Statistical independence, marginal distributions, and cumulants from hell

March 21, 2021

In general, two random observables x and y, with sample spaces X and Y, are said to be statistically independent if their joint 2-variate p.d.f. $f_{xy}(x,y)$ fully factorizes into the product of two marginal p.d.f.'s $f_x(x)$ and $f_y(y)$:

$$f_{xy}(x,y) = f_x(x) f_y(y)\,. \tag{1}$$

In the above equation, $f_x(x)$ and $f_y(y)$ are two marginal distributions defined as:

$$\begin{aligned} f_x(x) &= \int_Y f_{xy}(x,y)\,dy\,, \\ f_y(y) &= \int_X f_{xy}(x,y)\,dx\,. \end{aligned} \tag{2}$$

In general, $f_x(x)$ and $f_y(y)$ can have completely different functional forms.

Question 1: *Consider the following 2-variate p.d.f.'s: a)* $f_{xy}(x,y) = e^{-(x-4)^2}\log y$*, and b)* $f_{xy}(x,y) = e^{-(x-4)^2} + \log y$*. Demonstrate that in the first case observables* x *and* y *are statistically independent, and that in the second case they are not statistically independent.*

Hint #1: Implement in ROOT 2-variate p.d.f. as TF2 object, and sample observables x and y simultaneously from it. With all sampled pairs fill some 2D histogram (see e.g. the class TH2F). Normalize that 2D histogram, and then obtain two marginal histograms by making projections along x and y axis, i.e. schematically:

```
TH1F *marginalX = (TH1F*)f2DhistNormalized->ProjectionX();
TH1F *marginalY = (TH1F*)f2DhistNormalized->ProjectionY();
```

Note that if the starting 2D histogram is normalized, the two 1D marginal histograms obtained from it are automatically normalized. Finally, inspect a 2D ratio (e.g. introduce the new TH2F object for the ratio), between the starting 2D histogram, and the product of the two 1D marginal histograms. If this ratio is 1 everywhere, x and y are statistically independent.

Hint #2: In order to sample 2 observables from 2-variate p.d.f. simultaneously, the following code snippet might be helpful:

```
Double_t x = 0., y = 0.;
for(Int_t n = 0; n < nSamplings; n++)
{
 f2D->GetRandom2(x,y); // f2D was declared as TF2 object
 f2Dhist->Fill(x,y); // f2Dhist was declared as TH2F object
}
```

This procedure can be straightforwardly generalized for the 3D case (see the classes TF3 and TH3F). On the other hand, going to 4D or even higher dimensions is not that straightforward in ROOT. While the general n-dimensional histogram can be found in the class THnSparse (see `https://root.cern.ch/doc/master/classTHnSparse.html`), the new implementation of multidimensional sampling which would correspond to the general TFN object requires some (non-trivial) work.

Hint #3: In order to check whether a histogram is normalized or not, check out the value of:

```
hist->Integral();
```

This member function is the same also for 2D and 3D case.

By building up on top of this example, in real life for any 2D histogram obtained in an experiment we can deduce straightforwardly whether it was filled with two statistically independent or two correlated observables.

Cumulants from hell

If two random observables are not statistically independent (i.e. if they are correlated), without loss of generality we can always write:

$$f_{xy}(x,y) = f_x(x) f_y(y) + c_{xy}(x,y)\,. \tag{3}$$

Therefore, by construction any direct correlation between observables x and y is encoded solely in $c_{xy}(x,y)$, which is by definition the 2-particle cumulant, or the genuine 2-particle correlation. Cumulants cannot be measured directly, but from the above expressions we obtain trivially:

$$c_{xy}(x,y) = f_{xy}(x,y) - f_x(x) f_y(y)\,. \tag{4}$$

In terms of expectation values (averages), Eq. (4) reads:

$$\langle xy \rangle_c = \langle xy \rangle - \langle x \rangle \langle y \rangle\ . \tag{5}$$

This reasoning can be straightforwardly generalized for multivariate case. Without going into the derivation/details now, the final results for the relevant 2D and 3D cumulants, for a system consisting of 3 random observables x, y and z, are given by the following formulas:

$$\begin{aligned}
\langle xy \rangle_c &= \langle xy \rangle - \langle x \rangle \langle y \rangle \\
\langle xz \rangle_c &= \langle xz \rangle - \langle x \rangle \langle z \rangle \\
\langle yz \rangle_c &= \langle yz \rangle - \langle y \rangle \langle z \rangle \\
\langle xyz \rangle_c &= \langle xyz \rangle \\
&- \langle xy \rangle \langle z \rangle - \langle xz \rangle \langle y \rangle - \langle yz \rangle \langle x \rangle \\
&+ 2 \langle x \rangle \langle y \rangle \langle z \rangle
\end{aligned}$$

Question 2: *Please implement in ROOT the 3-variate p.d.f.* $f_{xyz}(x,y,z) \equiv e^{xy+xz+yz} = e^{xy} e^{xz} e^{yz}$ *by using the class TF3. Since this p.d.f. partially factorizes, it is clear that by design it describes only the genuine 2-particle correlations among x and y, x and z, and y and z, respectively. Demonstrate by developing the code in ROOT that all 2D cumulants in this example are not zero, while the 3D cumulant is zero.*

By building up on top of this example, in an experiment we can deduce whether there exists a genuine 3-body interaction in the system, which cannot be expressed as a superposition of lower two- and single-particle terms. Non-vanishing 3-particle cumulant undoubtedly points to the existence of genuine 3-body interaction.