Notes de travail
E. Tomasi-Gustafsson
A. Dbeyssi
IPN, 15/V/2014

Notes on statistics

Let us build a gaussian distribution on N events, randomly generated over a gaussian with mean value M and width σ (Fig. 1).

$$f(x,\bar{x},\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x-\bar{x}}{2\sigma^2}} \tag{1}$$

If $\bar{x} = 0$ and $\sigma = 1$ it is called a normal distribution. The full width at half maximum is

$$\Gamma = 2\sqrt{2\ln 2}\sigma = 2.355\sigma. \tag{2}$$

For an infinite sample, respectively 68.3%, 95.5% and 99.7% of the events fall within a range of 1σ, 2σ, and 3σ around the mean value.

A. Mean value

As the distributions are symmetrical, and the generation is random, the mean is strictly defined as the limit

$$\mu = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i \simeq \bar{x} \tag{3}$$

It is equivalent to the centroid or average value of the quantity x.

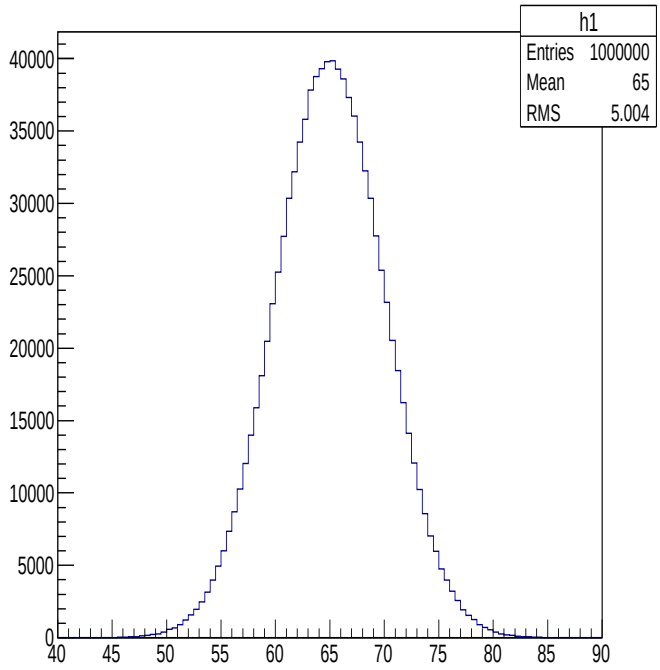

FIG. 1. Gaussian distribution.

B. Deviation

The deviation from a single measurement x_i from the mean is

$$d_i = x_i - \mu \tag{4}$$

The average of the deviations $\bar{d}$ for an infinite number of observations must vanish:

$$\lim_{N\to\infty} \bar{d} = \lim_{N\to\infty} \frac{1}{N} \sum x_i - \mu = 0 \tag{5}$$

In order to define a meaningful quantity, the average deviation has to be defined from absolute values:

$$\alpha = \lim_{N\to\infty} \left(\frac{1}{N} \sum |x_i - \mu| \right) \tag{6}$$

α is a measure of the dispersion of the expected observations about the mean. For statistical analysis purposes one define the variance: σ^2:

$$\sigma^2 = \lim_{N\to\infty} \left[\frac{1}{N} \sum (x_i - \mu)^2 \right] = \lim_{N\to\infty} \left(\frac{1}{N} \sum x_i^2 - \mu^2 \right) \tag{7}$$

and the standard deviation σ is the square root of the variance (root mean square of the deviations):

$$\sigma = \left[\frac{1}{N}\sum(x_i - \mu)^2\right] = \lim_{N\to\infty}\left(\frac{1}{N}\sum x_i^2 - \mu^2\right) \tag{8}$$

A better calculation of the variance is

$$\bar{\sigma}^2 = \sum \frac{(x_i - \bar{x})^2}{N-1} = \frac{\sum x_i^2 - 2\bar{x}\sum x_i + N\bar{x}^2}{N-1} = \frac{\sum x_i^2 - \bar{x}^2 \sum x_i/N}{N-1} \tag{9}$$

where we used eq. 3.

Note that:

- the median and the mean coincide if the distribution is simmetric.
- the mean specifies the probability distribution, has the same units as the true value and can be considered the best estimate of the true value.
- the variance and the standard deviation characterize the uncertainties associated with our experimental attempts to determine the true value
- the standard deviation is a measure of the uncertainty due to fluctuations in the observation.It gives the random error on an individual measurement. The error of the mean on several measurements is given below.

For the previous definitions we used a sample of N dimensions of quantities x_i. If we have a number M of these samples, we can calculate the individual mean values and standard deviations, as well as the general mean value. The question is how the stadard deviation of the mean is related to the standard deviations of the M individual samples. In case of gaussian distribution, the M mean values of the samples are also distributed along a gaussian, but with standard deviation σ^2/N.

The standard error on the mean is then:

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{N}}. \tag{10}$$

The uncertainty on the mean is given by the standard error of the mean and not the standard deviation. The larger is N, the estimate $\bar{x}$ becomes more and more precise.

C. Weighted mean

If the sets of measurements consists in data with their errors, the weighted mean is:

$$\bar{x} = \frac{\sum x_i/\sigma_i^2}{1/\sum 1/\sigma_i^2}, \tag{11}$$

and its error:

$$\sigma(\bar{x}) = \frac{1}{1/sum1/sigma_i^2}. \tag{12}$$

If the errors on the measurements are equal, Eqs. (11) and (12) reduce to Eqs. (3) and (10).

In a real experiment the statistics is limited. The physical information has to be extracted from one or a limited series of measurements.

When only one sample is available, $\sigma(\bar{x})$, Eq. 8 represents the precision of the instrument.

D. Application

Let us build 10 distributions $f_i, i = 1, 10$ randomly generated over a gaussian with the same mean value and width, but over 1000 events (Fig. 2). Let us collect the data (following Gaussian distributions) and compare the information from the different spectra.

The question is what is the reliability of the information from one of the ten spectra, i.e., compare the mean value μ_i and the standard deviation σ_i to μ and σ i.e., to a spectrum with "infinite" statistics.

The mean value of the histos of the means M_i and of the square deviations σ_i coincide with the mean and the square deviation of the distribution in Fig. 1.

In the present case, let us assume that we have ten independent measurements and the 'true values' being represented by the first gaussian ($\simeq N \to \infty$).

We have the numbers as in the table:

Note that the final result from the 10 combined sample of 1000 events is $\bar{x} \pm \sigma(\bar{x}) = 64.9869 \pm 5.039/\sqrt{(}1000)\sqrt{(}10) = 64.9869 \pm 0.05039$. to be compared to a gaussian hystogram of 10^4 events $\bar{x} \pm \sigma(\bar{x}) = 65.0332 \pm 0.0498$ and for 10^6 events $\bar{x} \pm \sigma(\bar{x}) = 65.000 \pm 0.005$.

E. Conclusions

- Collecting 10000 events distributed along a gaussian is strictly equivalent as composing 10 times 1000 events distributed along 10 gaussians from point of view of statistical

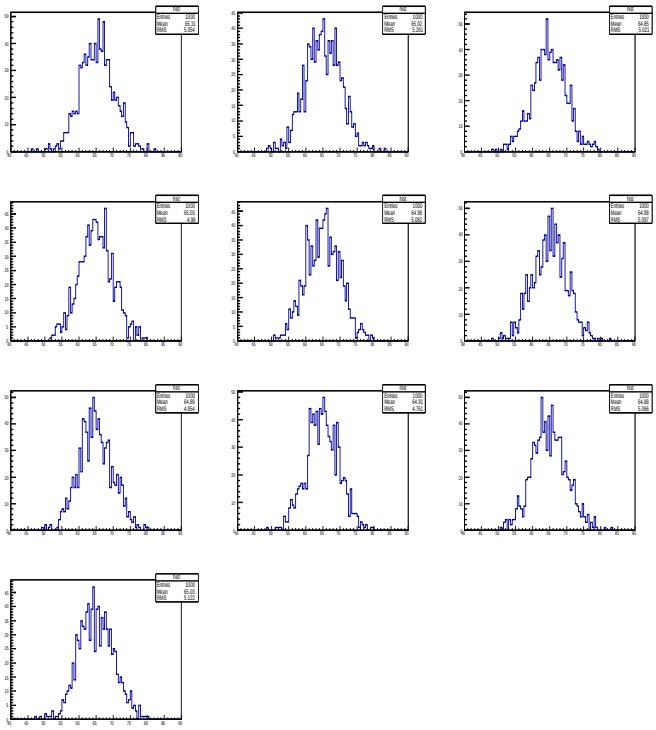

FIG. 2. Random gaussian distributions.

error

- The standard error of $\bar{x}$ is proportional to $1/\sqrt{N}$, that is, if one wishes to decrease the standard error by a factor of 10, one must increase the number of points in the sample by a factor of 100.

- The following considerations hold only in case of **gaussian, i.e., symmetric distributions**. Note that our previous studies on the FFs angular distribution showed that a quadratic fit (over R or R^2) gives symmetric distribution up to 8 GeV ^{2}, whereas a fit on A keeps symmetric up to $\simeq 14 GeV^2$.

N	$\bar{x}$	σ
1000000	65.0001	5.00427
10000	65.0332	4.98675
1000	65.3106	5.05357
	65.0171	5.26061
	64.8506	5.02085
	65.0288	4.9796
	64.9773	5.08238
	64.9798	5.09746
	64.893	4.95428
	64.9075	4.76148
	64.8765	5.06629
	65.0279	5.12246
	$\bar{x}$	$\bar{\sigma}$
	64.9869	5.039

TABLE I. Mean value $\bar{x}$ and deviation σ for samples of N events generated randomly along a gaussian distribution