Statistics

for data analysis

with examples from High Energy Physics

Peter H. Hansen

University of Copenhagen

Content

Basic concepts

Error propagation

Some basic pdf's

Monte Carlo generation

Statistical tests

Parameter estimation

Confidence levels

Preface

- The homepage for these lectures is `http://www.nbi.dk/~phansen/fys716`
- For proofs, examples and more subjects, see Glen Cowan: **Statistical data analysis**, Oxford Science Publications
- or – for an overview of everything, see `http://pdg.lbl.gov/`
- Numerical fortran programs are from `http:://cernlib.web.cern.ch/cernlib/` and C++ programs are from **CLHEP**: `http:://wwwasd.web.cern.ch/CLHEP/`.
- Other physics software links are found in `http:://www.hep.net/resources/software.html`.

What is probability?

Kolmogorov formulated the probability axioms in 1933. Consider a set, S, called "sample space", e.g. of all possible outcomes of a measurement. To each subset A of S, assign a real number $P(A)$, defined by the axioms:

- For all A, $P(A) \geq 0$.
- For disjoint subsets A and B: $P(A \cup B) = P(A) + P(B)$.
- For any S, $P(S) = 1$.

Frequentistic probability

- In the frequentistic view, the probability of some collection of events is

  the relative frequency of its occurence

  in a very large number of experiments, repeated

  under exactly the same circumstances

  .
- This is not always feasible, but in High Energy Physics you have a chance.
- However, you can *only* talk about the likelihood, $P(data|theory)$, that a given hypothesis will yield some measured data.

  You can *not* talk about $P(theory|data)$ – since there is no way to determine this by repeating. The theory is either never or always true!

Bayesian probability

- In the Bayseian view, probability is the degree of belief that a hypothesis will hold water in the future. We imagine to have an exclusive and exhaustive set of alternative hypothesis H_i available for explaining some measured data. In this picture, the probabilty for a given hypothesis can be assigned via Bayes' Theorem:

$$P(H_i|data) = \frac{P(data|H_i) \times P_0(H_i)}{\sum_j P(data|H_j) \times P_0(H_j)}$$

where $P_0(H_i)$ is the prior probability for hypothesis i, as obtained from the – admittedly subjective – knowledge prior to the measurement.

Probability density function

- The probability density function (p.d.f.) is defined by requiring the probability of finding a continous random variable x in the interval [x,x+dx] to be $f(x,\theta)dx$, where θ are possible fixed parameters.
- The p.d.f. must be normalized to unit area:

$$\int_{-\infty}^{\infty} f(x,\theta)dx = 1$$

.

- The cumulative distribution is $F(x) = \int_{-\infty}^{x} f(x)dx$ (or $F(x) = \sum_{x_i \leq x} P(x_i)$ for discrete random variables).

Joint p.d.f.

- For two random variables, the joint p.d.f., $f(x, y)$, is normalized to unit volume on the xy plane.
- Integrating over y, we obtain the marginal p.d.f., $f_x(x)$.
- The two variables are independent if
  $f(x, y) = f_x(x) f_y(y)$,
  so the knowledge of one variable does not change the p.d.f. of the other.

p.d.f. of functions of random variables

- A function $a(x)$ of a random variable with p.d.f. $f(x)$ has the p.d.f.:

$$g(a) = f(x(a)) \left| \frac{dx}{da} \right|$$

- Exercise: $F(x)$ is always uniformly distributed. Show it.
- If $\bar{a}$ are several functions of several random variables:

$$g(\bar{a}) = f(\bar{x}) \ |J| \, ,$$

where $|J|$ is the Jacobian determinant of $\frac{\delta x_i}{\delta a_j}$.

Moments

- For any function, $g(x)$, of a random variable with p.d.f. $f(x)$, the expectation value is:

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

- A p.d.f. may be characterized by its central moments:

$$m_n = \int_{-\infty}^{\infty} x^n f(x)dx$$

Mean and variance

- Some moments have English names, such as the mean:

$$\mu = \int_{-\infty}^{\infty} x f(x) dx = E[x]$$

The variance, σ^2, is defined as:

$$\sigma^2 = E[(x-\mu)^2] = E[x^2] - \mu^2$$

- The mean, μ, is often approximated by the sample mean: $\frac{1}{N}\sum x_i$, and the variance by the sample variance: $\frac{1}{N}\sum(x_i - \mu)^2$.

Correlations

The covariance of two random variables x and y with joint p.d.f. $f(x,y)$ is defined:

$$\begin{aligned} V_{xy} &= E[(x-\mu_x)(y-\mu_y)] \\ &= E[xy]-\mu_x\mu_y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy-\mu_x\mu_y \end{aligned}$$

More generally, for any two functions of n random variables $\bar{x}$, the covariance matrix is given by:

$$\begin{aligned} V_{ab} &= E[(a-\mu_a)(b-\mu_b)] \\ &= E[ab]-\mu_a\mu_b \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} a(\bar{x})b(\bar{x})f(\bar{x})d\bar{x}-\mu_a\mu_b \end{aligned}$$

Correlation coefficient

By construction V_{ab} is symmetric with positive diagonal: $V_{aa} = \sigma_{aa}^2$.
The degree of correlation is given by the correlation coefficient.

$$\rho_{ab} = \frac{V_{ab}}{\sigma_a \sigma_b}$$

taking values in the range $-1 \leq \rho_{ab} \leq 1$.
Notice that independent variables have $V_{ij} = 0,\ i \neq j$, while the converse is not neccessarily true.

Error propagation

Suppose we know the mean values μ_i and covariance matrix V_{ij} of some random variables x_i, but not the detailed p.d.f.

How do we then determine the mean and variance of some function $y(\bar{x})$? (the bar means vector!)

If a first-order Taylor expansion around μ_i is OK:

$$y(\bar{x}) \approx y(\bar{\mu}) + \sum_{i=1}^{n} \frac{\delta y}{\delta x_i}(x_i - \mu_i)$$

From this it easily follows (since $E[\bar{x} - \bar{\mu}] = 0$) that

$$\begin{aligned} E[y(\bar{x})] &\approx y(\bar{\mu}) \\ E[y^2(\bar{x})] &\approx y^2(\bar{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\delta y}{\delta x_i}\frac{\delta y}{\delta x_j}\right]_{\bar{x}=\bar{\mu}} V_{ij} \end{aligned}$$

Error propagation examples

- Example: $y = x1 + x2$.

$$\sigma_y^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 + 2V_{12}$$

- Example: $y = x_1 x_2$.

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_{x_1}^2}{x_1^2} + \frac{\sigma_{x_2}^2}{x_2^2} + 2\frac{V_{12}}{x_1 x_2}$$

General error propagation

Simlarly for m functions $y_1(\bar{x}), \cdots, y_m(\bar{x})$, we get the covariance matrix:

$$
\begin{aligned}
U_{kl} &= \mathrm{cov}[y_k, y_l] \approx \sum_{i,j=1}^{n} \left[\frac{\delta y_k}{\delta x_i} \frac{\delta y_l}{\delta x_j} \right]_{\bar{x}=\bar{\mu}} V_{ij} \\
U &= AVA^T \\
A_{kj} &= \left[\frac{\delta y_k}{\delta x_i} \right]_{\bar{x}=\bar{\mu}}
\end{aligned}
$$

where A^T is the transposed of A.
The above equation is the basis of error propagation, where errors on some random variables are propagated to functions of the variables. It is only exact for linear functions.

The uniform distribution

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \mu &= \frac{1}{2} \\ \sigma &= \frac{1}{\sqrt{12}} \end{aligned}$$

The uniform distribution - Examples

- Example: The energy of a photon from decay of a π^0 with energy E_π (easy to show).
- You could try out this simple fortran generator and compare with others such as RANMAR, RANLUX and RNDM

```
FUNCTION ZRAND()
PARAMETER(IA=205,IC=29573,IM=139968)
DATA LAST/4707/
LAST=MOD(IA*LAST+IC,IM)
IF(LAST.EQ.0) LAST=IC
ZRAND=FLOAT(LAST)/FLOAT(IM)
END
```

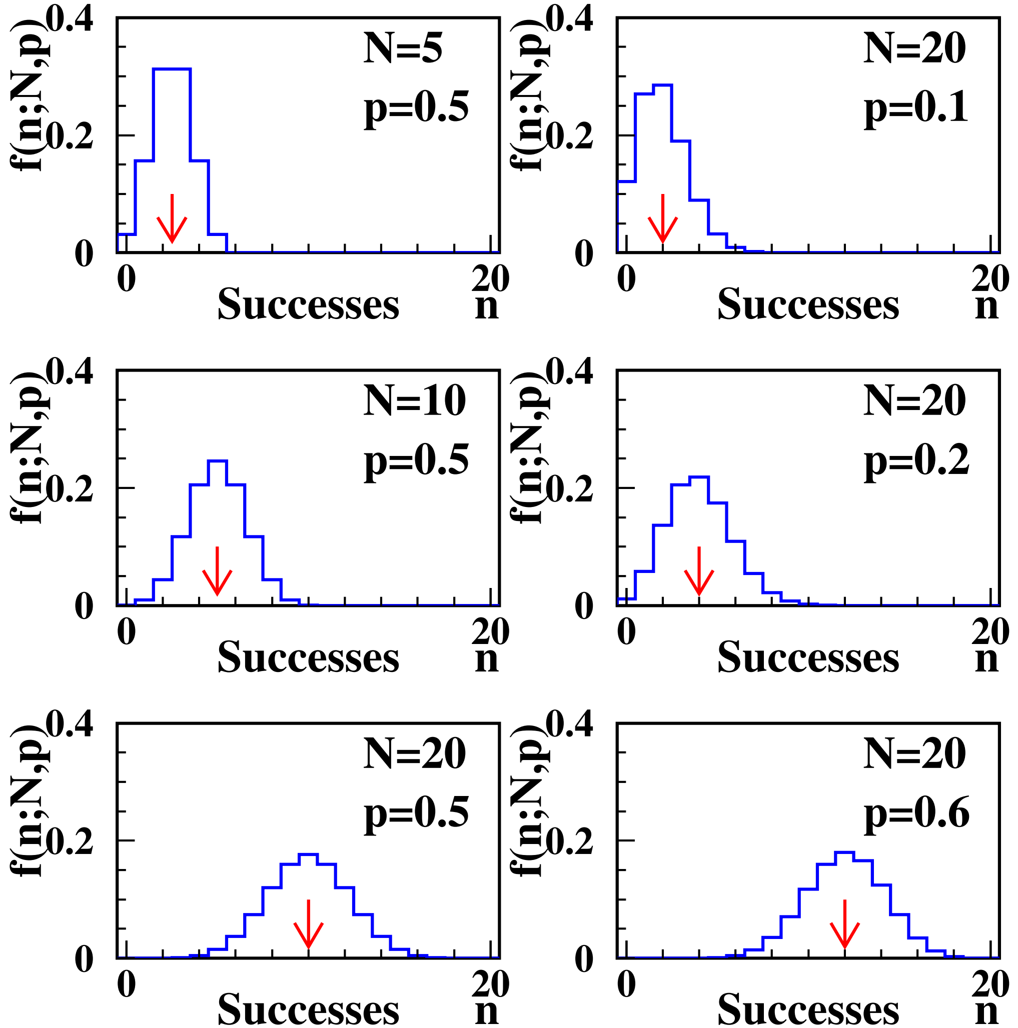
The Binomial distribution

Consider N independent trials with only two possible outcomes: success with probability p and failure with probability $1 - p$. The probability of n successes is:

$$
\begin{aligned}
f(n; N, p) &= \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \\
\mu &= E[n] = Np \\
\sigma &= \sqrt{Np(1-p)}
\end{aligned}
$$

An example is the occurrence of n triggers in N beam crossings.

The binomial distribution



The Poisson distribution

Consider some event happening with a fixed probability ν per time interval. Nothing else restricts the event rate. The probability of observing n events in such an interval is then:

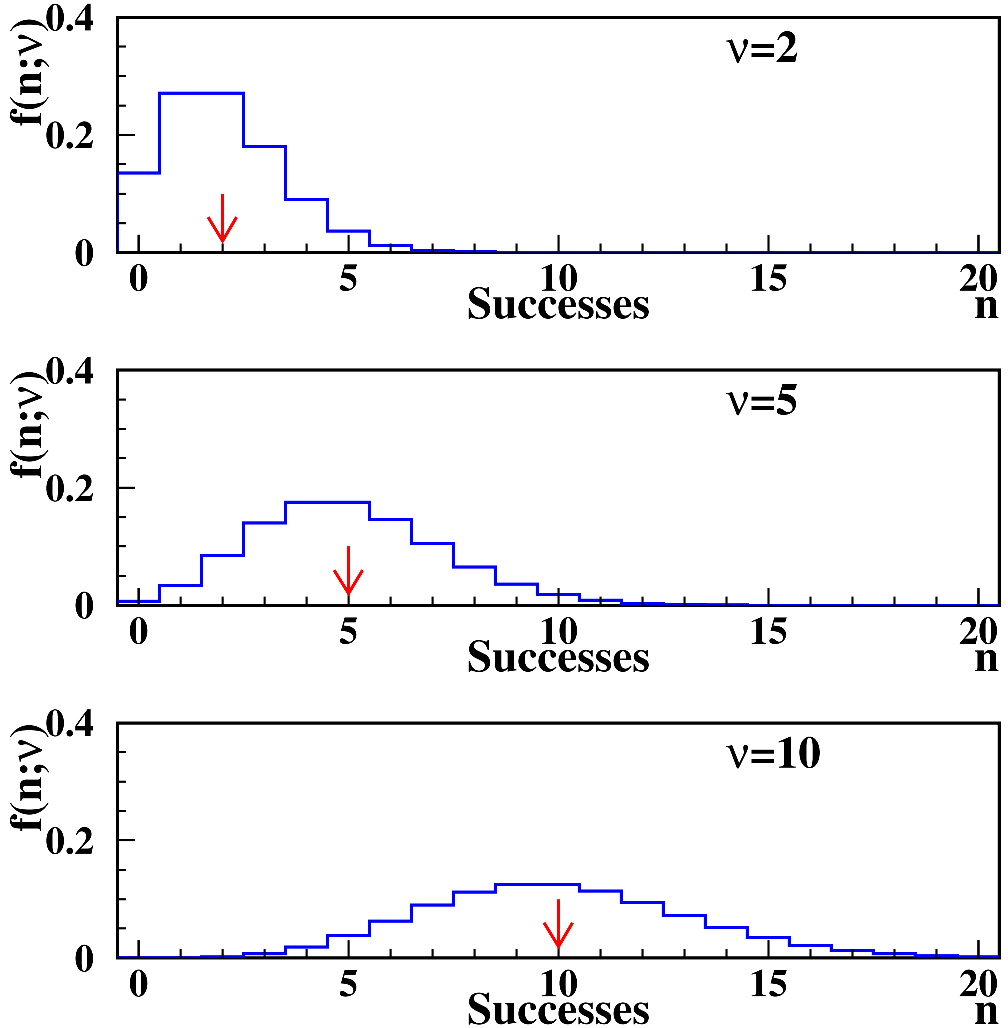
$$
\begin{aligned}
f(n;\nu) &= \frac{\nu^n}{n!} e^{-\nu} \\
\mu &= E[n] = \nu \\
\sigma &= \sqrt{\nu}
\end{aligned}
$$

The Poisson distribution is the binomial in the limit of large N and small p, keeping Np constant. It becomes Gaussian for large ν.
The distribution reflects " minimal information", given a certain mean value of integer counts.

The Poisson distribution - examples

- An example is the number of radioactive decays observed in a certain time and a certain amount of material.
- Another is the trigger-rate in a particle scattering experiment (large number of beam crossings, small cross-section).
- Another is the number of cars passing the institute per minute on working days from 2 to 3 PM.

The Poisson distribution



The exponential distribution

- This p.d.f. is defined on $0 \leq x < \infty$:

$$
\begin{aligned}
f(n;\xi) &= \frac{1}{\xi} e^{-x/\xi} \\
\mu &= E[x] = \xi \\
\sigma &= \xi
\end{aligned}
$$

- An example is the decay-time of an unstable particle.
Excercise: Show this. Show also that the time between two subsequent Poisson-distributed events is exponentially distributed.

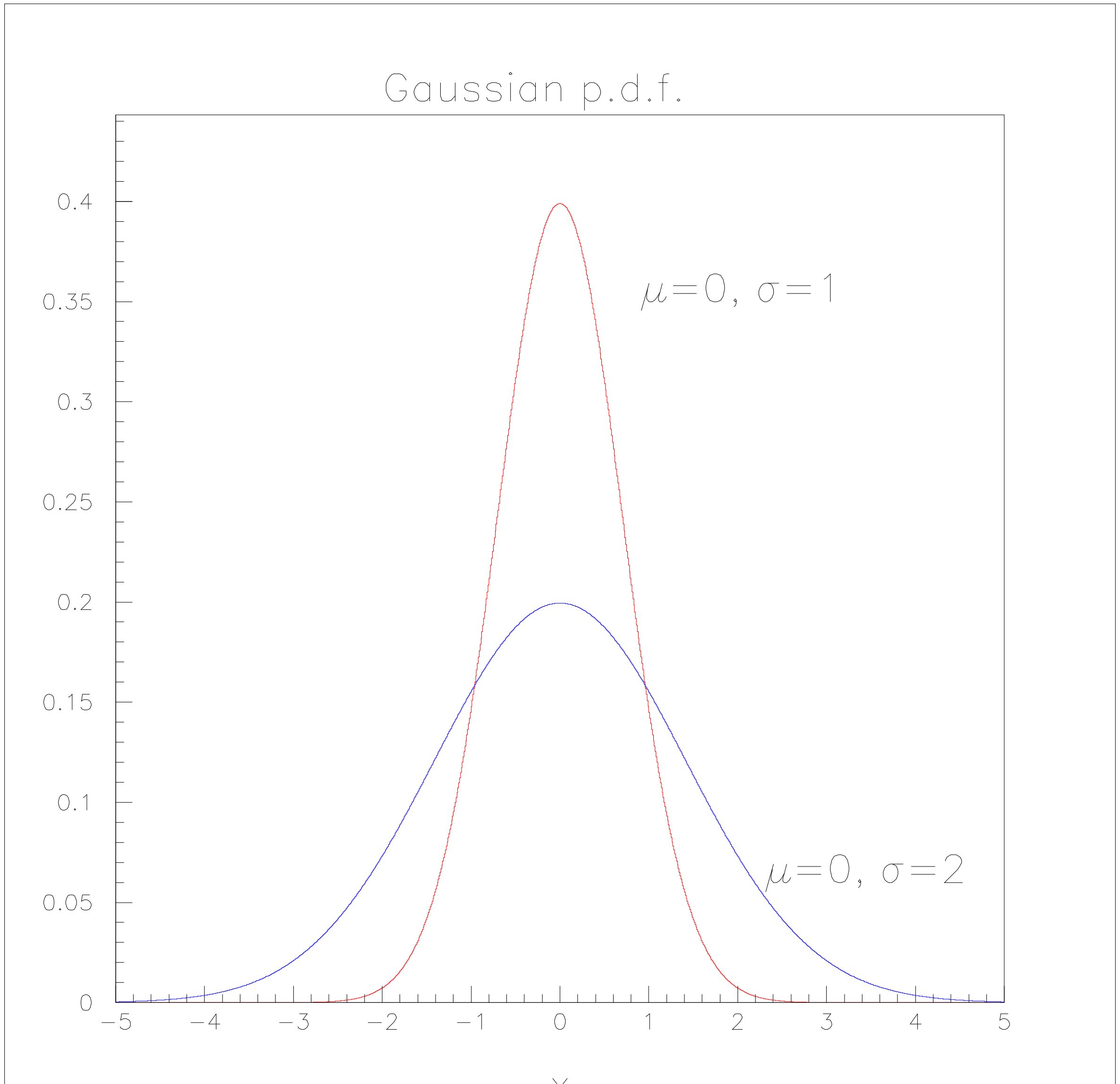
The Gaussian distribution

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

Its importance stems from the central limit theorem: The sum of n random variables x_i with any p.d.f. becomes Gaussian in the large n limit with $\mu = \sum \mu_i$ and $\sigma^2 = \sum \sigma_i^2$. Therefore measurement errors are treated as Gaussian random variables, holding them to be a large sum of small contributions.
The N-dimensional generalization of the Gaussian is:

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

The Gaussian distribution



The χ^2 distribution

$$
\begin{aligned}
f(x;\mu,\sigma) &= \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}\,,\ n = 1,2,\cdots \\
\mu &= E[z] = n \\
\sigma &= \sqrt{2n}
\end{aligned}
$$

where n is called the number of degrees of freedom (d.o.f.).

The χ^2 distribution

- Consider n independent Gaussian random measurements, x_i, with known means and variances. Then the variable

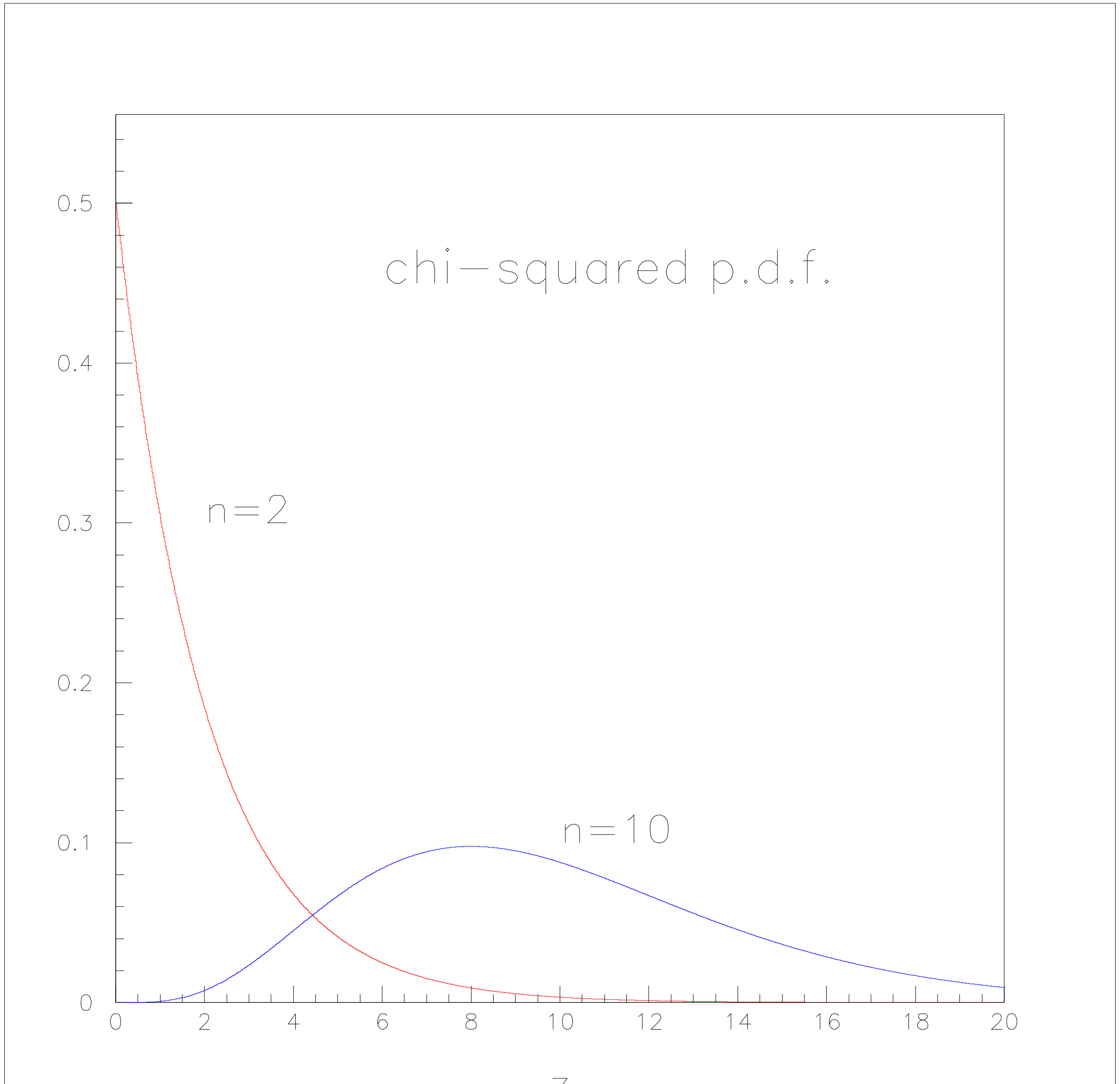
$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

  is χ^2-distributed for n d.o.f.

- More generally, if the x_i's are not independent, the χ^2 random variable is:

$$z = (\bar{x} - \bar{\mu})^T V^{-1} (\bar{x} - \bar{\mu})$$

The χ^2 distribution



The Breit-Wigner and Landau distributions

The cross-section for the production and decay of a resonance particle follows a Breit-Wigner distribution:

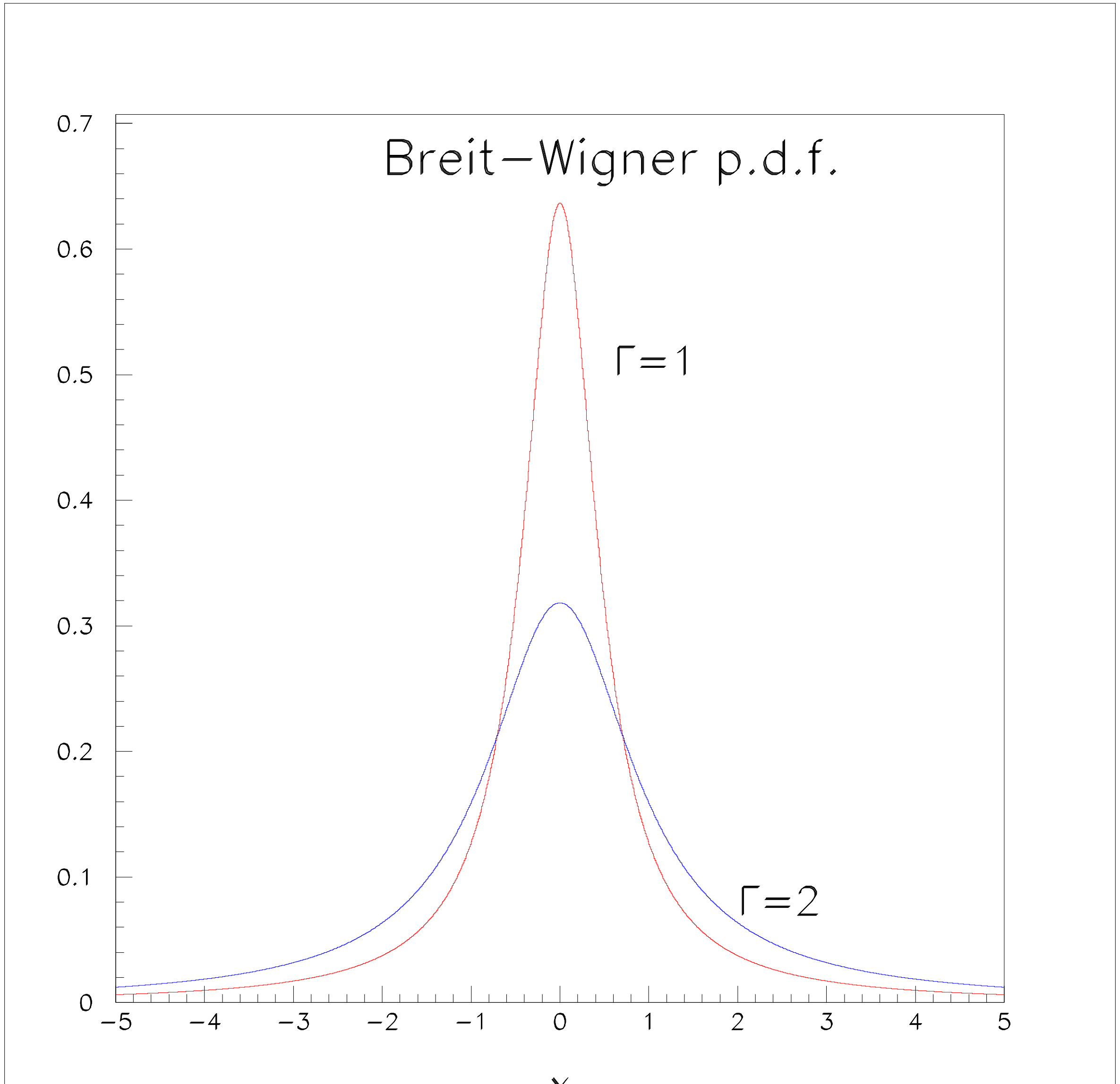
$$f(x; \Lambda, x_0) \quad = \quad \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}$$

with x_0 being the peak-position of the resonance and Γ the full-width at half maximum. (Γ is related to the resonance lifetime via Heisenbergs uncertainty relation $\Gamma = 1/\tau$ in natural units) .

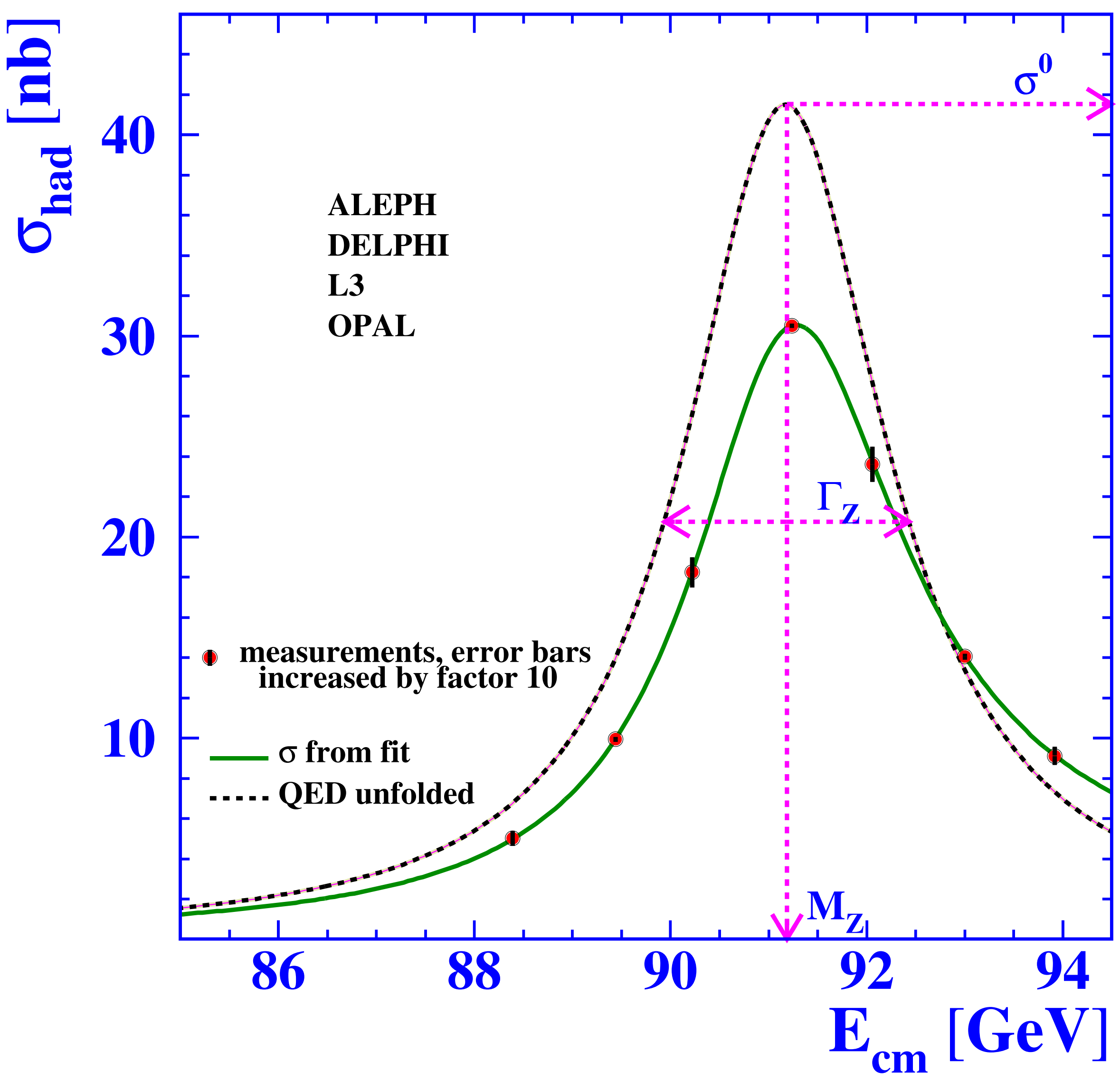
The tails are so large that the moments are divergent.

The same problem is present in the Landau distribution, because of hard scattered electrons (see Leo 2.6.3). Here, a truncated mean, with outliers excluded, is often used to determine the mean energy loss.

The Breit-Wigner distribution

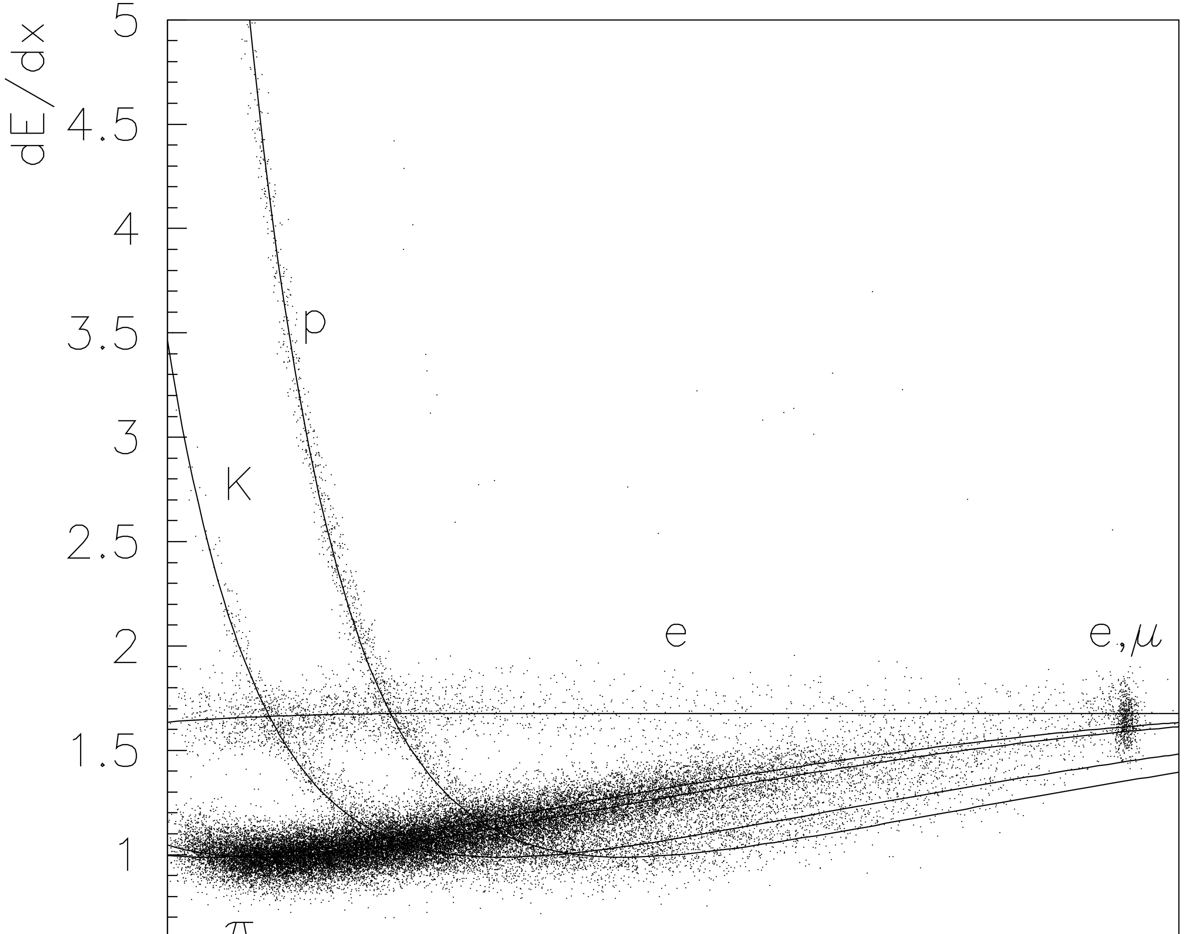


The Z lineshape



Landau smearing of the Bethe-Bloch curve

Example: dE/dx in the ALEPH TPC. Each point is a truncated mean over at least 150 measurements with the lowest 8 and the highest 40 excluded.



Monte Carlo methods

- For calculating a predicted marginal p.d.f. of some measured variable, the only practical means is often Monte Carlo integration, i.e. random sampling of simulated events.
- To this end we need to transform a sequence of uniform random numbers $r_1, r_2, \cdots$ to a sequence of x_i with some desired p.d.f. $f(x)$. Since $F(x)$ is uniformly distributed, we can do the job choosing $x_i = F^{-1}(r_i)$.
- In practice $F^{-1}(r)$ is known analytically for only very few functions, such as the exponential p.d.f:

$$\begin{aligned} F(x(r)) &= \int_0^{x(r)} \frac{1}{\xi} e^{-x/\xi} dx \\ x(r) &= -\xi \log(1-r) \end{aligned}$$

Acceptance-rejection method

Otherwise the acceptance-rejection method can be used. Assume the desired p.d.f. $f(x)$ can be boxed in the interval $x_1 < x < x_2$ and $0 < f(x) < y_{max}$. The following procedure:

1. Choose a random x between x_1 and x_2.
2. Choose a random y between 0 and y_{max}.
3. If $y < f(x)$, x is accepted.

gives a sequence of x distributed according to $f(x)$, since the probability for accepting x is always proportional to $f(x)$. For a sharply peaked $f(x)$ it is cheaper (in CPU-cycles) to envelope $f(x)$ under some simple function $g(x)$ tracing the peaked shape better than a flat line. In a first step, x is then chosen according to the (normalized) $g(x)$. Then use the procedure above with y_{max} replaced by $g(x)$.

Components of High Energy Physics MC

In a high-energy physics experiment, the simulation has very many steps, each carried out as above. To briefly mention a few of the steps:

- Smear the colliding beam momenta and the interaction point (according to Gaussian p.d.f.'s).
- Choose incoming partons according to structure functions
- Choose outgoing "partons" according to a matrix element
- Generate parton showers (e.g. with **PYTHIA**)
- Combine partons into hadrons (e.g. with **PYTHIA**)

Components of a High Energy Physics MC - cont'd

- Track particles (hadrons and leptons) through the apparatus in small steps. (with **GEANT**)
- At each little step choose whether they decay, scatter, react strongly, ionize, radiate or simply stop.
- In any case, track all the particles created in this step.
- Generate read-out signals according to some p.d.f.
- Superimpose noise according to some p.d.f.

Testing two alternative hypotheses

- Consider a set of data $x_1, \cdots, x_n$ relevant for the validity of two alternative hypotheses H_0 and H_1. To decide among the two, we construct a test statistic, $t(\bar{x})$, with the likelihoods $g(t|H_0)$ and $g(t|H_1)$.
- We may decide to reject H_0 if $t > t_{cut}$. The probability of making a mistake (for H_0 true) is

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

- The fraction of rejected H_1's are similarly:

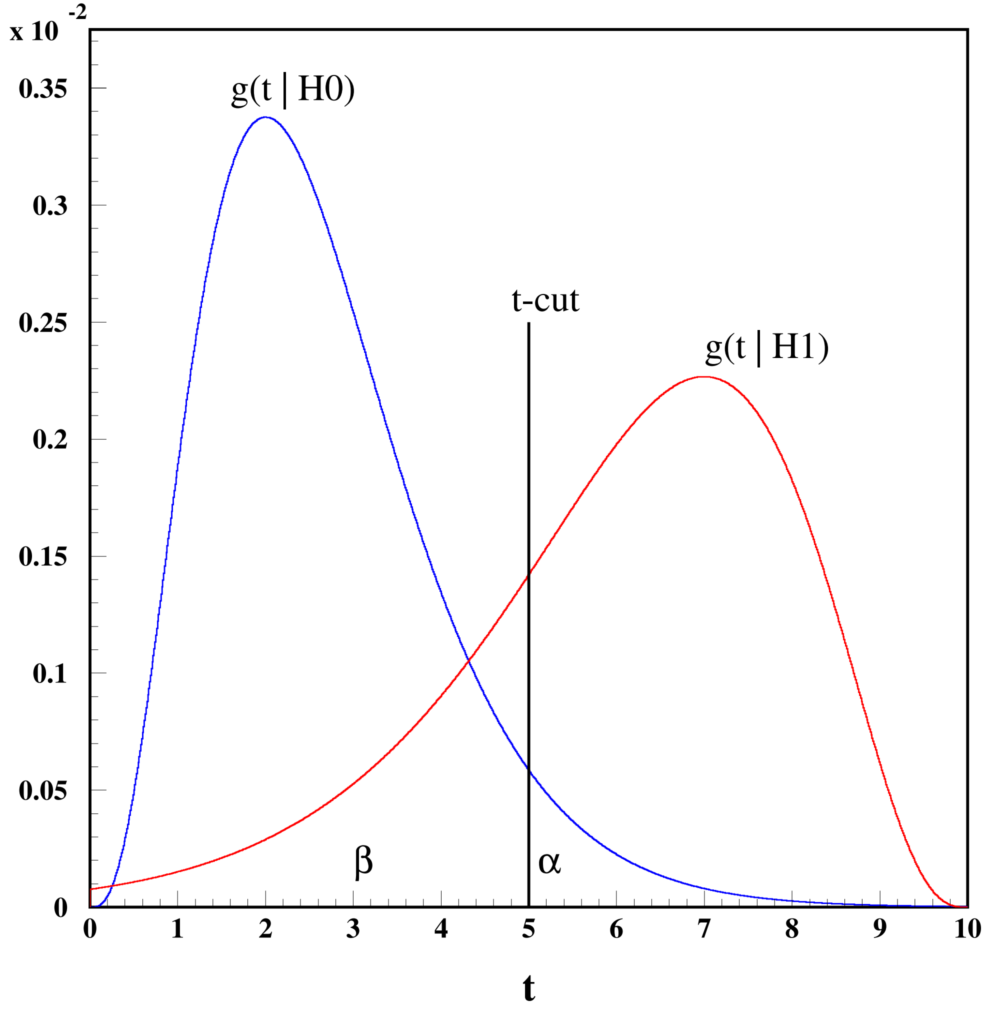
$$1 - \beta = \int_{t_{cut}}^{\infty} g(t|H_1)dt$$

The likelihood ratio

- The probability, α, for rejecting the "null hypothesis", H_0, is called the significance level of the test.
- Obviously, we want both α and β small in order to have good H_1 signal efficiency and small H_0 background above the cut, and vice versa below the cut. The Neyman-Pearson Lemma states that the optimal test statistics is the likelihood ratio:

$$t_{opt}(\bar{x}) = \frac{g(\bar{x}|H_0)}{g(\bar{x}|H_1)}$$

Test statistic example



Linear test statistics

- The likelihood ratio is often too demanding on computing ressources. A way out may be to assume a linear dependence between t_{opt} and $\bar{x}$:

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

- For each of the two alternative hypothesis, k, we have predicted means and sigmas of the test statistic:

$$\begin{aligned} \tau_k &= \bar{a}^T \bar{\mu}_k \\ \Sigma_k^2 &= \bar{a}^T V_k \bar{a} \end{aligned}$$

Fishers linear discriminant

- We want to maximize the separation:

$$J(\bar{a}) = \frac{(\tau_0 - \tau_1)^2}{(\Sigma_0^2 + \Sigma_1^2)}$$

- The result of this maximization is

$$\bar{a} \propto (V_0 + V_1)^{-1}(\mu_0 - \mu_1)$$

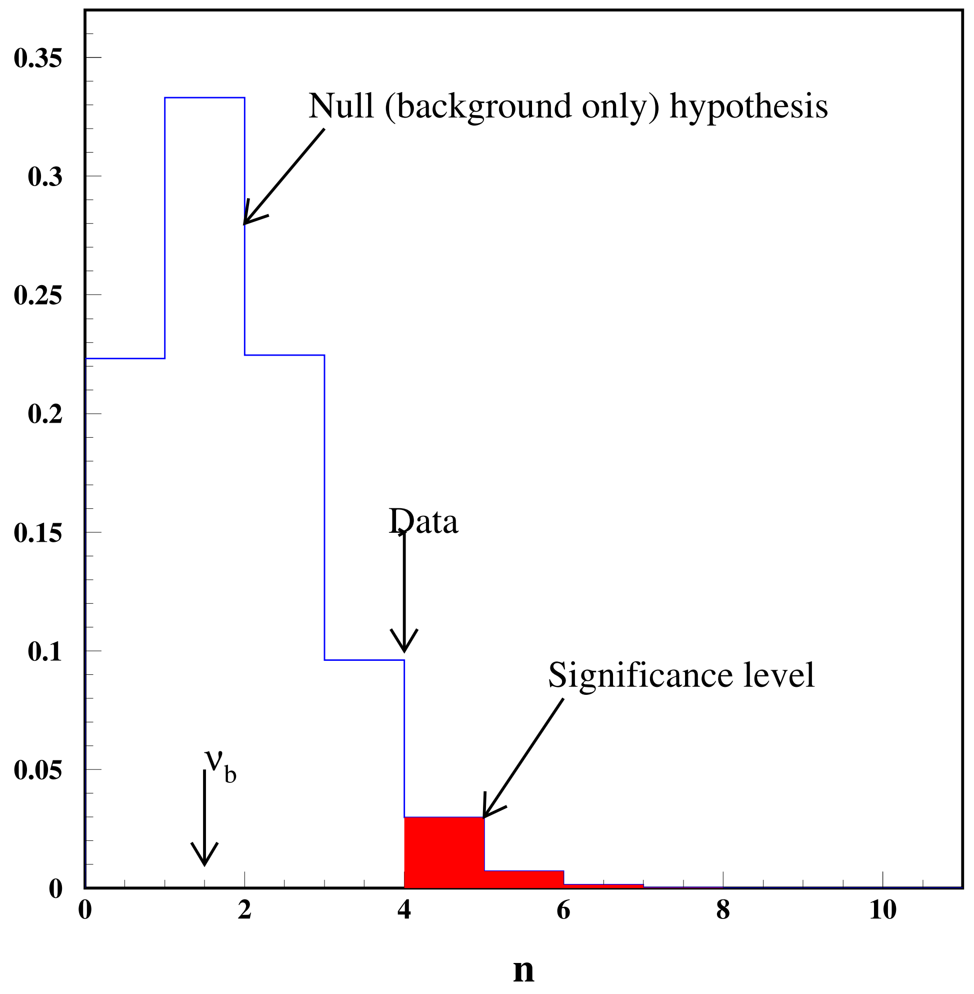
The corresponding $t(\bar{x}) = \bar{a}^T \bar{x}$ is Fishers linear discriminant. For Gaussian p.d.f's this is just as good as the likelihood ratio. For more distorted p.d.f.'s, a trained neural network is a good alternative.

Testing a single hypothesis

- To test a single hypothesis, H_0, we calculate the probability for measuring a data set which is less compatible than the actual observation with the hypothesis. This is the P-value or observed significance level.
- As an example, consider some evidence for a new signal in form of a much larger number of observed events n_{obs} than the expected mean number of background events ν_b. The probability to see n_{obs} events or more under the standard null hypothesis is

$$\begin{aligned} P(n \geq n_{obs}) &= \sum_{n_{obs}}^{\infty} f(n, \nu_b) \\ &= 1 - \sum_{0}^{n_{obs}-1} \frac{\nu_b^n}{n!} e^{-\nu_b}. \end{aligned}$$

Testing a single hypothesis



Some remarks on significance levels

- Note that P is not the probability of $\nu_{signal} = 0$. It is the probability of the data, assuming this hypothesis.
- Another pitfall is the "look-elsewhere" effect. If other new signals had been looked for and none found, the probability for accidentally finding one large excursion among all these possibilities should be quoted.
- Finally there is a possible systematic uncertainty in the estimation of ν_b.
- Thus, a very low P-value is required for a discovery, typically the probability for a Gaussian variable to be more than 5σ away from its expectation value.

Parameter estimation

- Consider a sample, $\bar{x} = (x_1, \cdots, x_n)$, of independent measurements of a single random variable with a p.d.f., $f(x;\theta)$, whose parameters, θ, are not known. We wish to construct an estimator, $\hat{\theta}(\bar{x})$.
- If $\hat{\theta}(\bar{x})$ converges to θ in the large n limit, the estimator is consistent.

bias and variance of estimators

- The estimator is itself a random variable with expectation value

$$E[\hat{\theta}] = \int \hat{\theta}(\bar{x}) f(x_1;\theta) \cdots f(x_n;\theta) dx_1 \cdots dx_n$$

The bias of the estimator is $b = E[\hat{\theta}] - \theta$.

- a measure of quality of the estimator is its mean squared error:

$$E[(\hat{\theta} - \theta)^2] = V[\hat{\theta}] + b^2$$

Estimator of the mean

- A consistent and unbiassed estimator of the mean μ is the sample mean:

$$< x > \quad = \quad \frac{1}{n} \sum_{i=1}^{n} x_i$$

The variance of this estimator is

$$V[< x >] \quad = \quad \frac{\sigma^2}{n}$$

Estimators of higher moments

- A consistent and unbiassed estimator of the variance μ is the sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - < x >)^2$$

The variance of this estimator (the "error on the error") is

$$V[s^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$$

where μ_4 is the fourth central moment ($= 3\sigma^4$ for Gauss).

- .. and so on for the covariance of two variables or higher moments.

Maximum likelihood Estimators

- The likelihood function is defined:

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

which is just the joint p.d.f. of the n measurements of x - now viewed as a function of the parameters θ. An obvious guess of θ are the values $\hat{\theta}$ that maximize the likelihood:

$$\left.\frac{\delta L}{\delta \theta_i}\right|_{\theta=\hat{\theta}} = 0$$

The Rao-Cramer-Frechet inequality

- It can be shown that the minimum possible variance is given by the Rao-Cramer-Frechet inequality:

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

- The ML estimator satisfies the equality in the large sample limit.
- The ML estimator is transformation invariant: $\widehat{g(\theta)} = g(\hat{\theta})$

A maximum likelihood estimator example

Example: Suppose a number of lifetimes, $(t_1 \cdots t_n)$, are measured for some sample of an unstable particle. Our hypothesis for the p.d.f. would be:

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

and our estimate of τ would be the value, $\hat{\tau}$, maximizing the log-likelihood:

$$\begin{aligned} \log L(\tau) &= \sum \left(\log \frac{1}{\tau} - \frac{t_i}{\tau} \right) \\ \hat{\tau} &= \frac{1}{n} \sum t_i \end{aligned}$$

Bias and variance - example

It is "easy" to show that the expectation value is $E[\hat{\tau}] = \tau$, so that the estimator is unbiased. Likewise it is "easy" to show that:

$$\begin{aligned} V[\hat{\tau}] &= E[\hat{\tau}^2] - (E[\tau])^2 \\ &= \frac{\tau^2}{n} \end{aligned}$$

in accordance with "the error on the mean", and saturating the RCF bound

Variance of ML estimators - by fitting

For the case of several parameters, the RCF bound (with no bias) is

$$\left(V^{-1}\right)_{ij} = E\left[-\frac{\delta^2 \log L}{\delta\theta_i \delta\theta_j}\right]$$

suggesting an estimator of the inverse covariance matrix:

$$\left(\widehat{V^{-1}}\right)_{ij} = -\left.\frac{\delta^2 \log L}{\delta\theta_i \delta\theta_j}\right|_{\theta=\hat{\theta}}$$

which for a single parameter becomes

$$\widehat{\sigma^2}_{\hat{\theta}} = \left.\left(-1 / \frac{\delta^2 \log L}{\delta\theta^2}\right)\right|_{\theta=\hat{\theta}}$$

Variance of ML estimators - by simulation

An alternative way is to simulate a large number of experiments, each with n measurements, to determine $\hat{\theta}$, $E[\hat{\theta}]$ and $V[\hat{\theta}]$. A Taylor expansion near the $\log L$ maximum gives:

$$
\begin{aligned}
\log L(\theta) &= \log L(\hat{\theta}) + \frac{1}{2!}\left[\frac{\delta^2 \log L}{\delta\theta^2}\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta})^2 + \cdots \\
\log L(\theta) &= \log L_{max} - \frac{(\theta-\hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}} \\
\log L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) &= \log L_{max} - \frac{1}{2}
\end{aligned}
$$

Hereby $\hat{\theta}$ and $\hat{\sigma}_{\hat{\theta}}$ can be determined by a graphical method using the two last equations.

The least squared method

Consider N measurements, y_i, supposedly given by some function $\lambda(x_i, \bar{\theta})$, where the variables x_i are known without error, but the parameters $\bar{\theta}$ are unknown. Suppose now the y_i's are Gaussian random variables centered around the value of λ. The joint p.d.f. is then a product of Gaussians, and its logarithm is:

$$\begin{aligned}\log L(\bar{\theta}) &= -\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\bar{\theta}))^2}{\sigma_i^2} \\ &= -\frac{1}{2}\chi^2(\bar{\theta})\end{aligned}$$

So maximizing $\log L$ corresponds to minimizing χ^2.
The method of least squares applies the latter procedure even to non-Gaussian variables.

Fitting non-independent data

In case the y_i are not independent, but described by an N-dimensional Gaussian with covariance matrix V, the quantity to be minimized is:

$$\chi^2(\bar{\theta}) = \sum_{i=1}^{N} (y_i - \lambda(x_i;\bar{\theta}))(V_{ij}^{-1})(y_j - \lambda(x_j;\bar{\theta}))$$

Goodness of fit

Consider again N measurements, y_i, and a hypothesis λ parametrized by m parameters. If the following conditions are fulfilled:

- The y_i's are independent Gaussian random variables
- λ is linear in the parameters
- λ has the correct functional form

then the χ^2 will follow the χ^2-distribution with $N - m$ degrees of freedom. The $P - value$

$$P = \int_{\chi^2}^{\infty} f(z; n_d) dz$$

may provide a subjective criteria for rejecting the hypothesis.

The least squared method (linear case)

Consider the case of a linear function of m parameters $\bar{\theta}$

$$\lambda(x_i, \bar{\theta}) = \sum_{j=1}^{m} a_j(x_i)\theta_j = \sum_{j=1}^{m} A_{ij}\theta_j$$

The functions a_j must be linearly independent.

χ^2 is then in matrix notation:

$$\chi^2 = (\bar{y} - A\bar{\theta})^T V^{-1} (\bar{y} - A\bar{\theta})$$

The least squared method (linear case)

The minimum χ^2 is given by

$$\nabla\chi^2 = -2(A^T V^{-1}\bar{y} - A^T V^{-1} A\bar{\theta}) = 0$$

which is solved by

$$\hat{\theta} = (A^T V^{-1} A)^{-1}(A^T V^{-1})\bar{y} \equiv B\bar{y}$$

Errors in the least squared method

By error propagation we get the covariance of the fitted parameters:

$$\begin{aligned} U &= BVB^T = (A^T V^{-1} A)^{-1} \\ &= \left(\frac{1}{2} \left[\frac{\delta^2 \chi^2}{\delta\theta_i \delta\theta_j} \right]_{\theta=\hat{\theta}} \right)^{-1} \end{aligned}$$

This corresponds to the RCF bound when the y_i's are Gaussian, in which case $\log L = -\chi^2/2$. Thus, changing one of the fitted parameters by one sigma will increase χ^2 to $\chi^2_{\min} + 1$.

Example: Straight line fit (1)

For the case of a straight-line hypothesis: $y(x) = \alpha_1 + \alpha_2 x$, one obtains for a sequence of independent measurements y_i the following parameter estimates,

$$
\begin{aligned}
\widehat{\alpha_1} &= (g_1 V_{22}^{-1} - g_2 V_{12}^{-1}) / D \, , \\
\widehat{\alpha_2} &= (g_2 V_{11}^{-1} - g_1 V_{12}^{-1}) / D \, ,
\end{aligned}
$$

where

$$
\begin{aligned}
\left(V_{11}^{-1}, V_{12}^{-1}, V_{22}^{-1}\right) &= \sum \left(1, x_i, x_i^2\right) / \sigma_i^2 \, , \\
(g_1, g_2) &= \sum (1, x_i) \, y_i / \sigma_i^2 \, ,
\end{aligned}
$$

respectively, and $D = V_{11}^{-1} V_{22}^{-1} - (V_{12}^{-1})^2$.

Example: Straight line fit (2)

The covariance matrix of the fitted parameters is

$$(V_{11}, V_{12}, V_{22}) \quad = \quad \left(V_{22}^{-1}, V_{12}^{-1}, V_{11}^{-1}\right) / D\,.$$

The estimated variance of an extrapolated value of y is

$$\sigma_y^2 \quad = \quad \frac{1}{V_{11}^{-1}} + \frac{V_{11}^{-1}}{D}\left(x - \frac{V_{12}^{-1}}{V_{11}^{-1}}\right)^2 \,.$$

Weighted averages in the least squared method

If we have several independent estimates y_i of the same quantity λ, but with different errors σ_i, we can combine these measurements using the formula for weighted average:

$$\begin{aligned}
\hat{\lambda} &= \frac{\sum y_i/\sigma_i^2}{\sum 1/\sigma_j^2} \\
V[\hat{\lambda}] &= \frac{1}{\sum 1/\sigma_j^2}
\end{aligned}$$

It becomes more complicated if the y_i are not independent (due to e.g. common systemic errors). Common errors should be separated out and added after the averaging.

The least squared method on binned data

Consider n observations of x presented in a histogram with N bins. We want to compare it with a hypothetical p.d.f. with probabilities $p_i(\theta)$ for each bin. For sufficiently large N, the number of entries in each bin, y_i, are approximately Poisson distributed. Thus, the parameters are found by minimizing:

$$\chi^2(\bar{\theta}) \quad = \quad -\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - np_i(\bar{\theta}))^2}{np_i(\bar{\theta})}$$

the Modified Least Squares method

- Often, the denominator is replaced by the measurement itself, y_i, for convenience. This is called the modified least squares method (MLS). It is the standard fit method in **PAW** and **ROOT**. But it is not ideal if some bins have very few entries. In this case use the Max Likelihood option.
- Notice also, that if the total number of entries is also left free in the fit, it will get a biassed estimate in general: $\hat{\nu}_{\mathrm{LS}} = n + \chi^2/2$ and $\hat{\nu}_{\mathrm{MLS}} = n - \chi^2$. So it is better to count the number of entries and fix it before the fit.

Numerical tools

- Fitting multi-parameter functions can be done by the **MINUIT** package, either interactively via its **PAW** or **ROOT** interface, or in a fortran program via the **HBOOK** interface or directly using **MINUIT**.
- In order to get started with these tools, it is essential to go through e.g. Troels Pedersens short tutorial. For bigger tasks you have to get hold of the **PAW** manual, or the **ROOT** manual, from the links listed in the beginning of theses lectures.

Confidence intervals

The standard way of reporting an estimate is $\hat{\theta} \pm \sigma_{\hat{\theta}}$. By this we mean that, should our experiment be repeated many times, the sample variance is estimated to be $\sigma^2_{\hat{\theta}}$. For a Gaussian p.d.f., $g(\hat{\theta};\theta)$, this means that 68.3% of these experiments should give an estimate within the standard error from the truth.

But if $g(\hat{\theta};\theta)$ is not Gaussian we need more precise ways of reporting a result. We need to indicate an interval of θ that is expected to cover the true value with some specified confidence level (CL).

Confidence intervals

- Let $u(\theta)$ be such that $P(\hat{\theta} > u(\theta)) = \alpha$, and let $v(\theta)$ be such that $P(\hat{\theta} < v(\theta)) = \beta$. Let the inverse of these functions be $a(\hat{\theta}) = u^{-1}(\hat{\theta})$ and $b(\hat{\theta}) = v^{-1}(\hat{\theta})$. Thus, regardless of the true value of θ we have:

$$
\begin{aligned}
P(v(\theta) < \hat{\theta} < u(\theta)) &= 1 - \alpha - \beta \\
P(a(\hat{\theta}) < \theta < b(\hat{\theta})) &= 1 - \alpha - \beta
\end{aligned}
$$

  where the last equation does not indicate the probability of θ, but rather the probability that the true value is "covered" by the interval. This probability is called the confidence level.
- For $\alpha = \beta = \gamma/2$ we talk about a central confidence interval. If e.g. $\beta = 0$, we talk about a one-sided confidence interval or limit.

Confidence intervals, Gaussian case

In case of a (multi)Gaussian p.d.f. for the estimated parameters, $g(\hat{\theta};\theta)$, the quantiles indicate how far away from $\hat{\theta}_{obs}$, measured in units of $\sigma_{\hat{\theta}}$, we need to go in order to obtain a certain confidence level. Here are some central and one-sided intervals:

sigma's	$1-\gamma$	*sigma's*	$1-\alpha$
1	0.6827	1	0.8413
2	0.9544	2	0.9772
3	0.9973	3	0.9987
1.645	0.90	1.282	0.90
1.960	0.95	1.645	0.95
2.576	0.99	2.326	0.99