Statistics for Data Analysis
Lecture 1

Flavour Physics at LHC School

CERN, Geneva

26-30 May 2025

https://indico.cern.ch/event/1508466/

Glen Cowan

Physics Department

Royal Holloway, University of London

`g.cowan@rhul.ac.uk`

`www.pp.rhul.ac.uk/~cowan`

Outline

→ Lecture 1:   Probability, Bayes vs. Frequentist

                                    Frequentist parameter estimation

                                    Hypothesis tests

Lecture 2: p-values

                                    Confidence limits

                                    Systematic uncertainties

                                    Bayesian parameter estimation

Lecture 3:   Significance, sensitivity

                                    Bayes factors

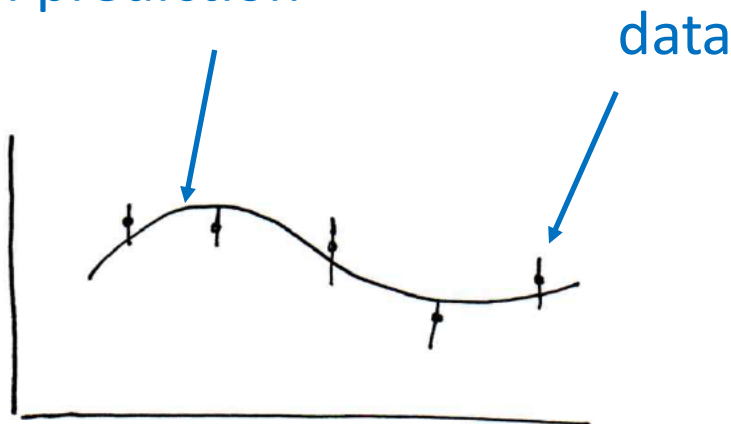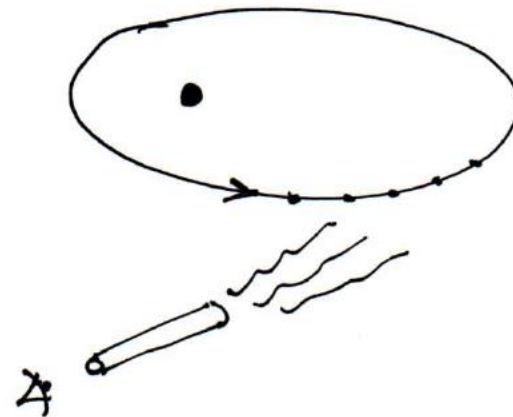   Models for anomalies

Theory ↔ Statistics ↔ Experiment

Theory (model, hypothesis):

$$F = -G \frac{m_1 m_2}{r^2} \quad , \quad \cdots$$

+ response of measurement apparatus

= model prediction

data

Experiment (observation):

Uncertainty enters on many levels

→ quantify with probability

A quick review of probability

Frequentist (A = outcome of repeatable observation)

$$P(A) = \lim_{n \to \infty} \frac{\text{outcome is in } A}{n}$$

Subjective (A = hypothesis)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

E.g. rolling a die, outcome n = 1,2,…,6:

$$P(n \leq 3 | n \text{ even}) = \frac{P((n \leq 3) \cap n \text{ even})}{P(n \text{ even})} = \frac{1/6}{3/6} = \frac{1}{3}$$

A and B are independent iff:

$$P(A \cap B) = P(A)P(B)$$

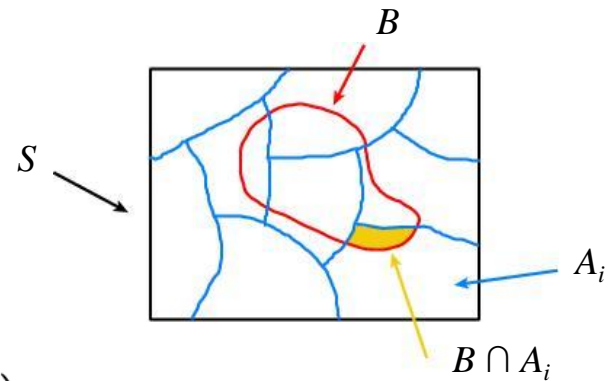I.e. if A, B independent, then

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Bayes' theorem

Use definition of conditional probability and $\quad P(A \cap B) = P(B \cap A)$

$$\rightarrow \quad P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{(Bayes' theorem)}$$

If set of all outcomes $S = \cup_i A_i$ with A_i disjoint, then law of total probability for $P(B)$ says

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

so that Bayes' theorem becomes $\quad P(A|B) = \dfrac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$

Bayes' theorem holds regardless of how probability is interpreted (frequency, degree of belief...).

Frequentist Statistics – general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations (shorthand: x).

Probability = limiting frequency

Probabilities such as

P (string theory is true),
P (0.117 < α_s < 0.119),
P (Biden wins in 2024),

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

Preferred theories (models, hypotheses, …) are those that predict a high probability for data "like" the data observed.

Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability for hypotheses:

probability of the data assuming hypothesis H (the likelihood)

prior probability, i.e., before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayes' theorem has an "if-then" character: If your prior probabilities were $\pi(H)$, then it says how these probabilities should change in the light of the data.

No general prescription for priors (subjective!)

Parameter estimation

The parameters of a pdf are any constants that characterize it,

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

i.e., θ indexes a set of hypotheses.

r.v.   parameter

Suppose we have a sample of observed values: $\boldsymbol{x} = (x_1, ..., x_n)$

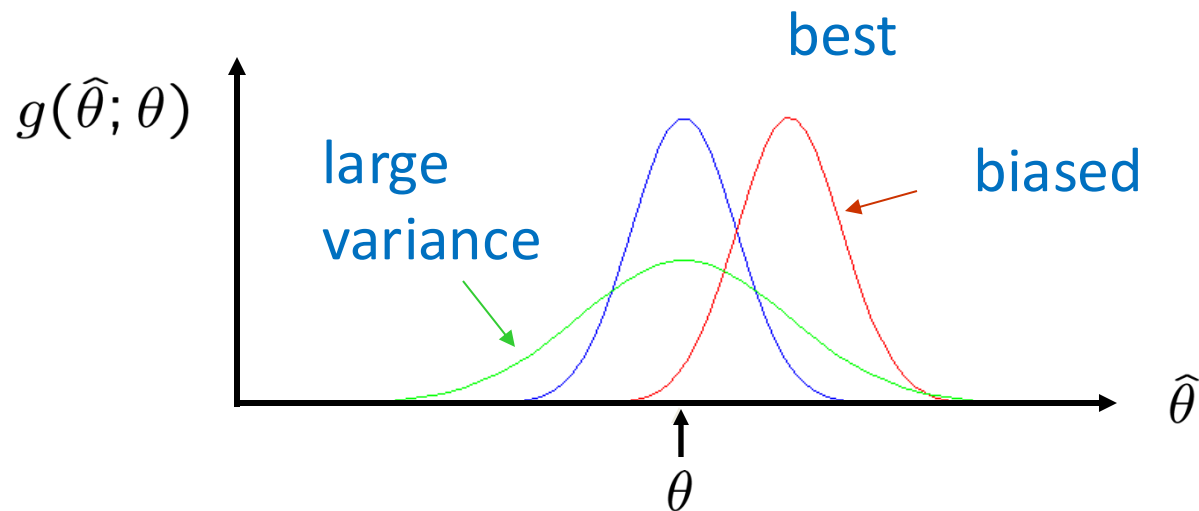We want to find some function of the data to estimate the parameter(s):

$$\widehat{\theta}(\vec{x})$$ ← estimator written with a hat

Sometimes we say 'estimator' for the function of $x_1, ..., x_n$; 'estimate' for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:

We want small (or zero) bias (systematic error): $b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

Hypothesis, likelihood

Suppose the entire result of an experiment (set of measurements) is a collection of numbers x.

A (simple) hypothesis is a rule that assigns a probability to each possible data outcome:

$$P(\mathbf{x}|H) \quad = \quad \text{the likelihood of } H$$

Often we deal with a family of hypotheses labeled by one or more undetermined parameters (a composite hypothesis):

$$P(\mathbf{x}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}) \qquad = \quad \text{the "likelihood function"}$$

Note:

1) For the likelihood we treat the data x as fixed.

2) The likelihood function $L(\boldsymbol{\theta})$ is not a pdf for $\boldsymbol{\theta}$.

The likelihood function for i.i.d.* data

* i.i.d. = independent and identically distributed

Consider n independent observations of x: x_1, ..., x_n, where x follows $f(x; \theta)$. The joint pdf for the whole data sample is:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$
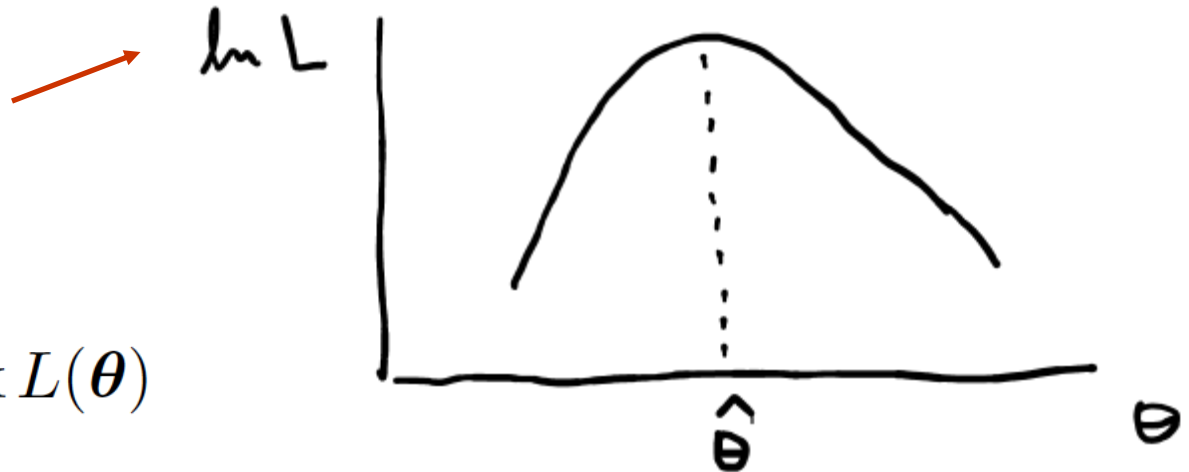
In this case the likelihood function is

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad (x_i \text{ constant})$$

Maximum Likelihood Estimators (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L equivalent to maximizing $\log L$

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

Could have multiple maxima (take highest).

MLEs not guaranteed to have any 'optimal' properties, (but in practice they're very good).

MLE example: parameter of exponential pdf

Consider exponential pdf, $\quad f(t; \tau) = \dfrac{1}{\tau} e^{-t/\tau}$

and suppose we have i.i.d. data, $t_1, \ldots, t_n$

The likelihood function is $\quad L(\tau) = \displaystyle\prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i; \tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

MLE example: parameter of exponential pdf (2)

Find its maximum by setting $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0$,

$$\rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Monte Carlo test:
   generate 50 values
   using $\tau = 1$:

We find the ML estimate:

$$\hat{\tau} = 1.062$$
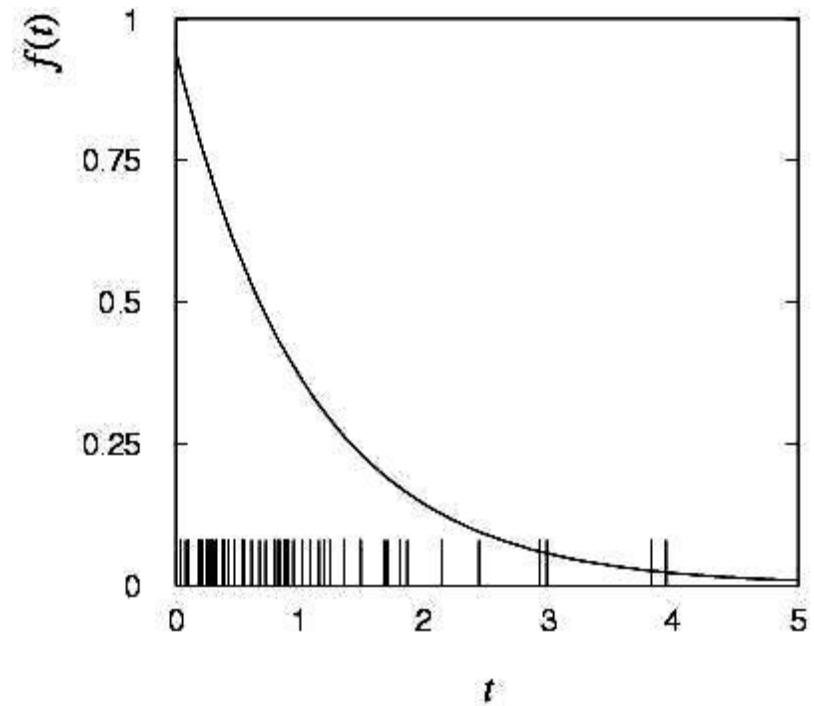
MLE example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$$

$$V[t] = \int_0^\infty (t - \tau)^2 \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$$

For the MLE $\quad \hat{\tau} = \dfrac{1}{n} \sum_{i=1}^n t_i \quad$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n}\sum_{i=1}^n t_i\right] = \frac{1}{n}\sum_{i=1}^n E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n}\sum_{i=1}^n t_i\right] = \frac{1}{n^2}\sum_{i=1}^n V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

Variance of estimators: Monte Carlo method

Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\hat{\sigma}_{\hat{\tau}} = 0.151$$

Note distribution of estimates is roughly Gaussian − (almost) always true for ML in large sample limit.
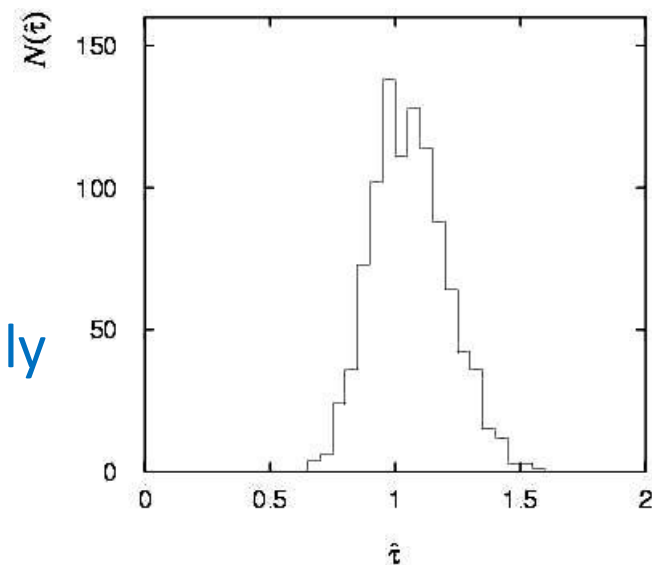
Variance of estimators from information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only ML):

Minimum Variance Bound (MVB)

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \bigg/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

$$(b = E[\hat{\theta}] - \theta)$$

Often the bias b is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\hat{\theta}] \approx -1 \bigg/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of $\ln L$ at its maximum:

$$\hat{V}[\hat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\bigg|_{\theta=\hat{\theta}}$$

Large-sample (asymptotic) properties of MLEs

Suppose we have an i.i.d. data sample of size n: $x_1,...,x_n$

In the large-sample (or "asymptotic") limit ($n \to \infty$) and assuming regularity conditions one can show that the likelihood and MLE have several important properties.

The regularity conditions include:

- the boundaries of the data space cannot depend on the parameter;

- the parameter cannot be on the edge of the parameter space;

- $\ln L(\theta)$ must be differentiable;

- the only solution to $\partial \ln L / \partial \theta = 0$ is $\hat{\theta}$.

In the slides immediately following, the properties are shown without proof for a single parameter; the corresponding properties hold also for the multiparameter case, $\boldsymbol{\theta} = (\theta_1,..., \theta_m)$.

log-likelihood becomes quadratic

The likelihood function becomes Gaussian in shape, i.e. the log-likelihood becomes quadratic (parabolic).
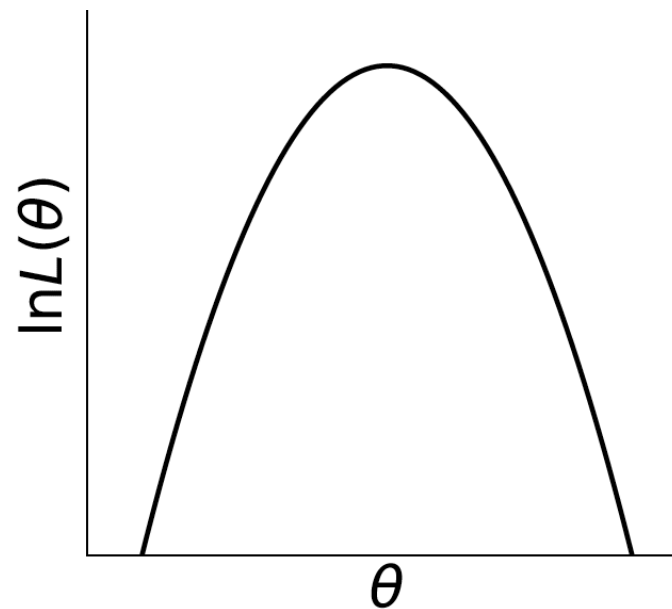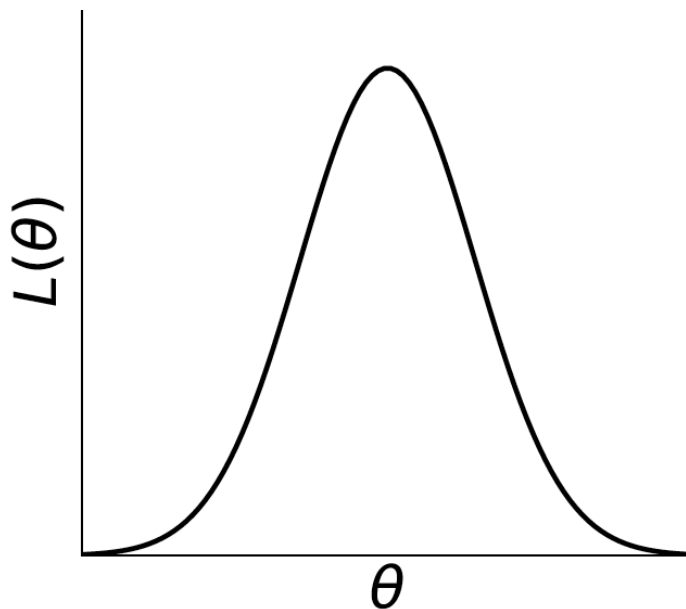
The MLE becomes increasingly precise as the (log)-likelihood becomes more tightly concentrated about its peak, but $L(\theta) = P(x|\theta)$ is the probability for x, not a pdf for θ.

The MLE converges to the true parameter value

In the large-sample limit, the MLE converges in probability to the true parameter value.

That is, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

The MLE is said to be *consistent*.

MLE is asymptotically unbiased

In general the MLE can be biased, but in the large-sample limit, this bias goes to zero:

$$\lim_{n \to \infty} E[\hat{\theta}] - \theta = 0$$

(Recall for the exponential parameter we found the bias was identically zero regardless of the sample size n.)

The MLE's variance approaches the MVB

In the large-sample limit, the variance of the MLE approaches the minimum-variance bound, i.e., the information inequality becomes an equality (and bias goes to zero):

$$\lim_{n \to \infty} V[\hat{\theta}] = -\frac{1}{E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

The MLE is said to be *asymptotically efficient*.

The MLE's distribution becomes Gaussian

In the large-sample limit, the pdf of the MLE becomes Gaussian,

$$f(\hat{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{\theta}}} e^{-(\hat{\theta}-\theta)^2/2\sigma_{\hat{\theta}}^2}$$

where $\sigma_{\hat{\theta}}^2$ is the minimum variance bound (note bias is zero).

For example, exponential MLE with sample size $n = 100$.

Note that for exponential, MLE is arithmetic average, so Gaussian MLE seen to stem from Central Limit Theorem.
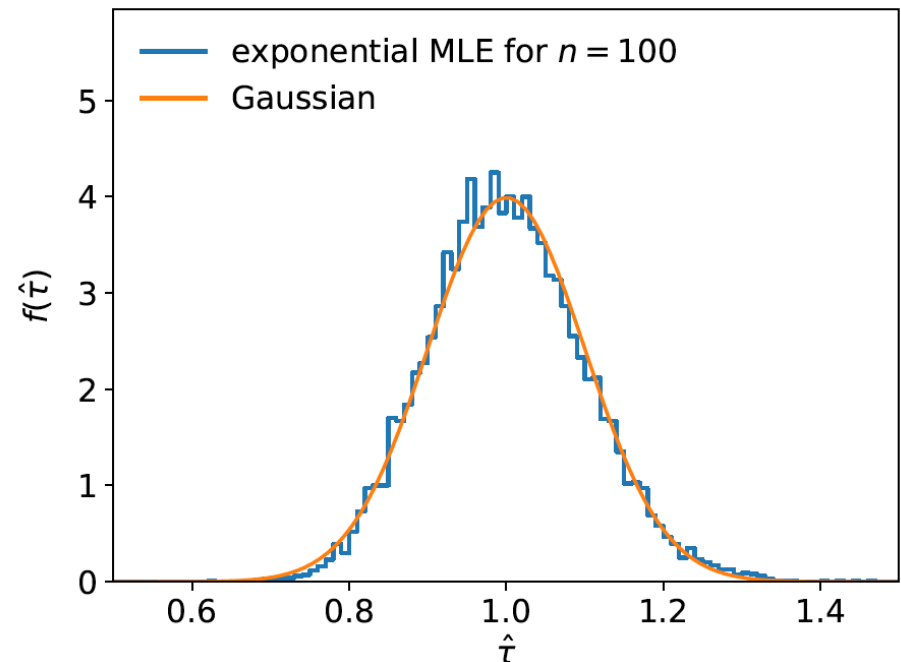
Distribution of MLE of exponential parameter
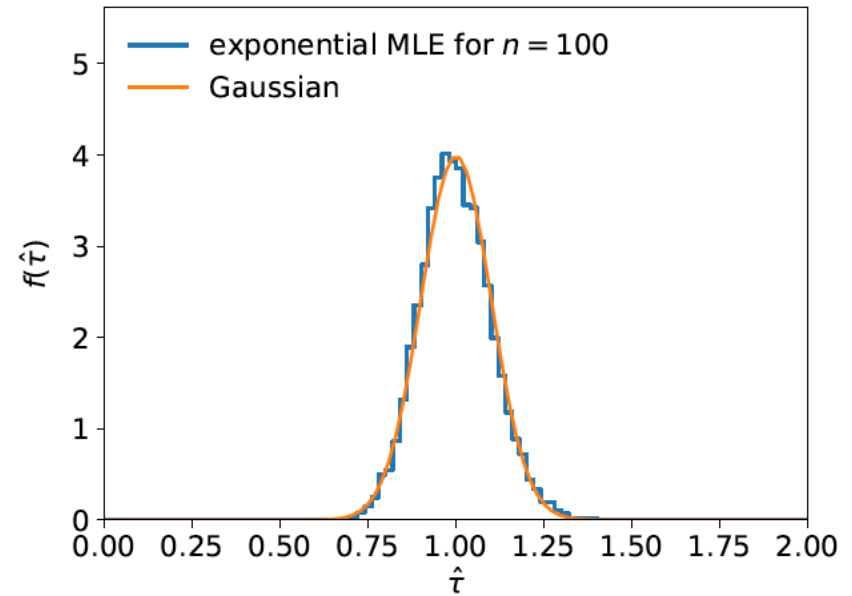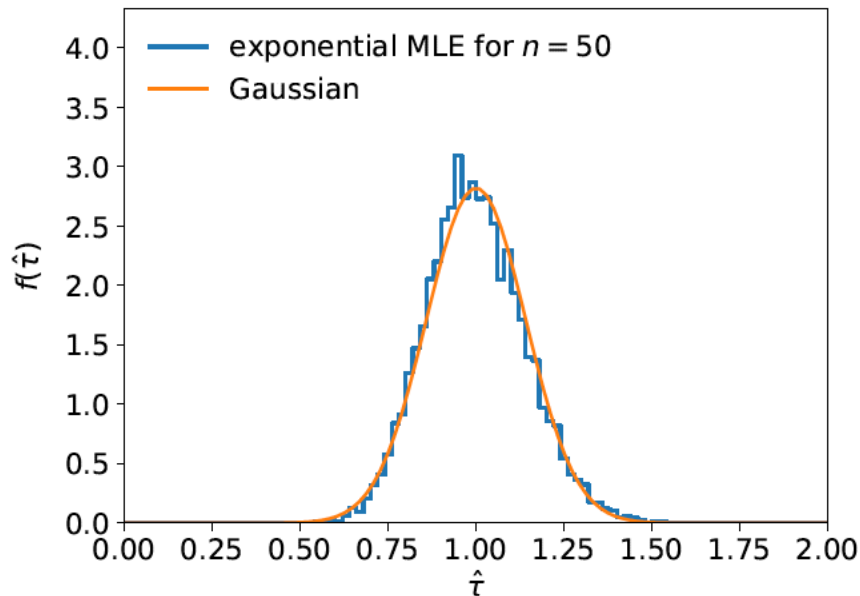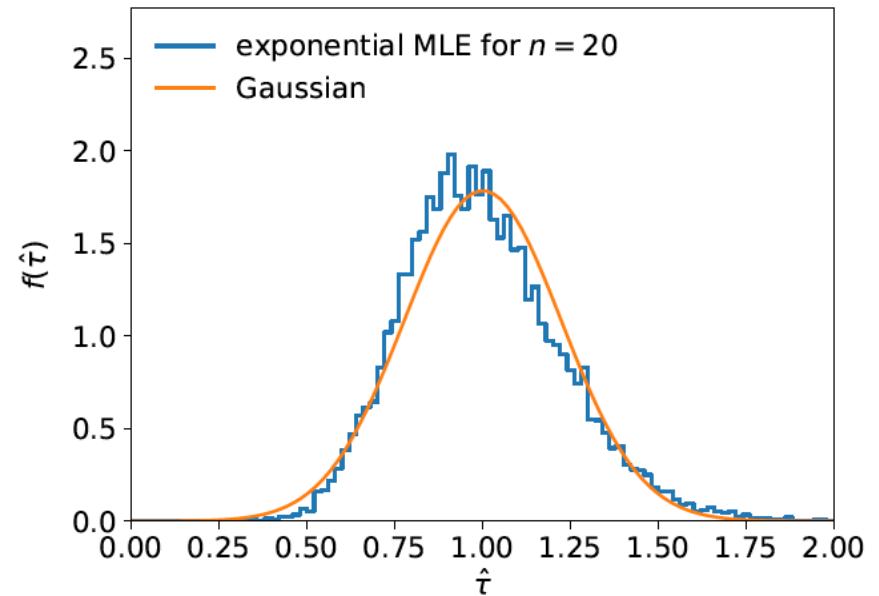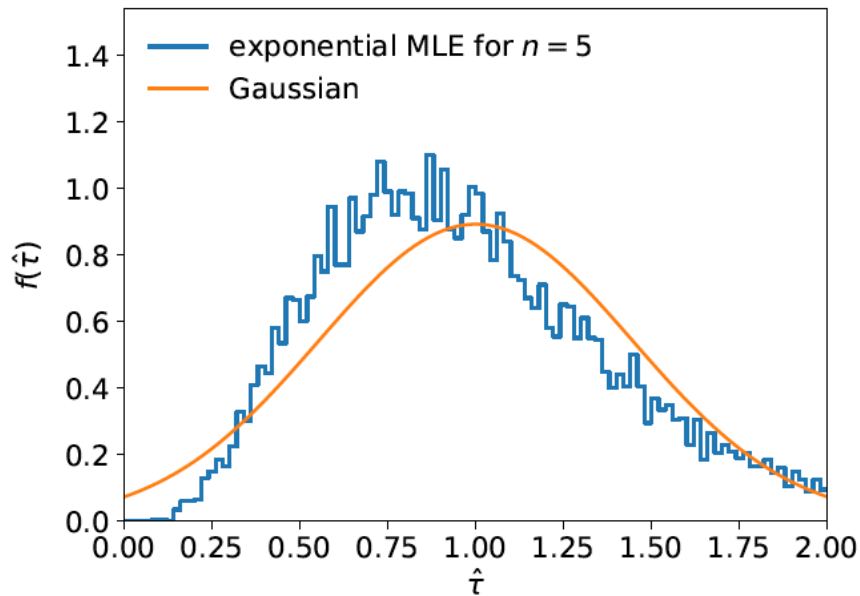
MVB for MLE of exponential parameter

Find $\quad \mathrm{MVB} = -\left(1 + \dfrac{\partial b}{\partial \tau}\right)^2 \Big/ E\left[\dfrac{\partial^2 \ln L}{\partial \tau^2}\right]$

We found for the exponential parameter the MLE $\quad \hat{\tau} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} t_i$

and we showed $b = 0$, hence $\partial b/\partial \tau = 0$.

We find $\quad \dfrac{\partial^2 \ln L}{\partial \tau^2} = \displaystyle\sum_{i=1}^{n}\left(\dfrac{1}{\tau^2} - \dfrac{2t_i}{\tau^3}\right)$

and since $E[t_i] = \tau$ for all i, $\quad E\left[\dfrac{\partial^2 \ln L}{\partial \tau^2}\right] = -\dfrac{n}{\tau^2}$,

and therefore $\mathrm{MVB} = \dfrac{\tau^2}{n} = V[\hat{\tau}]$. $\quad$ (Here MLE is "efficient").

Variance of estimators: graphical method

Expand $\ln L(\theta)$ about its maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!} \left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \dots$$

First term is $\ln L_{\max}$, second term is zero, for third term use information inequality (assume equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}}$$

i.e., $\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$

$\rightarrow$ to get $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until $\ln L$ decreases by 1/2.

Example of variance by graphical method

ML example with exponential:

$$\hat{\tau} \;=\; 1.062$$

$$\Delta\hat{\tau}_- \;=\; 0.137$$

$$\Delta\hat{\tau}_+ \;=\; 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \;\approx\; \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$
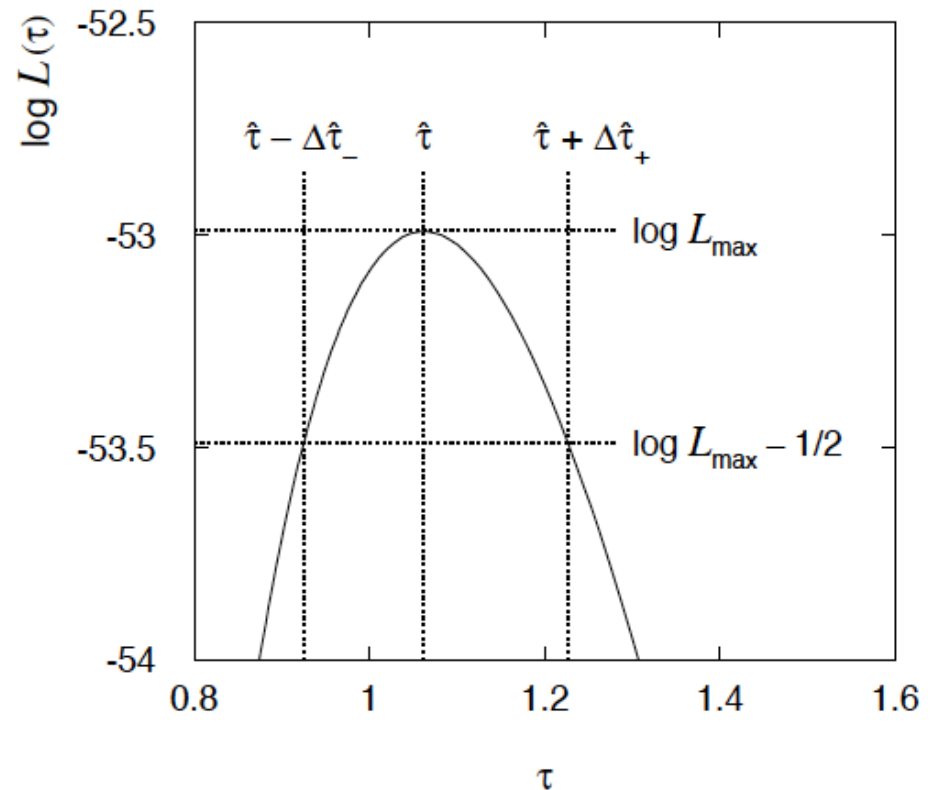
Not quite parabolic $\ln L$ since finite sample size (n = 50).

Information inequality for N parameters

Suppose we have estimated N parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_N)$

The *Fisher information matrix* is

$$I_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right] = -\int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} P(\mathbf{x}|\boldsymbol{\theta})\,d\mathbf{x}$$

and the covariance matrix of estimators $\hat{\boldsymbol{\theta}}$ is $V_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$

The information inequality states that the matrix

$$M_{ij} = V_{ij} - \sum_{k,l}\left(\delta_{ik} + \frac{\partial b_i}{\partial \theta_k}\right) I_{kl}^{-1}\left(\delta_{lj} + \frac{\partial b_j}{\partial \theta_l}\right)$$

is positive semi-definite:

$z^{\mathrm{T}} M z \geq 0$ for all $z \neq 0$, diagonal elements ≥ 0

Information inequality for N parameters (2)

In practice the inequality is ~always used in the large-sample limit:

bias $\rightarrow 0$

inequality $\rightarrow$ equality, i.e, $M = 0$, and therefore $V^{-1} = I$

That is, $\quad V_{ij}^{-1} = -E \left[\dfrac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right]$

This can be estimated from data using $\quad \widehat{V}_{ij}^{-1} = -\dfrac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \bigg|_{\hat{\boldsymbol{\theta}}}$

Find the matrix V^{-1} numerically (or with automatic differentiation), then invert to get the covariance matrix of the estimators

$$\widehat{V}_{ij} = \widehat{\mathrm{cov}}[\hat{\theta}_i, \hat{\theta}_j]$$

Multiparameter graphical method for variances

Expand $\ln L(\boldsymbol{\theta})$ to 2nd order about MLE:

$$\ln L(\boldsymbol{\theta}) \approx \ln L(\hat{\boldsymbol{\theta}}) + \sum_i \frac{\partial \ln L}{\partial \theta_i}\bigg|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i) + \frac{1}{2!} \sum_{i,j} \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\bigg|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$
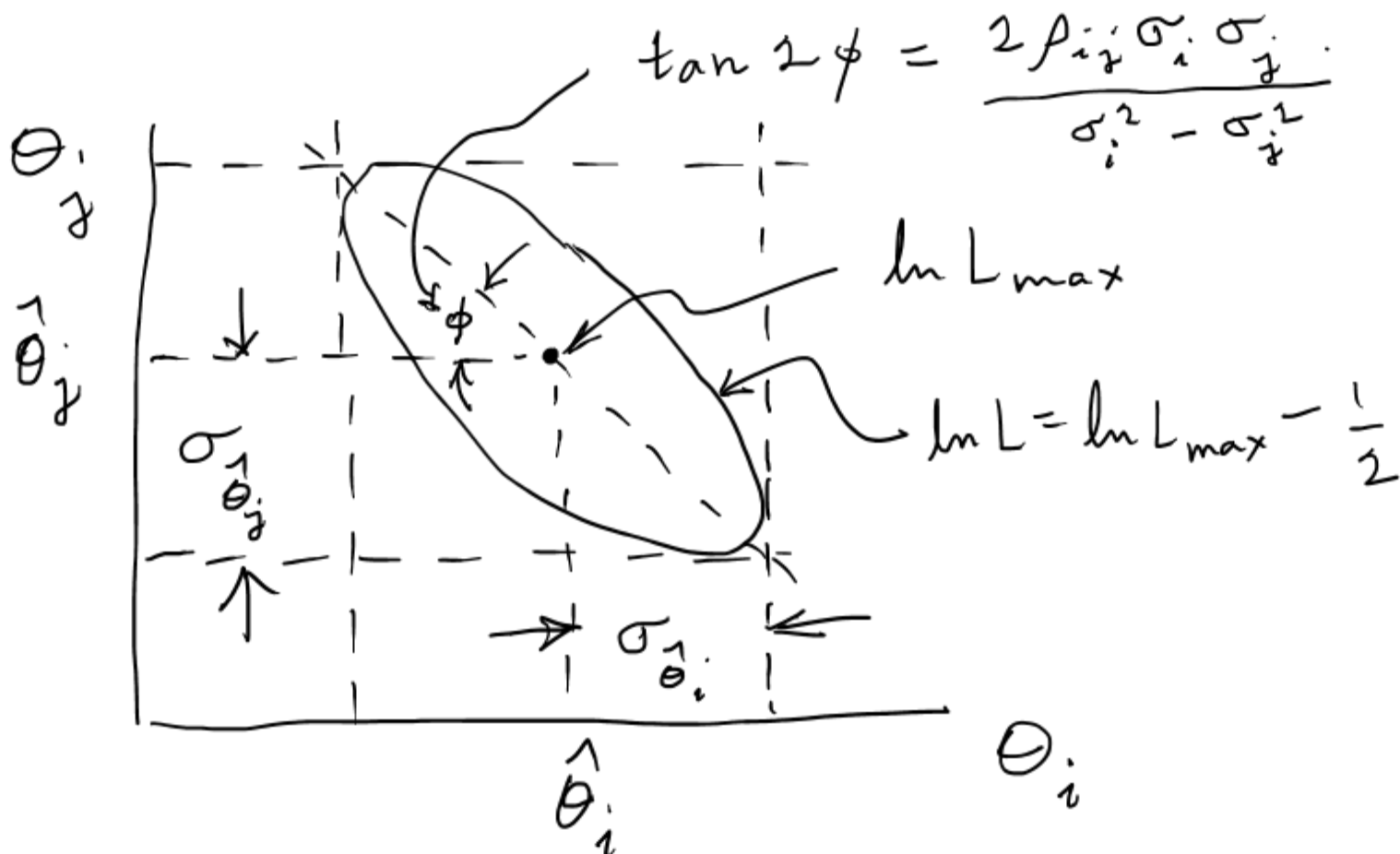
$\ln L_{\mathrm{max}}$ zero

relate to covariance matrix of MLEs using information (in)equality.

Result: $\ln L(\boldsymbol{\theta}) = \ln L_{\mathrm{max}} - \frac{1}{2} \sum_{i,j} (\theta_i - \hat{\theta}_i) V_{ij}^{-1} (\theta_j - \hat{\theta}_j)$

So the surface $\ln L(\boldsymbol{\theta}) = \ln L_{\mathrm{max}} - \frac{1}{2}$ corresponds to

$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T V^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = 1$, which is the equation of a (hyper-) ellipse.

Multiparameter graphical method (2)

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2}$$

$\ln L_{max}$

$\ln L = \ln L_{max} - \frac{1}{2}$

Distance from MLE to tangent planes gives standard deviations.

Frequentist hypothesis tests

Suppose a measurement produces data x; consider a hypothesis H_0 we want to test and alternative H_1

H_0, H_1 specify probability for x: $P(x|H_0), P(x|H_1)$

A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in w \mid H_0) \leq \alpha$$

Need inequality if data are discrete.

α is called the size or significance level of the test.

If x is observed in the critical region, reject H_0.

data space Ω
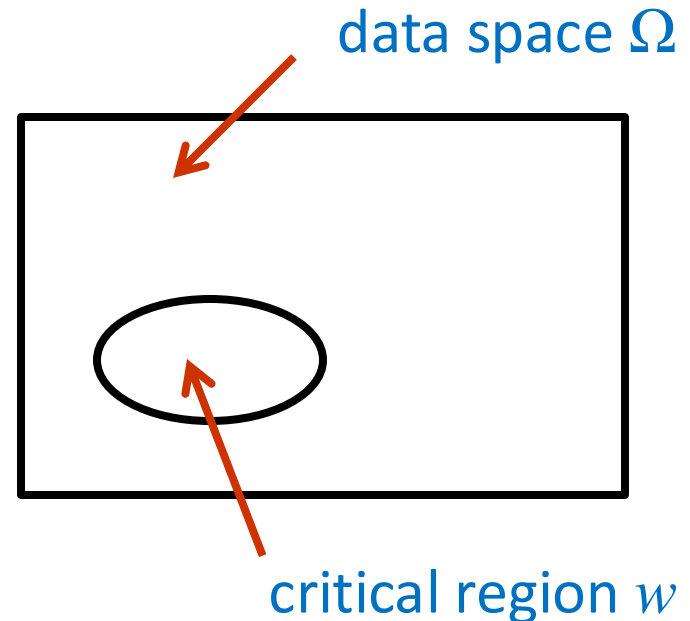
critical region w

Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same size α.

Use the alternative hypothesis H_1 to motivate where to place the critical region.

Roughly speaking, place the critical region where there is a low probability (α) to be found if H_0 is true, but high if H_1 is true:
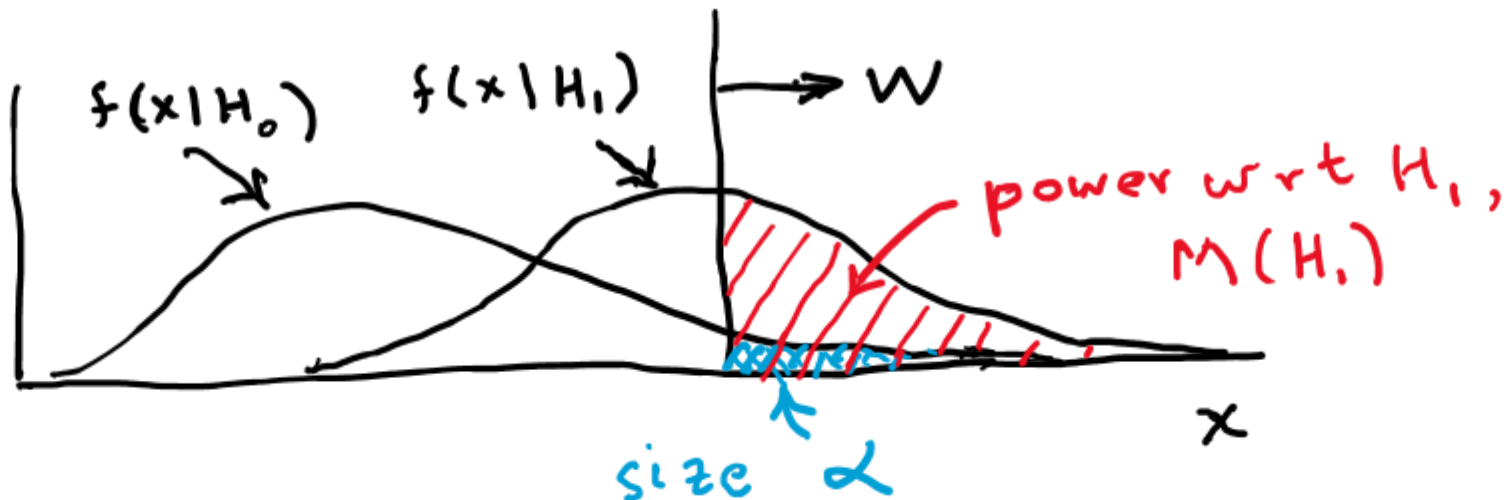
Classification viewed as a statistical test

Suppose events come in two possible types:

s (signal) and b (background)

For each event, test hypothesis that it is background, i.e., H_0 = b.

Carry out test on many events, each is either of type s or b, i.e., here the hypothesis is the "true class label", which varies randomly from event to event, so we can assign to it a frequentist probability.

Select events for which where H_0 is rejected as "candidate events of type s". Equivalent Particle Physics terminology:

background efficiency
$$\varepsilon_{\rm b} = \int_W f(\mathbf{x}|H_0)\, d\mathbf{x} = \alpha$$

signal efficiency
$$\varepsilon_{\rm s} = \int_W f(\mathbf{x}|H_1)\, d\mathbf{x} = 1 - \beta = \text{power}$$

Example of a test for classification

Suppose we can measure for each event a quantity x, where

$$f(x|s) = 2(1 - x)$$

$$f(x|b) = 4x^3$$

with $0 \leq x \leq 1$.
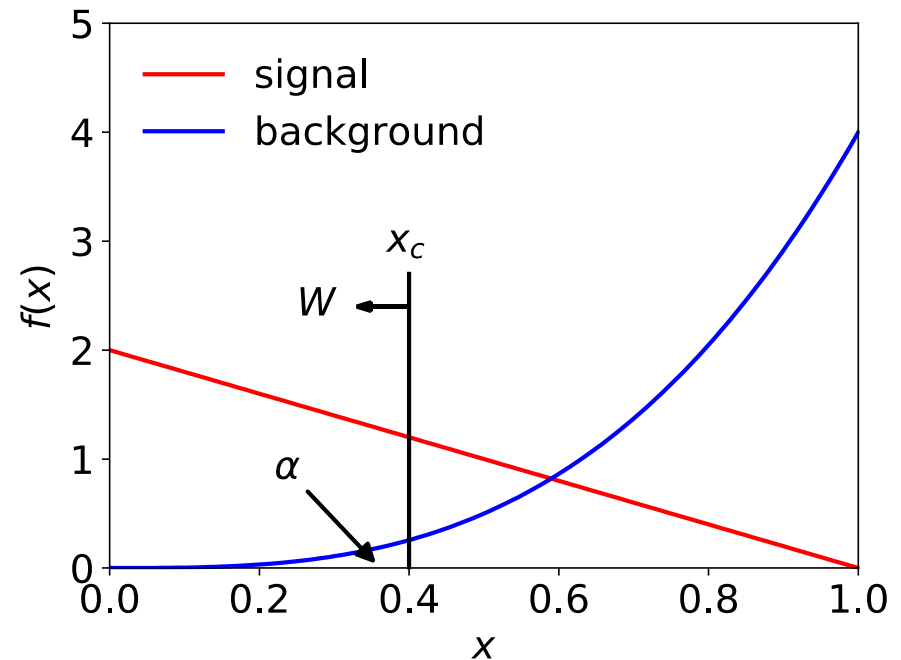
For each event in a mixture of signal (s) and background (b) test

H_0 : event is of type b

using a critical region W of the form: $W = \{ x : x \leq x_c \}$, where x_c is a constant that we choose to give a test with the desired size α.

Classification example (2)

Suppose we want $\alpha = 10^{-4}$. Require:

$$\alpha = P(x \leq x_{\mathrm{c}}|\mathrm{b}) = \int_0^{x_{\mathrm{c}}} f(x|b)\, dx = \left. \frac{4x^4}{4} \right|_0^{x_{\mathrm{c}}} = x_{\mathrm{c}}^4$$

and therefore $x_{\mathrm{c}} = \alpha^{1/4} = 0.1$

For this test (i.e. this critical region W), the power with respect to the signal hypothesis (s) is

$$M = P(x \leq x_{\mathrm{c}}|\mathrm{s}) = \int_0^{x_{\mathrm{c}}} f(x|\mathrm{s})\, dx = 2x_{\mathrm{c}} - x_{\mathrm{c}}^2 = 0.19$$

Note: the optimal size and power is a separate question that will depend on goals of the subsequent analysis.

Classification example (3)

Suppose that the prior probabilities for an event to be of type s or b are:

$$\pi_{\mathrm{s}} = 0.001$$
$$\pi_{\mathrm{b}} = 0.999$$

The "purity" of the selected signal sample (events where b hypothesis rejected) is found using Bayes' theorem:

$$P(\mathrm{s}|x \leq x_{\mathrm{c}}) = \frac{P(x \leq x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}}}{P(x \leq x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}} + P(x \leq x_{\mathrm{c}}|\mathrm{b})\pi_{\mathrm{b}}}$$

$$= 0.655$$

Particle Physics context for a hypothesis test

A simulated SUSY event ("signal"):

high p_T muons

high p_T jets of hadrons

p

p

missing transverse energy
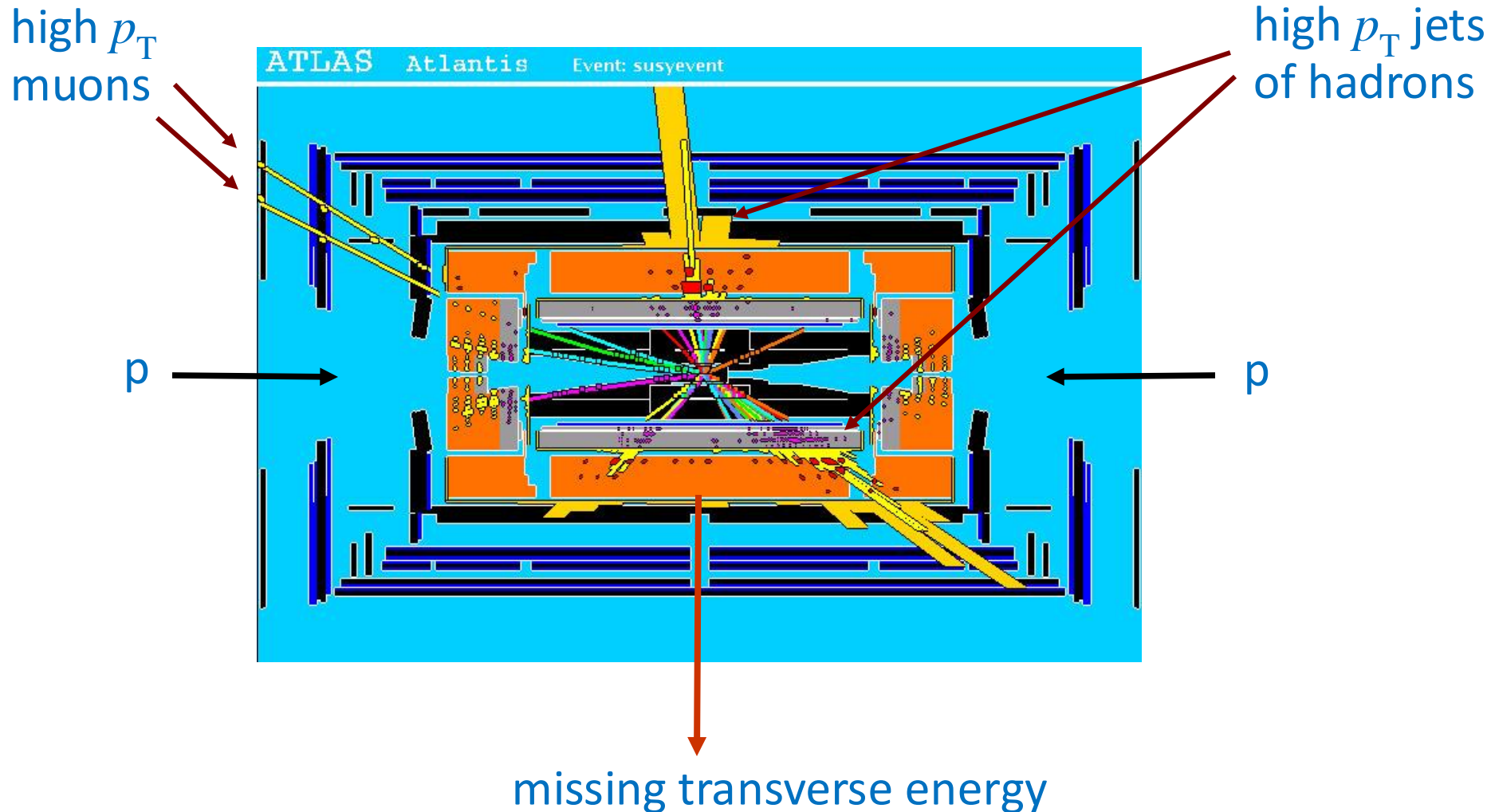
Background events
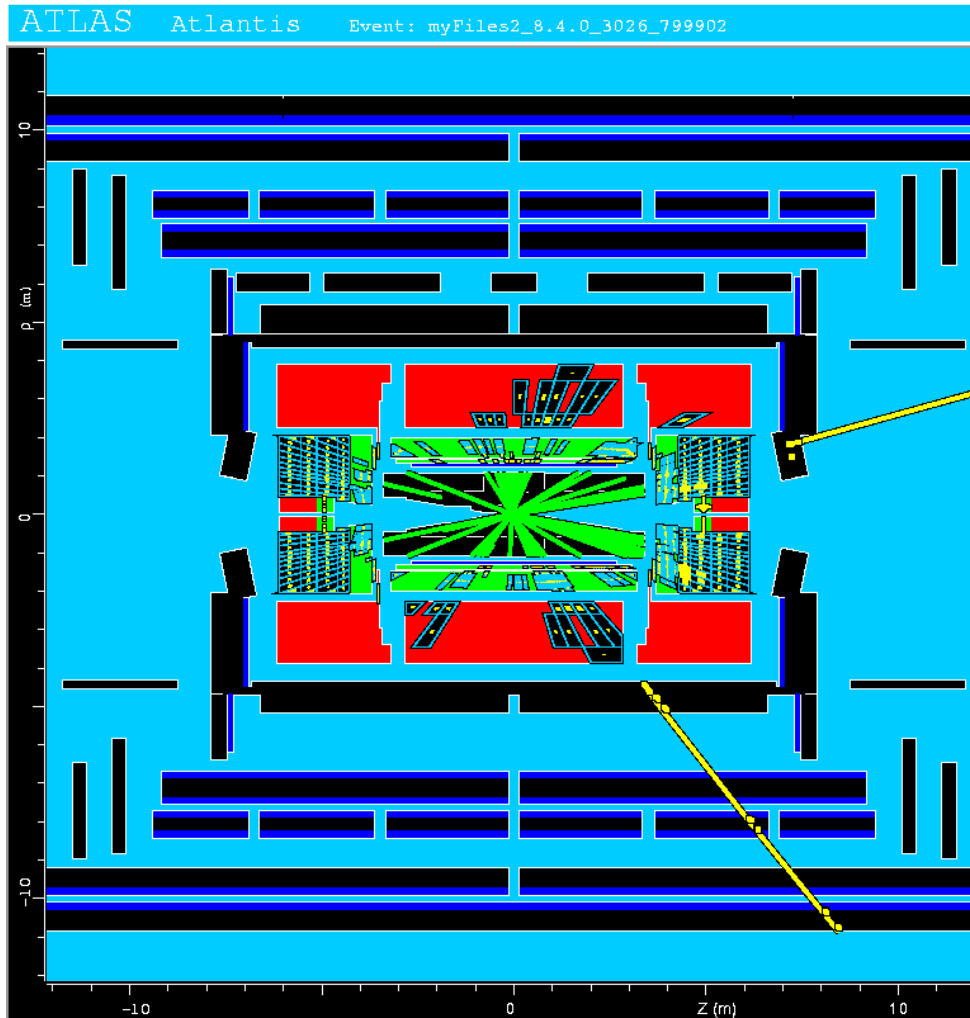
This event from Standard Model ttbar production also has high p_T jets and muons, and some missing transverse energy.

$\rightarrow$ can easily mimic a
    signal event.

Classification of proton-proton collisions

Proton-proton collisions can be considered to come in two classes:

signal (the kind of event we're looking for, $y = 1$)
background (the kind that mimics signal, $y = 0$)

For each collision (event), we measure a collection of features:

x_1 = energy of muon x_4 = missing transverse energy
x_2 = angle between jets x_5 = invariant mass of muon pair
x_3 = total jet energy x_6 = ...

The real events don't come with true class labels, but computer-simulated events do. So we can have a set of simulated events that consist of a feature vector $\boldsymbol{x}$ and true class label y (0 for background, 1 for signal):

$$(\boldsymbol{x}, y)_1, (\boldsymbol{x}, y)_2, ..., (\boldsymbol{x}, y)_N$$

The simulated events are called "training data".

Distributions of the features

If we consider only two features $x = (x_1, x_2)$, we can display the results in a scatter plot (red: $y = 0$, blue: $y = 1$).

For real events, the dots are black (true type is not known).

For each real event test the hypothesis that it is background.

(Related to this: test that a sample of events is *all* background.)

The test's critical region is defined by a "decision boundary" – without knowing the event type, we can classify them by seeing where their measured features lie relative to the boundary.
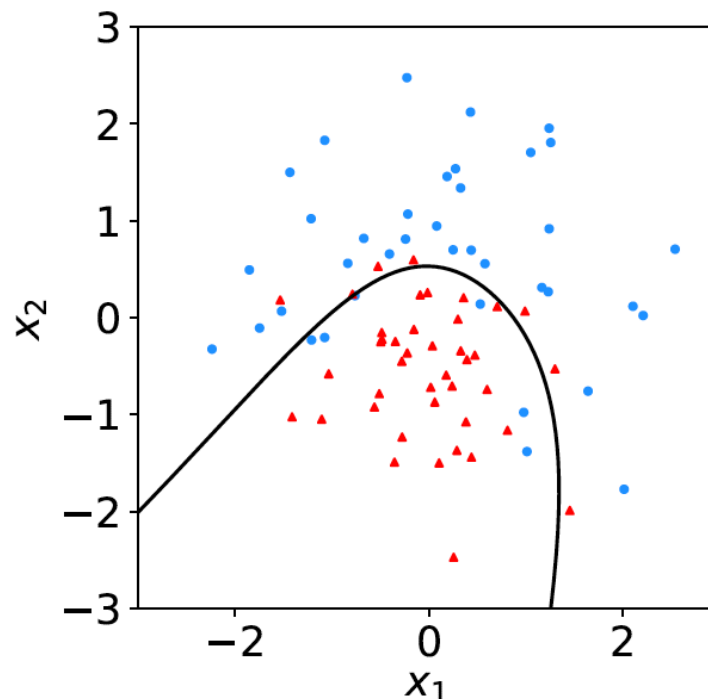
Decision function, test statistic

A surface in an n-dimensional space can be described by

$$t(x_1, \ldots, x_n) = t_c$$

scalar function

constant

Different values of the constant t_c result in a family of surfaces.

Problem is reduced to finding the best decision function or test statistic $t(\boldsymbol{x})$.
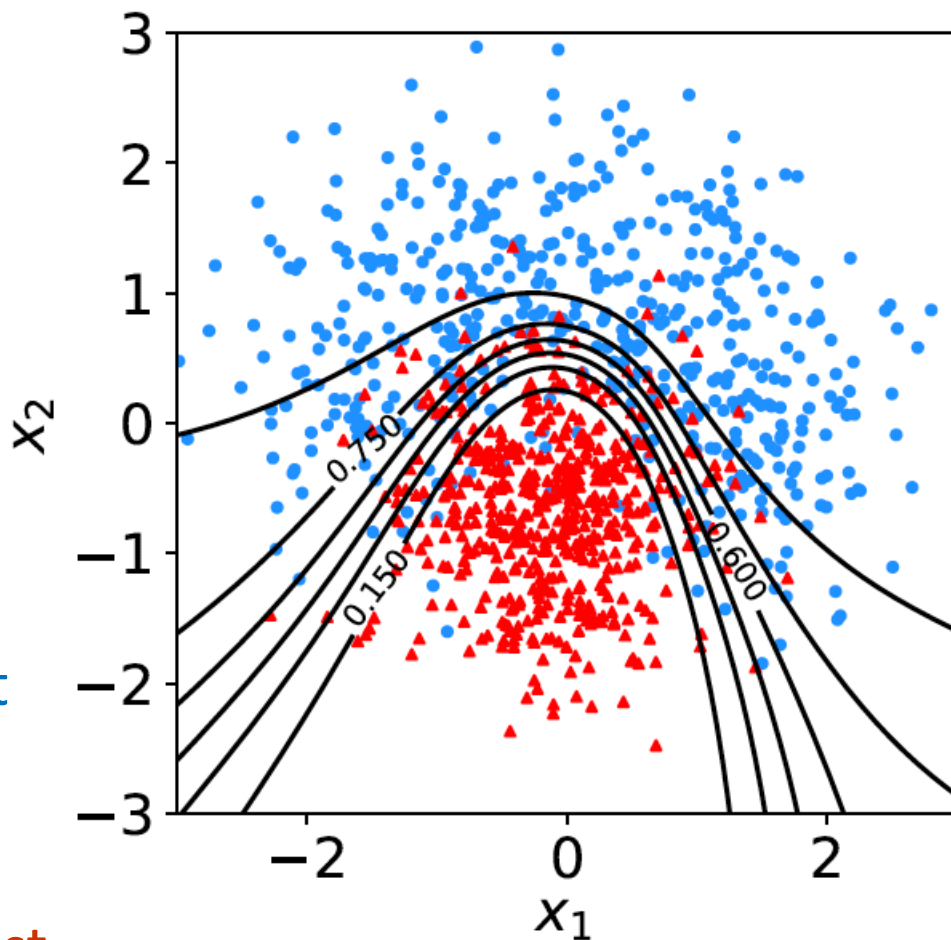
Distribution of $t(\boldsymbol{x})$

By forming a test statistic $t(\boldsymbol{x})$, the boundary of the critical region in the n-dimensional $\boldsymbol{x}$-space is determined by a single single value t_{c}.
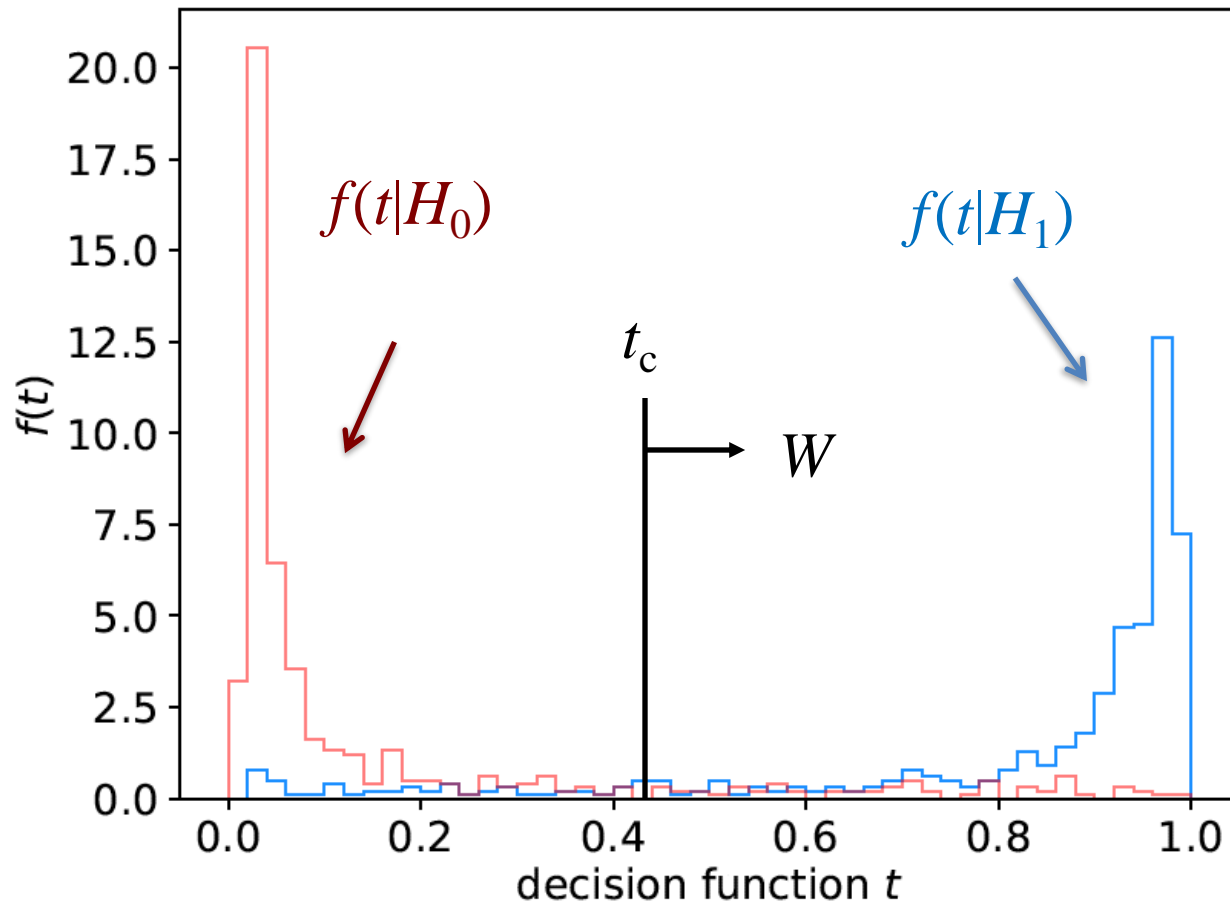
Types of decision boundaries

So what is the optimal boundary for the critical region, i.e., what is the optimal test statistic $t(x)$?

First find best $t(x)$, later address issue of optimal size of test.

Remember x-space can have many dimensions.
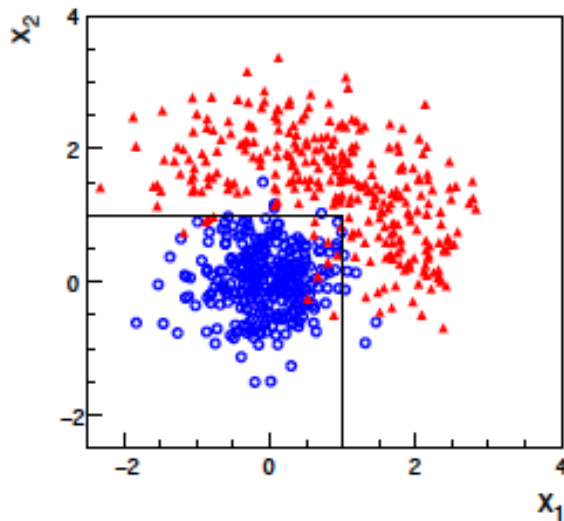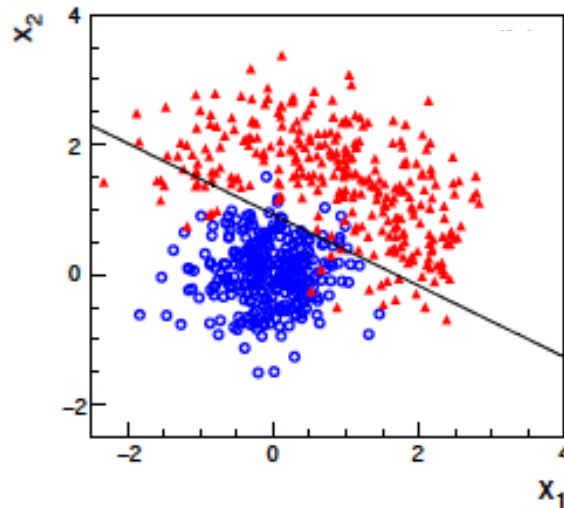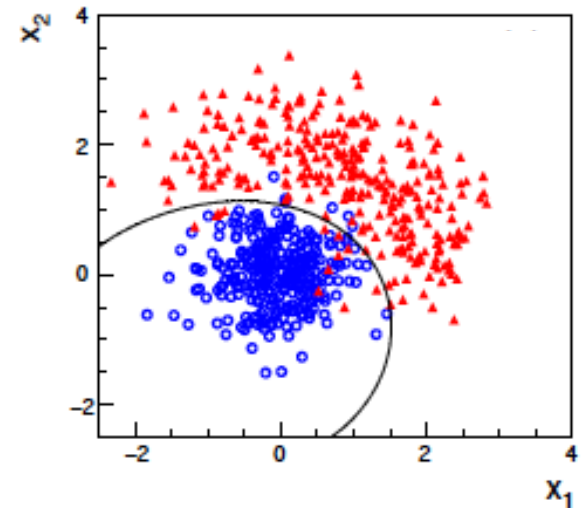
"cuts"   linear   non-linear

Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way', in particular if the data space is multidimensional?

Neyman-Pearson lemma states:

For a test of H_0 of size α, to get the highest power with respect to the alternative H_1 we need for all x in the critical region W

"likelihood ratio (LR)"

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \geq c_\alpha$$

inside W and $\leq c_\alpha$ outside, where c_α is a constant chosen to give a test of the desired size.

Equivalently, optimal scalar test statistic is

$$t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$$

N.B. any monotonic function of this is leads to the same test.

Neyman-Pearson doesn't usually help

We usually don't have explicit formulae for the pdfs $f(x|s), f(x|b),$ so for a given x we can't evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}$$

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data:

generate $x \sim f(x|s) \quad \rightarrow \quad x_1,..., x_N$

generate $x \sim f(x|b) \quad \rightarrow \quad x_1,..., x_N$

This gives samples of "training data" with events of known type.

- Use these to construct a statistic that is as close as possible to the optimal likelihood ratio ($\rightarrow$ Machine Learning).

Approximate LR from histograms

Want $t(x) = f(x|s)/f(x|b)$ for x here
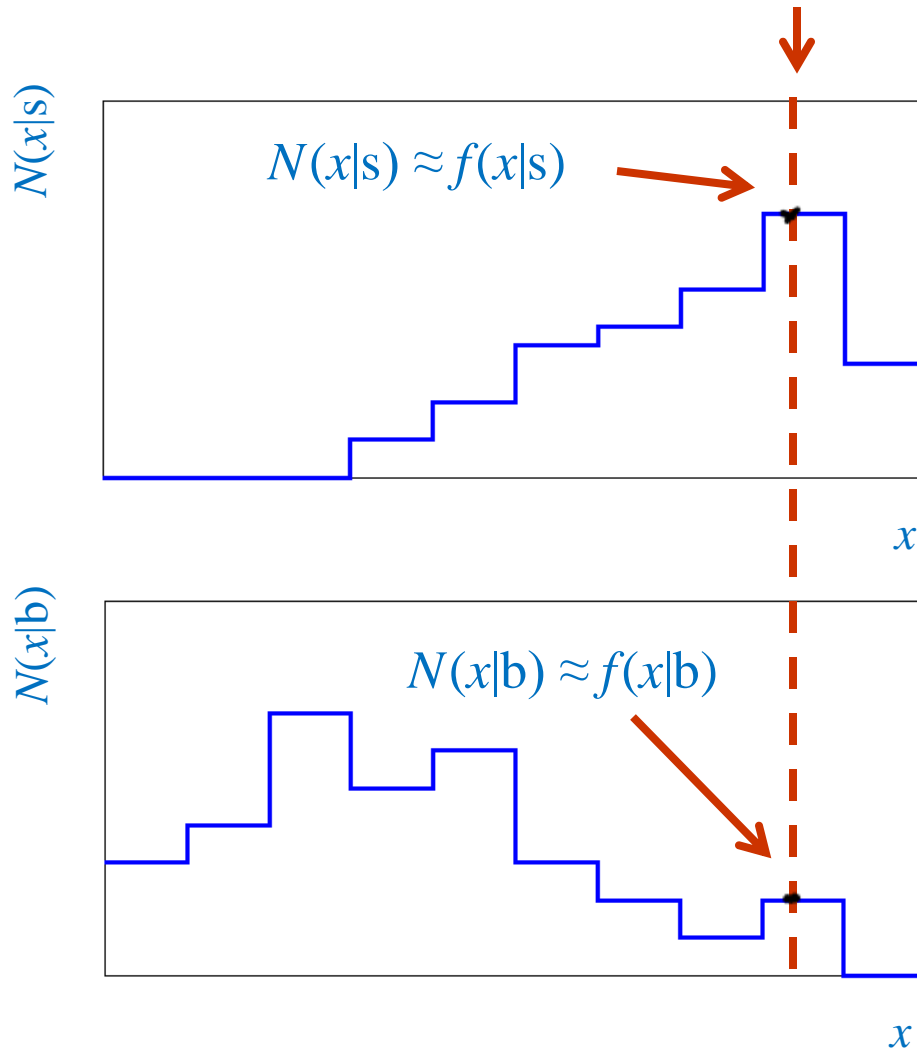
One possibility is to generate MC data and construct histograms for both signal and background.

Use (normalized) histogram values to approximate LR:

$$t(x) \approx \frac{N(x|s)}{N(x|b)}$$

Can work well for single variable.

Approximate LR from 2D-histograms

Suppose problem has 2 variables. Try using 2-D histograms:

signal

back-
ground
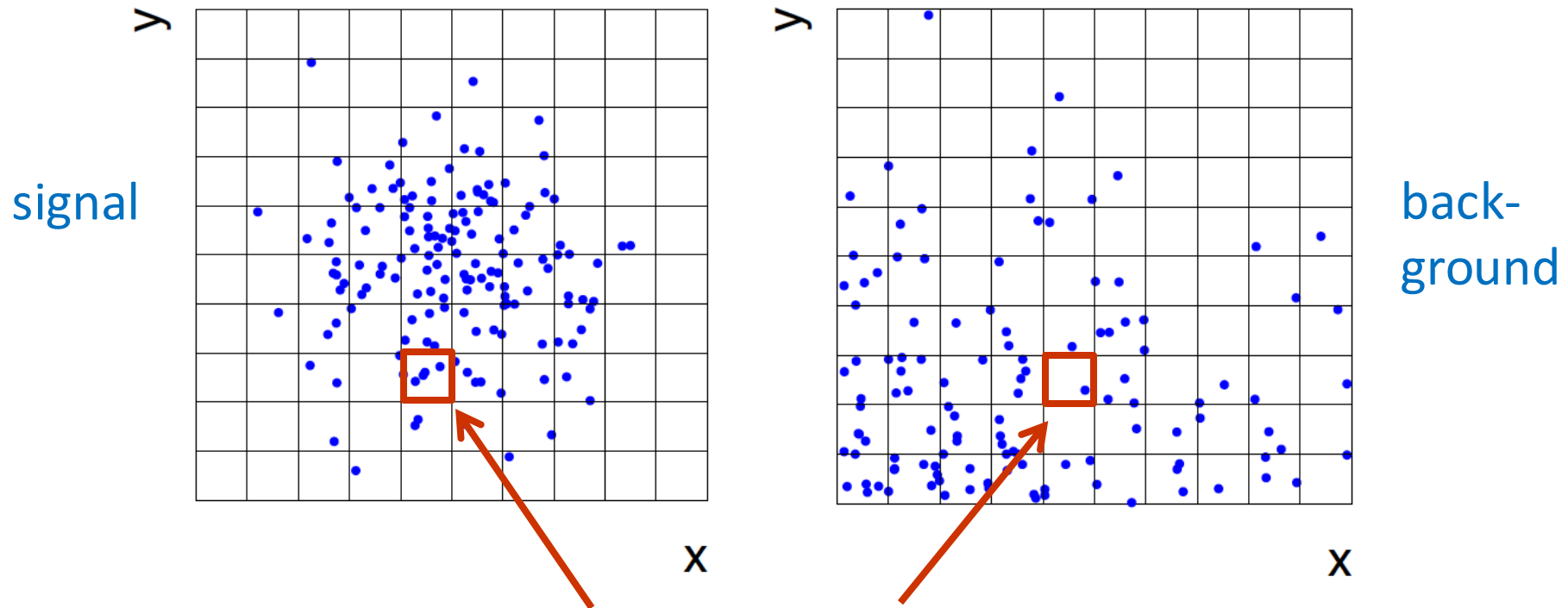
Approximate pdfs using $N(x,y|\text{s})$, $N(x,y|\text{b})$ in corresponding cells.

But if we want M bins for each variable, then in n-dimensions we have M^n cells; can't generate enough training data to populate.

$\rightarrow$ Histogram method usually not usable for $n > 1$ dimension.

Strategies for multivariate analysis

Neyman-Pearson lemma gives optimal answer, but cannot be used directly, because we usually don't have $f(\boldsymbol{x}|\mathrm{s}), f(\boldsymbol{x}|\mathrm{b})$.

Histogram method with M bins for n variables requires that we estimate M^n parameters (the values of the pdfs in each cell), so this is rarely practical.

A compromise solution is to assume a certain functional form for the test statistic $t(\boldsymbol{x})$ with fewer parameters; determine them (using MC) to give best separation between signal and background.

Alternatively, try to estimate the probability densities $f(\boldsymbol{x}|\mathrm{s})$ and $f(\boldsymbol{x}|\mathrm{b})$ (with something better than histograms) and use the estimated pdfs to construct an approximate likelihood ratio.

Multivariate methods   (Machine Learning)

Many new (and some old) methods:

Fisher discriminant

(Deep) Neural Networks

Kernel density methods

Support Vector Machines

Decision trees

Boosting

Bagging

Extra slides

Some statistics books, papers, etc.

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998

R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989

Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014.

Luca Lista, *Statistical Methods for Data Analysis in Particle Physics*, Springer, 2017.

L. Lyons, *Statistics for Nuclear and Particle Physics*, CUP, 1986

F. James., *Statistical and Computational Methods in Experimental Physics*, 2nd ed., World Scientific, 2006

S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998.

R.L. Workman et al. (Particle Data Group), Prog. Theor. Exp. Phys. 083C01 (2022); `pdg.lbl.gov` sections on probability, statistics, MC.

Some distributions

Distribution/pdf	Example use in Particle Physics
Binomial	Branching ratio
Multinomial	Histogram with fixed N
Poisson	Number of events found
Uniform	Monte Carlo method
Exponential	Decay time
Gaussian	Measurement error
Chi-square	Goodness-of-fit
Cauchy	Mass of resonance
Landau	Ionization energy loss
Beta	Prior pdf for efficiency
Gamma	Sum of exponential variables
Student's t	Resolution function with adjustable tails

Binomial distribution

Consider N independent experiments (Bernoulli trials):

outcome of each is 'success' or 'failure',

probability of success on any given trial is p.

Define discrete r.v. n = number of successes ($0 \le n \le N$).

Probability of a specific outcome (in order), e.g. 'ssfsf' is

$$pp(1-p)p(1-p) = p^n(1-p)^{N-n}$$

But order not important; there are $\dfrac{N!}{n!(N-n)!}$

ways (permutations) to get n successes in N trials, total probability for n is sum of probabilities for each permutation.

Binomial distribution (2)

The binomial distribution is therefore

$$f(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

random
variable
   parameters

For the expectation value and variance we find:

$$E[n] = \sum_{n=0}^{N} n f(n; N, p) = Np$$

$$V[n] = E[n^2] - (E[n])^2 = Np(1-p)$$

Binomial distribution (3)

Binomial distribution for several values of the parameters:
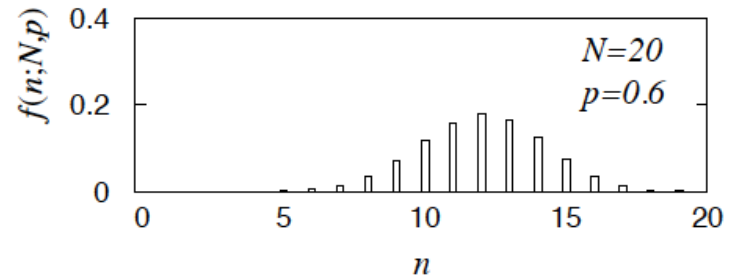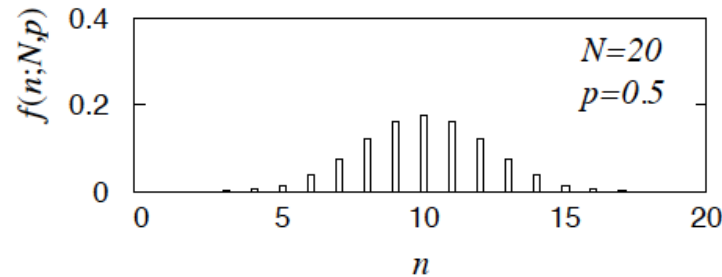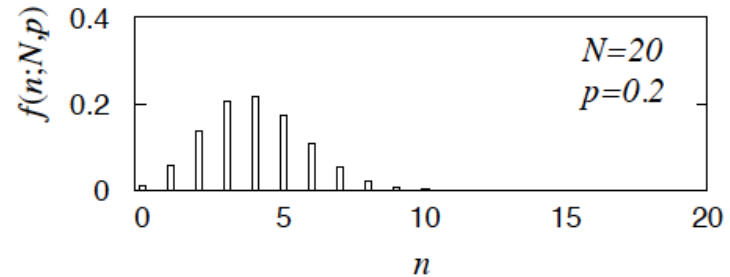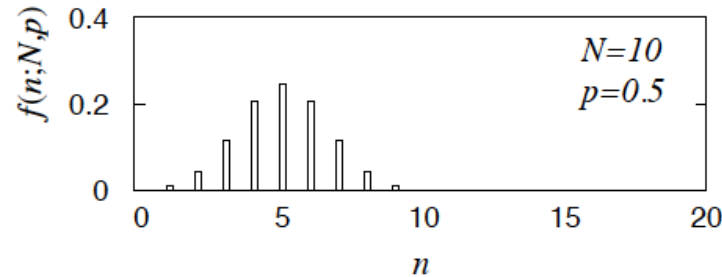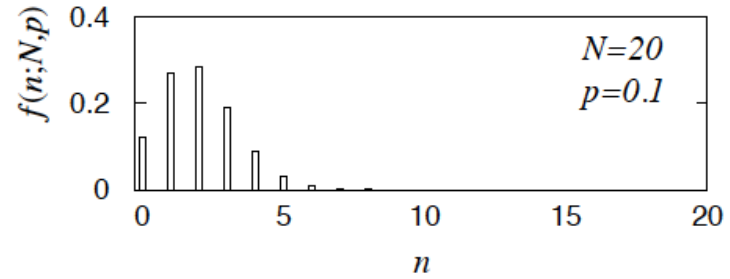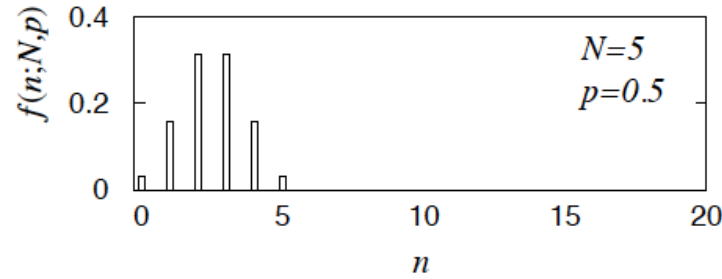
Example: observe N decays of $W^{\pm}$, the number n of which are $W \rightarrow \mu\nu$ is a binomial r.v., p = branching ratio.

Multinomial distribution

Like binomial but now m outcomes instead of two, probabilities are

$$\vec{p} = (p_1, \ldots, p_m), \quad \text{with} \quad \sum_{i=1}^{m} p_i = 1 .$$

For N trials we want the probability to obtain:

n_1 of outcome 1,
n_2 of outcome 2,
$\vdots$
n_m of outcome m.

This is the multinomial distribution for $\vec{n} = (n_1, \ldots, n_m)$

$$f(\vec{n}; N, \vec{p}) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$$

Multinomial distribution (2)

Now consider outcome i as 'success', all others as 'failure'.

$\rightarrow$ all n_i individually binomial with parameters N, p_i

$$E[n_i] = Np_i, \quad V[n_i] = Np_i(1 - p_i) \quad \text{for all } i$$

One can also find the covariance to be

$$V_{ij} = Np_i(\delta_{ij} - p_j)$$

Example: $\vec{n} = (n_1, \ldots, n_m)$ represents a histogram

with m bins, N total entries, all entries independent.

Poisson distribution

Consider binomial n in the limit

$$N \to \infty, \qquad p \to 0, \qquad E[n] = Np \to \nu \,.$$

$\to$ n follows the Poisson distribution:

$$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu} \quad (n \geq 0)$$

$$E[n] = \nu \,, \quad V[n] = \nu \,.$$

Example: number of scattering events
n with cross section $\int$ found for a fixed
integrated luminosity, with $\nu = \sigma \int L \, dt$.
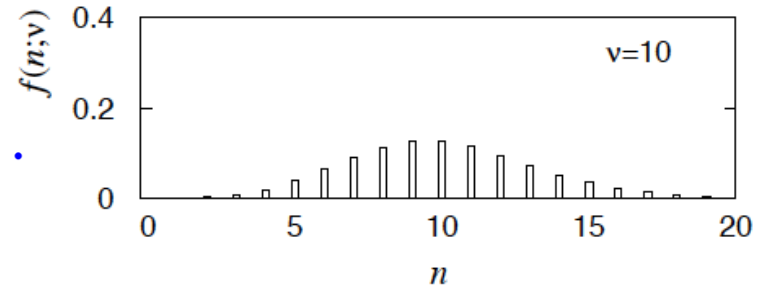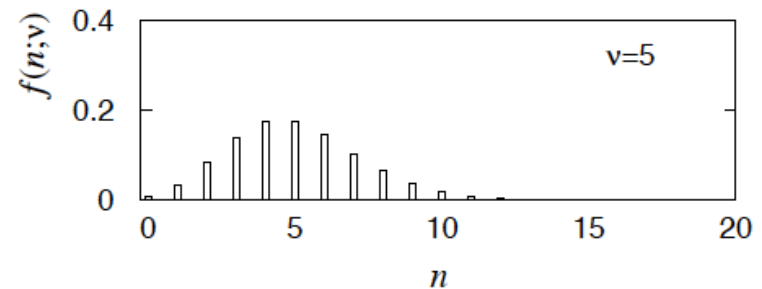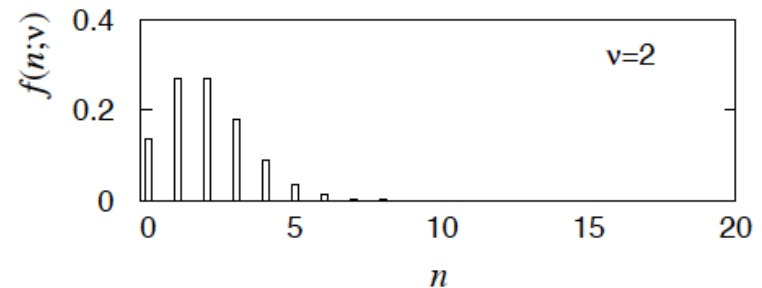
Uniform distribution

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \frac{1}{2}(\alpha + \beta)$$

$$V[x] = \frac{1}{12}(\beta - \alpha)^2$$
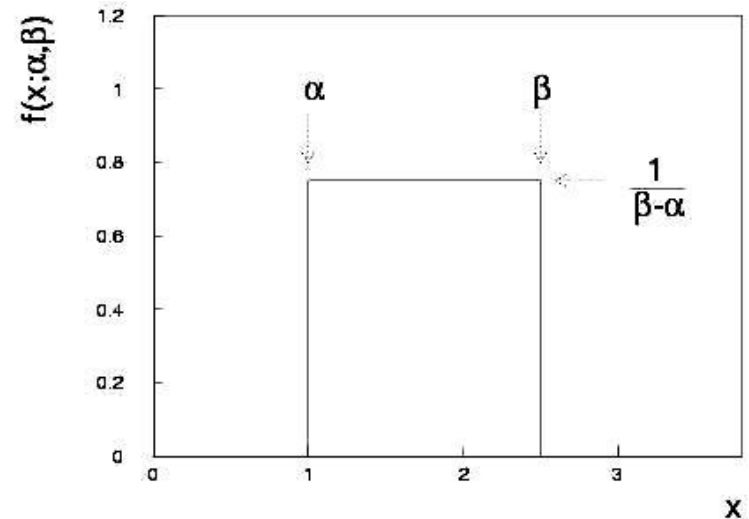
Notation: x follows a uniform distribution between α and β

write as: $x \sim \mathrm{U}[\alpha,\beta]$

Exponential distribution

The exponential pdf for the continuous r.v. x is defined by:

$$f(x; \xi) = \begin{cases} \frac{1}{\xi} e^{-x/\xi} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \xi$$

$$V[x] = \xi^2$$
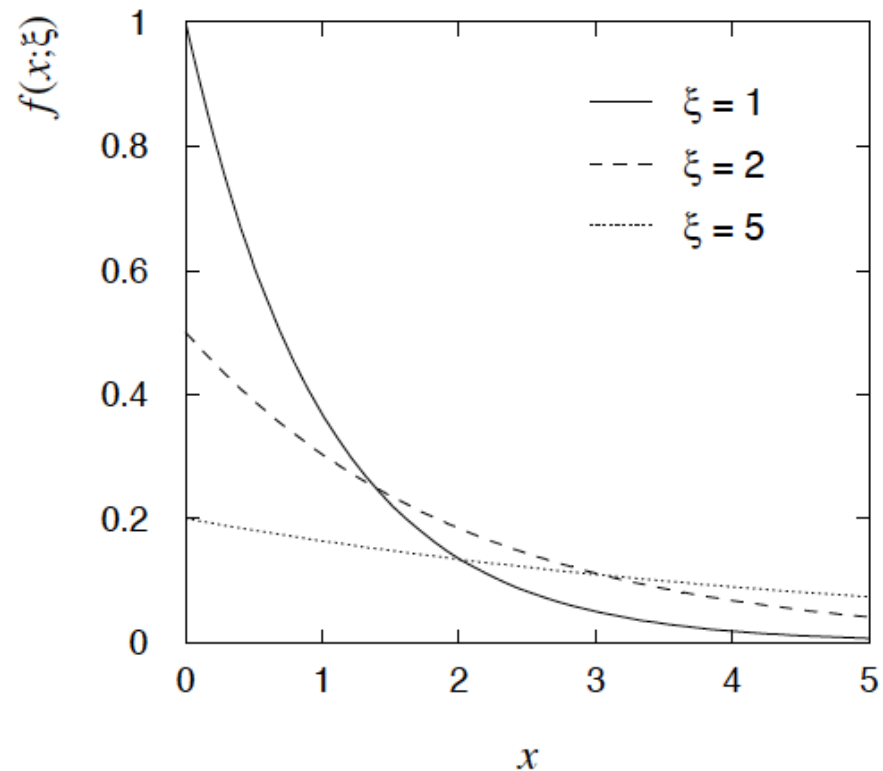
Gaussian (normal) distribution

The Gaussian (normal) pdf for a continuous r.v. x is defined by:

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

$$E[x] = \mu$$

$$V[x] = \sigma^2$$

N.B. often μ, μ^2 denote mean, variance of any r.v., not only Gaussian.
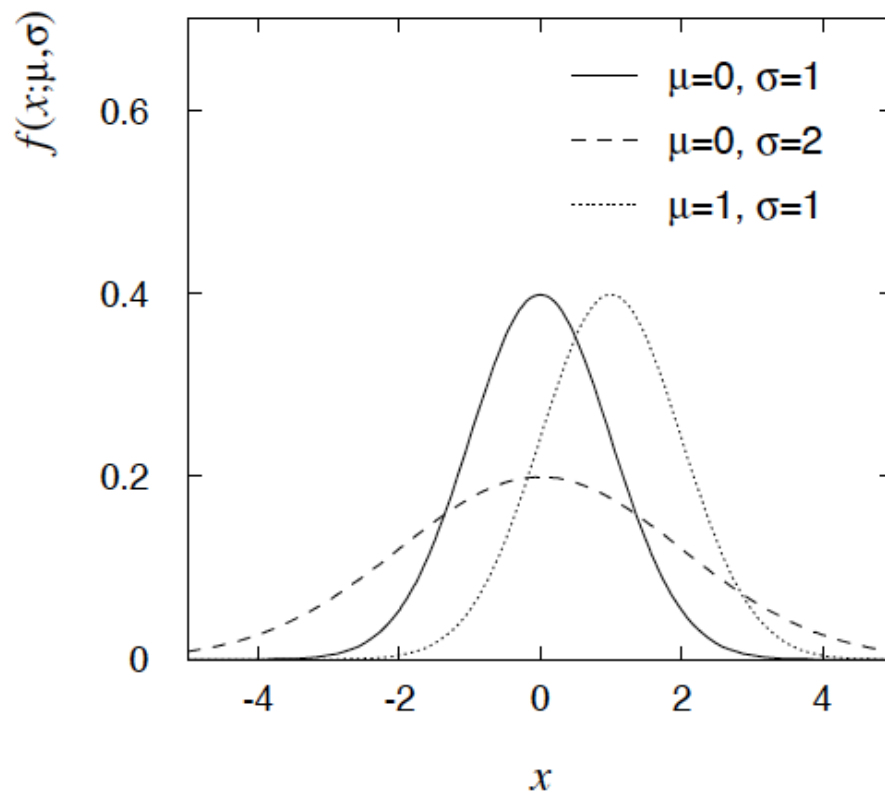
Multivariate Gaussian distribution

Multivariate Gaussian pdf for the vector $\vec{x} = (x_1, \ldots, x_n)$:

$$f(\vec{x}; \vec{\mu}, V) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T V^{-1} (\vec{x} - \vec{\mu})\right]$$

$\vec{x}, \ \vec{\mu}$ are column vectors, $\vec{x}^T, \ \vec{\mu}^T$ are transpose (row) vectors,

$$E[x_i] = \mu_i, , \qquad \text{cov}[x_i, x_j] = V_{ij} .$$

Marginal pdf of each x_i is Gaussian with mean μ_i, standard deviation $\sigma_i = \sqrt{V_{ii}}$.

Two-dimensional Gaussian distribution

$$f(x_1, x_2, ; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

where $\rho = \mathrm{cov}[x_1, x_2]/(\sigma_1\sigma_2)$
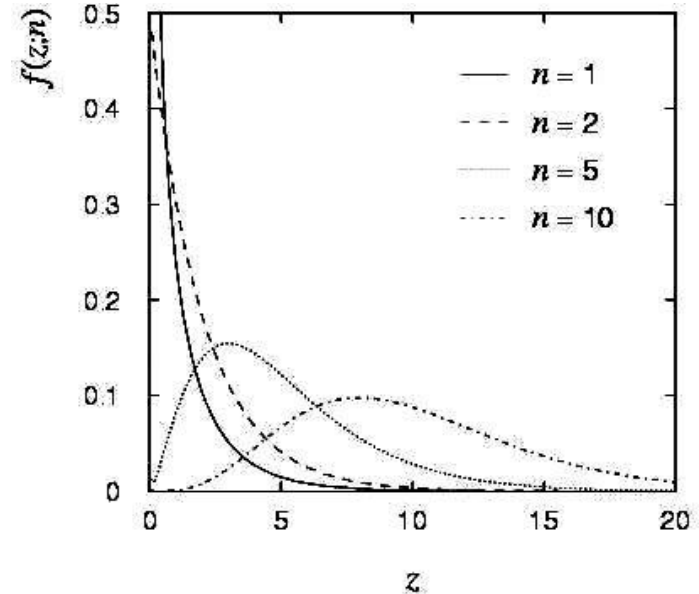is the correlation coefficient.
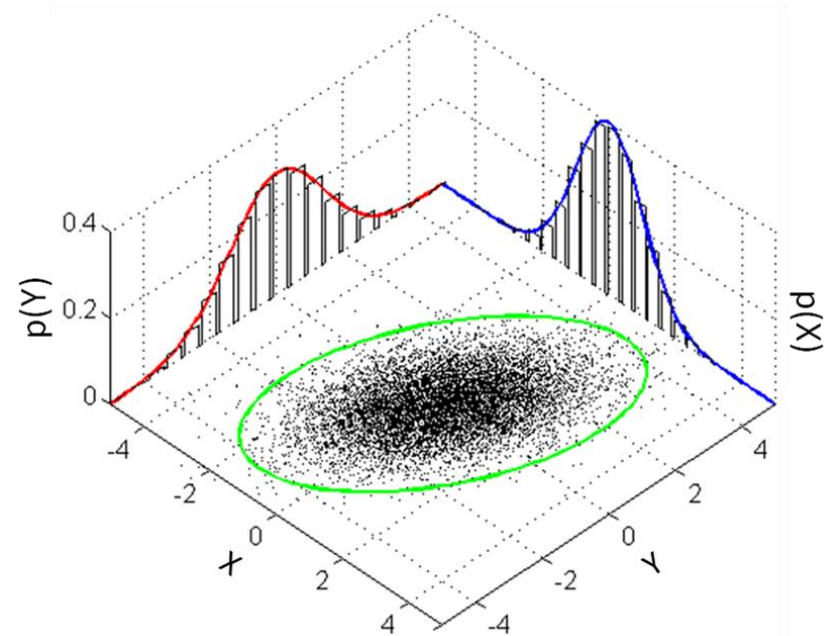
Chi-square (χ^2) distribution

The chi-square pdf for the continuous r.v. z ($z \geq 0$) is defined by

$$f(z; n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$$

n = 1, 2, ... = number of 'degrees of freedom' (dof)

$$E[z] = n, \quad V[z] = 2n.$$

For independent Gaussian x_i, $i = 1, \ldots, n$, means μ_i, variances σ_i^2,

$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \quad \text{follows } \chi^2 \text{ pdf with } n \text{ dof.}$$

Example: goodness-of-fit test variable especially in conjunction with method of least squares.

Cauchy (Breit-Wigner) distribution

The Breit-Wigner pdf for the continuous r.v. x is defined by

$$f(x; \Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}$$

($\Gamma = 2$, $x_0 = 0$ is the Cauchy pdf.)

$E[x]$ not well defined, $\quad V[x] \to \infty$.

$x_0 = $ mode (most probable value)

$\Gamma = $ full width at half maximum
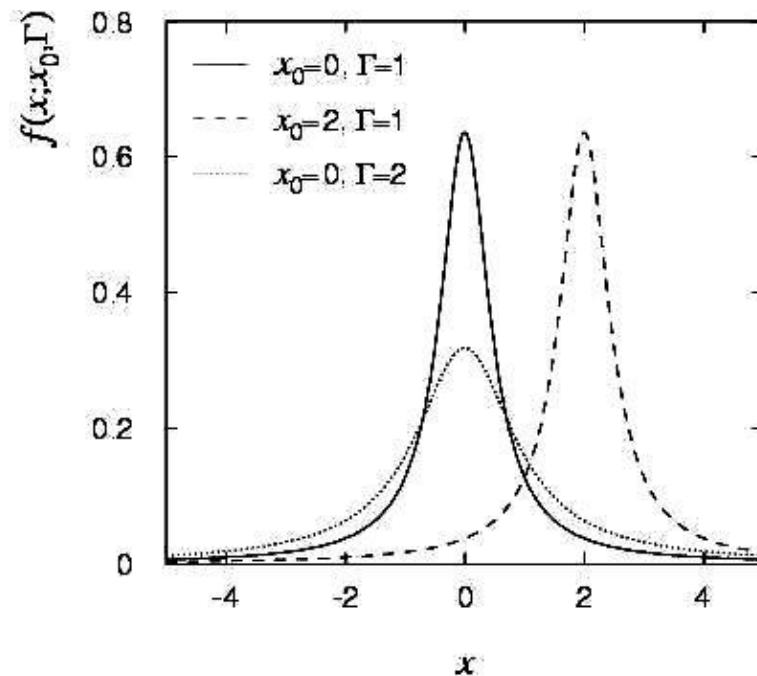
Example: mass of resonance particle, e.g. ρ, K^*, φ^0, …

$\Gamma = $ decay rate (inverse of mean lifetime)

Landau distribution

For a charged particle with $\beta = v/c$ traversing a layer of matter of thickness d, the energy loss Δ follows the Landau pdf:

$$f(\Delta; \beta) = \frac{1}{\xi}\phi(\lambda) \, ,$$

$$\phi(\lambda) = \frac{1}{\pi}\int_0^\infty \exp(-u\ln u - \lambda u)\sin \pi u \, du \, ,$$

$$\lambda = \frac{1}{\xi}\left[\Delta - \xi\left(\ln\frac{\xi}{\epsilon'} + 1 - \gamma_{\mathsf{E}}\right)\right] \, ,$$

$$\xi = \frac{2\pi N_{\mathsf{A}}e^4 z^2 \rho \sum Z}{m_{\mathsf{e}}c^2 \sum A}\frac{d}{\beta^2} \, , \qquad \epsilon' = \frac{I^2 \exp\beta^2}{2m_{\mathsf{e}}c^2\beta^2\gamma^2} \, .$$
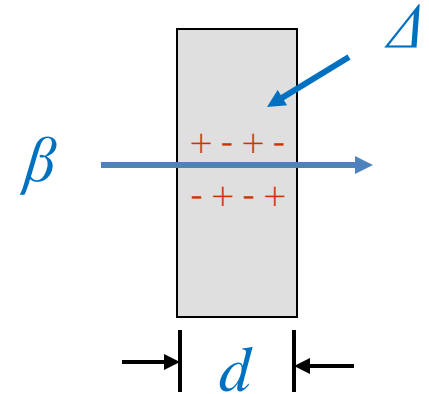
L. Landau, J. Phys. USSR **8** (1944) 201; see also
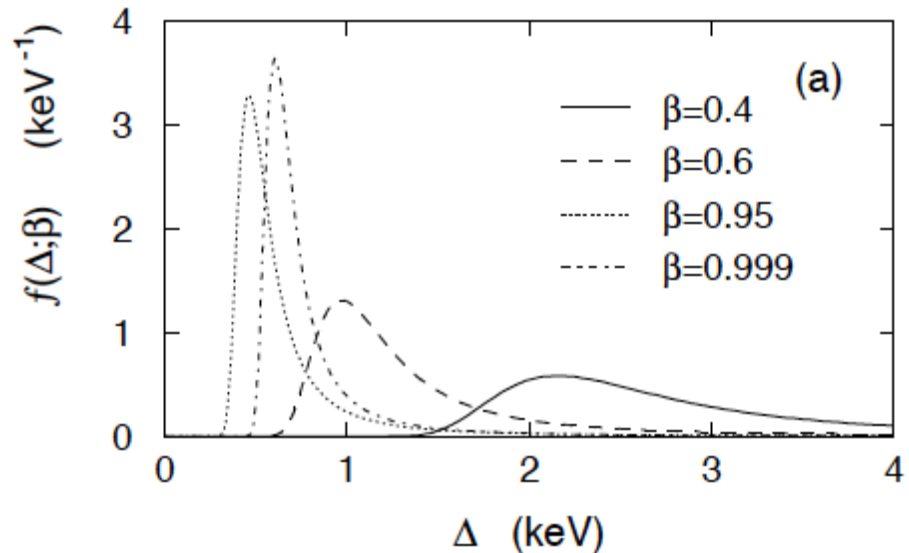W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. **30** (1980) 253.

Landau distribution (2)

Long 'Landau tail'

    $\rightarrow$ all moments ∞

Mode (most probable value) sensitive to β ,
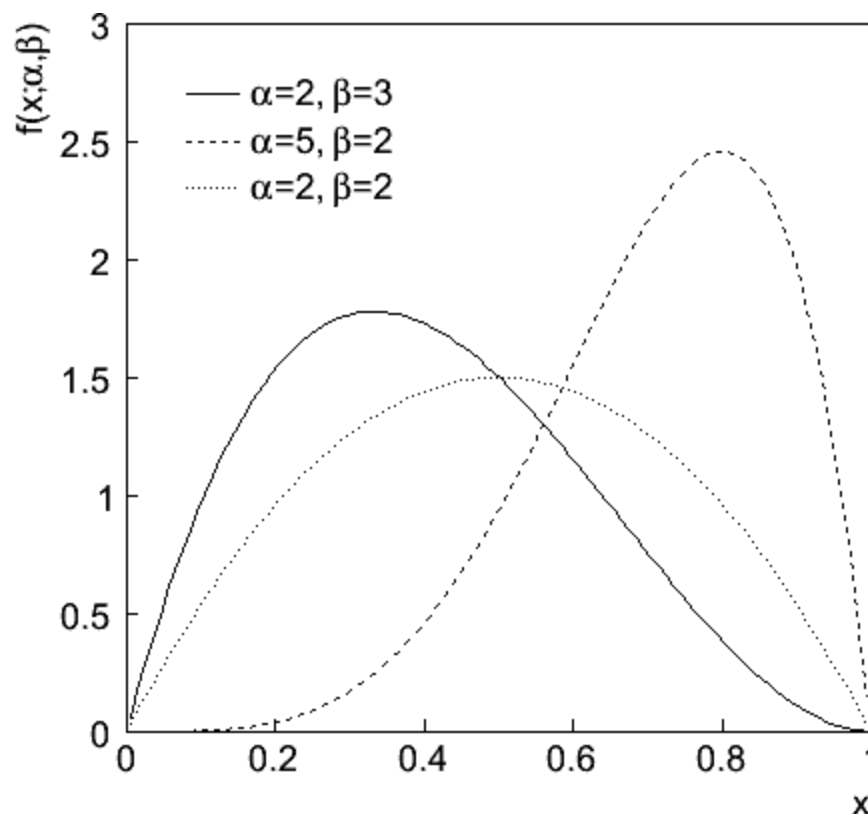
    $\rightarrow$ particle i.d.
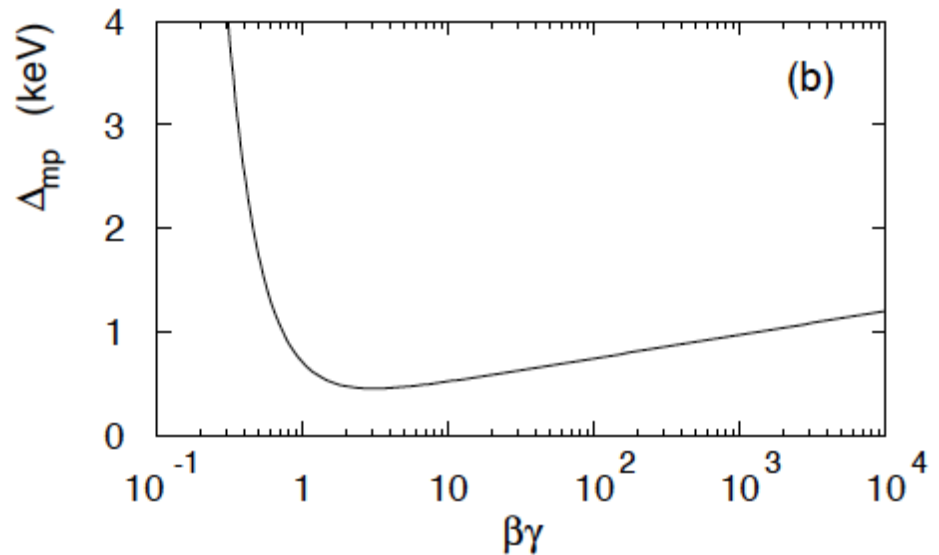
Beta distribution

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$E[x] = \frac{\alpha}{\alpha + \beta}$$

$$V[x] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Often used to represent pdf
of continuous r.v. nonzero only
between finite limits, e.g.,
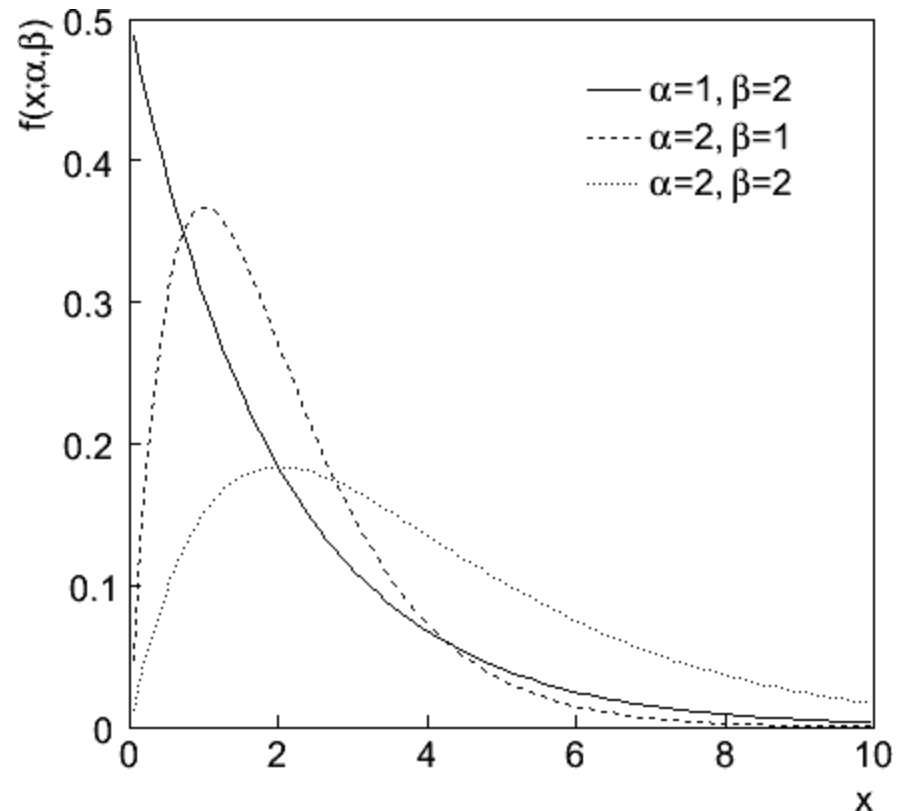$$y = a_0 + a_1 x, \quad a_0 \leq y \leq a_0 + a_1$$

Gamma distribution

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

$$E[x] = \alpha\beta$$

$$V[x] = \alpha\beta^2$$

Often used to represent pdf of continuous r.v. nonzero only in $[0,\infty]$.

Also e.g. sum of n exponential r.v.s or time until nth event in Poisson process ~ Gamma

Student's t distribution

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

$$E[x] = 0 \quad (\nu > 1)$$

$$V[x] = \frac{\nu}{\nu - 2} \quad (\nu > 2)$$

ν = number of degrees of freedom
    (not necessarily integer)
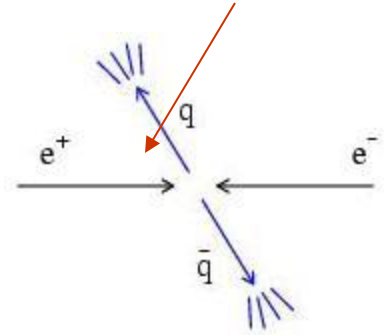
$\nu = 1$ gives Cauchy,

$\nu \to \infty$ gives Gaussian.
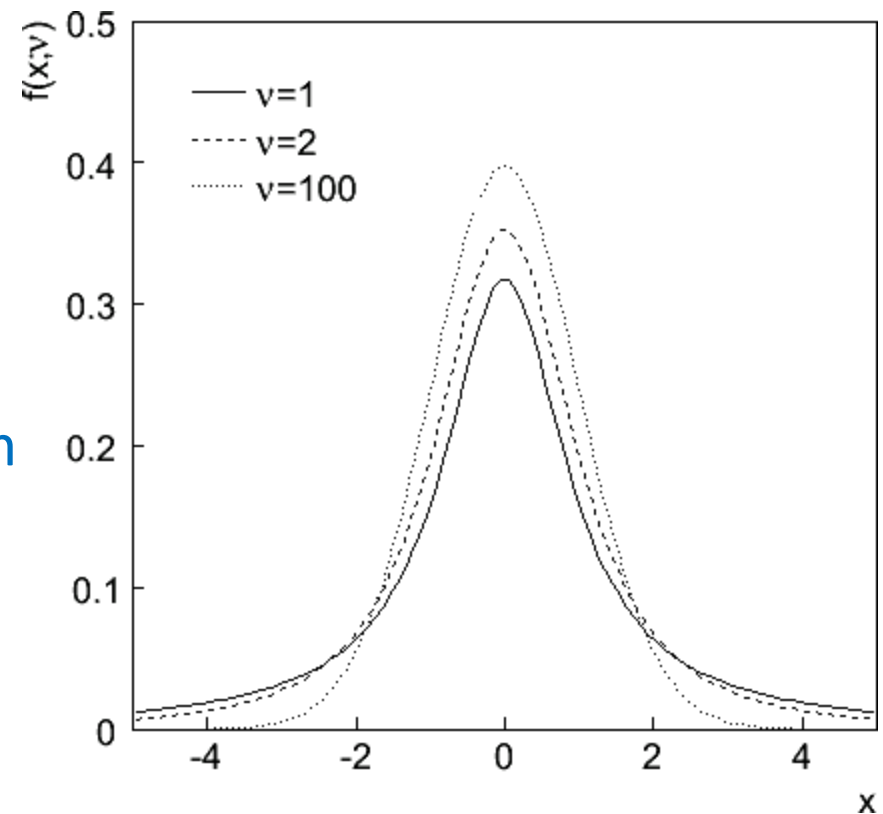
Example of ML with 2 parameters

Consider a scattering angle distribution with $x = \cos\theta$,

$$f(x; \alpha, \beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta/3}$$

or if $x_{\min} < x < x_{\max}$, need to normalize so that

$$\int_{x_{\min}}^{x_{\max}} f(x; \alpha, \beta)\,dx = 1 .$$

Example: $\alpha = 0.5$, $\beta = 0.5$, $x_{\min} = -0.95$, $x_{\max} = 0.95$, generate $n = 2000$ events with Monte Carlo.

$$\ln L(\alpha, \beta) = \sum_{i=1}^{n} \ln f(x_i; \alpha, \beta) \quad \longleftarrow \quad$$
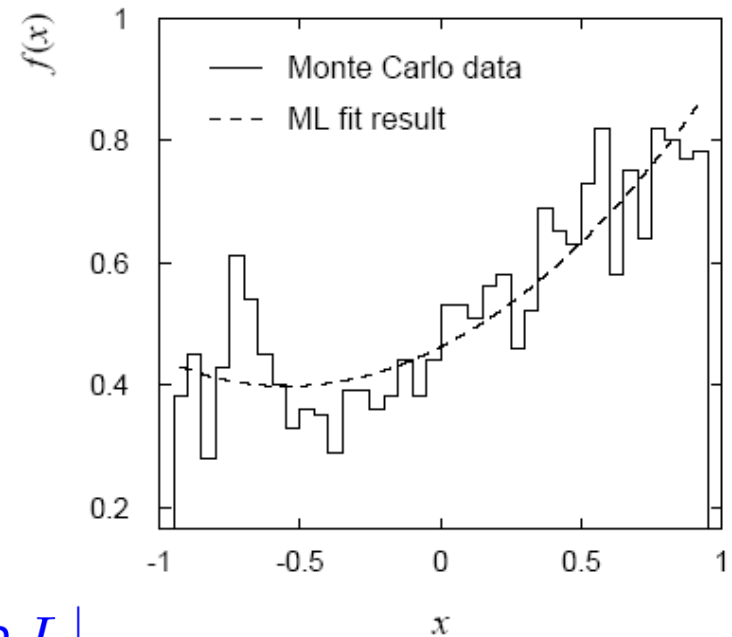
need to find maximum numerically

Example of ML with 2 parameters: fit result

Finding maximum of $\ln L(\alpha, \beta)$ numerically gives

$$\hat{\alpha} \ = \ 0.508$$
$$\hat{\beta} \ = \ 0.47$$

N.B. No binning of data for fit, but can compare to histogram for goodness-of-fit (e.g. 'visual' or p).

(Co)variances from $\quad (\widehat{V^{-1}})_{ij} = -\dfrac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\bigg|_{\vec{\theta}=\hat{\vec{\theta}}}$

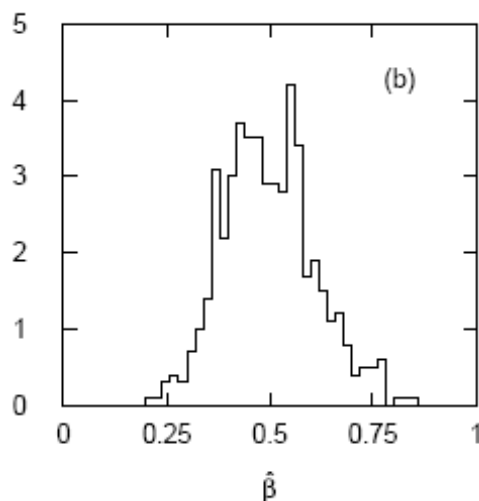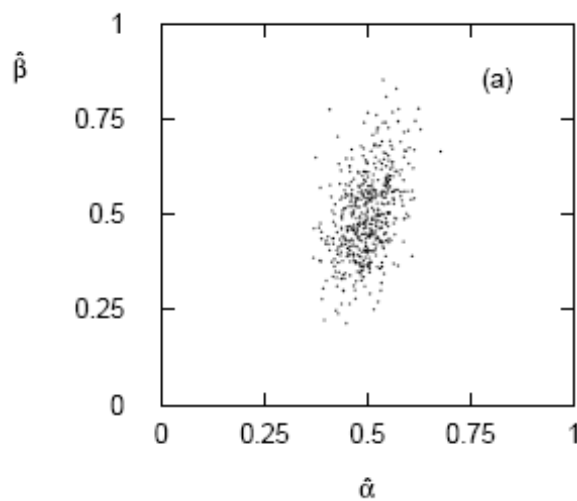$$\hat{\sigma}_{\hat{\alpha}} \ = \ 0.052 \qquad \mathrm{cov}[\hat{\alpha}, \hat{\beta}] \ = \ 0.0026$$
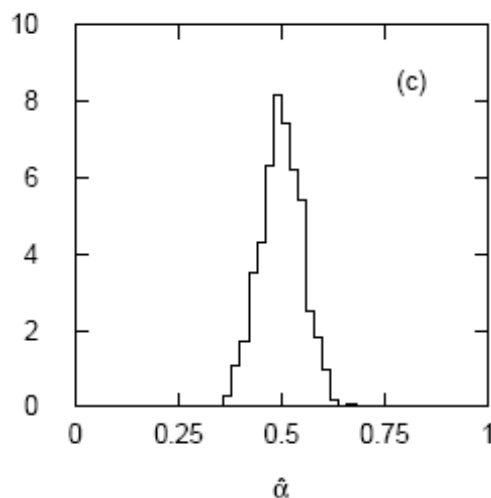$$\hat{\sigma}_{\hat{\beta}} \ = \ 0.11 \qquad\qquad r \ = \ 0.46 \ = \ \text{correlation coefficient}$$

Two-parameter fit: MC study

Repeat ML fit with 500 experiments, all with $n = 2000$ events:

$$\overline{\hat{\alpha}} = 0.499$$

$$s_{\hat{\alpha}} = 0.051$$

$$\overline{\hat{\beta}} = 0.498$$

$$s_{\hat{\beta}} = 0.111$$

$$\widehat{\text{cov}}[\hat{\alpha}, \hat{\beta}] = 0.0024$$

$$r = 0.42$$

Estimates average to ~true values; (Co)variances close to previous estimates; marginal pdfs approximately Gaussian.

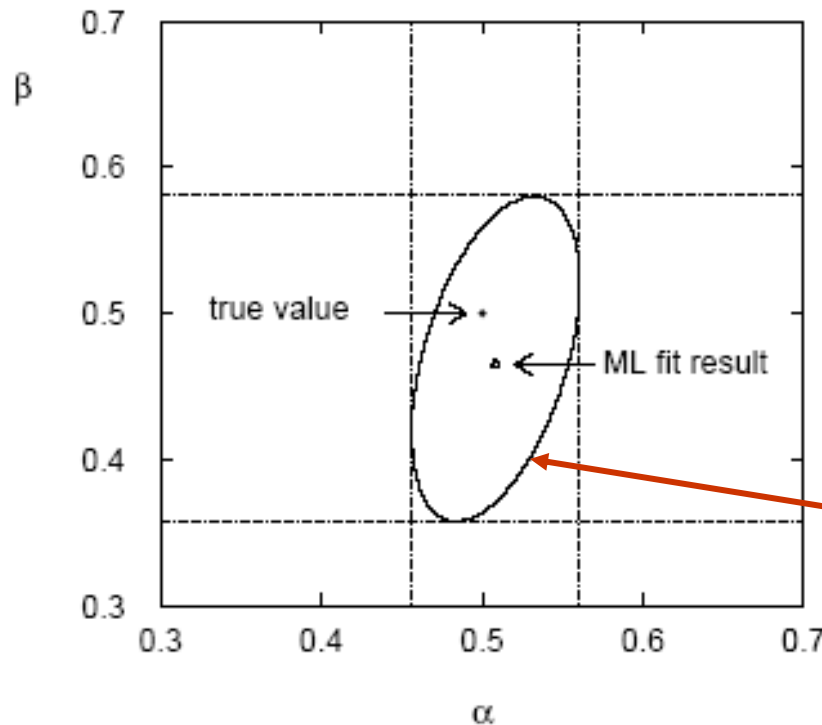# The $\ln L_{\mathrm{max}} - 1/2$ contour for two parameters

For large n, $\ln L$ takes on quadratic form near maximum:

$$\ln L(\alpha, \beta) \approx \ln L_{\mathrm{max}}$$

$$-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)^2 + \left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)^2 - 2\rho\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)\left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)\right]$$

The contour $\ln L(\alpha, \beta) = \ln L_{\mathrm{max}} - 1/2$ is an ellipse:

$$\frac{1}{(1-\rho^2)}\left[\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)^2 + \left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)^2 - 2\rho\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)\left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)\right] = 1$$

(Co)variances from ln L contour

The α, β plane for the first MC data set

$$\ln L(\alpha, \beta) = \ln L_{\max} - 1/2$$

$\rightarrow$ Tangent lines to contours give standard deviations.

$\rightarrow$ Angle of ellipse φ related to correlation: $\tan 2\phi = \dfrac{2\rho\sigma_{\widehat{\alpha}}\sigma_{\widehat{\beta}}}{\sigma_{\widehat{\alpha}}^2 - \sigma_{\widehat{\beta}}^2}$

Proof of Neyman-Pearson Lemma

Consider a critical region W and suppose the LR satisfies the criterion of the Neyman-Pearson lemma:

$$P(x|H_1)/P(x|H_0) \geq c_\alpha \text{ for all } x \text{ in } W,$$
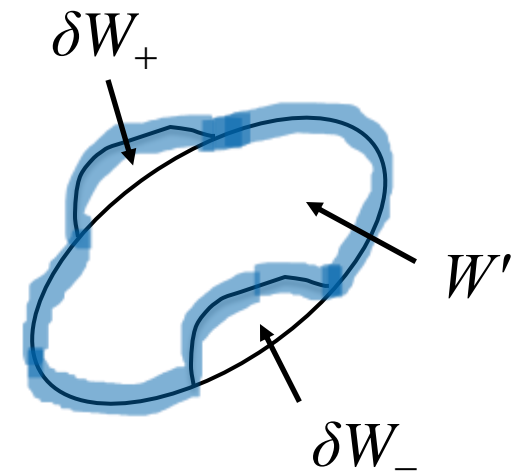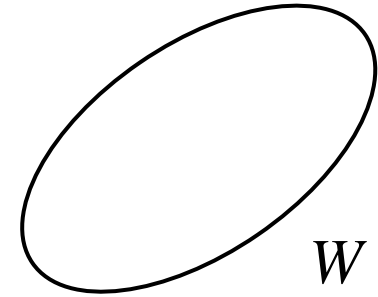$$P(x|H_1)/P(x|H_0) \leq c_\alpha \text{ for all } x \text{ not in } W.$$

Try to change this into a different critical region W' retaining the same size α, i.e.,

$$P(\mathbf{x} \in W'|H_0) = P(\mathbf{x} \in W|H_0) = \alpha$$

To do so add a part δW_+, but to keep the size α, we need to remove a part δW_-, i.e.,
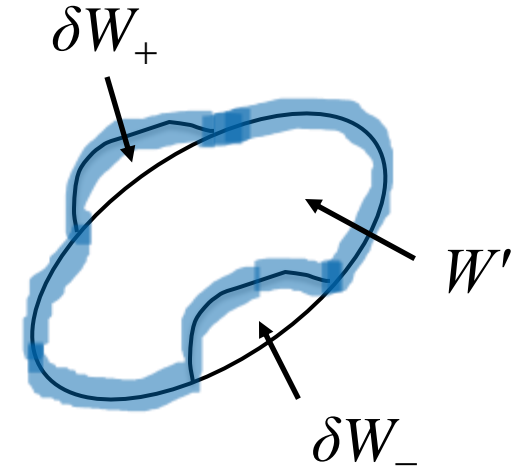
$$W \to W' = W + \delta W_+ - \delta W_-$$

$$P(\mathbf{x} \in \delta W_+|H_0) = P(\mathbf{x} \in \delta W_-|H_0)$$

Proof of Neyman-Pearson Lemma (2)

But we are supposing the LR is higher for all x in δW_- removed than for the x in δW_+ added, and therefore

$$P(\mathbf{x} \in \delta W_+ | H_1) \leq P(\mathbf{x} \in \delta W_+ | H_0) c_\alpha$$

$$P(\mathbf{x} \in \delta W_- | H_1) \geq P(\mathbf{x} \in \delta W_- | H_0) c_\alpha$$

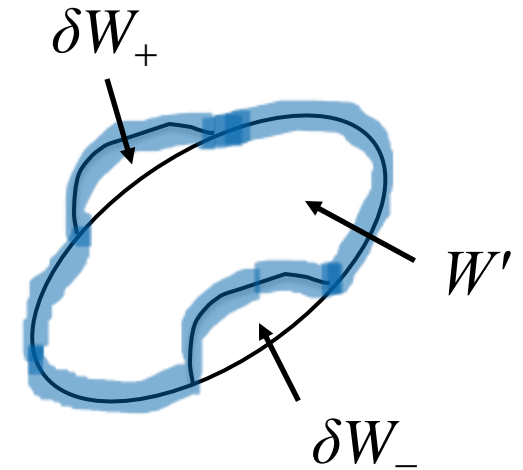The right-hand sides are equal and therefore

$$P(\mathbf{x} \in \delta W_+ | H_1) \leq P(\mathbf{x} \in \delta W_- | H_1)$$

Proof of Neyman-Pearson Lemma (3)

We have

$$W \cup W' = W \cup \delta W_+ = W' \cup \delta W_-$$

Note W and δW_+ are disjoint, and W' and δW_- are disjoint, so by Kolmogorov's 3rd axiom,

$$P(\mathbf{x} \in W') + P(\mathbf{x} \in \delta W_-) = P(\mathbf{x} \in W) + P(\mathbf{x} \in \delta W_+)$$

Therefore

$$P(\mathbf{x} \in W'|H_1) = P(\mathbf{x} \in W|H_1) + P(\mathbf{x} \in \delta W_+|H_1) - P(\mathbf{x} \in \delta W_-|H_1)$$

≤ 0

Proof of Neyman-Pearson Lemma (4)

And therefore

$$P(\mathbf{x} \in W'|H_1) \leq P(\mathbf{x} \in W|H_1)$$

i.e. the deformed critical region W' cannot have higher power than the original one that satisfied the LR criterion of the Neyman-Pearson lemma.