DRAFT 0.0

Glen Cowan
4 May, 2021

Obtaining a pdf from MC

Suppose an MC model is able to generate a vector of data $\mathbf{x}$ for one "experiment", which can be summarized in a histogram $\mathbf{n} = (n_1, \ldots, n_N)$. The components of $\mathbf{x}$ could be the dilepton invariant mass values for each event and $\mathbf{n}$ is the corresponding histogram. Let the expected number of entries in the histogram be $\boldsymbol{\nu} = E[\mathbf{n}] = (\nu_1, \ldots, \nu_N)$; this is referred to as a "template".

Further, we define a test statistic $y(\mathbf{x})$ (or equivalently $y(\mathbf{n})$) that takes as input the outcome of the experiment and produces a single scalar value y. The goal is to determine the distribution of y under the assumption of different hypotheses (Standard Model, Clockwork Gravity, ...).

Suppose for the moment that we regard the data $\mathbf{x}$ as continuous and we work with it rather than the histogram $\mathbf{n}$. Formally we can write for the pdf of y

$$f(y|\boldsymbol{\mu}, \boldsymbol{\theta}) = \int \delta(y - y(\mathbf{x}))\, p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\theta})\, d\mathbf{x}\,, \tag{1}$$

where $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\theta})$ is the joint pdf for the data outcome $\mathbf{x}$ given values of parameters of interest $\boldsymbol{\mu}$ and nuisance parameters $\boldsymbol{\theta}$. Here $y(\mathbf{x})$ represents the statistic as a function of $\mathbf{x}$ and y written with no argument represents a given value that the function could take on. The parameters of interest could be, e.g., $\boldsymbol{\mu} = (k, M_5)$ of a Clockwork Gravity model and $\boldsymbol{\theta}$ could involve parameters related to the detector response.

Suppose that the nuisance parameters are characterized by a prior pdf $\pi(\boldsymbol{\theta})$. Then the distribution of y averaged with respect to this prior is

$$g(y|\boldsymbol{\mu}) = \int f(y|\boldsymbol{\mu}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int\int \delta(y - y(\mathbf{x}))\, p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\mathbf{x} \tag{2}$$

As a consequence of the average over $\boldsymbol{\theta}$, the pdf $g(y|\boldsymbol{\mu})$ will become broader than $f(y|\boldsymbol{\mu}, \boldsymbol{\theta})$, and the p-value that one finds from integrating the pdf to the right or left of a given y_{obs} is correspondingly larger.

If we want to treat more formally the fact that the histogram data are discrete, we can write

$$g(y|\boldsymbol{\mu})\, dy = \sum_{\mathbf{n} \in d\boldsymbol{\Omega}} \int P(\mathbf{n}|\boldsymbol{\mu}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{3}$$

where $d\boldsymbol{\Omega} = \{\mathbf{n} : y(\mathbf{n}) \in [y, y + dy]\}$. But this formal distinction is of no consequence when we determine the pdf using Monte Carlo.

In practice we do not use Eq. (2) or (3) explicitly, but rather we determine the pdf of y using Monte Carlo. First, we choose a point in $\boldsymbol{\mu}$ space to test, and then iterate the following steps:

1. Generate $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.
2. Determine the mean values for the mass distribution $\nu_i(\boldsymbol{\theta})$ for all bins $i = 1, \ldots, N$.
3. Generate independent Poisson distributed data $\mathbf{n}$ according to $n_i \sim \text{Poisson}(\nu_i(\boldsymbol{\theta}))$, i.e., one "toy experiment".
4. Evaluate the statistic $y(\mathbf{n})$, e.g., by computing the power spectrum of the mass distribution $\mathbf{n}$.
5. Record the value of y, e.g., in a histogram

By iterating these steps one builds up the distribution $g(y|\boldsymbol{\mu})$. This must be repeated for all points in $\boldsymbol{\mu}$ space that one wants to test.

An alternative procedure would be to sample $\boldsymbol{\theta}$ as in step 1 above but then to use this same $\boldsymbol{\theta}$ for some number of repetitions of steps 2 through 5. In the large sample limit this should also lead to the same pdf $g(y|\boldsymbol{\mu})$.

If one repeats steps 2 through 5 multiple times with the same $\boldsymbol{\theta}$, then the generated values of y are no longer statistically independent, and this will influence the statistical uncertainty of estimated p-values. For example, suppose one finds 100 values of y with $y \geq y_{\text{obs}}$ out of 2000 toy experiments and thus the estimated p-value is 0.05. If the y values are all independent, then this has a relative statistical uncertainty of $\sim 1/\sqrt{100} = 10\%$, and on this basis one might decide that 2000 toy experiments is enough. But if the y values are not independent, then the relative statistical error on the p-value will be larger and one might want more MC data.