Normalizing and fitting histograms in ROOT

March 19, 2021

In this exercise our starting random observable is the azimuthal angle φ whose sample space is the interval $[0, 2\pi)$ and whose probability density function (p.d.f.) can always be expanded in the Fourier series:

$$f(\varphi) = \frac{1}{2\pi}\left[1 + 2\sum_{n=1}^{\infty} v_n \cos[n(\varphi - \Psi_n)]\right]. \tag{1}$$

In the context of flow analyses, v_n are *anisotropic flow harmonics*, and Ψ_n corresponding *symmetry planes*. One of the most important objects in anisotropic flow analysis is the so-called Q-vector (also sometimes called flow vector) evaluated in harmonic n:

$$Q_n \equiv \sum_{i=1}^{M} e^{in\varphi_i}, \tag{2}$$

where M is total number of particles in an event (multiplicity). Closely related observable is q_n, the so-called modulus of reduced Q-vector:

$$q_n \equiv \frac{|Q_n|}{\sqrt{M}}. \tag{3}$$

The requirement for the factor $1/\sqrt{M}$ in definition (3) can easily be understood as follows. In case of a data sample consisting of uncorrelated (i.e. randomly sampled) particles, the modulus $|Q_n|$ grows as $\sqrt{M}$. In this case the Q-vector is nothing but the sum of random unit steps in a 2D plane, and the problem is completely equivalent to the famous "random walk problem in 2D", for which it is known that the distance from the origin grows as a $\sqrt{\text{number of steps}}$. In the context of the Q-vector definition, the "new step" is made by adding a new particle to the Q-vector, hence the total number of steps is the multiplicity M, and the "distance from the origin" is $|Q_n|$. This means that, as defined in Eq. (3), the modulus of reduced Q-vector, q_n, will not exhibit any trivial dependence on multiplicity, and as a consequence its distribution will not be systematically biased by trivial event-by-event multiplicity fluctuations, i.e. q_n is much cleaner observable than Q_n.

Question 1: *Please implement p.d.f. from Eq. (1) as a ROOT's TF1 object in a ROOT's standalone macro. Assume for simplicity that all harmonics except elliptic flow v_2 are zero, and treat v_2 and Ψ_2 as parameters in your implementation (i.e. the only variable is azimuthal angle). Make 10000 events 'on-the-fly' in computer's memory with fixed input anisotropic flow of $v_2 = 0.05$ in each event and random orientation of Ψ_2 in the interval $[0, 2\pi)$ in each event. Sample in each event from such p.d.f. 500 particles, calculate modulus q_n of reduced Q-vector in each event, and fill the histogram.*

At the end of the day, you have obtained a histogram with the very nice distribution of q_n. In this exercise, you will learn how to normalize and fit such distributions in ROOT manually (i.e. not using some predefined member functions of histogram classes). In general, if n is total number of entries in a histogram, n_i number of entries in the ith bin, and Δx_i is the width of the ith bin, then the histogram can be normalized with the following transformation:

$$n_i \to \frac{n_i}{n\Delta x_i}, \tag{4}$$

since then the histogram area is

$$\text{area} = \sum_i n_i \Delta x_i \to \sum_i \frac{n_i}{n\Delta x_i}\Delta x_i = \sum_i \frac{n_i}{n} = \frac{n}{n} = 1. \tag{5}$$

This procedure works also if histogram bins do not have equal width.

Question 2: *Please implement the code snippet in ROOT which will normalize the histogram with q_n distribution obtained in previous step.*

Now let's see how we can fit that distribution in ROOT. To leading order, the distribution of q_n can be fitted with the following Bessel-Gaussian p.d.f.

$$f(q_n) = \frac{q_n}{\sigma_n^2} \exp\left(-\frac{v_n^2 M + q_n^2}{2\sigma_n^2}\right) I_0\left(\frac{q_n v_n \sqrt{M}}{\sigma_n^2}\right) . \tag{6}$$

In the above equation, I_0 is a modified Bessel function of the first kind, and M is the multiplicity. We see that the flow harmonic v_n appears as one of the parameters in the above expression, and experimentally can be obtained by fitting the measured $f(q_n)$ distribution. Another parameter appearing in Eq. (6) is σ_n^2, which quantifies various sources of fluctuations and systematic biases. In the ideal case, when only flow correlations are present in a data sample, $\sigma_n^2 = \frac{1}{2}$.

Question 3: *Please fit the histogram holding the normalized distribution of q_n with Bessel-Gaussian p.d.f. in Eq. (6), by treating only q_n as variable, and all other quantities (v_2, σ_2) as parameters. Since M is constant, you could fit for the product $v_n^2 M$, and then just divide out the constant M value. What is the value for parameter v_2 obtained from the fit, and its error? Do you reproduce the fixed input value of v_2? What is the χ^2 of this fit?*

Hint: Implement Bessel-Gaussian p.d.f. as TF1 and then inspect how the member function Fit(...) of histogram class works.

Finally, to make it sure that the procedure to transfer and interpret any histogram as probability density function (p.d.f.) is completely under control, there is the final exercise. In this exercise we demonstrate in ROOT how the correspondence between histograms and p.d.f.'s can be established, both for histograms of equal and non-equal width. As a concrete example, we consider the following univariate p.d.f. defined over the interval $[0, 100)$:

$$f(x) \equiv \frac{3x}{5000}\left(1 - \frac{x}{100}\right) . \tag{7}$$

The prefactor 3/5000 is just a normalization constant, so that:

$$\int_0^{100} f(x)\, dx = 1 . \tag{8}$$

Question 4: *Book in ROOT two TH1F histograms in the interval [0,100), but:*
a) the first histogram has 100 bins with the same width;
b) the second histogram has 10 bins with the following boundaries 0.,20.,30.,40.,45.,50.,55.,60.,70.,80.,100.
Implement p.d.f. from Eq. (7) in ROOT as a TF1 object, sample from it 10M entries, and fill the sampled values in both histograms booked in a) and b).

The main question now is how to cross-check whether the values in the both histograms describe correctly the theoretical p.d.f. from Eq. (7). Since the p.d.f. is normalized, in order to make the comparison, we also need to normalize the histograms. For that sake, please follow the prescription from Eqs. (4) and (5).

Question 5: *Demonstrate that both histograms correctly describe the theoretical input p.d.f. For instance, introduce two canvases, on the first canvas plot the normalized histogram from a) and the TF1 object corresponding to p.d.f. in Eq. (7) on top of it, and analogously on the second canvas for the normalized histogram from b).*

Hint 1: In order to get total number of entries of histogram (also when non-unit weights are used), use

```
Double_t nEntries = hist->Integral();
```

Hint 2: In order to get the histogram area (a.k.a. 'normalization constant'), use:

```
Double_t norm = hist->Integral("width");
```

In experiment we do the opposite: We typically do not know what is the p.d.f. of the problem in question, but from the measurement (i.e. from the obtained histogram), and after normalizing it, we can fit and get the p.d.f. Ideally, we have an idea for the functional form of p.d.f. based on some underlying physics, and then from the fit we just extract the unknown parameters. If we are completely blank about the underlying physics, as the last resort we can always perform the polynomial fit.