Learning goals of the week:

- the logic behind error propagation uncertainties
- what is a probability density function
- parent vs. sampling distributions
- expectation value variance and their estimators

Week 2

Error propagation (I)

Propagation of *statistical* uncert.

(Most of the times, one cannot use standard error propagation for *systematic uncertainties*! We will come back to this later.)

We start with the problem of propagating the statistical uncertainties of two measurements which we can assume to be *uncorrelated* (the first measurement does not affect in any way the second measurement)

Let's begin from the easiest example to get the logic:
suppose you have two measurements x ± δx and y ± δy and you want to sum them (e.g. two volumes of liquid): x+y

The goal is to convey the uncertainty on x+y. First ask yourself what you mean by x±δx.

We assume here that when quoting x+ δx we want to convey that the true (unknown) value of the quantity we are measuring is in the interval [x-δx, x+δx].

This way we assume the true value is an unknown fixed value, while the measured quantities are uncertain numbers (random variables), and the interval [x-δx, x+δx] contains the true value with some confidence (see later).

Frequentist vs. Bayesian digression

The opposite approach is taken in Bayesian probability / statistics-inference.

The Frequentist^ perspective is that the true parameter value θ is fixed but unknown, while the measured θ is a random variable a function of the dataset (which is seen as random). The Bayesian perspective uses probability to reflect "degrees of certainty" or "states of knowledge". The dataset is directly observed and so is not random. On the other hand, the true parameter θ is unknown or uncertain and thus is represented as a random variable.

This simple difference leads to profoundly different schools of thought. We will come back with the Bayes theorem later.

In this classes we will use the frequentist approach. The "justification" to do this is that for the purpose of the experiments you will encounter here you should never end up in a situations where priors can have an impact on the outcome of the measurement. (e.g. you will always have large statistics at hand and your values will always be away from physical boundaries). The Frequentist/Likelihood approach is justified.

It is important that once in your life you understand the two approaches.
Suggested reading: "Bayesian Reasoning in Data Analysis : A Critical Introduction"
G. D'Agostini
General public: "The theory that would not die" Sharon Bertsch McGrayne

Propagation of statistical uncert.

Now that we agree on what we mean by $x+\delta x$, let's see how to propagate it.

The uncertainty on $x+y$ can be estimated as the half the difference between the largest and the smallest value that the results can take:

$(x+y)\mathrm{max} = x+\delta x + y+\delta y$
$(x+y)\mathrm{min} = x -\delta x + y -\delta y$

$1/2 * [(x+\delta x + y+\delta y) - (x-\delta x + y-\delta y)] =$
$1/2 * [2\, \delta x + 2\, \delta y] = \delta x+\delta y$

i.e. $x+y \pm(\delta x+\delta y)$

What's wrong with this ? This is an overestimate of the uncertainty.

Because you assume the δx covers a range of possible measurements outcomes (we haven't said yet how those outcome are distributed) quoting $\delta x+\delta y$ assumes the extreme fluctuations of the measurement in both x and y.

Gaussian limit

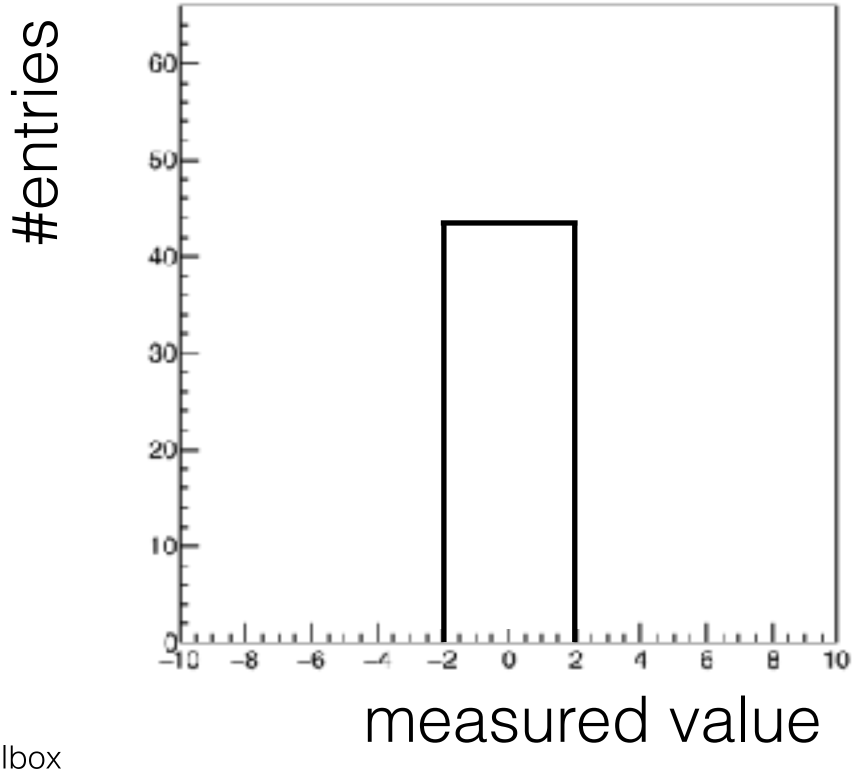
Let’s look at where δx come from.

Typically each measurement is affected by several sources of uncertainties which all sum up to the final value of δx.

Example: measure the weight of an object with a scale.
Sources of uncertainty: movements, parallax, rounding, etc...
Example: measure the momentum of a charged particle by fitting a curved track.
Sources of uncertainty: single hit resolution, multiple scattering losses, etc…

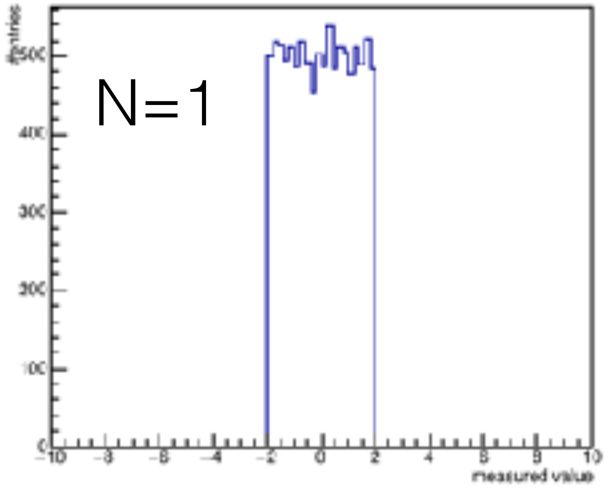
Imagine you could switch off all sources of uncertainties but one and repeat the measurement several times.

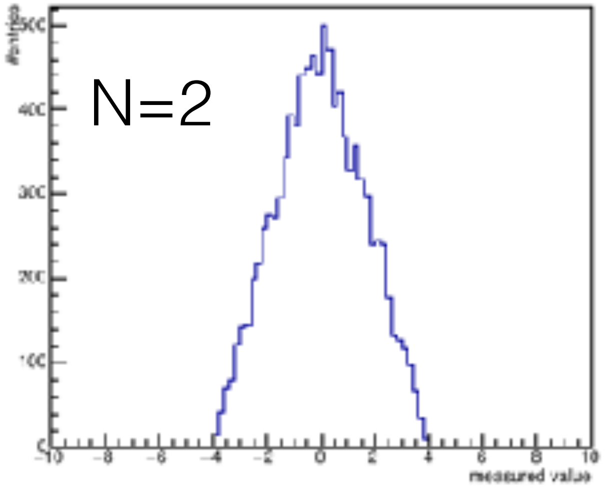
We don’t have any good reason to assume *how* the measurement would be distributed, so let’s assume for example that they all are flat distributed (try to take any other shape).



Gaussian limit

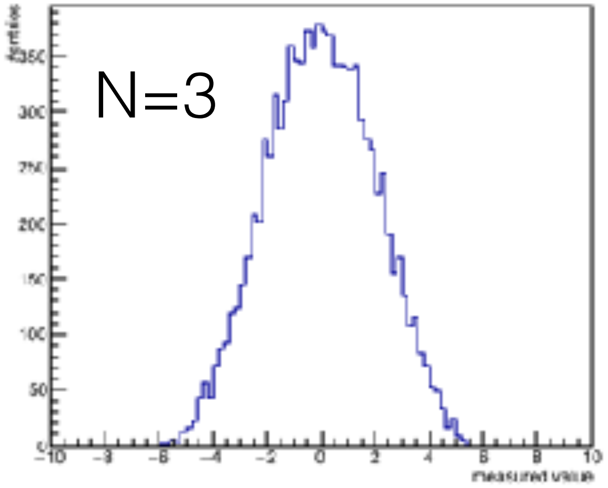
What happens if I sum N uncorrelated sources of uncertainties, each of which would produce a flat distribution of the measured values ?

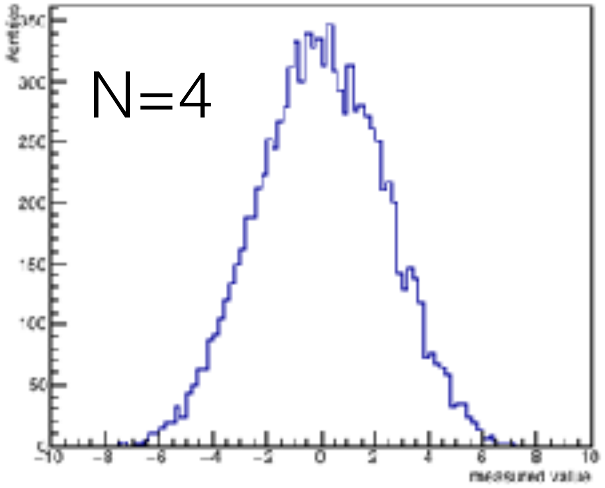




It approaches a Gaussian distribution !

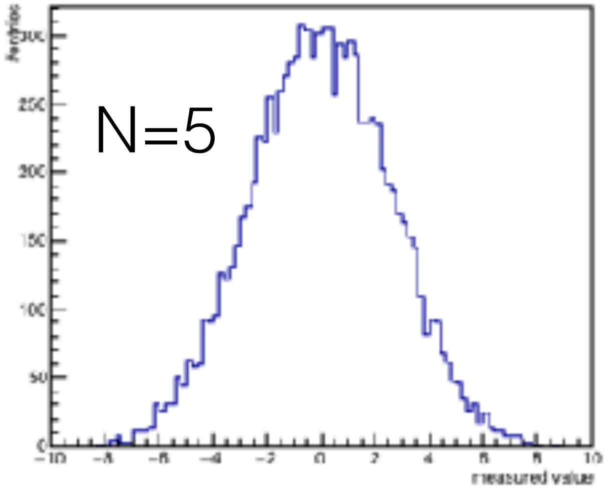
Central Limit Theorem (we'll see it later with pdf limits)

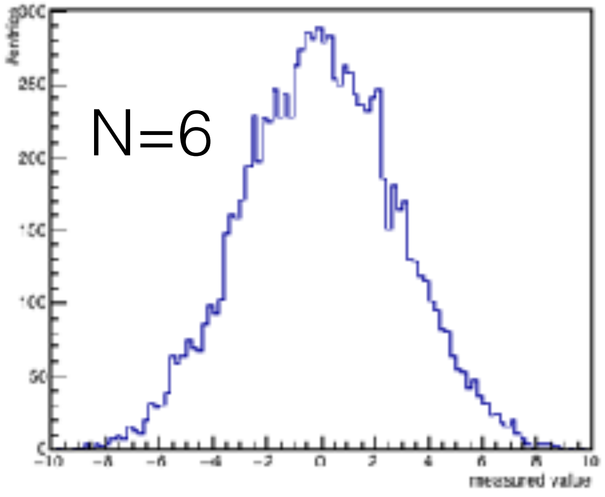




The convention is to quote as uncertainty "δx" the 1 standard deviation or "1σ" gaussian interval
(or 68% confidence interval).

There is nothing special behind using 68%. It's just conveniently corresponding to 1σ.

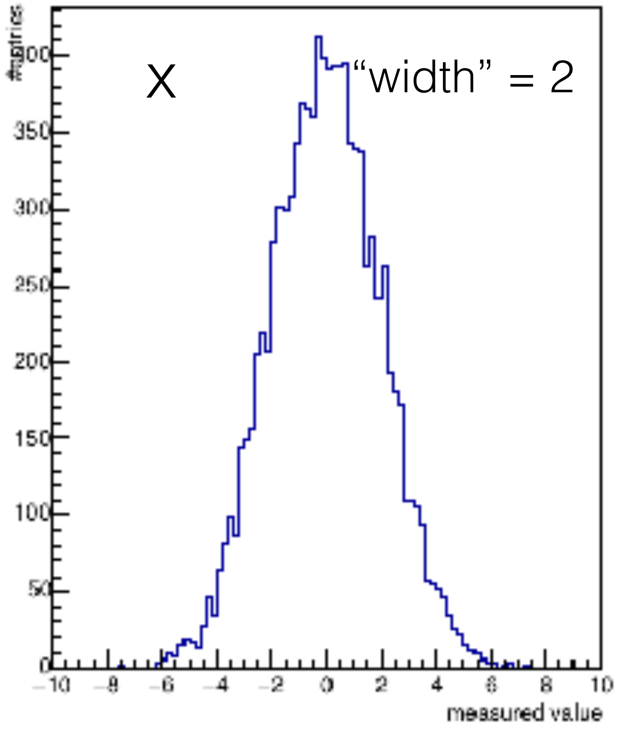


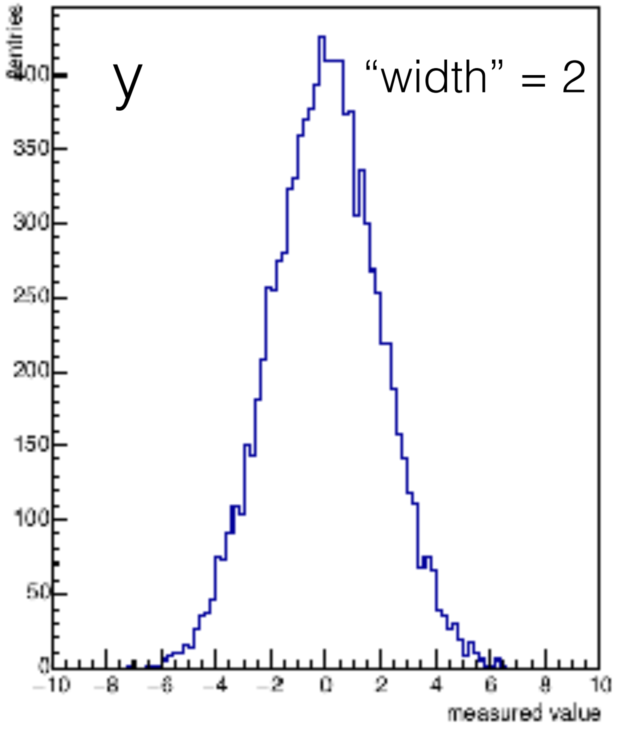


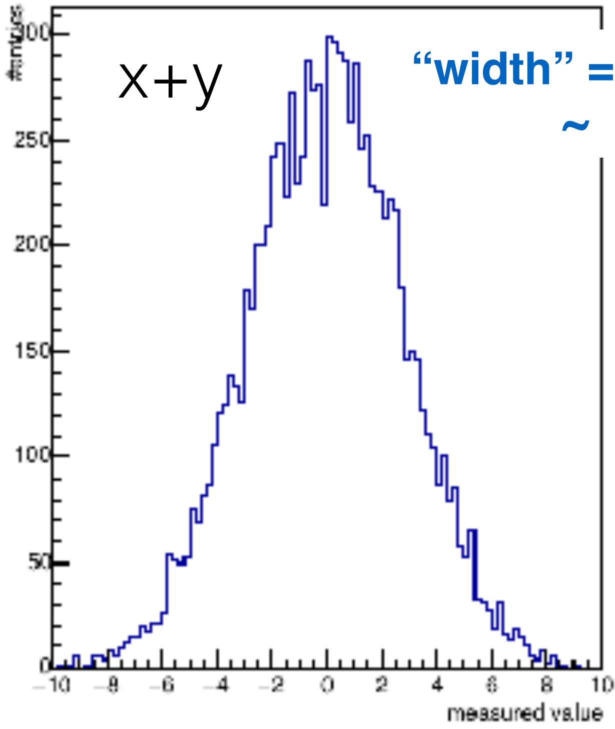
Propagation as convolution

Knowing this, we use/interpret the δx as the “width of the gaussian”.
Now what happens when you add x±δx + y±δy ?

Let’s pretend we’re repeating the measurements of x and y several times, add them and fill a histogram with the result. (Assume a gaussian resolution effect on the measured values, and again x and y uncorrelated)







“width” = sqrt($2^2 + 2^2$)
~ 2.8

This is the reason why we quote the uncertainty on the sum as the sum in quadrature of δx and δy = sqrt($\delta x^2 + \delta y^2$) = $\delta x \oplus \delta y$

This should’t come as a surprise: you are convoluting two gaussians.

The same reasoning leads to the same result ($\delta x \oplus \delta y$) for the difference x-y. (try it yourself)

General formula

Let's take two gaussian-distributed variables x, y centred at x_0 and y_0 respectively with standard deviations σ_x, σ_y, but now let's compute how a generic function of x and y is distributed (not just their sum). Expand around (x_0, y_0) and see how "wide" it is:

$$f(x,y) \propto \underbrace{f(x_0,y_0)}_{\text{constant}} + \underbrace{\left.\frac{\partial f}{\partial x}\right|_{x=x_0}}_{\text{constant}} \underset{\substack{\uparrow \\ \text{Gaussian distributed} \\ \text{centred at zero}}}{(x-x_0)} + \underbrace{\left.\frac{\partial f}{\partial y}\right|_{y=y_0}}_{\text{constant}} \underset{\substack{\uparrow \\ \text{Gaussian distributed} \\ \text{centred at zero}}}{(y-y_0)}$$

First term can be ignored: adding a constant k to x shifts the centre of the gaussian to k, but we're interested in the width σ (the uncertainty) unchanged.

Second and third terms: multiplying x by a constant k, shift the centre to kx_0, (here the gaussians are centred at zero so they don't move) and increase the width as kσ_x

here k = $\left.\frac{\partial f}{\partial x}\right|_{x=x_0}$

The effect on the width of the gaussians are $\left.\frac{\partial f}{\partial x}\right|_{x=x_0} \sigma_x$ and $\left.\frac{\partial f}{\partial y}\right|_{y=y_0} \sigma_y$

Propagation of statistical uncert.

Putting everything together the width of f(x,y) is

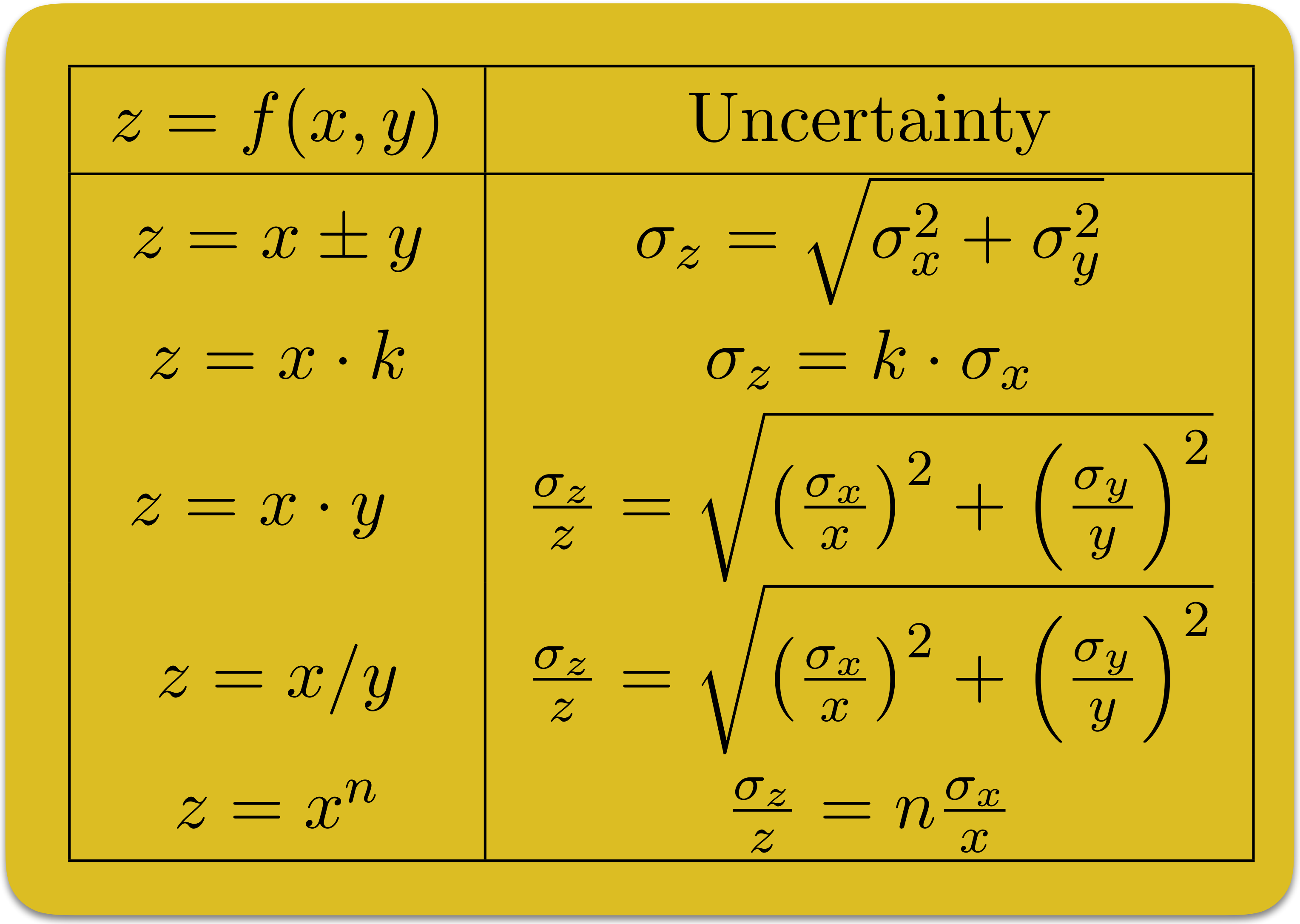
$$\sigma_f = \sqrt{\left(\left.\frac{\partial f}{\partial x}\right|_{x=x_0} \sigma_x \right)^2 + \left(\left.\frac{\partial f}{\partial y}\right|_{y=y_0} \sigma_y \right)^2}$$

Which trivially generalizes to the case of N-variables $f(x_1, \ldots, x_N)$ as:

$$\boxed{\sigma_f = \sqrt{\sum_{i=1}^{N} \left(\left.\frac{\partial f}{\partial x_i}\right|_{x_i=x_i^0} \sigma_{x_i} \right)^2}}$$

Propagation of statistical uncert.

From the general formula you can verify you can find the usual results:



$z = f(x,y)$	Uncertainty
$z = x \pm y$	$\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2}$
$z = x \cdot k$	$\sigma_z = k \cdot \sigma_x$
$z = x \cdot y$	$\frac{\sigma_z}{z} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2}$
$z = x/y$	$\frac{\sigma_z}{z} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2}$
$z = x^n$	$\frac{\sigma_z}{z} = n\frac{\sigma_x}{x}$

x+y, x-y
Sum in quadrature of the
absolute uncertainties

xy, x/y
Sum in quadrature of the
relative uncertainties

Bibliography

Error propagation:

Taylor: Chapters 1-5

Probability Density Functions

PDF: Probability Density Function

We define a random variable as any function of the data.

The event space is the set of all possible values that the random variable can take.
The event space can be finite or infinite.

A random variable (or better the event space) and the data themselves can be discrete or continuous.

The distribution f(x) describing the random variable x is called probability density function (or pdf for short)

The probability for the random variable x to be in the interval [x, x+dx] is f(x)dx
NB: f(x) is a probability density (with dimensions $[x]^{-1}$), f(x)dx is a probability

Properties:

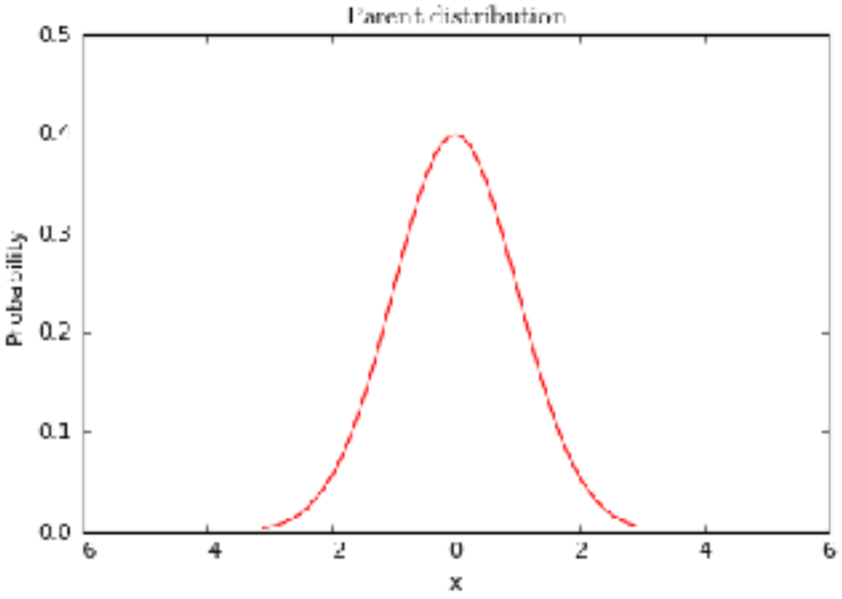
- f(x) ≥0 over the event space
- f(x) is normalized to unity over the event space $\int f(x)dx = 1$

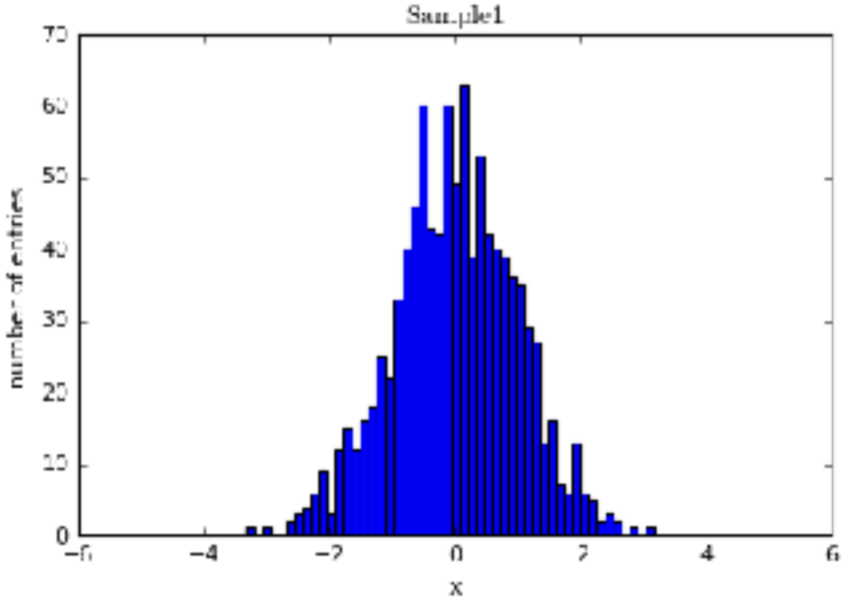
Parent vs. sampling distribution

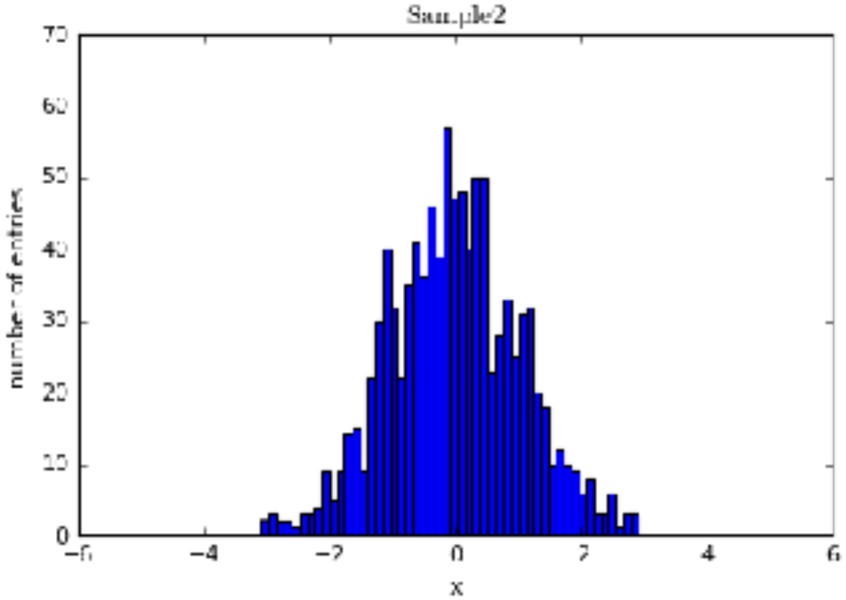
We will call:

“parent distribution” the (unknown) true distribution of the data
“sampling distribution” any dataset obtained by sampling the parent distribution

Different sampling of the same parent distribution will result in different datasets.
(another name for sample distribution is realization)







Characterize a dataset

Expectation value

The expectation value of a random variable x (or mean or first moment) is defined as:

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

continuous

$$< r >= \sum r_i P(r_i) = E[x]$$

discrete

i.e. it’s the probability-weighted average of all possible values.

Example: rolling a die E[x] = 1*1/6 + 2*1/6 + 3*1/6 + 4*1/6 + 5*1/6 + 6*1/6 = 3.5

Example: in quantum mechanics the eigenvalues (outcomes of a measurement) are weighted on their probability to occur

$$\langle A \rangle_\psi = \sum_j a_j |\langle \psi | \phi_j \rangle|^2$$

The expectation value is a linear operator:

$$< a \cdot g(x) + b \cdot h(x) >= a < g(x) > + b < h(x) >$$

Notice that $< fg > \neq < f >< g >$ unless the f and g are independent

Variance and standard deviation

The variance is a measure of the spread of the data around the mean μ of the pdf

$$V(x) =< (x-\mu)^2 >= \int_{x_{min}}^{x_{max}} (x-\mu)^2 f(x) dx =< x^2 > -\mu^2$$

A useful property: $V(a + bx) = b^2 V(x)$

an offset to the variable doesn't change how the data are distributed around their mean; scaling the variable by a constant b increases the variance by b^2

The standard deviation is defined as the square root of the variance $\sigma = \sqrt{V(x)}$

NB: these definitions rely on the knowledge of the parent distribution

Estimators for mean and variance

In general the true values of the mean and the standard deviation of the parent distribution are not known and they need to be estimated.

Un unbiased estimator of the true mean μ is given by the expectation value <x> (we will come back to the meaning of "unbiased" when we will talk about parameter estimation)

Given a dataset {Xi} i=1..N, the arithmetic mean of the dataset (or average) is:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

The mean of a function of the data is

$$\bar{f} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

When values are classified by their frequency (e.g. a histogram with m bins and n_j entries per bin) we can rewrite the mean as

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{m} n_j X_j$$

Estimators for mean and variance

When we know the true mean μ, the variance can then be estimated as:

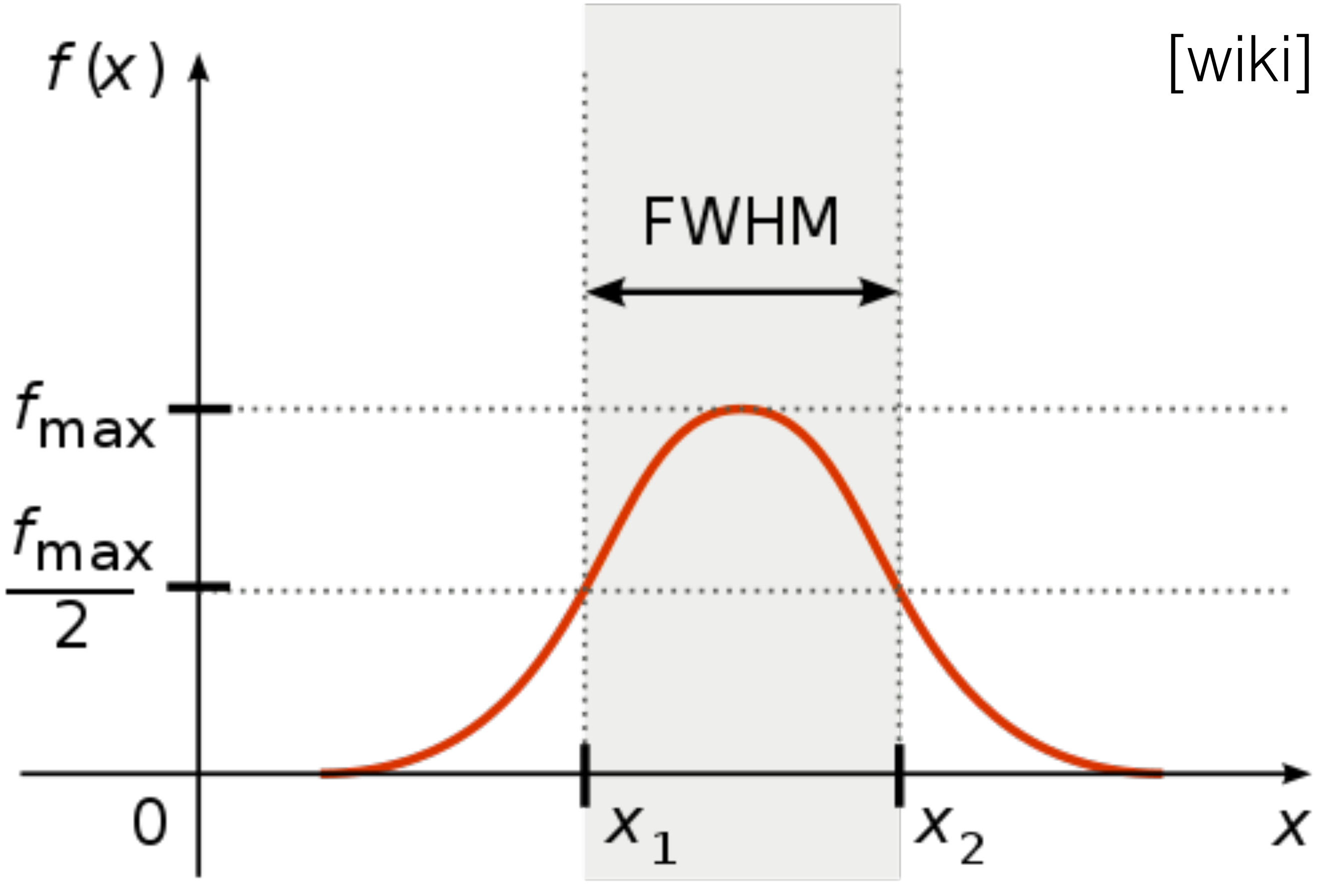
$$V(x) =< (x-\mu)^2 >$$

otherwise the true mean μ has to be replaced by the estimated one and the variance becomes:

$$s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2 = \frac{1}{N-1}\sum_i \left(x_i - \frac{1}{N}\sum_i x_i \right)^2$$

Full width at half maximum

Another way to characterize the spread of the data if to compute the FWHM.



[wiki]

For a gaussian distribution (see later) the relation between FWHM and the standard deviation is

$$\mathrm{FWHM} = 2\sqrt{2\ln 2}\;\sigma \approx 2.355\;\sigma.$$

Other means

Weighted mean: typically used to combine measurements with different resolutions given a dataset {Xi} i=1..N and their (event by event) uncertainties {σ_i} i=1..N, the weighted mean is defined as:

$$X_{weighted} = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

where $w_i = 1/\sigma^2_i$ is called the weight of the event.

Example: x = 1±0.1 ; y = 2±0.1 $x_{weitghted}$ = 1.5
x = 1±0.1 ; y = 2±0.05 $x_{weitghted}$ = 1.8
x = 1±0.1 ; y = 2±0.001 $x_{weitghted}$ = 1.99

Geometric mean: used to characterize the mean of a geometric sequence ($a, \ ar, \ ar^2, \ ar^3, \ ar^4, \ \ldots$) (defined only for values of the same sign)

$$\mu_g = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$$

The geometric mean of two numbers, a and b, is the length of one side of a square whose area is equal to the area of a rectangle with sides of lengths a and b.

Example: A population of bacteria grows from 2000 to 9000 in 3 days.
What is the daily grow (assuming a constant rate r) ? (same for interest rates in finance)

1st day: n1 = 2000 + 2000 r
2nd day: n2 = n1 + n1 r = 2000 $(1+r)^2$
3rd day: n3 = n2 + n2 r = 2000 $(1+r)^3$ = 9000 ⇒ $1 + r = \sqrt[3]{4.5} = 65.1\%$

Other means

$$\mu_h = \frac{N}{\sum_i 1/x_i}$$

Harmonic mean: used to characterize the mean value of a harmonic sequence ($\frac{1}{a}, \frac{1}{a+d}, \frac{1}{a+2d}, \frac{1}{a+3d}, \ldots$)

Example: A car travels at 80 km/h for the half of the trip and 100 km/h for the second half. What's his average speed? 2/(1/80+1/100) = 89 km/h
(here you are averaging over periods of time)

Moments summary

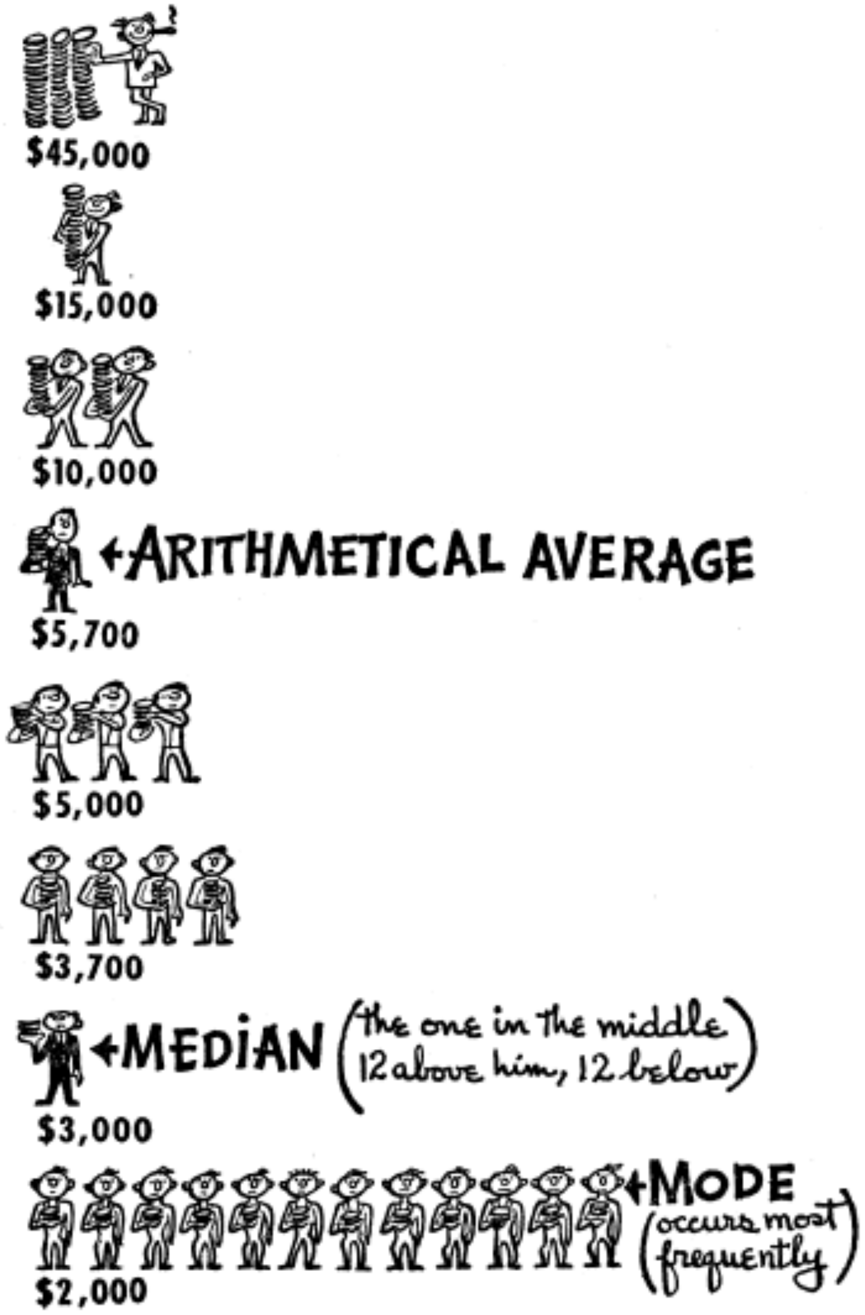
	Definition	Mechanics
sum / total probability	$P = \int_{-\infty}^{+\infty} f(x)dx$	total mass
mean / expectation value	$\mu = \int_{-\infty}^{+\infty} x f(x)dx$	centre of mass
variance	$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$	moment of inertia
skewness	$\gamma_1 = \int_{-\infty}^{+\infty} \frac{(x-\mu)^3}{\sigma^3} f(x)dx$	-
kurtosis	$Kurt[X] = \int_{-\infty}^{+\infty} \frac{(x-\mu)^4}{\sigma^4} f(x)dx - 3$	-

Other properties: median, most probable

The median (or 50% quantile) of a dataset is the value which splits the frequency distributions into two sized halves

$$\int_{-\infty}^{x_{median}} f(x')dx' = \int_{x_{median}}^{+\infty} f(x')dx' = 0.5$$

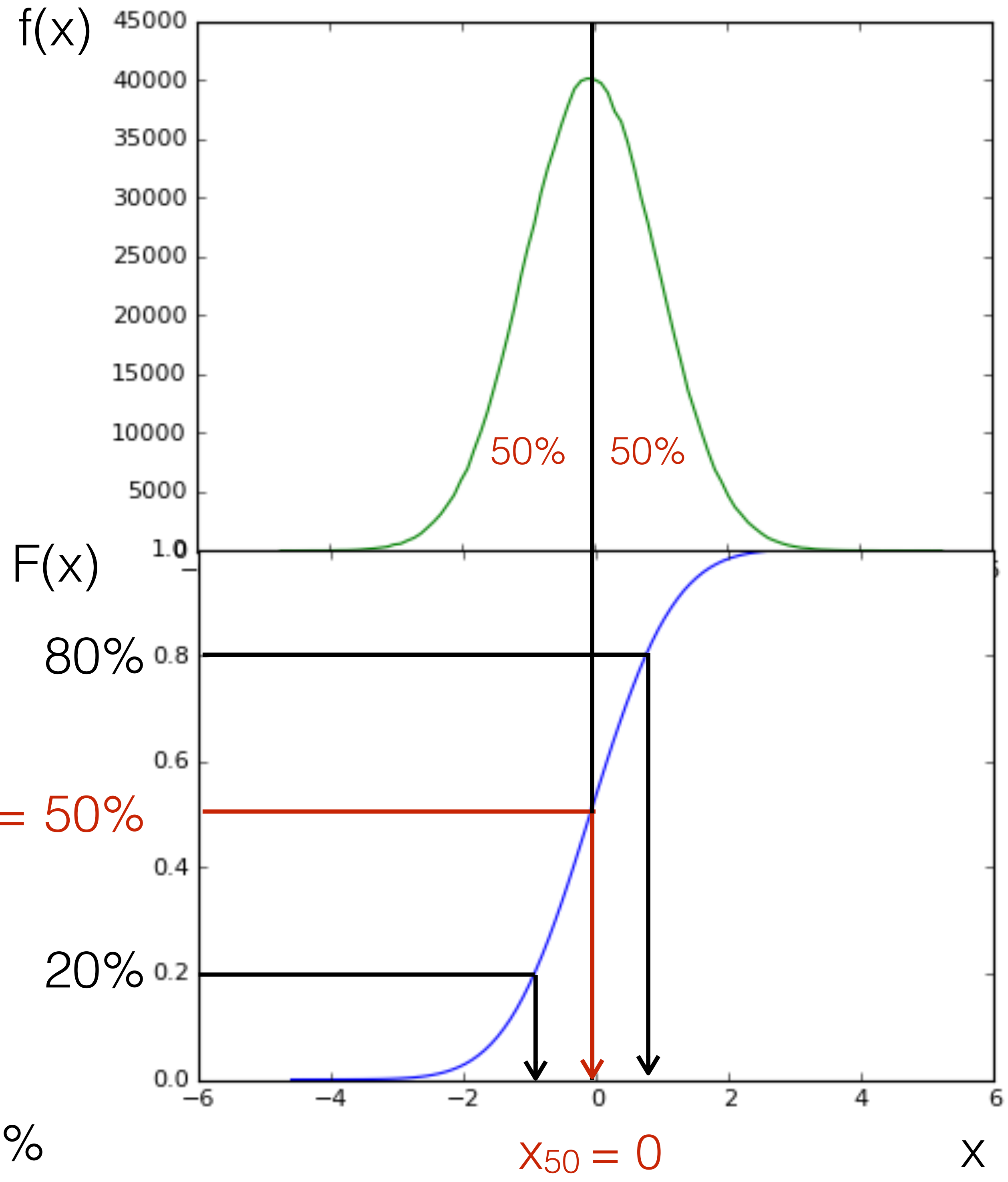
The most probable value (mode) of a dataset is the value that occurs more frequently


$45,000
$15,000
$10,000
←ARITHMETICAL AVERAGE
$5,700
$5,000
$3,700
←MEDIAN (the one in the middle 12 above him, 12 below)
$3,000
←MODE (occurs most frequently)
$2,000

Cumulative and quantiles

x_q = quantile of order q (or q-point), with $0 \leq q \leq 1$, of a distribution is the value of x such that $F(x_q) = q$

The quantile is just the inverse of the cdf
$x_q = F^{-1}(q)$



Quantiles with a name:
quartiles = 0%, 25%, 50%, 75%, 100%
percentiles = 0%,1%,2%,…, 98%, 99%, 100%

Error propagation (II)

Uncertainty on the mean

This is probably the most common case you will encounter.

Suppose you have n measurements of a single quantity x. You estimate the best value of x as the average of the measurements $\bar{x} = \sum_i x_i / N$

The difference between $\bar{x}$ and μ (the true *unknown* quantity) is, assuming the CLT, described by a gaussian distribution with variance:

$$\begin{aligned} Var(\bar{x}) &= \langle (\bar{x} - \mu)^2 \rangle \\ &= \langle \left(\frac{1}{n} \sum_i x_i - \mu \right)^2 \rangle \\ &= \frac{1}{n} n \langle x^2 \rangle + \frac{n(n-1)}{n^2} \langle x_i x_j \rangle_{i \neq j} - 2\mu \langle \bar{x} \rangle + \mu^2 \\ &= \frac{\langle x^2 \rangle}{n} + \frac{n-1}{n} \mu^2 - \mu^2 \\ &= \frac{\langle x^2 - \mu^2 \rangle}{n} = \frac{\sigma^2}{n} \end{aligned}$$

The standard deviation of the mean falls like 1/√n

This is the reason why if you want to improve the uncertainty on your measurement by a factor of 2 you need 4 times more statistics

Uncertainty on the mean

Example Take a photo-detector with an energy resolution of 50 keV. If a mono-energetic photon (coming e.g. from a certain nuclear decay) is registered, its energy is only known to a precision of 50 keV. If 100 (mono-energetic) photons are measured (all coming from the same nuclear decay), then the uncertainty of the mean energy is only $50/\sqrt{100} = 5$ keV. For a resolution of 1 keV we need 2500 events. So, to double the precision, you need four times more photons. □

Uncertainty on the weighted mean

Intuition: n measurements x_i with different uncertainties σ_i: the measurements with large uncertainties will “matter” less than measurements with small uncertainties.

Example You have two measurements of x: 10 ± 0.1 and 8 ± 5. In this case the second measurement will have basically no weight in your knowledge about x. □

The correct estimation of the uncertainty is obtained from the weighted average of the measurements:

$$\bar{x} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2} \qquad \sigma_{\bar{x}}^2 = \frac{1}{\sum 1/\sigma_i^2}$$

Example: compute the best estimate of the Higgs mass from:

- ATLAS: m_H = 125.36 ± 0.41 GeV
- CMS: m_H = 125.02 ± 0.30 GeV

We find: m_H = 125.14 ± 0.24 GeV.

Uncertainty on the weighted mean

Comments:

- weighted mean collapses to the arithmetic mean when all uncertainties are equal

- “Few measurements with small uncertainties are better than many measurements with large uncertainties”. Let the uncertainty of a first set of n_1 measurements of the quantity x be σ_1. The uncertainty on the mean is $\sigma_{\bar{x}} = \sigma_1/\sqrt{n_1}$. If we have a second set of n_2 measurements with uncertainty σ_2 and $\sigma_2 > \sigma_1$ then to get to the same precision you need to collect more data as:

$$n_2 = n_1 \left(\frac{\sigma_2}{\sigma_1}\right)^2$$

- must be taken with a grain of salt if the individual results and their uncertainty’s deviate too much from each other.

Example: An experiment measures in one hour 100 ± 10 events, and another experiment measures in one hour only 1 ± 1 events. From the weighted mean we would have 2 ± 1 events. But the (unweighted) mean would give 50.5±5.
Don’t blindly quote the mean or the weighted mean: go back and understand why you get such different outcomes (it might be a problem of some parameters of the data taking, some faulty equipment, some trivial mistake etc...).

In case you can’t find any reason for that, it would be wise to give the full information at hand and preset both results

Bibliography

Weighted averages

Taylor: Chapter 7