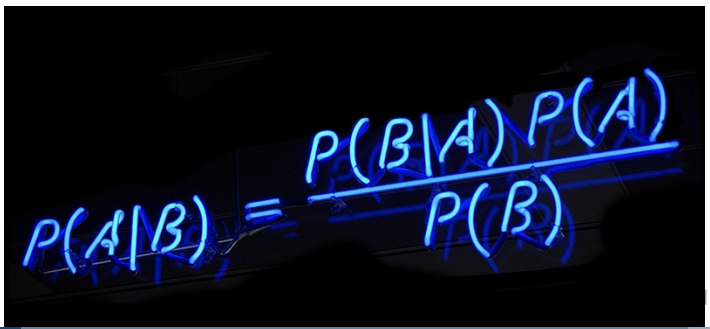
Statistics in Data Analysis

All you ever wanted to know about statistics but never dared to ask

part 4

Pawel Brückman de Renstrom
(pawel.bruckman@ifj.edu.pl)

March 26, 2025



Question from the previous lecture

Consider the exponential p.d.f.,

$$f(x;\tau) = \frac{1}{\tau} e^{-x/\tau}, \qquad x \geq 0.$$

1. Show that the corresponding cumulative distribution is given by

$$F(x;\tau) = 1 - e^{-x/\tau}$$

2. Show that the conditional probability to find a value $x < x_0 + x'$ given that $x > x_0$ is equal to the (unconditional) probability to find x less than x', i.e.

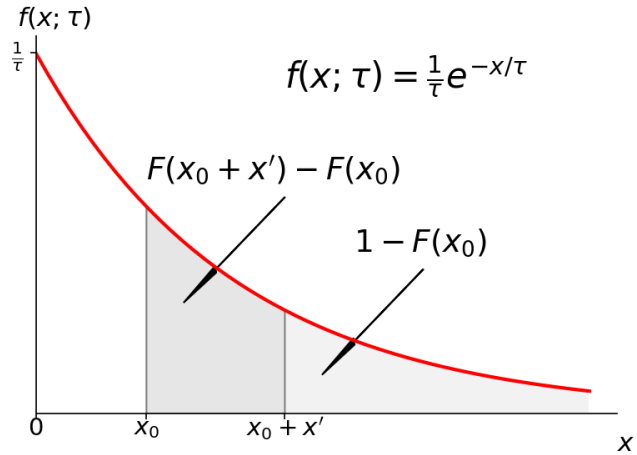
$$P(x < x_0 + x' | x \geq x_0) = P(x \leq x').$$

Solution to be sent to me before the next lecture

Solution

1 We find the cumulative function by a simple integration:

$$F(x;\tau) = \int_0^x \frac{1}{\tau} e^{-y/\tau} \mathrm{d}y = \left| \frac{-\tau}{\tau} e^{-y/\tau} \right|_0^x = 1 - e^{-x/\tau}$$



2 For the second part, we use the definition of the *conditional probability*:

$$P(x < x_0 + x' | x \geq x_0) = \frac{P(x_0 \leq x < x_0 + x')}{P(x \geq x_0)}$$

$$= \frac{F(x_0 - x') - F(x_0)}{1 - F(x_0)} = \frac{1 - e^{-(x_0+x')/\tau} - 1 + e^{-x_0/\tau}}{1 - 1 + e^{-x_0/\tau}}$$

$$= 1 - e^{-x'/\tau} = P(x \leq x') \qquad \therefore$$

No matter where you start your observation the (properly normalised) remainder of the exponent looks the same!

Statistical tests

elementary notions

Goal: make a statement about how well the observed data stand in agreement with the assumed p.d.f., i.e. a **hypothesis**.

- Hypothesis under test is called the **null hypothesis**, H_0.
- Hypothesis which uniquely determines the $f(x)$ is called **simple**
- otherwise, when free parameters are involved, $f(x;\theta)$ is called **composite**.
- Statement on hypothesis validity often involves **alternative hypotheses**, $H_1, H_2....$
- For $\mathbf{x} = (x_1, x_2, ..., x_n)$, the joint p.d.f. for a hypothesis H is given by the conditional probability: $f(\mathbf{x}|H)$.
- We define a **test statistic** $t(\mathbf{x})$ (can be multi-dimensional) which, in turn, is characterised by its p.d.f.'s: $g(t|H)$.

Statistical tests

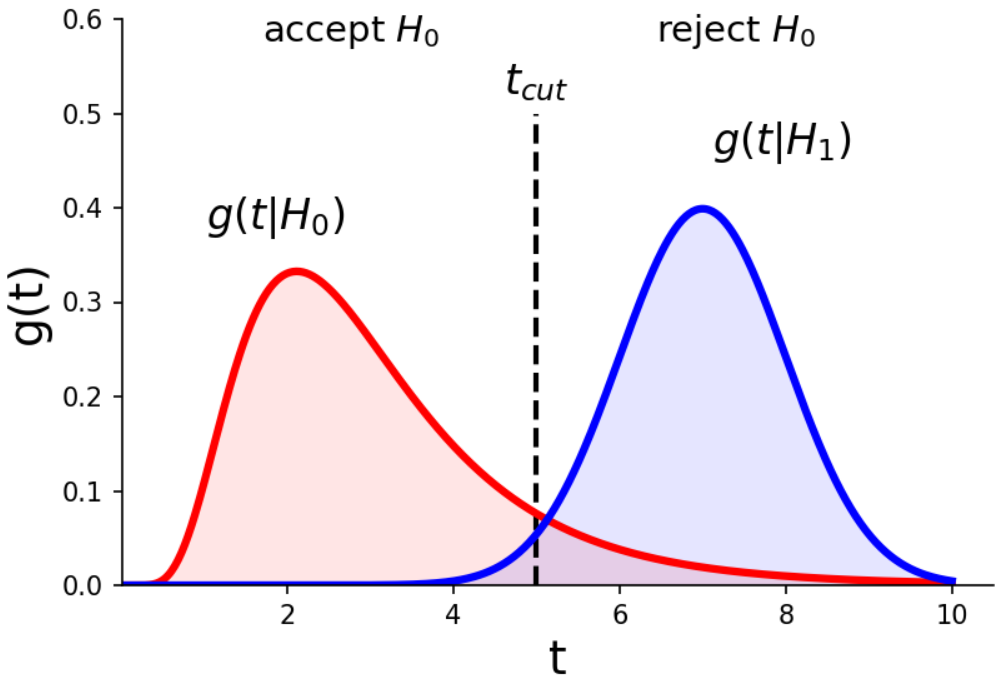
simple hypotheses

- **acceptance region**: $t < t_{\text{cut}}$,
- **critical region**: $t > t_{\text{cut}}$,
- **error of the first kind**:

$$\alpha = \int_{t_{\text{cut}}}^{\infty} g(t|H_0)dt,$$

- **error of the second kind**:

$$\beta = \int_{-\infty}^{t_{\text{cut}}} g(t|H_1)dt,$$



- α is also known as **significance level** of the test.
- $1 - \beta$ is called the **power** of the test.
- $\varepsilon = 1 - \alpha$ is the **efficiency** of the test.
- If N_0 and N_1 denote total yields of the two hypotheses (signal/background) **purity** is given as $\frac{\varepsilon N_0}{\varepsilon N_0 + \beta N_1}$.

Neyman-Pearson lemma

For a scalar t, the relationship between *efficiency* and *purity* is uniquely determined.
However, if $\mathbf{t} = (t_1, t_2, ..., t_n)$ optimal choice of the multidimensional acceptance region is not straightforward.

Likelihood ratio

The **Neyman-Pearson lemma**[a] states that the acceptance region giving the highest power (highest purity) for a given efficiency is the region of $\mathbf{t}$-space such that:

$$r = \frac{g(t|H_0)}{g(t|H_1)} > c. \tag{1}$$

The ratio r is known as the **likelihood ratio**.

[a] We shall leave it without a proof.

Linear test statistic

Fisher discriminant

$$t(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i = \mathbf{a}^T \mathbf{x} \tag{2}$$

We can express the mean and the variance for t under a hypothesis ($k = 0, 1$):

$$E_k[t] = \int t g(t|H_k) dt = \mathbf{a}^T E_k[\mathbf{x}], \tag{3}$$

$$V_k[t] = \int (t - E_k[t])^2 g(t|H_k) dt = \mathbf{a}^T V_k[\mathbf{x}]\ \mathbf{a}. \tag{4}$$

The coeficients $\mathbf{a}$ of the Fisher discriminant are obtained by maximising the expression:

$$J(\mathbf{a}) = \frac{(E_0[t] - E_1[t])^2}{V_0[t] + V_1[t]}, \tag{5}$$

which leads to:

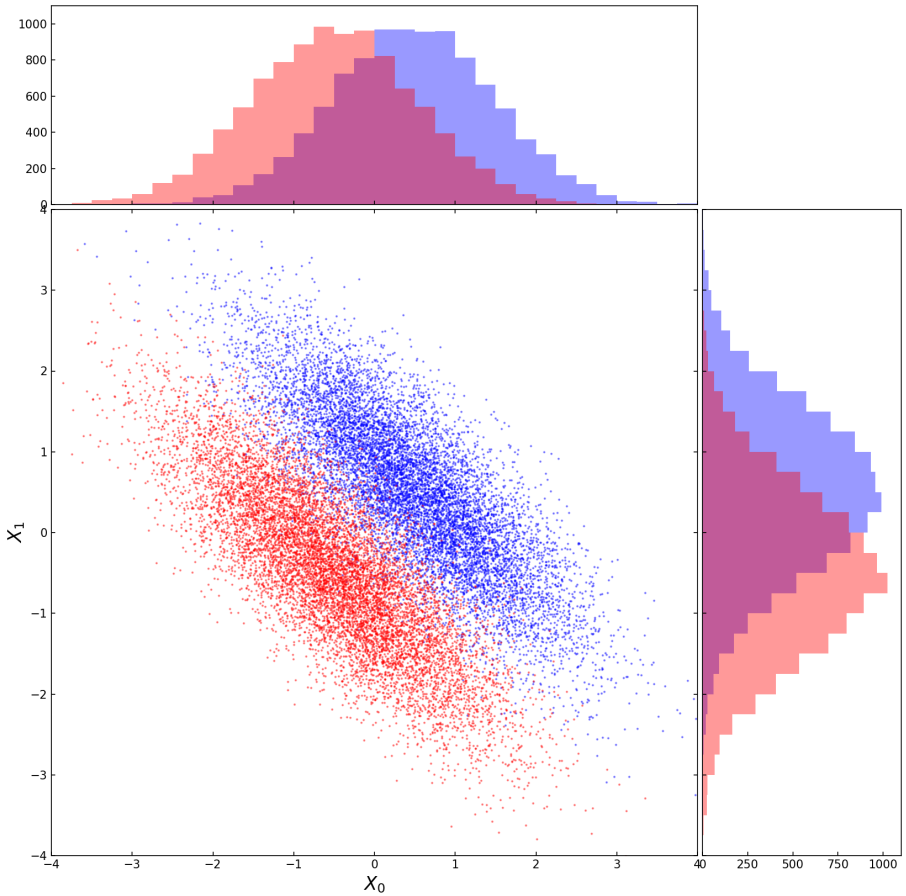
$$\mathbf{a} \propto (V_0[\mathbf{x}] + V_1[\mathbf{x}])^{-1} (E_0[\mathbf{x}] - E_1[\mathbf{x}]). \tag{6}$$

The **Fisher's linear discriminant function** is given by just the mean and covariance of the data $\mathbf{x}$, but without knowledge of the actual joint p.d.f.'s ($f(\mathbf{x}|H_0)$ and $f(\mathbf{x}|H_1)$).

Fisher discriminant

at work...

$E_0[X_0] = E_0[X_1] = 0.5,$
$E_1[X_0] = E_1[X_1] = -0.5,$
$\sigma_0 = \sigma_1 = 1,\ V_{0,1} = -0.8$



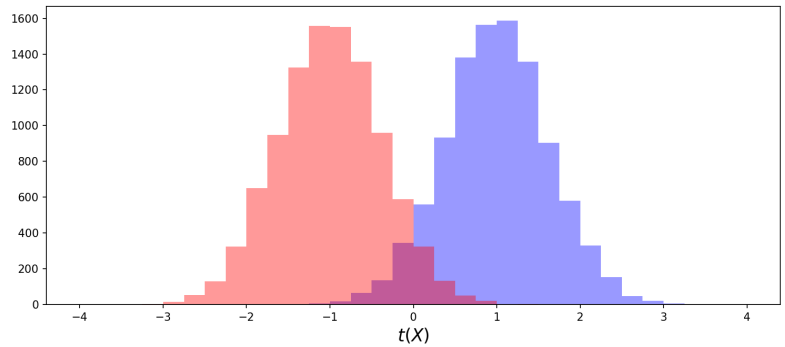
$$E_0 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \quad E_1 = \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}$$

$$V_0 = V_1 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$

$$(V_0 + V_1)^{-1} = \begin{pmatrix} \frac{2}{1.44} & \frac{1.6}{1.44} \\ \frac{1.6}{1.44} & \frac{2}{1.44} \end{pmatrix}$$

$$\mathbf{a} \propto (V_0 + V_1)^{-1}(E_0 - E_1) = \begin{pmatrix} \frac{3.6}{1.44} \\ \frac{3.6}{1.44} \end{pmatrix}$$

$$t(\mathbf{X}) = \mathbf{a}^T \mathbf{X} = X_0 + X_1$$



Fisher discriminant

properties

- For 1-D case, the Fisher discriminant is trivially given by the x itself.
- For a special case when $f(\mathbf{x}|H_0)$ and $f(\mathbf{x}|H_1)$ are both multi-dimensional Gaussians with common covariance matrix ($V_0 = V_1 = V$):

$$t(\mathbf{x}) = a_0 + (\mu_0 - \mu_1)^T V^{-1} \mathbf{x},$$

the likelihood ratio becomes a monotonic function of t:

$$r = \exp[-\tfrac{1}{2}(\mathbf{x}-\mu_0)^T V^{-1}(\mathbf{x}-\mu_0) + \tfrac{1}{2}(\mathbf{x}-\mu_1)^T V^{-1}(\mathbf{x}-\mu_1)] \propto \exp[(\mu_0-\mu_1)^T V^{-1}\mathbf{x}]$$

$$r \propto e^t \qquad (t = \ln(r) + \text{const.})$$

This means that the Fisher discriminant is just as good as the likelihood ratio.

- Posterior probability of H_0 takes a particularly simple form:

$$P(H_0|\mathbf{x}) = \frac{f(\mathbf{x}|H_0)P_0}{f(\mathbf{x}|H_0)P_0 + f(\mathbf{x}|H_1)P_1} = \frac{1}{1+\frac{P_1}{P_0 r}} = \frac{1}{1+e^{-t}} \equiv s(t) \qquad (7)$$

s is a **logistic sigmoid**.

$$a_0 = -\tfrac{1}{2}E_0[\mathbf{x}]^T V^{-1} E_0[\mathbf{x}] + \tfrac{1}{2}E_1[\mathbf{x}]^T V^{-1} E_1[\mathbf{x}] + \ln\left(\frac{P_0}{P_1}\right)$$

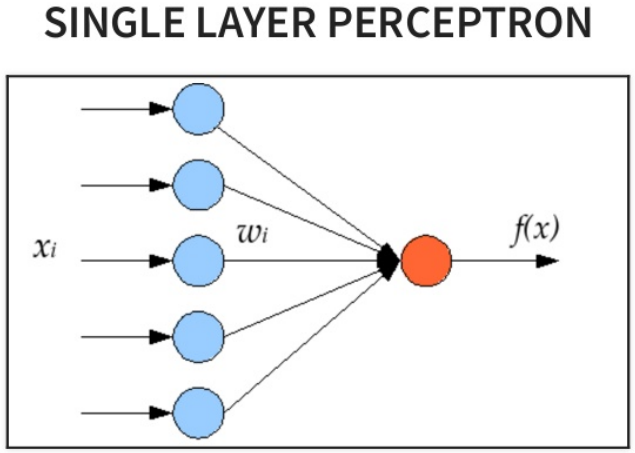
Non-linear test statistics

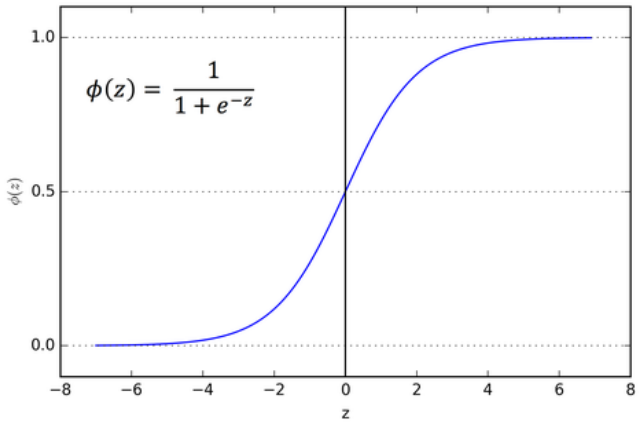
artificial neural networks

If the distributions $\neq$Gaussian or covariances are different, Fisher discriminant is not optimal any longer. Let us take:

$$t(\mathbf{x}) = s\left(a_0 + \sum_{i=1}^{n} a_i x_i\right), \qquad (8)$$

where s called the **activation function** is an arbitrary monotonic function (e.g. the logistic sigmoid). Such a test statistic is called a **single-layer perceptron**. The vector of inputs represents a set of **nodes** called the **input layer** while the test statistic $t(\mathbf{x})$ is called the **output node** (can be more than one if t is a vector).

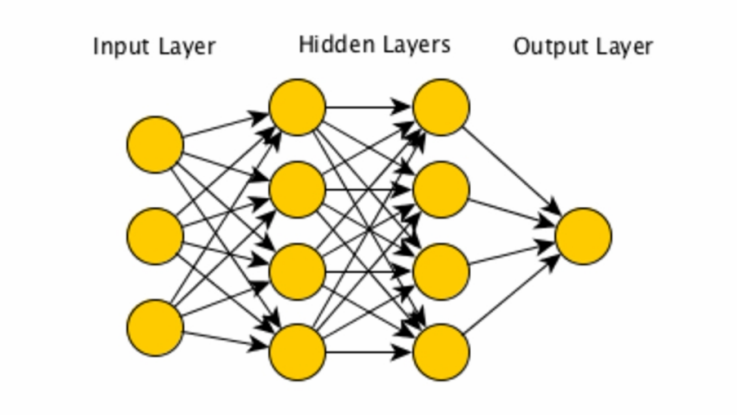




Non-linear test statistics

artificial neural networks

An arbitrary number of **hidden layers** containing any number of nodes can be combined into a **feed-forward network**:



$$h_i^{k+1}(\mathbf{x}) = s\left(w_0^k + \sum_{i=1}^{n} w_i^k h_i^k(\mathbf{x})\right). \qquad (9)$$

The network is parameterised by the **weights** w_i^k attributed to each connection.
Optimization of weights is based on minimisation of an error function, such as:

$$\Delta = E_0[(t - t^{(0)})^2] + E_1[(t - t^{(1)})^2], \qquad (10)$$

in analogy to sum of variances minimized for the Fisher discriminant. The optimization is typically achieved by means of **training**.
A popular method is e.g. the **error back-propagation**.

Learn all about neural networks and other MVA methods from the lecture series by Marcin Wolter.

Goodness-of-fit

Goodness-of-fit is used to assess the compatibility with the model without resorting to any alternative ones.

Recall tossing a coin. N trials results in n_h heads and $n_t = N - n_h$ tails. Let us take $N = 20$ and $n_h = 17$. How is this result compatible with the "fair coin" hypothesis?

$$P(n_h; N) = \frac{N!}{n_h!(N - n_h)!} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^{N-n}, \quad E[n_h] = 10$$

The P-**value** is the probability, under the assumption of H_0, of obtaining a result as compatible or less with H_0 than the actually observed.
In our example this means sum of probabilities for $n_h = 0, 1, 2, 3, 17, 18, 19, 20$ and yields $P(n_h|H_0) = 0.0026$.
This is NOT the probability of H_0 to be true! All we have assessed here is $P(n_h|H_0)$ which, in the frequentist approach, gives the fraction of times one would obtain a result as compatible with H_0 or less so if the identical experint (20 tosses of a fair coin) was repeated many times.
Probability of H_0 under the observation ($P(H_0|n_h)$) requires the prior probability for H_0 - the probability that the coin is fair before having seen the outcome of the experiment, as well as for the alternative hypotheses - recall the Bayes theorem.

Significance of the signal

In a counting experiment, we expect ν_b background events and ν_s signal ones. The total number of events observed, $n = n_b + n_s$, is therefore a Poisson variable with mean $\nu = \nu_b + \nu_s$. The probability to observe n events is:

$$f(n; \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)}. \tag{11}$$

In an experiment we observed $n_{\rm obs}$ events. We can quantify the confidence of observation of the signal ($\nu_s \neq 0$) by computing likelihood of the outcome under the background-only hypothesis:

$$P(n \geq n_{\rm obs}) = \sum_{n=n_{\rm obs}}^{\infty} f(n; \nu_s = 0, \nu_b) = 1 - \sum_{n=0}^{n_{\rm obs}-1} f(n; \nu_s = 0, \nu_b) = 1 - \sum_{n=0}^{n_{\rm obs}-1} \frac{\nu_b^n}{n!} e^{-\nu_b}. \tag{12}$$

E.g, if we expect $\nu_b = 0.5$ background events and have observed $n_{\rm obs} = 5$, the P-value is 1.7×10^{-4}.

Note: this is NOT probability of $\nu_s = 0$!

It is tempting to say "we observed $5 \pm \sqrt{5}$ which is only two standard deviations from the expected 0.5 background events". This would lead us to a wrong conclusion that we are compatible with the background-only at 5% level.

Significance of the signal

set a @95% **confidence level** (CL)

In a counting experiment, we expect ν_b background events and ν_s signal ones. The total number of events observed, $n = n_b + n_s$, is therefore a Poisson variable with mean $\nu = \nu_b + \nu_s$. The probability to observe n events is:

$$f(n; \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)}. \tag{13}$$

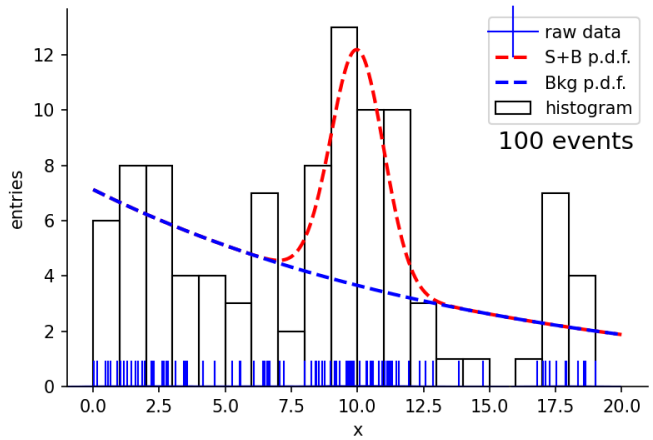
We expect $\nu_b = 0.5$ background events and have observed $n_{\text{obs}} = 5$.
What statements can be made about the signal, then?

1. We can put an *upper limit* on the yield of the signal:
   $n_{\text{signal}} < 10$ @95% CL. $\quad P(n \leq 5) = \sum_{n=0}^{5} f(n; \nu_s = 10, \nu_b = 0.5) = 0.05$
2. We can put a *lower limit* on the yield of the signal:
   $n_{\text{signal}} > 1.5$ @95% CL. $\quad P(n \geq 5) = \sum_{n=5}^{\infty} f(n; \nu_s = 1.5, \nu_b = 0.5) = 0.05$

Note: We get all different results depending on the question asked!
In statistics it is essential (and sometimes difficult) to adequately formulate the problem.

Pearson's χ^2 test

Let us consider a continuous random variable x resulting from an experiment. How can we make a statement about compatibility with background-only hypothesis? Or claim observation of signal on top of the background?



Usual procedure is to histogram observed data $\mathbf{n} = (n_1, ..., n_N)$ and construct the **Pearson's** χ^2 statistic:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}, \qquad \text{note} : \nu_i = (\sqrt{\nu_i})^2 = \sigma^2 \tag{14}$$

where ν_i is the expected number of entries under the assumed hypothesis (model). If number of events in a bin is not too small the statistic will follow a χ^2 distribution for N DoF's. This holds regardless of the distribution of x. χ^2 test is said to be **distribution free**.

Pearson's χ^2 test

The P-value is obtained from the integral of the χ^2 distribution from the observed value to infinity:

$$P = \int_{\chi^2}^{\infty} f_{\chi^2}(x, N_{\mathrm{DoF}})dx \tag{15}$$

Note: Expectation value is N_{DoF}, so χ^2/N_{DoF} should be distributed around unity with mean 1. This quantity is often quoted as a measure of goodness-of-fit. However, it has to be kept in mind that the resulting P-value depends on N_{DoF}, not only the ratio:

$$\chi^2 = 15, \quad N_{\mathrm{DoF}} = 10 \Longrightarrow P = 0.13 \quad \text{while} \quad \chi^2 = 150, \quad N_{\mathrm{DoF}} = 100 \Longrightarrow P = 9 \times 10^{-4}$$

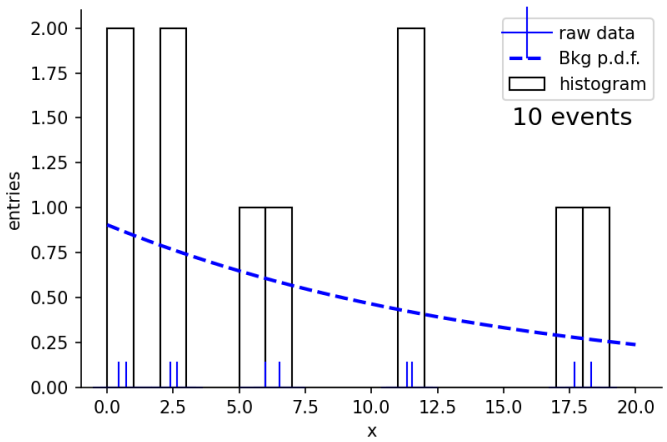
If the normalisation is known, i.e. $\sum n_i = \sum \nu_i$ in Eq. 14, the test statistic follows χ^2 distribution with $N_{\mathrm{DoF}} = N - 1$. More generally, if m parameters of the model are estimated from data, the χ^2 statistic will obey $N_{\mathrm{DoF}} = N - m$.

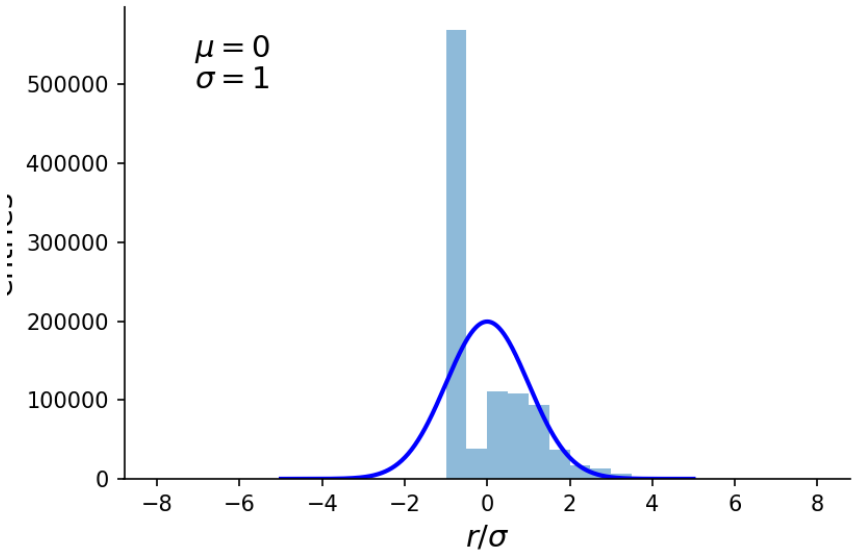
In our example: $\chi^2 = 76.6$, $N_{\mathrm{DoF}} = 19$, which gives $P-\mathrm{value} = 7 \times 10^{-9}$.

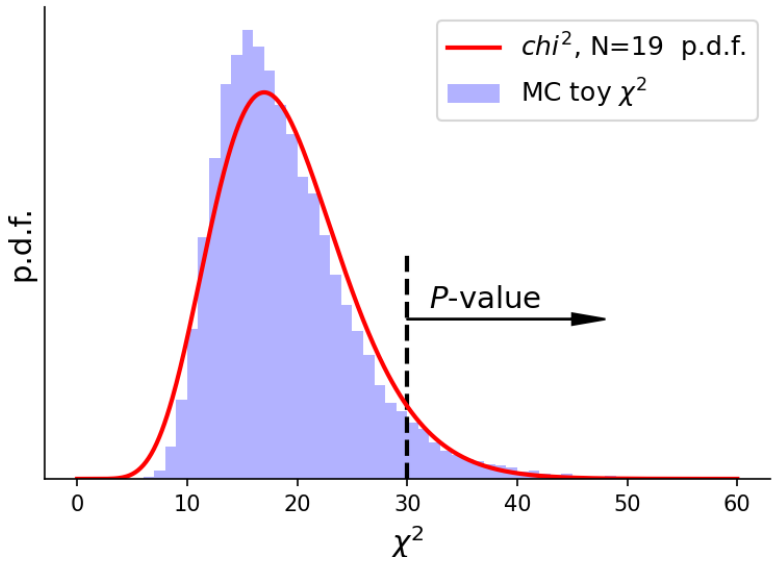
Pearson's χ^2 test

validity of χ^2 statistic (10 events)

Let us inspect the residual pull $((n_i - \nu_i)/\sqrt{\nu_i})$ dependence on the event number as well as compare the actual statistic distribution from Monte Carlo toy generation to the predicted χ^2 shape.



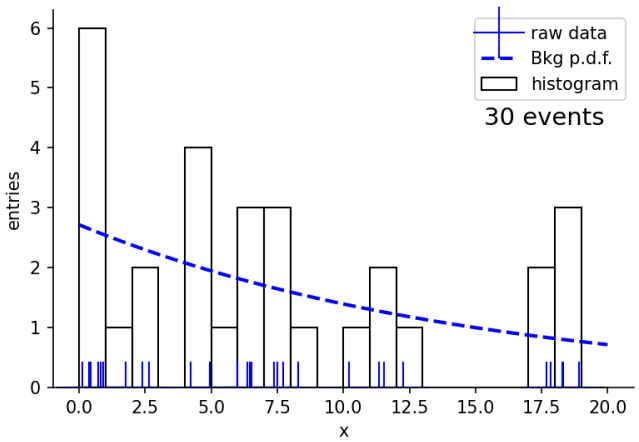


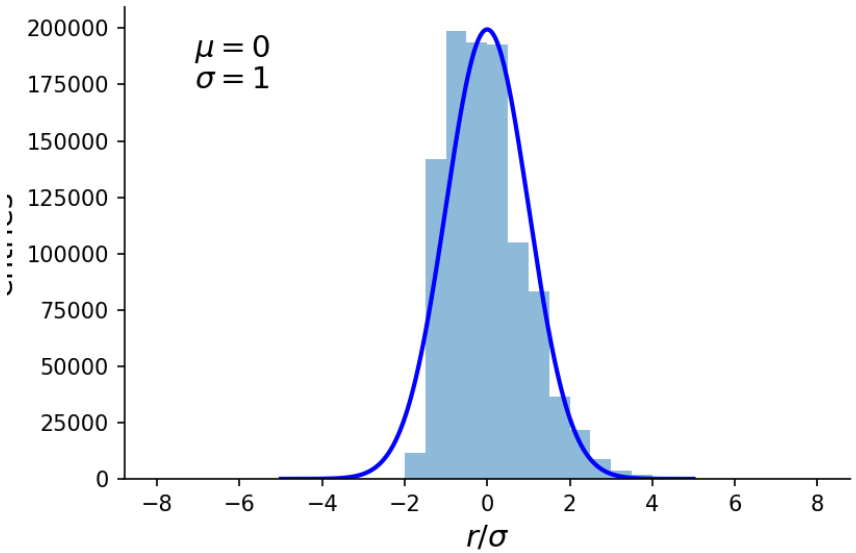


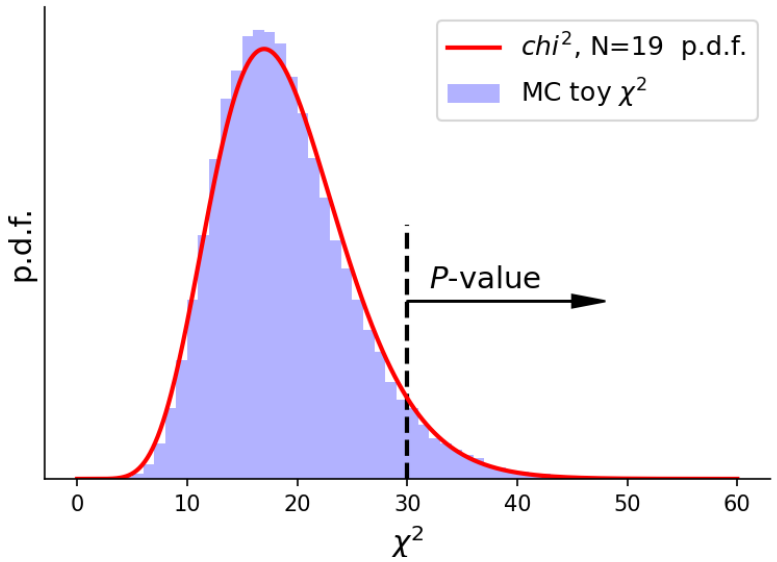
Pearson's χ^2 test

validity of χ^2 statistic (30 events)

Let us inspect the residual pull $((n_i - \nu_i)/\sqrt{\nu_i})$ dependence on the event number as well as compare the actual statistic distribution from Monte Carlo toy generation to the predicted χ^2 shape.



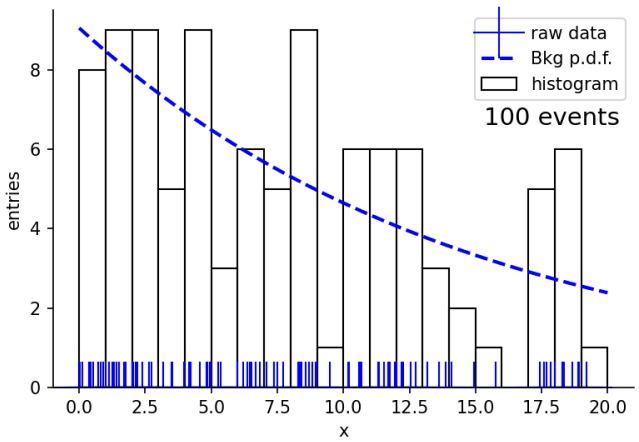


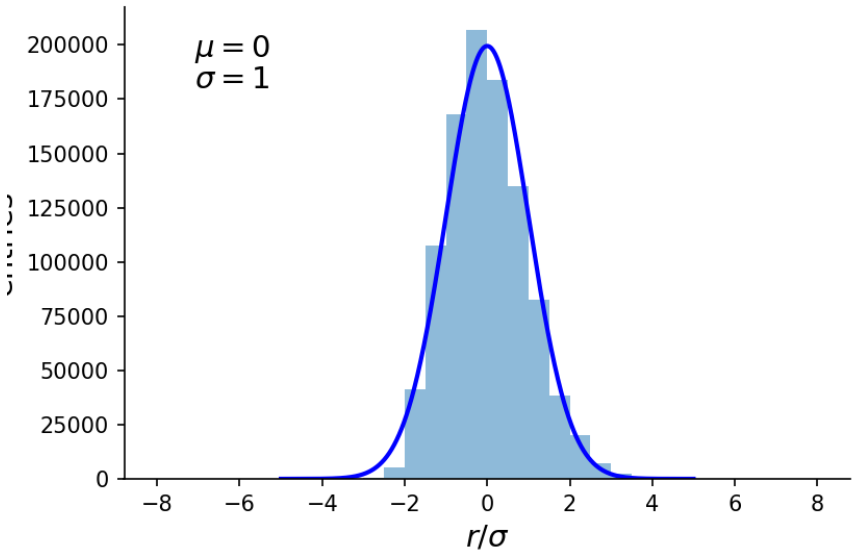


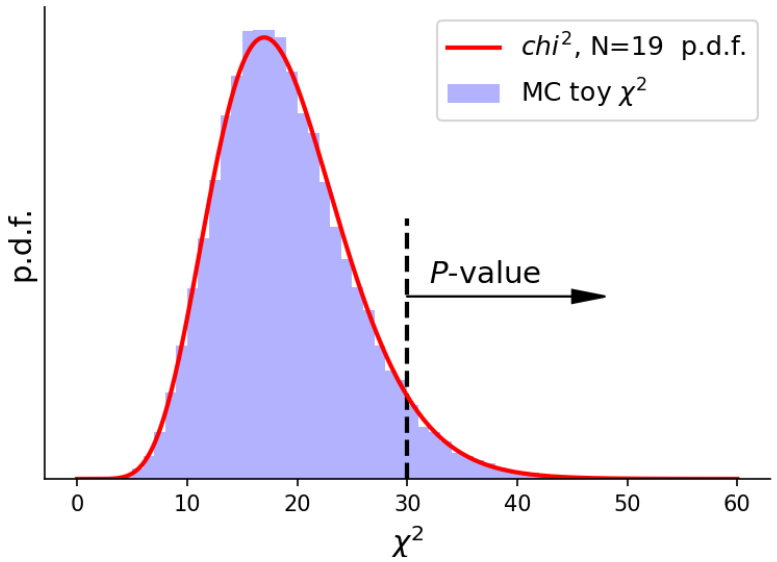
Pearson's χ^2 test

validity of χ^2 statistic (100 events)

Let us inspect the residual pull $((n_i - \nu_i)/\sqrt{\nu_i})$ dependence on the event number as well as compare the actual statistic distribution from Monte Carlo toy generation to the predicted χ^2 shape.



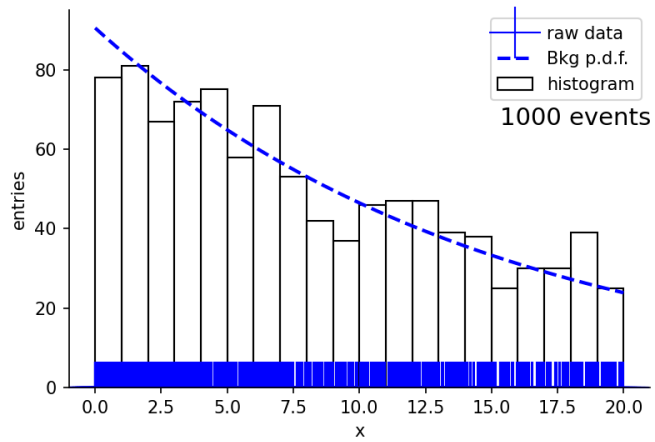


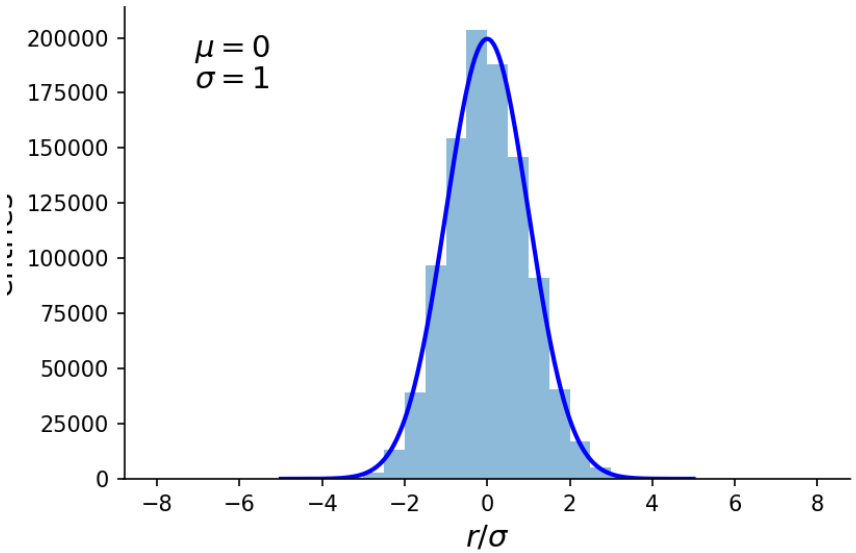


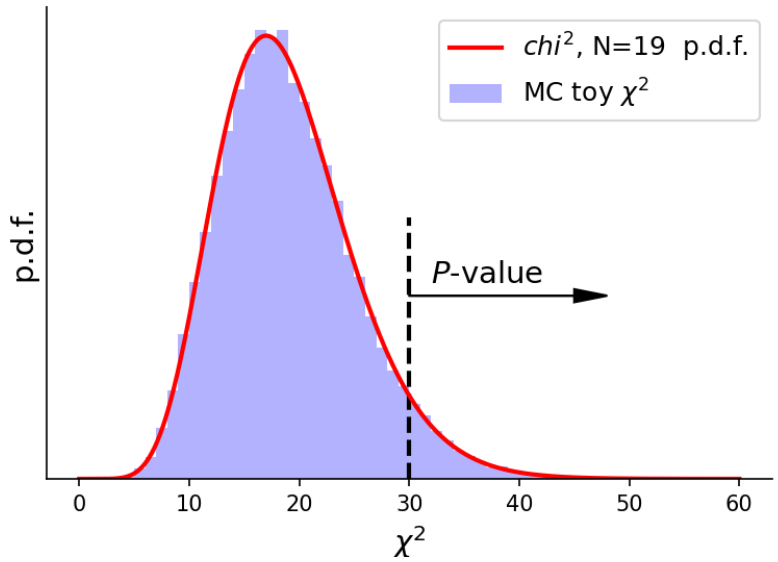
Pearson's χ^2 test

validity of χ^2 statistic (1000 events)

Let us inspect the residual pull $((n_i - \nu_i)/\sqrt{\nu_i})$ dependence on the event number as well as compare the actual statistic distribution from Monte Carlo toy generation to the predicted χ^2 shape.







Parameter estimation

General concepts

A **sample** of n independent observations from the sample space x described by the p.d.f. $f(x)$ is described by the joint p.d.f. given by:

$$f_{\text{sample}} = f(x_1)f(x_2)...f(x_n) \tag{16}$$

Let us consider n measurements of x whose p.d.f. is not known.
The central problem of statistics is to infer properties of $f(x)$ based on finite number of observations $x_1, ...x_n$.

- Often, the hypotesis for p.d.f. $f(x; \boldsymbol{\theta})$ depends on unknown parameters $\boldsymbol{\theta} = (\theta_1, ...\theta_m)$.
- A function of $\mathbf{x} = (x_1, ...x_n)$ free of unkown parameters is called a **statistic**.
- Statistic used to estimate a property of p.d.f. is called an **estimator**. An estimator for θ will be denoted $\widehat{\theta}$.

Estimator

basic properties

- If for any $\epsilon > 0$: $\lim_{n\to\infty} P(|\widehat{\theta} - \theta| > \epsilon) = 0$ the setimator is **consistent**.
- The procedure of estimating a parameter's value given the data $\mathbf{x}$ is called **parameter fitting**.
- The estimator $\widehat{\theta}(\mathbf{x})$ is itself a random variable with p.d.f. $g(\widehat{\theta};\theta)$
- The p.d.f. of a statistic is called a **sampling distribution**.
- Expectation value of an estimator $\widehat{\theta}$:

$$E[\widehat{\theta}(\mathbf{x})] = \int \widehat{\theta} g(\widehat{\theta};\theta) d\widehat{\theta} = \int ... \int \widehat{\theta}(\mathbf{x}) f(x_1;\theta)...f(x_n;\theta) dx_1...dx_n \quad (17)$$

- **Bias** of an estimator: $b = E[\widehat{\theta}(\mathbf{x})] - \theta$.
  Note: Bias does not depend on a specific sample but rather on $f(x)$, the estimator and sample size.

Estimator

basic properties

- An estimator for which $b = 0$ independently on the sample size is said **unbiased**.
- Note: A consistent estimator can be biased.
- For obvious reasons unbiased estimators are preferred.
- A useful measure of the quality of an estimator is **mean squared error** (MSE):

$$MSE = E[(\widehat{\theta} - \theta)^2] = E[(\widehat{\theta} - E[\widehat{\theta}])^2] + (E[\widehat{\theta} - \theta])^2 = V[\widehat{\theta}] + b^2, \quad (18)$$

Show the above. Recall, θ is the true value (const.!)
which can be interpreted as the squared sum of statistical and systematic errors (*adding statistical and systematic uncertainties in quadrature*).

Estimators

mean, variance and covariance

- An estimator for the expectation value:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{19}$$

- An estimator for the variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}(\bar{x^2} - \bar{x}^2) \tag{20}$$

Note: If the true mean μ is known then:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \bar{x^2} - \mu^2 \tag{21}$$

- An estimator for the covariance:

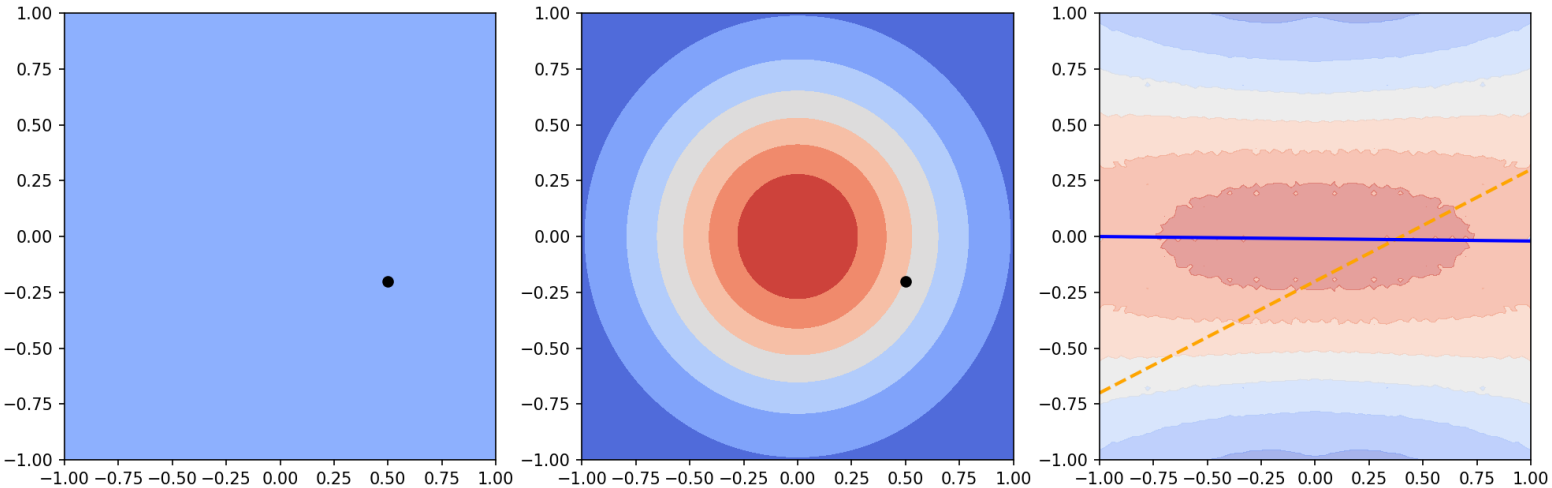
$$\widehat{V}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{n}{n-1}(\bar{xy} - \bar{x}\bar{y}) \tag{22}$$

When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.1$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(-1, 1)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 0 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)



Bayessian fit

How to interpret what's on the plots...

1. The rightmost plots show the considered data points plus the probability contour to find the fitted line in the (x, y) plane.
2. The leftmost plots show the 2D p.d.f. contours in the (p_1, p_0) plane corresponding to a single measurement. As $y = p_0 + p_1 x$ we get: $p_0 = y - x p_1$. So ignoring the error, a single (x,y) point corresponds to a straight line in the (p_1, p_0) plane. When the Gaussian uncertainty on the y measurement is taken into account, this translates into a band with a Gaussian cross-section.
3. The middle plots show our current best knowledge about p_0 and p_1 as the (p_1, p_0) contour after all previous measurements have been considered. First slide shows just our assumed PRIOR (0.0 ± 0.5 for both parameters). The following show how our knowledge about the two parameters builds up POSTERIOR.

$$P\left((p_0, p_1)|\text{meas}_k\right) \propto P\left(\text{meas}_k|(p_0, p_1)\right) P\left((p_0, p_1)\right)$$

In order to add subsequent measurement, one simply multiplies the p.d.f. coming from the measurement (left surface) by the current p.d.f. for the fitted line parameters (middle surface). This is done point-by-point over the 2D surface. Finally, one needs to take care of proper POSTERIOR normalisation.

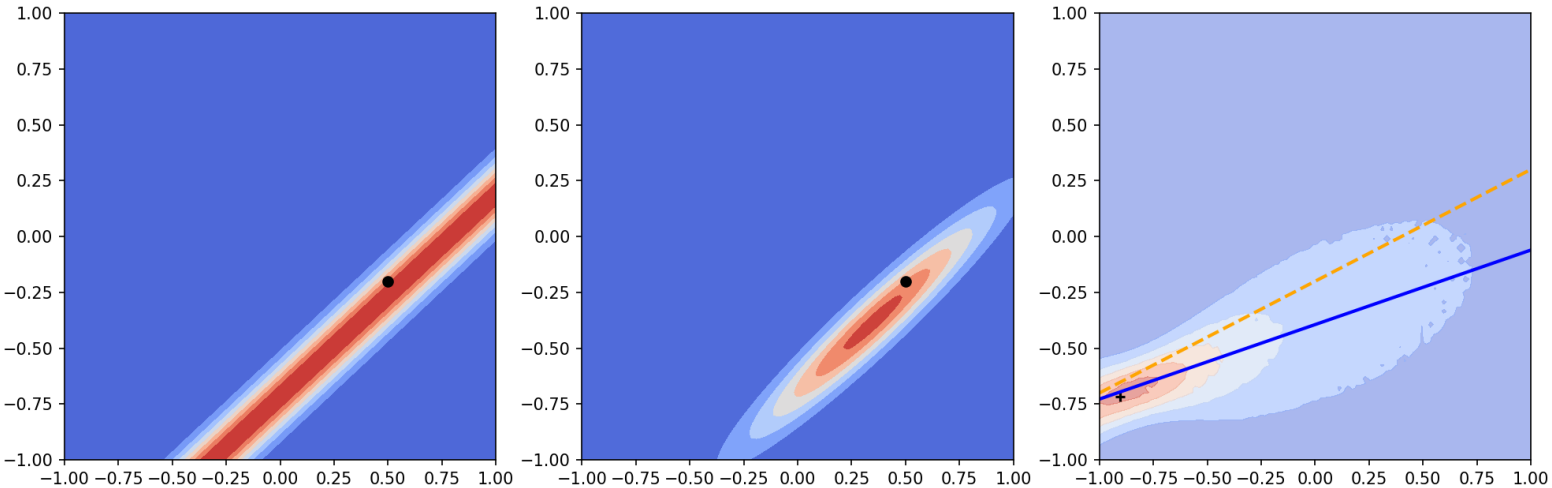
NOTE: The black point in left and middle plots indicates the value of the true p_0 and p_1 parameters $(-0.2, 0.5)$. It is added there to guide your eye and show that the fit converges to the actual parameters.

When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.1$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(-1, 1)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 1 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)

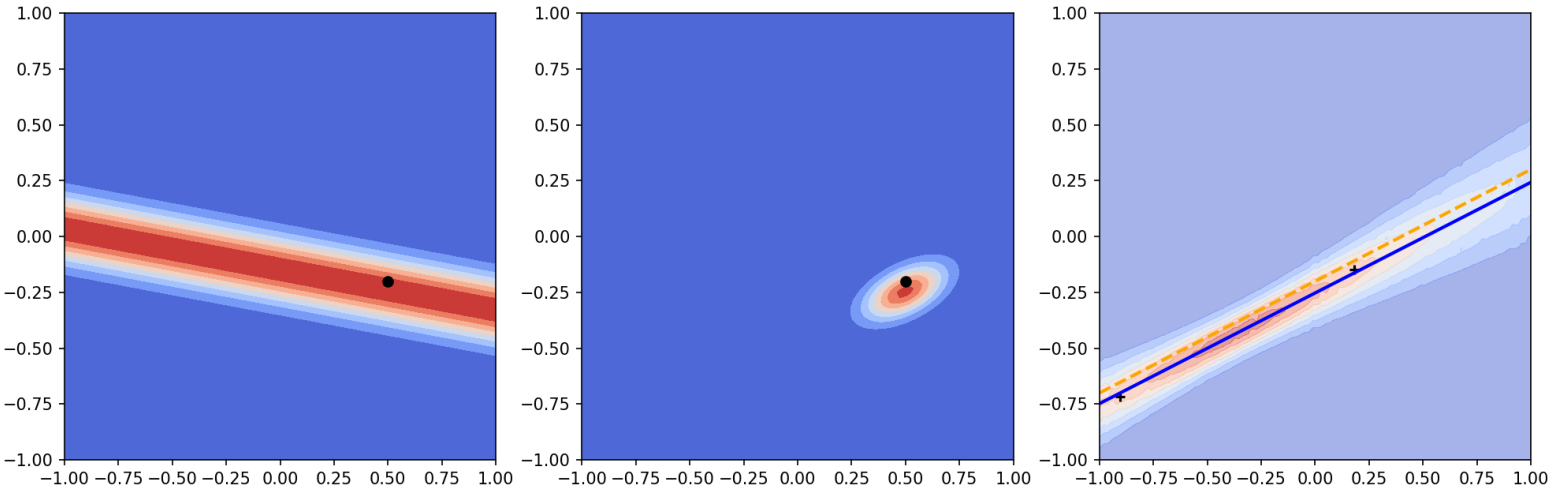


When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.1$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(-1, 1)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 2 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)

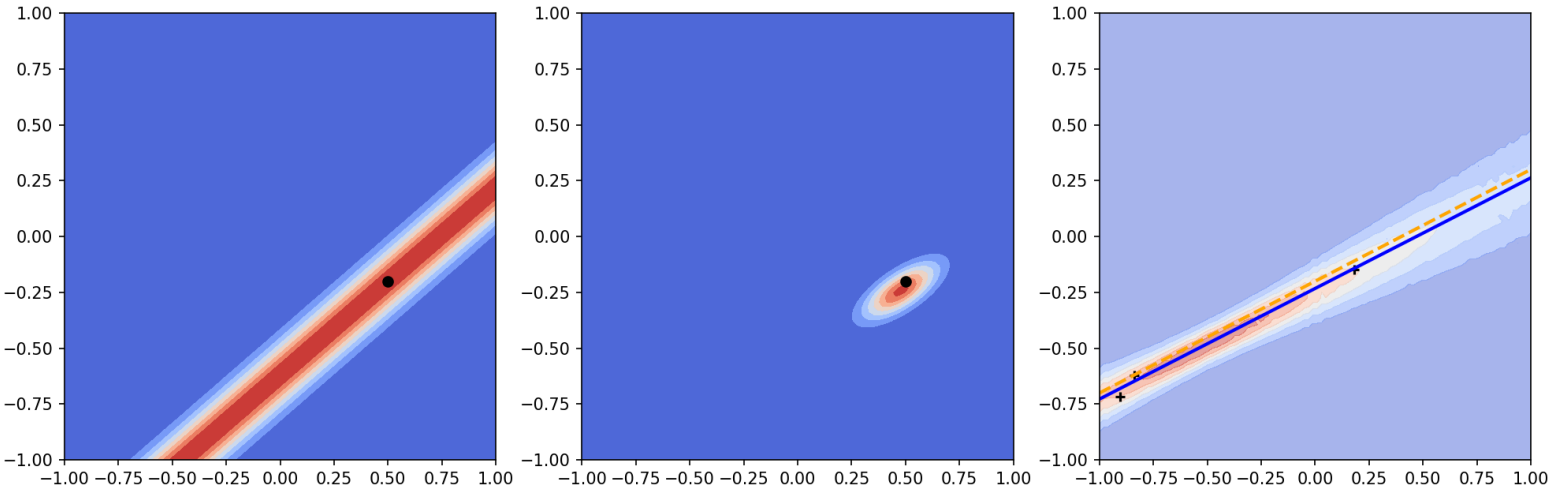


When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.1$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(-1, 1)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 3 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)

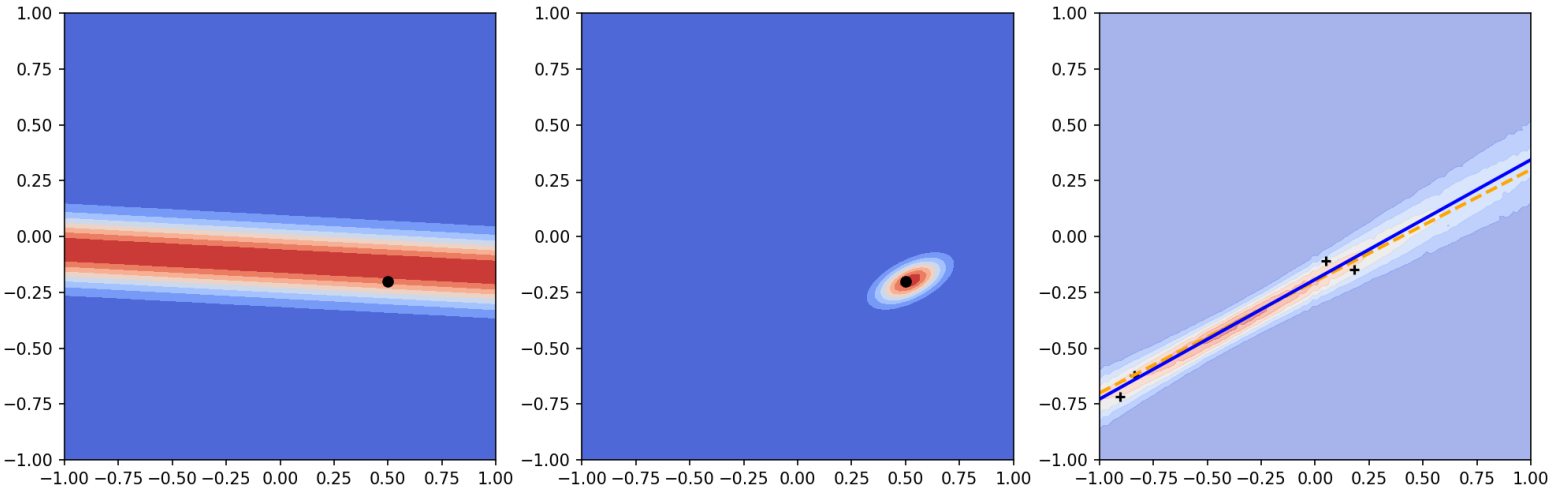


When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.1$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(-1, 1)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 4 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)

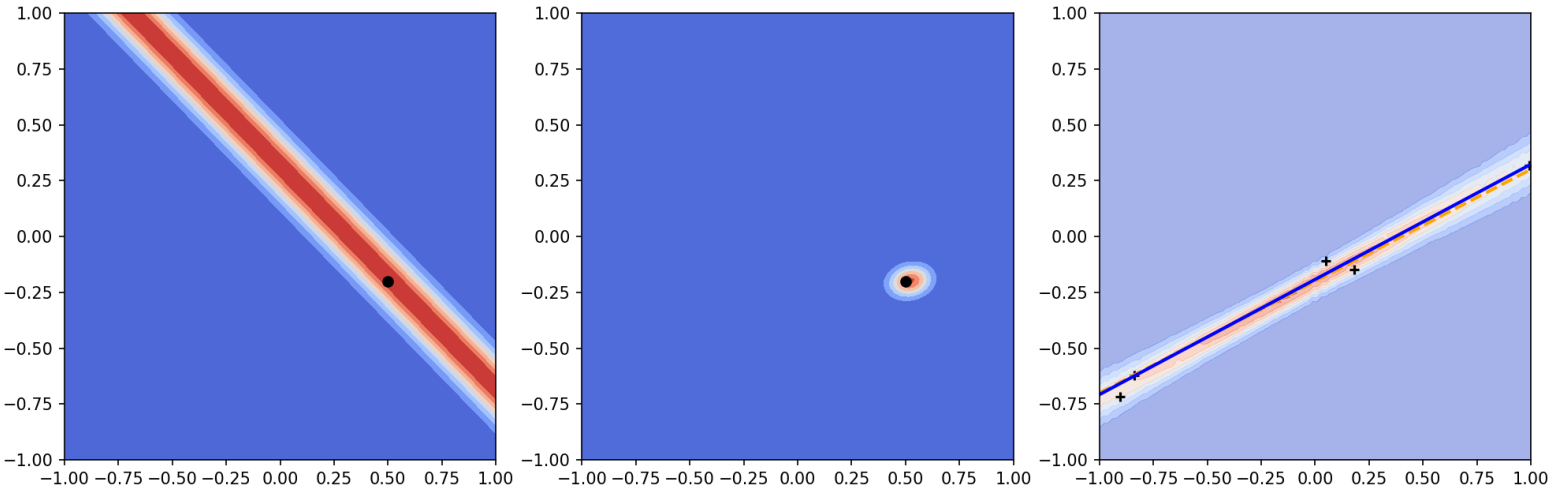


When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.1$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(-1, 1)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 5 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)

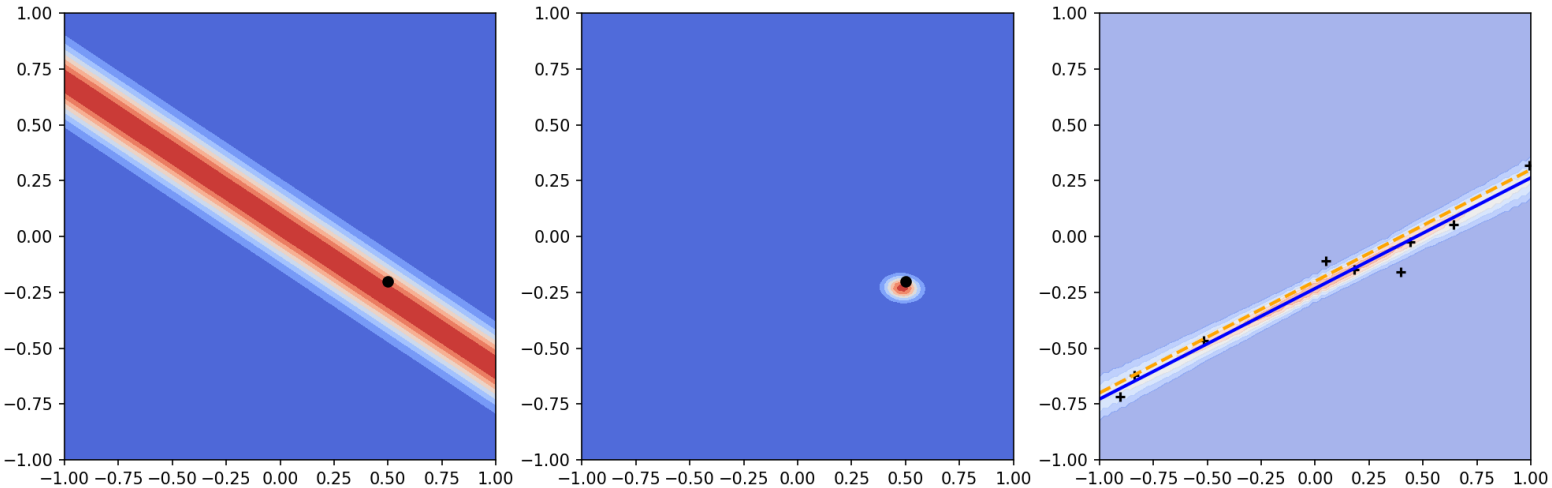


When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.1$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(-1, 1)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 9 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)

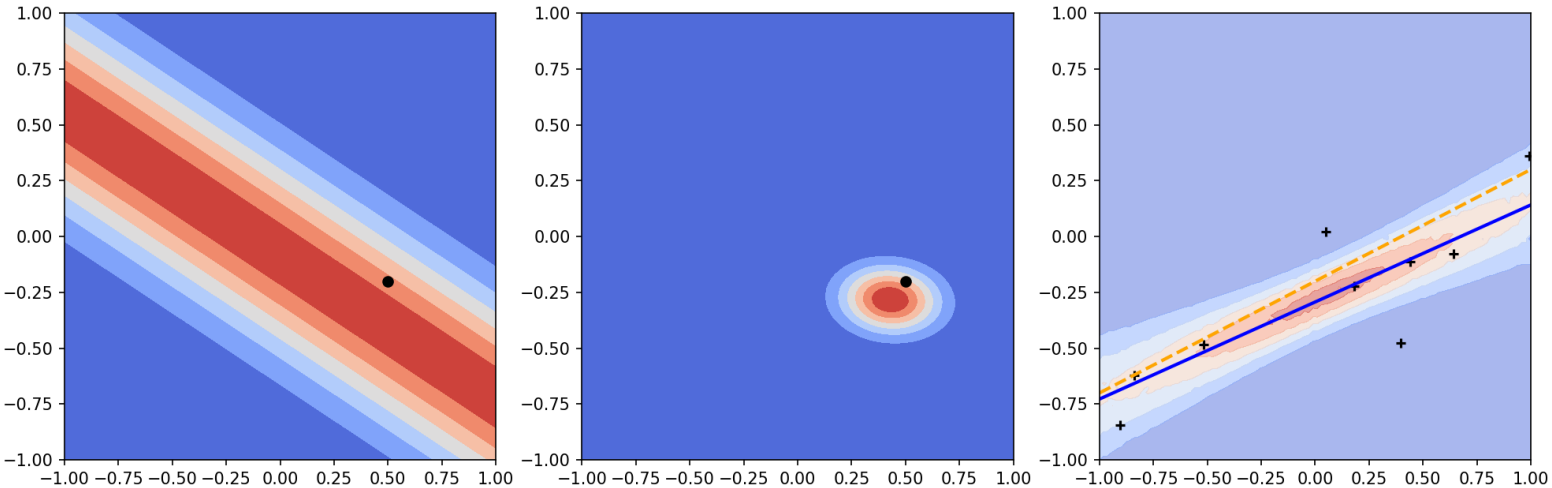


When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.3$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(-1, 1)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 9 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)

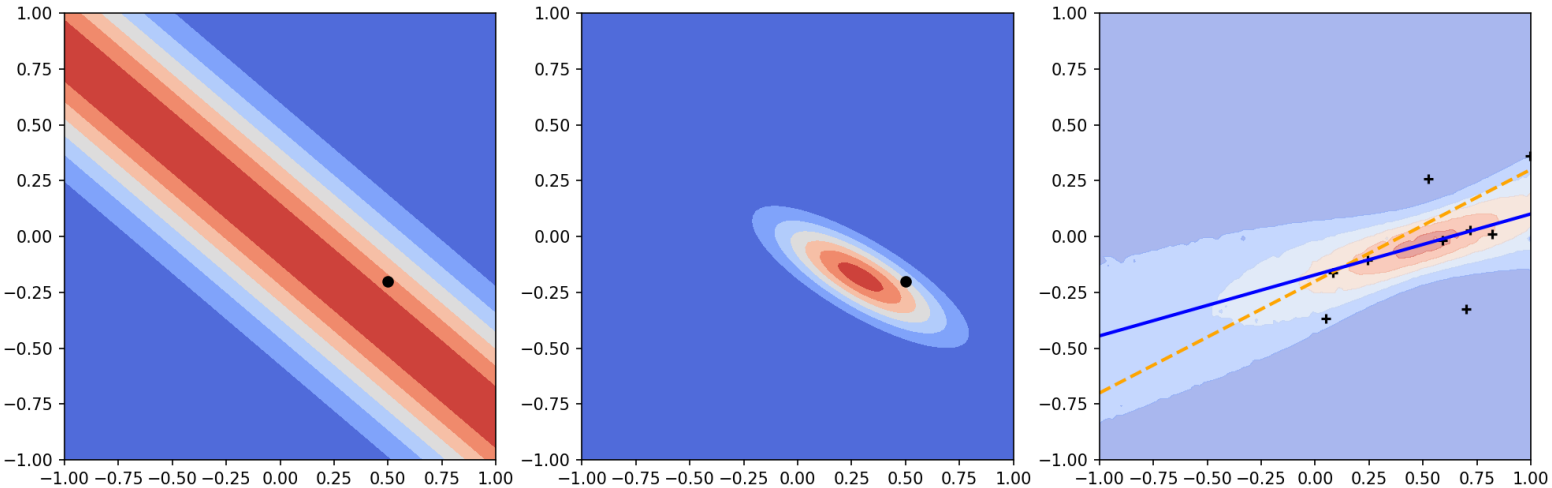


When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.3$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(0, 1)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 9 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)

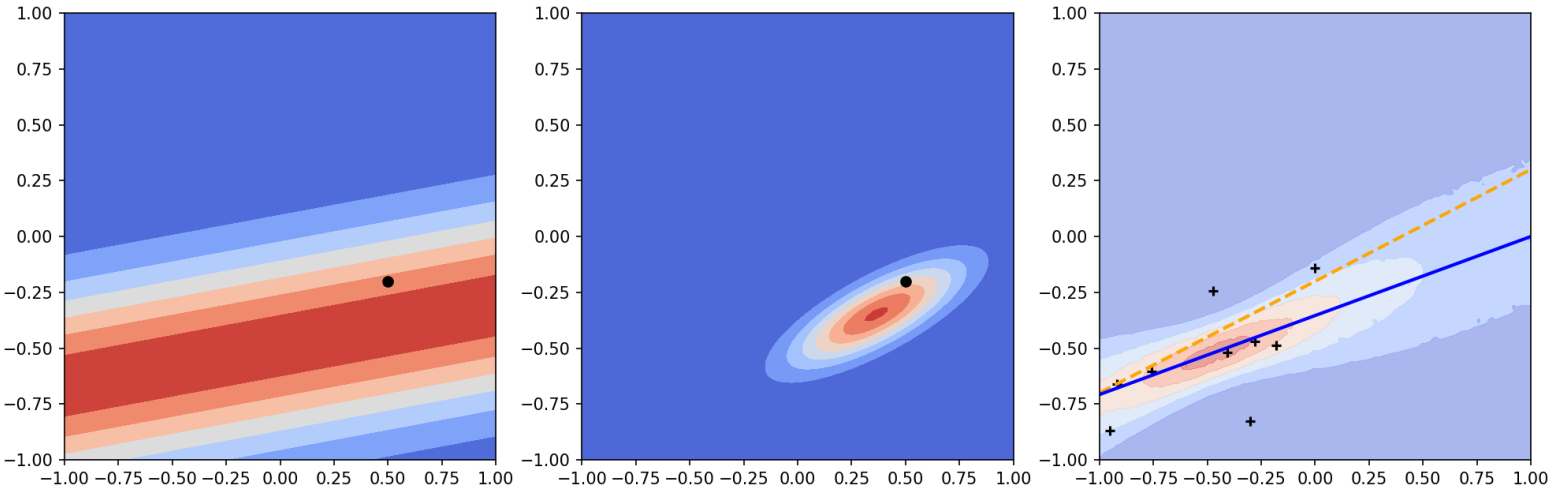


When the p.d.f. is known...

Bayesian iterations...

- We want to estimate parameters $p0, p1$ of a straight line, $y = p0 + p1x$, from the measurements (x_i, y_i) where x_i is assumed acurately known and y_i bares a Gaussian uncertainty with a known standard deviation σ.
- The true values are $p0 = -0.2, \quad p1 = 0.5, \quad \sigma = 0.3$
- A measurement provides a 2D p.d.f.: $f(p0, p1) = G(\mu = p0 + p1x_i - y_i, \sigma)$
- Measurements are randomly chosen from the range $(-1, 0)$.
- We start with a vague Gaussian prior: $p0_{\text{pri}} = 0.0 \pm 0.5, \quad p1_{\text{pri}} = 0.0 \pm 0.5.$
- Posterior is a product of the prior and measuremnt 2D p.d.f.'s.

MEASUREMENT n^o 9 $f(p0, p1)$ **PRIOR/POSTERIOR** $f(p0, p1)$ **FIT RESULT** (x, y)



Maximum Likelihood estimator

Let $f(x;\theta)$ is a p.d.f. of a known form but unknown parameter θ (more generally $\boldsymbol{\theta} = (\theta_1, ...\theta_m)$). Let $x_1, x_2, ..., x_n$ be a sample of n events drawn from the above p.d.f. Generally, $\boldsymbol{x_i}$ may be a multidimensional vector. We define:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{23}$$

called the **likelihood function**.

- L is technically a joint p.d.f. of x but, assuming a fixed data sample, represents a function of θ.
- The **maximum likelihood** (ML) estimator $\widehat{\boldsymbol{\theta}}$ is given by:

$$\frac{\partial L}{\partial \theta_i} = 0, \qquad i = 1, ..., m. \tag{24}$$

- **Log-likelihood function** is commonly used:

$$\log L(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta) \tag{25}$$

Maximum Likelihood estimator

lifetime example

An experiment measures n decays of the same particle (or follows the lifespan of n limousines) which are drawn from the exponential p.d.f.:

$$f(x;\tau) = \frac{1}{\tau} e^{-x/\tau} \tag{26}$$

- The ML estimator $\widehat{\tau}$ is given by:

$$\log L(\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{x}{\tau} \right) \qquad \frac{\partial \log L}{\partial \tau} = 0 \Longrightarrow \widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{27}$$

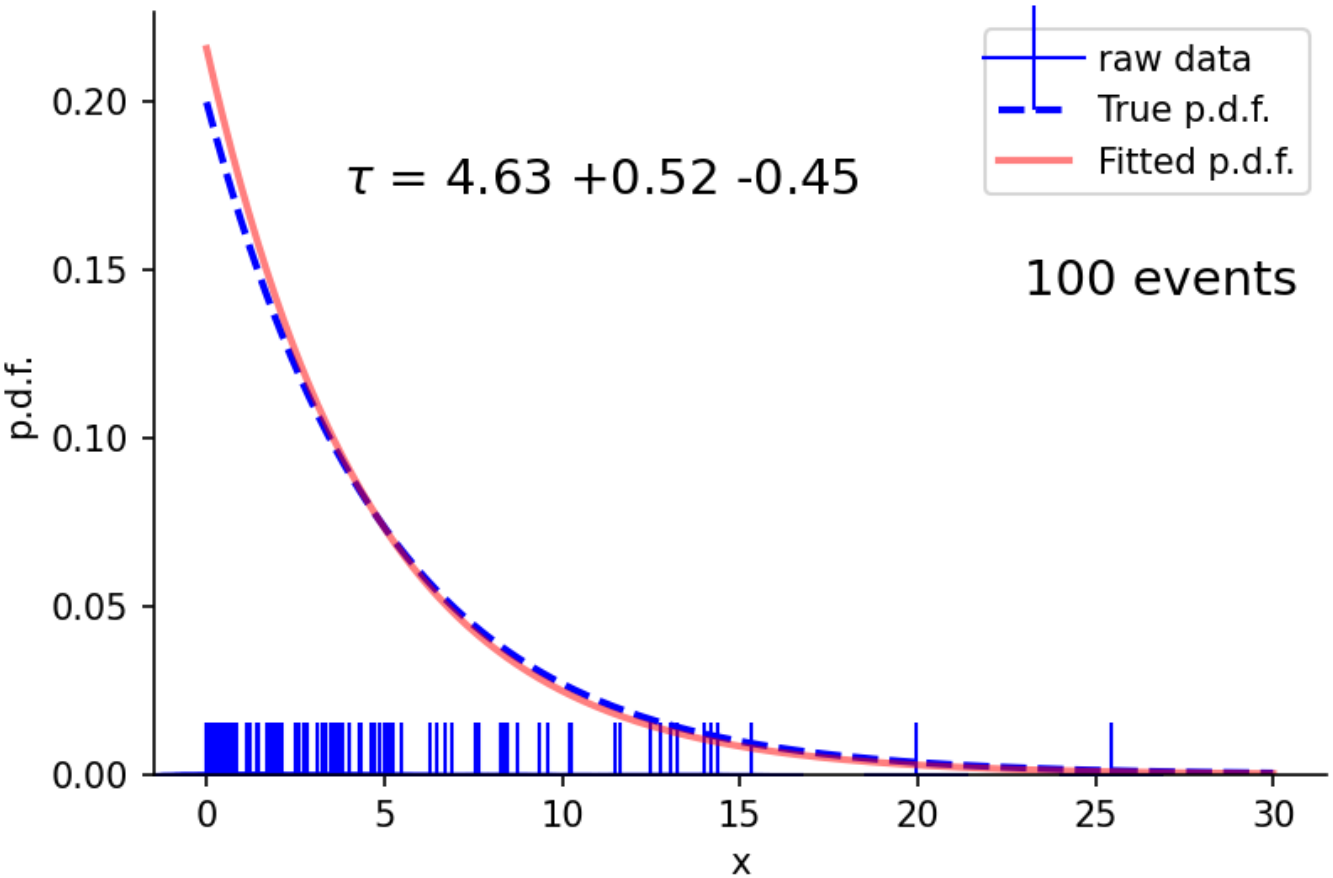
- and its expectation value is:

$$E[\widehat{\tau}] = \int \ldots \int \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) \frac{1}{\tau} e^{-x_1/\tau} \ldots e^{-x_n/\tau} dx_1 \ldots dx_n = \tau \tag{28}$$

- $\widehat{\tau}$ is an *unbiased* estimator!

Example of ML lifetime estimator

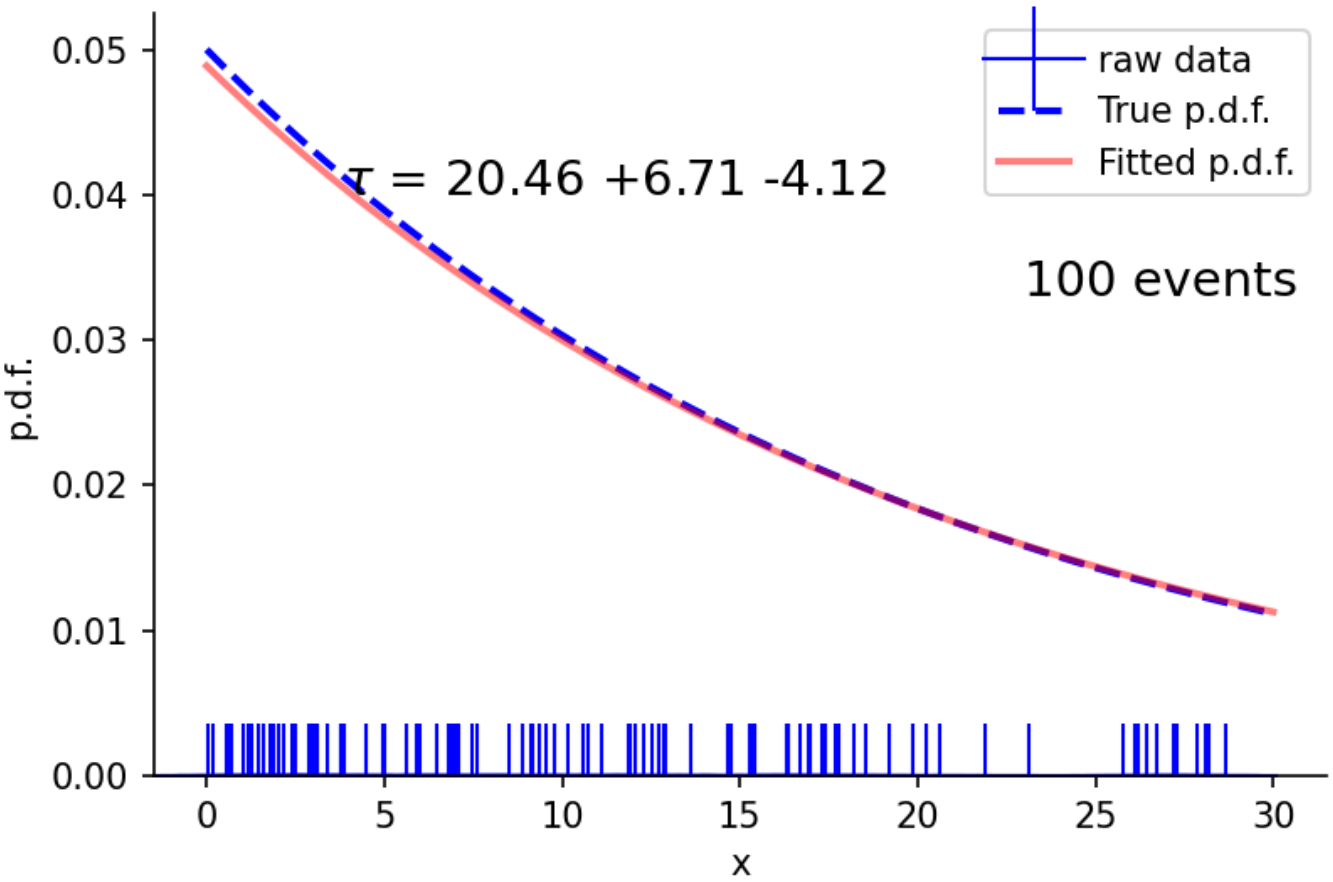
Example of 100 Monte Carlo generated observations of an exponential random variable x with mean $\tau = 5$.



The fitted ML $\widehat{\tau} = 4.63$ while the $\bar{\tau} = 4.58$.

Example of ML lifetime estimator

Example of 100 Monte Carlo generated observations of an exponential random variable x with mean $\tau = 20$.



The fitted ML $\widehat{\tau} = 20.46$ while the $\bar{\tau} = 11.46$. ! why?

Maximum Likelihood estimator

Gaussian distribution

An experiment performs n measurements of the Gaussian-distributed random variable x with unknown μ and σ^2. The log-likelihood function is:

$$\log L(\mu, \sigma^2) = \sum_{i=1}^{n} \left(\ln \frac{1}{2\pi} + \frac{1}{2} \ln \frac{1}{\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \tag{29}$$

$$\frac{\partial \log L}{\partial \mu} = 0 \Longrightarrow \widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \frac{\partial \log L}{\partial \sigma^2} = 0 \Longrightarrow \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{\mu})^2. \tag{30}$$

- $E[\widehat{\mu}] = \mu$, so $\widehat{\mu}$ is an *unbiased* estimator.
- $E[\widehat{\sigma^2}] = \frac{n-1}{n}\sigma^2$, so $\widehat{\sigma^2}$ is *biased*. Nonetheless, it is still a *consistent* estimator.

(Recall: s^2 is an unbiased estimator of σ^2.)

Variance of ML estimator

Central question of the parameter estimation: What is the uncertainty (variance) of our estimate? Let's take the exponential decay example for which ML estimator is $\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} x_i$. For independent x_i, x_j $(V_{ij} = 0)$ we have:

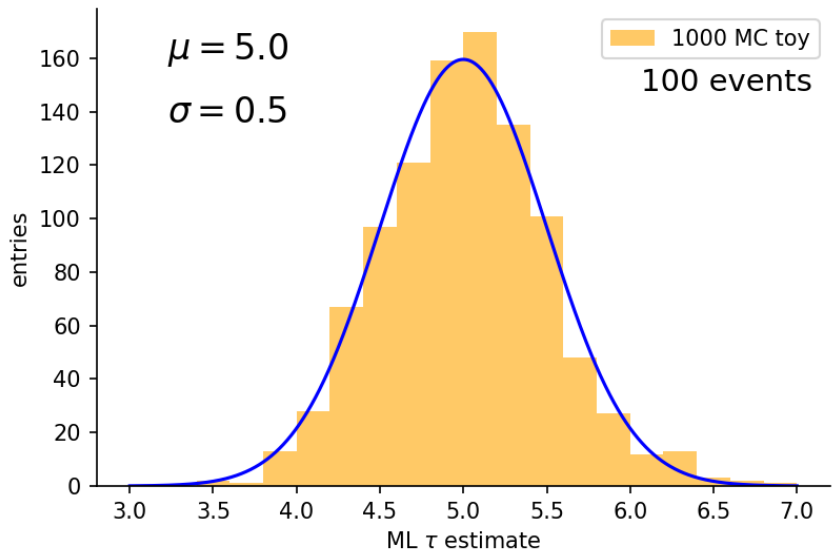
$$V[\widehat{\tau}] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \left(E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\right)^2 = \frac{\tau^2}{n} \tag{31}$$

- In practice, reported is rather $\frac{\widehat{\tau}^2}{n}$.
- Indeed, let us check it using toy Monte Carlo approach using our exponential decay example.

$$\tau = \sigma = 5$$

We take 1000 toy each performing a ML fit with 100 events.

- We expect $\sigma_{\widehat{\tau}} = \sigma/\sqrt{100} = 0.5$.



Variance of ML estimator

Central question of the parameter estimation: What is the uncertainty (variance) of our estimate? Let's take the exponential decay example for which ML estimator is $\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

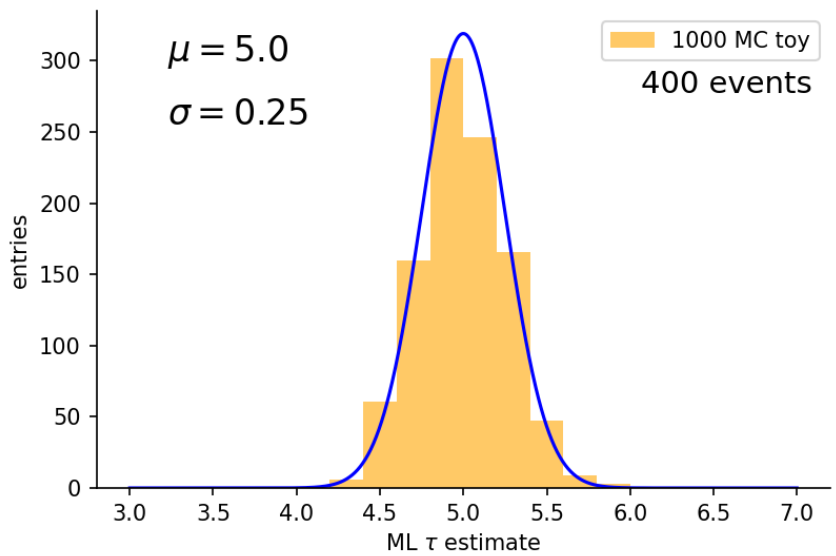
$$V[\widehat{\tau}] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \left(E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\right)^2 = \frac{\tau^2}{n} \tag{32}$$

- In practice, reported is rather $\frac{\widehat{\tau}^2}{n}$.
- Indeed, let us check it using toy Monte Carlo approach using our exponential decay example.

$$\tau = \sigma = 5$$

We take 1000 toy each performing a ML fit with 400 events.

- We expect $\sigma_{\widehat{\tau}} = \sigma/\sqrt{400} = 0.25$.



Variance of ML estimator

Central question of the parameter estimation: What is the uncertainty (variance) of our estimate? Let's take the exponential decay example for which ML estimator is $\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

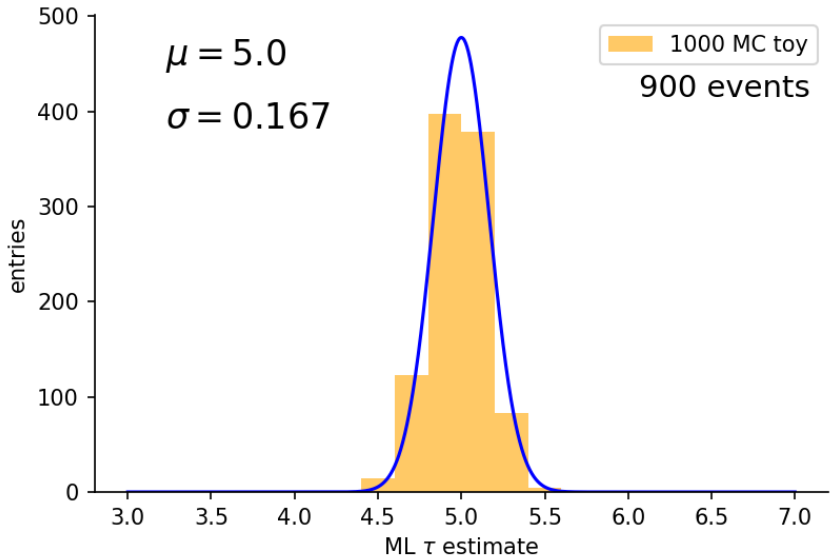
$$V[\widehat{\tau}] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \left(E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\right)^2 = \frac{\tau^2}{n} \tag{33}$$

- In practice, reported is rather $\frac{\widehat{\tau}^2}{n}$.
- Indeed, let us check it using toy Monte Carlo approach using our exponential decay example.

$$\tau = \sigma = 5$$

We take 1000 toy each performing a ML fit with 900 events.

- We expect $\sigma_{\widehat{\tau}} = \sigma/\sqrt{900} = 0.16(6)$.



Variance of the mean

For independent x_i, x_j $(V_{ij} = 0)$ we have:

$$V[\widehat{\tau}] = E[\widehat{\tau}^2] - (E[\widehat{\tau}])^2 = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \left(E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\right)^2 =$$

$$\frac{1}{n^2}\int \ldots \int \left(\sum_{i=1}^{n} x_i^2 + \sum_{i\neq j}^{n} x_i x_j\right) \frac{1}{\tau} e^{-x_1/\tau} \ldots \frac{1}{\tau} e^{-x_n/\tau} dx_1 \ldots dx_n - \tau^2 =$$

$$= \frac{1}{n^2}\left(2n\tau^2 + n(n-1)\tau^2\right) - \tau^2 = \frac{\tau^2}{n}. \tag{34}$$

More generally:

$$V[\bar{x}] = E[\bar{x}^2] - (E[\bar{x}])^2 = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \left(E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\right)^2 =$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} E[x_i^2] + \sum_{i\neq j}^{n} E[x_i x_j]\right) - E[x_i]^2 =$$

$$= \frac{1}{n^2}\left(nV[x_i] + nE[x_i]^2 + n(n-1)E[x_i]^2\right) - E[x_i]^2 = \frac{V[x_i]}{n}, \tag{35}$$

where in both cases we used: $V[x_i] = E[x_i^2] - E[x_i]^2 \Rightarrow E[x_i^2] = V[x_i] + E[x_i]^2$,
$V_{ij} = E[x_i x_j] - E[x_i]E[x_j]$ so $V_{ij} \overset{i\neq j}{=} 0 \Rightarrow E[x_i x_j] = E[x_i]E[x_j]$.

Questions

Suppose a beam of particles is known to consist of charged pions and muons. For each particle in the beam we measure a variable t, whose distribution for pions (π) and muons (μ) is

$$f(t;\pi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu_\pi)^2/2\sigma^2}, \qquad f(t;\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu_\mu)^2/2\sigma^2},$$

where $\mu_\pi = 0$, $\mu_\mu = 2$ and $\sigma = 1$. For each particle we want to test the hypothesis H_0 that it is a pion against the alternative H_1 that it is a muon. The critical region of the test is given by $t > t_c$ where t_c is a given constant.

1. Suppose we want the significance of the test to be $\alpha = 0.05$. Illustrate where the critical region lies and what α means on a sketch of the p.d.f.s $f(t|\pi)$ and $f(t|\mu)$ and show that t_c is numerically about 1.64.
2. Suppose a sample of particles is known to consist of 99% pions and 1% muons. What is the purity of the muon sample selected by $t > t_c$? Here, purity means the probability to be a muon given that the particle had $t > t_c$ (i.e., it was rejected as a pion and thus selected as a muon candidate).

Solutions to be sent to me before the next lecture

Thank you

Back-up

Expectation value for the variance estimators s^2 and S^2

$$E[s^2] = \frac{1}{n-1}\sum_i E[(x_i - \bar{x})^2] = \frac{1}{n-1}\sum_i E[x_i^2 - 2x_i\bar{x} + \bar{x}^2] =$$

$$= \frac{1}{n-1}\sum_i \left(E[x_i^2] - \frac{2}{n} E\left[x_i \sum_j x_j\right] + \frac{1}{n^2} E\left[\sum_k x_k \sum_j x_j\right]\right) =$$

$$= \frac{1}{n-1}\sum_i \left(E[x_i^2] - \frac{2}{n}\sum_j E[x_i x_j] + \frac{1}{n^2}\sum_{k,j} E[x_k x_j]\right) =$$

$$=^* \frac{1}{n-1}\sum_i \left(\mu^2 + \sigma^2 - \frac{2}{n}(\mu^2 + \sigma^2 + (n-1)\mu^2) + \frac{1}{n^2}\left[(n^2-n)\mu^2 + n(\mu^2+\sigma^2)\right]\right) =$$

$$= \frac{1}{n-1}\sum_i \left(0 \times \mu^2 + \frac{n-1}{n}\sigma^2\right) = \frac{1}{n-1} n \frac{n-1}{n}\sigma^2 = \sigma^2, \quad \square \tag{36}$$

$$E[S^2] = \frac{1}{n}\sum_i E[(x_i - \mu)^2] = \frac{1}{n}\sum_i E[x_i^2 - 2x_i\mu + \mu^2] =^* \frac{1}{n}\sum_i \left(\mu^2 + \sigma^2 - 2\mu^2 + \mu^2\right) =$$

$$= \frac{1}{n} n\sigma^2 = \sigma^2, \quad \square \tag{37}$$

* by virtue of identities used in (38).

Estimators

mean, variance and covariance

- Given the estimator $\widehat{\theta}$, one can compute its variance $V[\widehat{\theta}] = E[\widehat{\theta}^2] - (E[\widehat{\theta}])^2$, which gives a measure of the uncertainty.
- Most commonly used is the variance of the sample mean $\bar{x}$:

$$V[\bar{x}] = E[\bar{x}^2] - (E[\bar{x}])^2 = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)\right] - \mu^2 =$$

$$= \frac{1}{n^2}\sum_{i,j=1}^{n} E[x_i x_j] - \mu^2 = \frac{1}{n^2}[(n^2-n)\mu^2 + n(\mu^2+\sigma^2)] - \mu^2 = \frac{\sigma^2}{n}, \quad (38)$$

where σ^2 is the variance of $f(x)$ and we used the fact that $E[x_i x_j] = \mu^2$ for $i \neq j$ and $E[x_i^2] = \mu^2 + \sigma^2$.