Statistics and Data Analysis

Christian Bohm
University of Stockholm

The aim is to introduce and discuss some key statistical concepts

Classical measurements

- **The determination of the size or magnitude of some physical entity**

or

- **Comparison of an unknown quantity with some known quantity of the same kind**

The physical entity to be measured can be assed to be **fixed**. It can, of course, vary with time but at each instant it has a fixed value.

The measurements are affected by **fluctuating** parameters that cause their exact value to be undetermined, but their variations can be described by a **probability distribution function (pdf)**

The distribution can not be determined by one measurement

you need **several measurements,**

infinitely many for a exact determination, but

a smaller number, **a sample,** can be sufficient for an **approximate determination**

However, sometime the distribution can be deduced from physics

Thus in a measurement, it is important to **identify the pdf** of the result

If this is not possible - one must at least **characterize** the pdf as well as possible

The most important parameter is **position**, then **width**, **skewness**, etc.
(these parameters can be determined with good precision from a smaller amount of data)

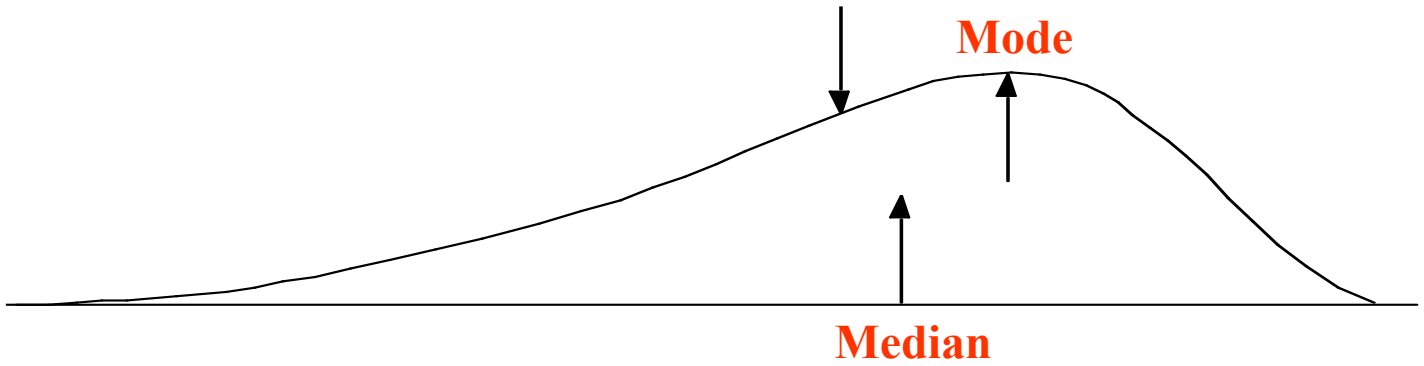
The position of the distribution

The expectation value of x

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

f's 1:st moment
(center of gravity)

Population mean $\mu = E(x)$



$$\int_{-\infty}^{m} f(x) dx = 0.5$$

Choice of position parameter depend on the type of measurement

Mean most common

The width of the distribution

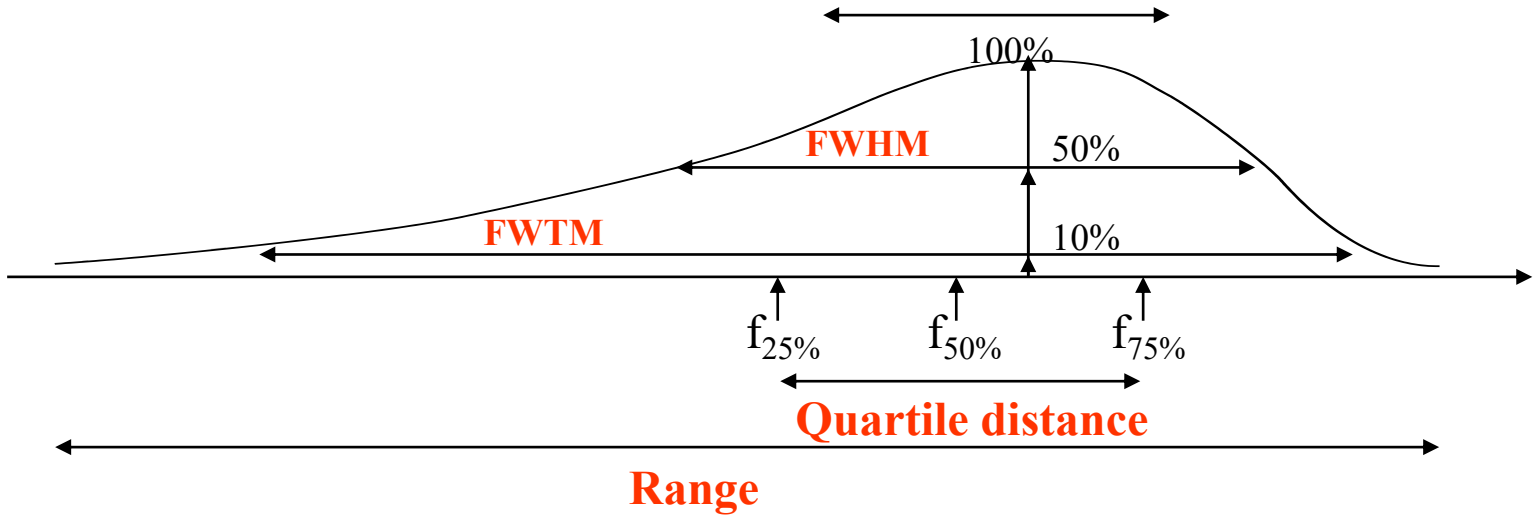
Population variance

$$Var(x) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = E((x-\mu)^2) = E(x^2) - 2\mu E(x) + \mu^2 = E(x^2) - \mu^2$$

f's 2:nd central moment

f's 2:nd moment

σ = standard deviation



Choice of parameter depend on the type of measurement

Standard deviation and **F**ull **W**idth **H**alf **M**aximum (**FWHM**) most common

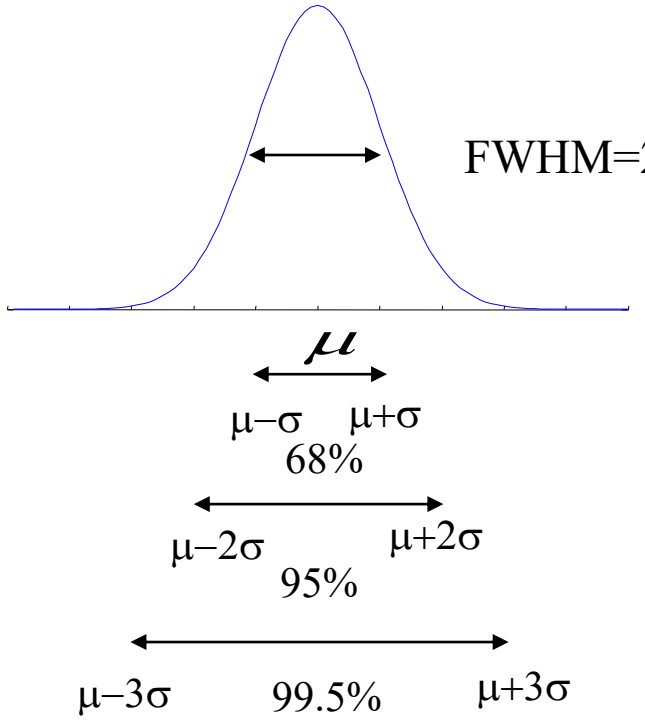
The Normal Distribution

Variable x, real number

Parameters σ, μ standard deviation, mean

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

$N(\mu,\sigma^2)$ denotes a normal distributed parameter with mean μ and standard deviation σ



A 100x100 image with Gaussian data will contain about 500 2σ points and 50 3σ points.
Finding a 5σ point is not so spectacular
The number of fake signals increase in large data sets

Also called the **Gaussian** distribution

Central limit theorem

If the independent events X_i have the mean μ and the variance σ^2

then

$\sum_{1 \le i \le N} X_i$ has the mean $N\mu$ and the variance $N\sigma^2$

→ $\frac{1}{N} \sum_{1 \le i \le N} X_i$ has the mean μ and the variance σ^2/N

→ $\frac{\overline{X} - \mu}{\sigma / \sqrt{N}}$ has the mean 0 and the variance 1

The central limit theorem claims that $\frac{\overline{X} - \mu}{\sigma / \sqrt{N}}$ is **normally distributed**

i.e. it has the limit distribution **N(0,1)**

This is why the normal distribution is so important
Quadrupling the number of measurements -> halves the statistical error

The Binomial Distribution

Repeating independent elementary binary events (succeed – fail) each with the probability p

E.g.

Tossing coins	elementary event – coin toss
Drawing tickets with replacement	elementary event – draw
Radioactive decay	elementary event – decay of a nucleus
Monte Carlo simulations	elementary event – one case

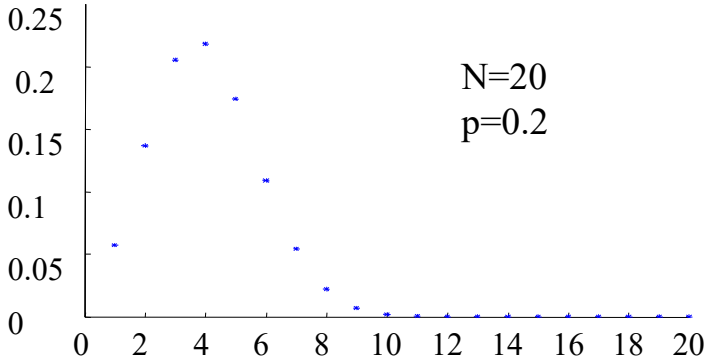
Parameters $0 \le p \le 1$ probability
N>0 number of trails

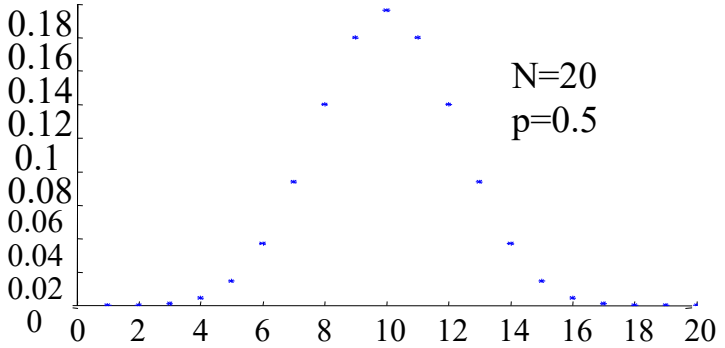
Variable r

Probability distribution $p(r) = \binom{N}{r} p^r (1-p)^{N-r}$

Mean $E(r) = Np$

Variance $V(r) = Np(1-p)$



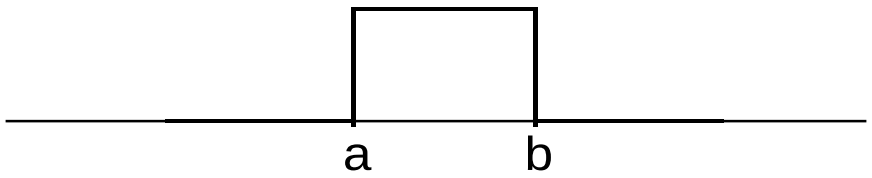


The uniform distribution

E.g.

Pseudo random number in a computer, truncation errors and digitization errors are uniformly distributed

Variable	x, real number
Parameters	a, b, a < b
Probability distribution function	$p(x) = 1/(b-a)$; $a \le x \le b$
Mean	$E(x) = (a+b)/2$
Variance	$V(x) = (b-a)^2/12$



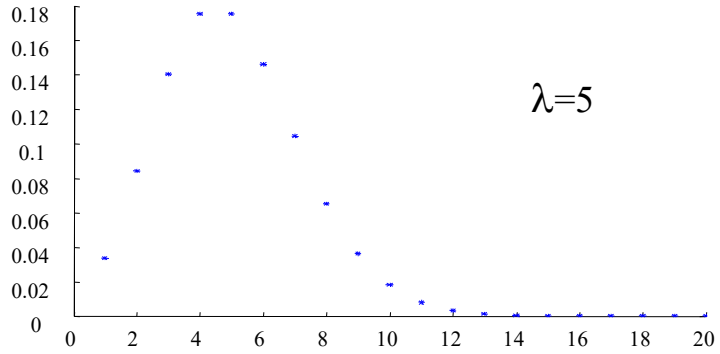
The Poisson distribution

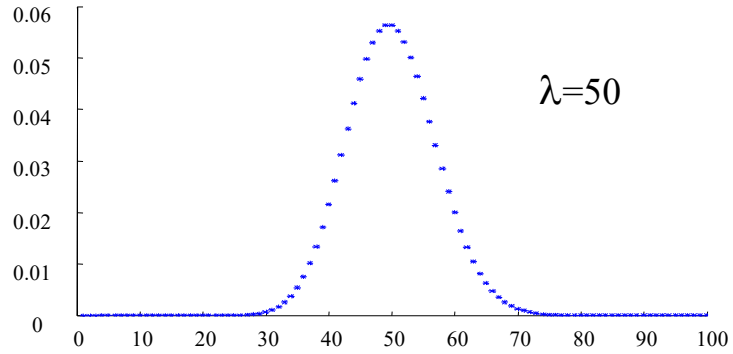
The **probability for a certain number of events during a time period** if the **probability per time unit** for such a event is **constant** (λ) and **independent of what happened before**.
One can say that the process has **no memory**

Parameter	0< λ, events/time unit
Variabel	r≥0, the number of events
Probability distribution	$p(r) = \frac{\lambda^r e^{-\lambda}}{r!}$
Mean	$E(r) = \lambda$
Variance	$V(r) = \lambda$

Binomial distribution --> Poisson distribution with $\lambda = Np$ if $N \to \infty$ and $p \to 0$ while Np=const

Radioactive decays (approx. Poisson)
Histograms with many events (approx Poisson)





Assuming Poisson instead of Binomial when p is large leads to over estimation of variance. In semiconductor detectors this is compensated by using the Fano factor

Poisson distribution with n>50 looks like a normal distribution

Stochastic variables

a variable **x** which can assume different values with the probability density function f_x

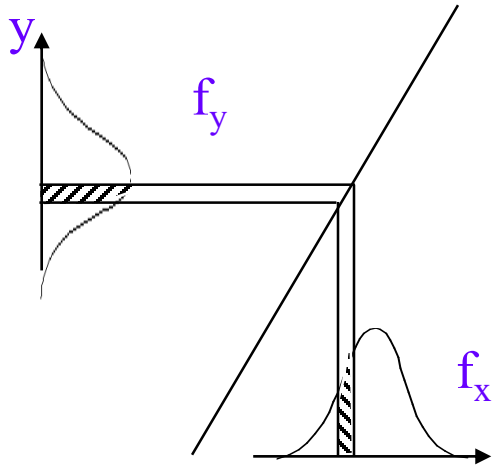
x is therefore **completely defined by f**

You can multiply sv:s with coefficients, add or subtract them

y=2**x** has a probability density as well and is thus also a stochastic variable. Here the distribution just expanded with a factor 2.

Linear transforms of stochastic variables

The stochastic variable **y**=2**x** has the probability density $f_x(y/2)/2$



y=2x

The areas correspond to the same event

Probabilities for the selected event:

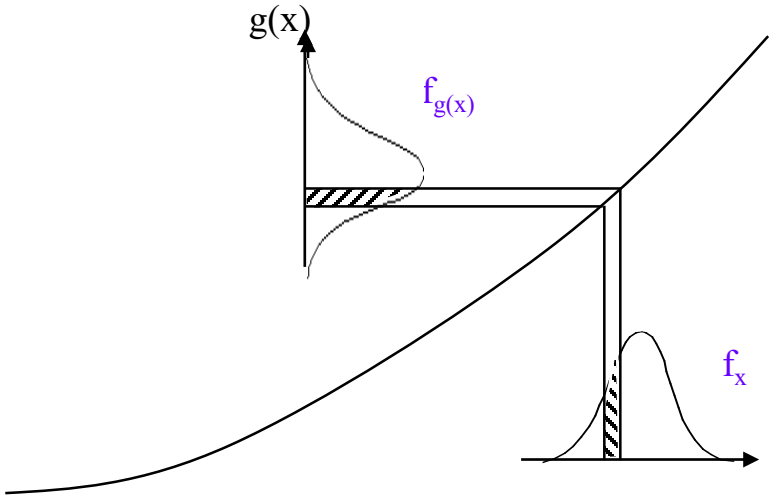
$f_y(y)\,dy = f_x(x)\,dx$

but since: $dy = y'dx = 2dx$

$f_y(y) = f_x(x)/2 = f_x(y/2)/2$

Non-linear transform of stochastic variables

The stochastic variable y=g(x) has the probability density $f_x(x) / |g'(x)|$



The areas correspond to the same event

Probabilities for the selected event:
$f_{g(x)}(g(x))\, dg(x) = f_x(x)\, dx$

but since: $dg(x) = g'(x)\, dx$
$f_{g(x)}(g(x)) = f_x(x) / |g'(x)|$ provided g(x) is one-one

If g(x) is non-linear the pdf will be deformed

But if g(x) is approximatively linear in the main range of x
then g(x) will, apart from a coefficient, have the similar pdf as x

Monte Carlo simulations

- are consists of many choices of sv:s with different distribution functions

- A uniformly distributed sv is easily generated in a computer.
- You start with a **seed number** to initialize the random number genertor which then generate a sequence of numbers, one for each function call, a sequence of approximately uniformely distributed values.
- This is called **pseudo random variable.** If you start with the same seed you get the same sequence (very useful if you are developing a program).

- The **pseudo random variable** can then be transformed into any desired distribution

Statistics

A **statistic** is a **function of stochastic variables**

$T_N = f(X_1, X_2, , X_N)$ is a **statistic** ;where X is a measurement

The calculation {X} --> T_N implies a **data reduction**

The task is to perform data reduction without loosing information

Estimators

Let us use the statistics T_N to **estimate** the physical parameter θ

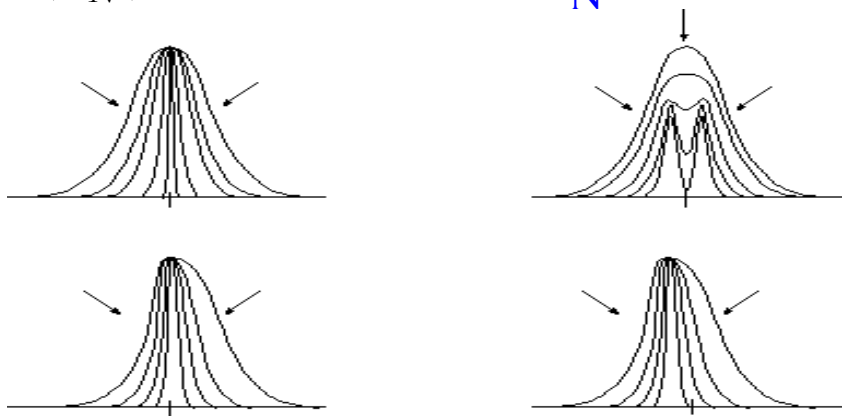
T_N is the called an **estimator**

An infinitely large sample should give the true value

If $\lim_{N\to\infty} T_N = \theta$ then T_N is **consistent**

The mean of a large number of small sample estimators should give the true value

If $E(T_N) = \theta$ for all N then T_N is **unbiased**



unbiased

biased

consistent **inconsistent**

Estimators

If T_N uses the information well it is **effective**

If T_N is not sensitive to small variations in the distribution then T_N is **robust**

A lack of consistency correspond to the presence of **systematical errors**

Samples

If you have a sample with N measured values x_i then
The sample mean is

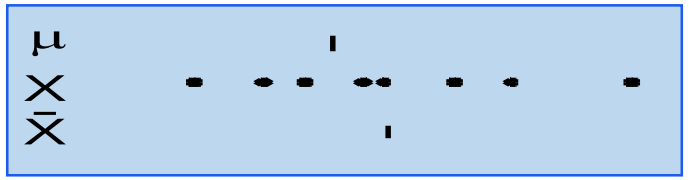
$$\bar{x} = \frac{1}{N}\sum_i x_i$$

It is a consistent estimator of the population mean μ
One also easily show that it is unbiased,
since the mean of many small samples is the same as the mean
of one large sample

$$s^2 = \frac{1}{N}\sum_i (x_i - \bar{x})^2 \quad \text{and} \quad s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$$

are both consistent estimators of σ^2 but
only the right one is unbiased

Why N-1?



$\bar{x}$ is more central in the sample than μ thus

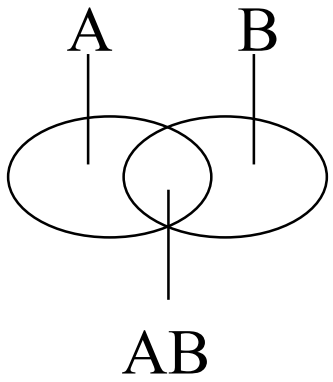
$$\sum_i (x_i - \bar{x})^2 \le \sum_i (x_i - \mu)^2$$

N-1 compensates for the under estimation

Bayes' theorem*

P(A) is the probability that A will occur

P(A|B) is the conditional probability that A will occur if B has occured



$P(B|A) = P(AB)/P(A)$ →

$P(AB) = P(A)P(B|A) = P(A|B)P(B)$ Bayes' theorem

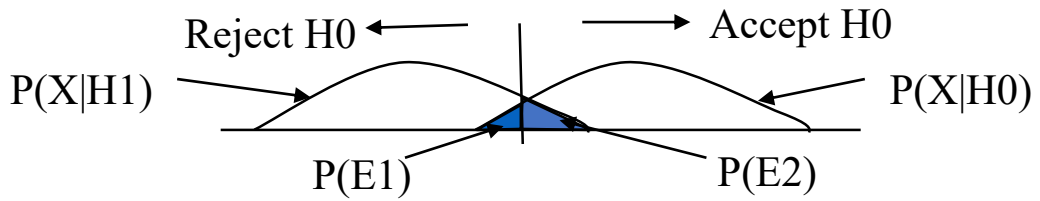
Bayesian methods
P(A) Prior distribution
P(AB) Posterior distribution

Tests of hypotheses

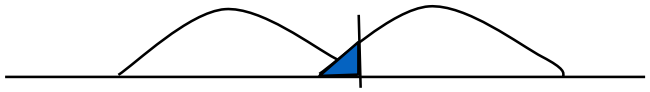
H0 the null-hypothesis the hypothesis you want to test - e.g. there is a pulse
H1 an alternative hypothesis – there was no pulse, just noise

Error of the first kind (E1): Erroneous rejection of the null-hypothesis – loosing the pulse, inefficiency

Error of the second kind (E2): Erroneous rejection of the alternate hypothesis – reading noise as a pulse



Choose a limit so that P(E2) becomes sufficiently small – below a significance level 5% is common.
If P(E2) becomes too large improve the data (more measurements)



Difficult to optimize choice of two conflicting parameters

Find a **cost function** which includes the probabilities and the cost caused by the different errors
Choose a limit that minimizes the cost function

We need many measurements to claim a discovery

To determine if **N** observed events include a new type of events that would constitute a discovery we must determine if the data could be produced by combinations of well-understood events.

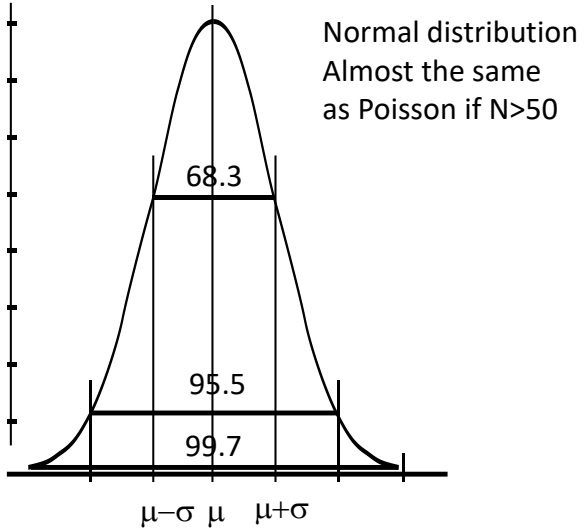
The probability for such events is the background **B** and its standard deviation σ(**B**)=sqrt(B) (Poisson statistics)

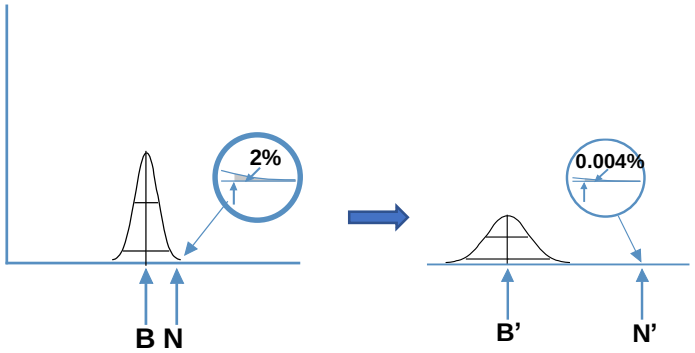
For **N** to contain a discovery, **N** must be significantly larger than **B**
For example if **N** is 80 and **B** is 64 then σ(**B**) is 8
N is 2σ above **B** i.e. 2% probability that N is just random noise

If we measure 4 times as long **N´** will be at 320 and **B´** 256 making σ(**B´**) 16, i.e. the difference is then about 4σ corresponding to 0.004% that the measurement is just random noise. This is a much smaller probability but it is not enough to claim a discovery.

5σ (0.00002% it is random noise) is often required for discovery.

However, this argument assumes no systematic errors. Increasing the number of events means that the relative influence of the systematic errors increase.





Fitting parametrized expressions to data sequences*

To fit a theoretical expression $f(k,\theta)$ that depends on the parameter θ to a set of data $y(k)$ where $k=1$ to n, you need a figure of merit to optimize.

The f.o.m. expresses how close f and y are.

Since y contains noise the f.o.m. is a statistic with a pdf even if f is correct.

One such methods is the Least Square Method (**LSM**) where the f.o.m. is the sum of the square normalized distance between the f and y points. Normalize means that you divide the square distance with the variance of the distance.

The sum of the squares are **χ^2-distributed**

The degrees of freedom is the number of data points minus the the number of parameters fitted.

Another method is to calculate the probability (Likelyhood) for getting the measurement results y if f is the correct description. The parameter values are choosen that maximizes the Likelyhood (**ML-method**)

The χ^2-distribution*

If x_i are N different N(0,1) normal distributed variables then the sum

$$q = \sum_{1 \le i \le N} x_i^2$$

is χ^2-distributed with N degrees of freedom

Variable — q, positive real number

Parameter — N degrees of freedom

Probability distribution

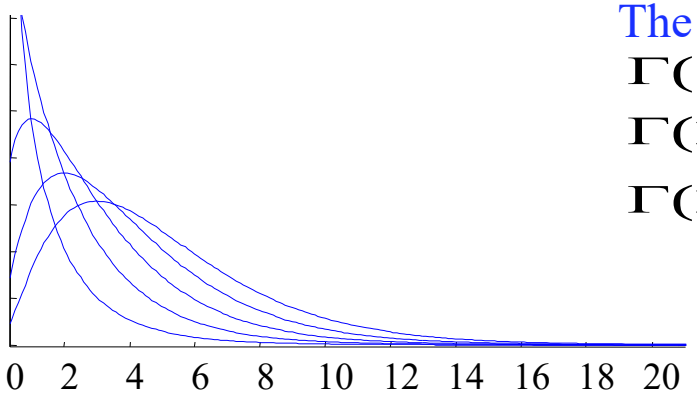
$$f(q) = \frac{(q/2)^{N/2-1} e^{-q/2}}{2\Gamma(N/2)}$$

Mean

$$E(q) = N$$

Variance

$$V(q) = 2N$$



The Γ-function is defined as:

$$\Gamma(n) = (n-1) \cdot \Gamma(n-1)$$

$$\Gamma(1) = 1$$

$$\Gamma(1/2) = \sqrt{2\pi}$$

Thank you for your attention