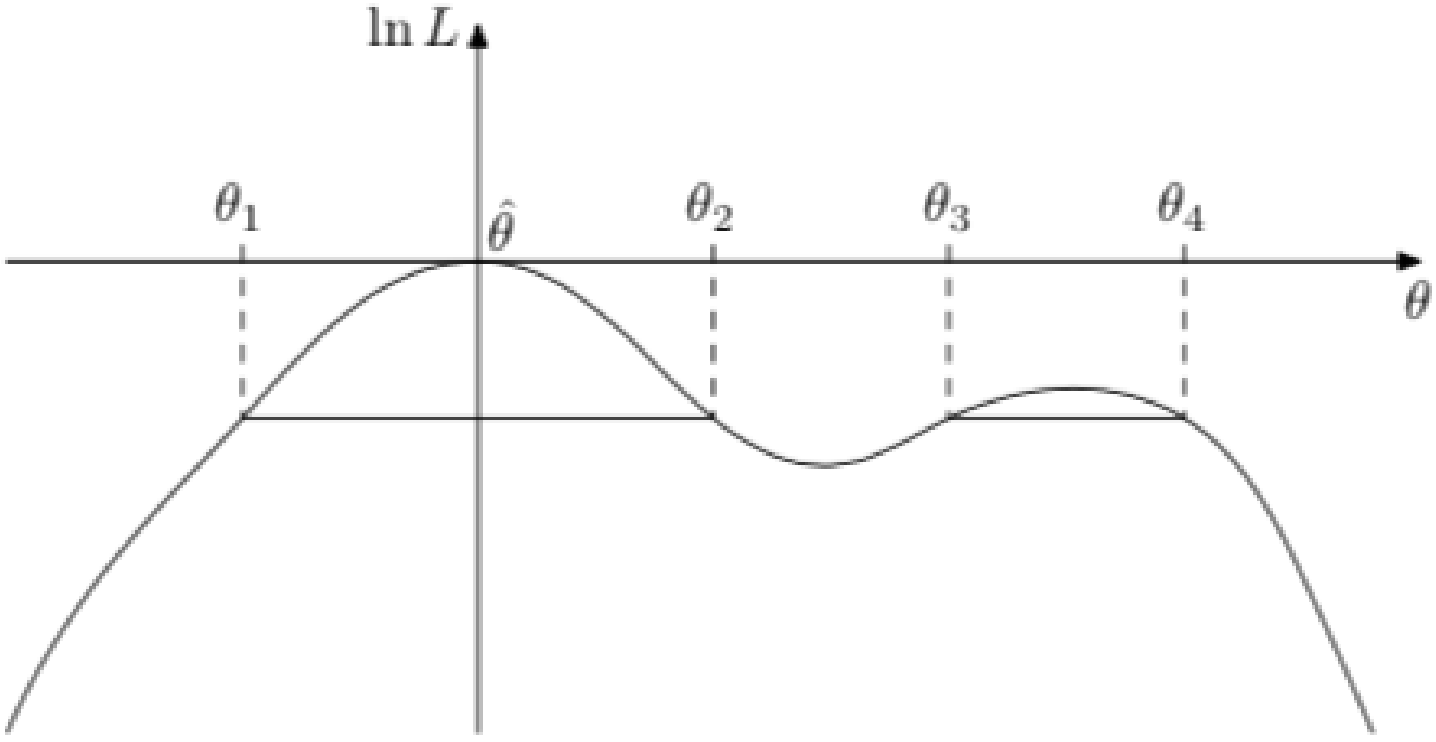
In the previous slides, we have seen that the normal theory allows one to get honest confidence intervals for quantities distributed according to Gauss. This result can be read differently: if the likelihood is parabolic, then we can honestly calculate the confidence intervals.

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 with such a construction, there is a chance to find a second peak (and use it in CL definition).

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:

$$g(x_k) = \ max\ limits_{x_i, i\ nek}\ lnL(X).$$

  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).

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;
- classic way: the task becomes very multidimensional;
- mixed way: let's pretend that we are Bayesians, marginalize, and then use as a classic conclusion.
- make a combination of the profiling and classic ways.

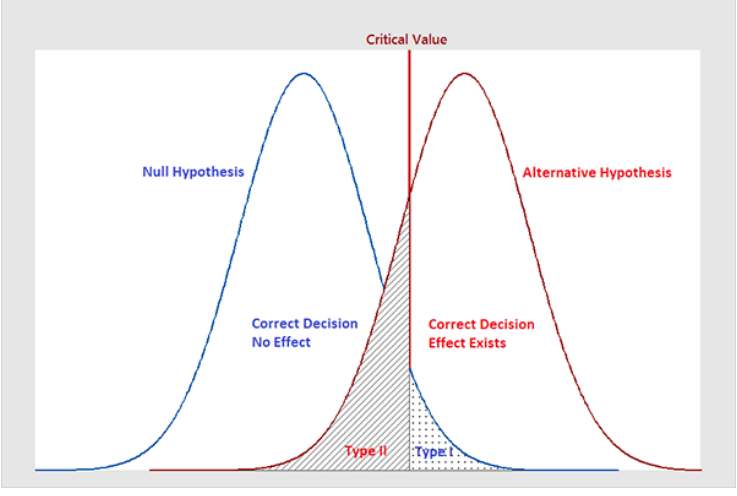
Hypotheses testing

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($\mathbb{P} = \alpha$)	OK
result	reject	False Positive	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.

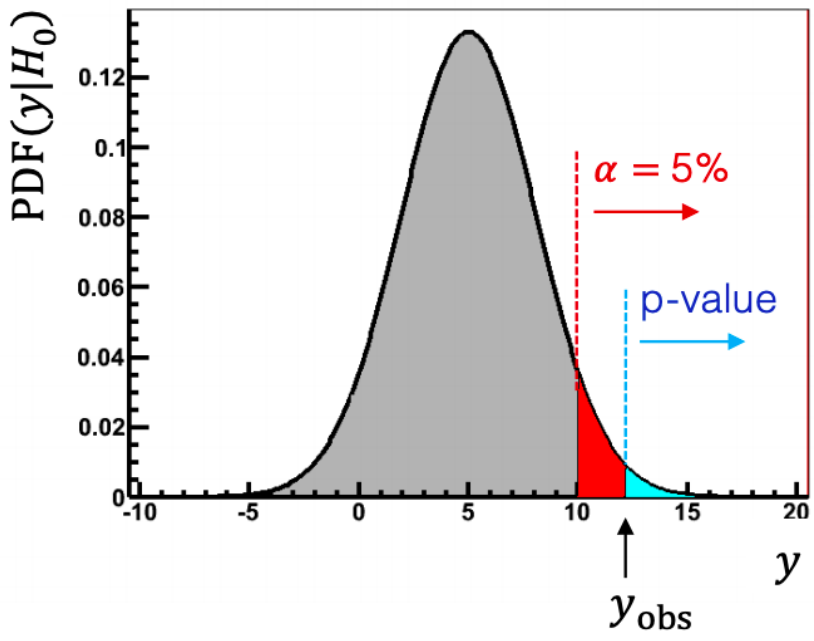
p-value

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 ith 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



$$= \int_{y_{obs}}^{\infty} \mathrm{PDF}(y|H_0)\, dy$$

NB: α must be predefined!

NB2: p-value does not say anything about significance of your answer!

NB3: p-value does not say anything about probability of your hypothesis.

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.