DRAFT 0.2

Glen Cowan
4 August 2017

Obtaining a test for new physics from a classifier

Statistical tests can be used in two distinct ways in HEP analyses: for classification (event selection) and for testing the properties of an entire sample events, e.g., a test of the hypothesis that it only contains background. This note examines the relationship between these two tasks. Similar results are derived in a more formal way in Ref. [1].

1 Testing each event (signal selection)

Consider two types of events, signal (s) and background (b), and suppose that the set of numbers $\mathbf{x}$ that we can measure for each event follows either $p(\mathbf{x}|s)$ or $p(\mathbf{x}|b)$ depending on its type. Suppose we want to find the critical region w of a test with a given size α of the hypothesis that the event is of type b. From the Neyman-Pearson lemma, we know that the boundary of w that gives the maximum power with respect to the hypothesis of type s is given by a surface of constant likelihood ratio

$$r(\mathbf{x}) = \frac{p(\mathbf{x}|s)}{p(\mathbf{x}|b)} \ . \tag{1}$$

Once we have the statistic $r(\mathbf{x})$, it can be used to test individual events to classify them as signal or background. For example, if for an individual event we test the hypothesis that it is of type b, then the critical region w of the test is chosen to be the region $r(\mathbf{x}) \geq c_\alpha$, where the constant c_α is adjusted to give the desired size α.

Thus all of the information about whether an event is accepted or rejected by the test is contained in the scalar value $r(\mathbf{x})$, and once this function has been fixed we can determine in principle its distributions $p(r|s)$ and $p(r|b)$. In this way the multi-dimensional problem in $\mathbf{x}$-space is reduced to a single dimensional problem in r-space.

Furthermore, if we use a monotonic function of $r(\mathbf{x})$, say, $y(r)$, as a test statistic, then this must lead to the same critical region, since the region of data space where $r(\mathbf{x}) \geq c_\alpha$ is the same as the region where $y(r(\mathbf{x})) \geq y(c_\alpha)$ (here and in the following we take $y(r)$ to be monotonically increasing).

In general we do not have formulas for the pdfs $p(\mathbf{x}|s)$ and $p(\mathbf{x}|b)$, so we are unable to evaluate the likelihood ratio $r(\mathbf{x})$ at an arbitrary point in $\mathbf{x}$-space. Instead we have Monte Carlo models that allow us to generate events of both types. These events can be used as training data to determine a function that approximates $r(\mathbf{x})$ or a monotonic function thereof that we will write as $y(\mathbf{x})$; in practice this could for example be the output from a multivariate algorithm such as a neural network.

In a manner equivalent to using a critical region, we can define p-values of the s and b hypotheses using the test statistic as the boundary of the regions deemed to have equal or lesser compatibility. For example, the p-value of the b hypothesis for a given event is

$$p_b = P(y(\mathbf{x}) \geq y(\mathbf{x}_{\text{obs}})|b) \ , \tag{2}$$

and the critical region of a test of size α is simply the region of data space that would result in a p-value of α or less.

If the test results in rejecting the background hypothesis then we may choose to accept the event as signal. The probabilities to reject/accept the background hypothesis are

$$P(\mathbf{x} \in w|b) = \varepsilon_b = \alpha \tag{3}$$

$$P(\mathbf{x} \in w|s) = \varepsilon_s = M_s \; . \tag{4}$$

That is, the background efficiency ε_b is the same as the size of the test α and the signal efficiency ε_s is the power of the test with respect to the signal hypothesis, M_s.

If the original sample contains a mixture of signal and background with relative abundances π_s and π_b (the prior probabilities), then the purity of the selected signal sample is found from Bayes' theorem as

$$P(s|\mathbf{x} \in w) = \frac{P(\mathbf{x} \in w|s)\pi_s}{P(\mathbf{x} \in w|s)\pi_s + P(\mathbf{x} \in w|b)\pi_b} \; . \tag{5}$$

2 Test for presence of signal

Suppose one has a sample of n events, each of which is characterized by a set of numbers $\mathbf{x}$, i.e., the data consist of the values $\mathbf{x}_1, \ldots, \mathbf{x}_n$. From these data we want to test the hypothesis:

$$H_0 : \text{all events are of the background type}$$

versus the alternative

$$H_1 : \text{the event sample contains a mixture of signal and background} \; .$$

Let us suppose that the number of events n follows a Poisson distribution with a mean value $\mu s + b$, i.e.,

$$P(n|\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \; , \tag{6}$$

where b here refers to the expected number of background events and s is the expected number of signal events from some nominal model.[1]

Further we will suppose that values $\mathbf{x}$ measured for each event follow the pdfs $p(\mathbf{x}|s)$ and $p(\mathbf{x}|b)$. For a hypothesized signal strength μ the pdf is a mixture,

$$p(\mathbf{x}|\mu) = \frac{\mu s}{\mu s + b} p(\mathbf{x}|s) + \frac{b}{\mu s + b} p(\mathbf{x}|b) \; . \tag{7}$$

From the full event sample of n events the likelihood is thus the Poisson probability to find n events multiplied by the joint pdf for the values $\mathbf{x}_1, \ldots, \mathbf{x}_n$, (the extended likelihood):

[1] That is, with little danger of ambiguity we take s and b to be both labels for the signal and background processes as well as the expected numbers of signal and background events.

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \prod_{i=1}^{n} \left[\frac{\mu s}{\mu s + b} p(\mathbf{x}_i|s) + \frac{b}{\mu s + b} p(\mathbf{x}_i|b) \right] . \quad (8)$$

Here we will take the expected event numbers s and b to be known and the strength parameter μ is thus the parameter whose value we want to test. If we reject $\mu = 0$ (the background-only hypothesis), then we call this "discovery" of the signal process. Even in the absence of a discovery we can test nonzero values of μ and report the results as a confidence interval.

Note that taking H_1 to mean that some events of signal type are simply mixed in with the background events is not the most general alternative hypothesis. In general a new physics process may result in altered probabilities for n and $\mathbf{x}$ for all events. But the idea of a mixture of signal and background is a good approximation in many searches for new signal processes and we will assume here that it is a valid one.

Suppose when we test H_0 we want to have a maximum power with respect to the alternative of some nonzero value of μ. The Neyman-Pearson lemma says that the boundary of the optimal critical region is a surface of constant likelihood ratio,

$$\frac{L(\mu)}{L(0)} = e^{-\mu s} \prod_{i=1}^{n} \left(1 + \frac{\mu s}{b} \frac{p(\mathbf{x}_i|s)}{p(\mathbf{x}_i|b)} \right) , \quad (9)$$

or equivalently of the its logarithm,

$$\ln \frac{L(\mu)}{L(0)} = -\mu s + \sum_{i=1}^{n} \ln \left(1 + \frac{\mu s}{b} \frac{p(\mathbf{x}_i|s)}{p(\mathbf{x}_i|b)} \right) . \quad (10)$$

As the term $-\mu s$ does not depend on the data it can be dropped and we can define the critical region using the test statistic

$$Q = \sum_{i=1}^{n} \ln \left(1 + \frac{\mu s}{b} \frac{p(\mathbf{x}_i|s)}{p(\mathbf{x}_i|b)} \right) . \quad (11)$$

The test statistic Q as defined here is not, however, directly usable as we do not in general have the joint pdfs $p(\mathbf{x}|s)$ and $p(\mathbf{x}|b)$, so we cannot evaluate it for arbitrary $\mathbf{x}$. Rather, we have Monte Carlo models that can be used to generate events according to the two hypotheses and these can be used to train a classifier function $y(\mathbf{x})$ using, e.g., a neural network or boosted decision tree. As noted above, for an optimal classification of events one would like $y(\mathbf{x})$ to be a monotonic function of the likelihood ratio $r(\mathbf{x}) = p(\mathbf{x}|s)/p(\mathbf{x}|b)$.

The pdf $p(r)$ of the likelihood ratio $r(\mathbf{x})$ is related to that of the data $\mathbf{x}$ by considering a region ω of $\mathbf{x}$-space inside which the likelihood ratio takes on values in an interval $[r, r + dr]$. Integrating the pdf of $\mathbf{x}$ over this region thus gives $p(r)dr$. Doing this for both the signal and background distributions gives

$$p(r|s)\, dr = \int_{\omega} p(\mathbf{x}|s)\, d\mathbf{x} , \quad (12)$$

$$p(r|b)\, dr = \int_{\omega} p(\mathbf{x}|b)\, d\mathbf{x} . \quad (13)$$

The ratio of pdfs of r for s and b events is therefore

$$\frac{p(r|s)}{p(r|b)} = \frac{\int_\omega p(\mathbf{x}|s)\, d\mathbf{x}}{\int_\omega p(\mathbf{x}|b)\, d\mathbf{x}} \ . \tag{14}$$

In the numerator of the right-hand side we can substitute $p(\mathbf{x}|s) = p(\mathbf{x}|b)r(\mathbf{x})$. Furthermore in the infinitesial region ω where r is found in the interval $[r, r+dr]$, $r(\mathbf{x})$ is constant and can be pulled outside of the integral. The integrals in numerator and denominator then cancel and one finds

$$\frac{p(r|s)}{p(r|b)} = r(\mathbf{x}) = \frac{p(\mathbf{x}|s)}{p(\mathbf{x}|b)} \ . \tag{15}$$

In this way we can rewrite the test statistic Q from Eq. (11) using the ratio of pdfs for the likelihood ratio $r(\mathbf{x})$:

$$Q = \sum_{i=1}^{n} \ln \left(1 + \frac{\mu s}{b} \frac{p(r_i|s)}{p(r_i|b)} \right) \ . \tag{16}$$

Recall that the classifier $y(\mathbf{x})$ that gives the optimal performance will be a monotonic function of the likelihood ratio $r(\mathbf{x})$. If $y(r)$ is monotonic then the pdf of y is related to that of r (here using the argument to label the function) by

$$p(y) = p(r)\,|J| \ , \tag{17}$$

where the absolute value of the Jacobian is $|J| = |dr/dy|$, and with similar formulas holding for both signal and background events. Therefore the Jacobian factor will cancel in a ratio of probabilities and we have

$$\frac{p(y|s)}{p(y|b)} = \frac{p(r|s)}{p(r|b)} \ . \tag{18}$$

Thus we can write the statistic Q in terms of the pdfs of a statistic $y(\mathbf{x})$ that is a monotonic function of the likelihood ratio r as

$$Q = \sum_{i=1}^{n} \ln \left(1 + \frac{\mu s}{b} \frac{p(y_i|s)}{p(y_i|b)} \right) \ . \tag{19}$$

This result (also derived in Ref. [1]) shows that we can obtain a test statistic usable for the entire set of events by first training a multivariate classifier $y(\mathbf{x})$ to separate signal and background events. By doing this to obtain optimal separation (in the Neyman-Pearson sense) between the two event types, $y(\mathbf{x})$ must turn out to be a monotonic function of the likelihood ratio $r(\mathbf{x})$. Once such a $y(\mathbf{x})$ is found, Monte Carlo can be used to find the pdfs $p(y|s)$ and $p(y|b)$. And from these one can determine the value of the statistic Q using Eq. (19), with which we can carry out tests of different values of μ.

3 Distribution of the test statistic

To carry out a test using Q we need to know how this statistic is distributed under assumption of different values of μ, i.e., $f(Q|\mu)$, and in particular to test H_0 we need $f(Q|0)$. Figure 1 shows schematically the distribution of Q under $\mu = 0$ and $\mu = 1$. The vertical line in the plot illustrates the single value of the static observed from the real experiment, $Q_{\rm obs}$.

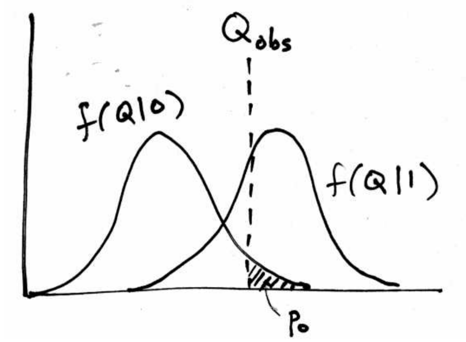

Figure 1: Distributions of the statistic Q assuming $\mu = 0$ and $\mu = 1$. The vertical line indicates the observed value $Q_{\rm obs}$ and the shaded area indicates the p-value of the $\mu = 0$ hypothesis.

The p-value of H_0, p_0, is given by the area under $f(Q|0)$ to the right of $Q_{\rm obs}$, i.e.,

$$p_0 = \int_{Q_{\rm obs}}^{\infty} f(Q|0)\, dQ \ . \tag{20}$$

The important point about the test statistic Q defined in Eq. (11) is that it requires only the values s and b and the distribution of the test statistic y. That is, we can construct the test for discovery by solving the problem of event classification, which defines the statistic $y(\mathbf{x})$, and we then use Monte Carlo to determine the pdfs $p(y|s)$ and $p(y|b)$. Then from the total cross sections of signal and background processes we can find the expected total numbers of signal and background events s and b, and in this way one can find for the observed event sample the value of Q.

To obtain p-values one needs the distribution of Q and finding this can be a challenging task. In the way it has been defined above one sees that Q is a sum of terms each of which follows the same pdf as determined from $p(\mathbf{x}|\mu)$. Therefore the pdf of Q is the convolution of the pdfs of the individual terms, and one may use Fourier transform methods to find $f(Q|\mu)$, as done in Ref. [2]. But in general finding this pdf can be difficult, especially when the problem includes nuisance parameters corresponding to systematic uncertainties.

4 Binned analysis

As above we assume that data $\mathbf{x}$ can be generated according to the signal or background processes, and by evaluating the statistic $y(\mathbf{x})$ with the resulting events one can construct a histogram with N bins of y for both the s and b hypotheses. We can find in this way, for a data sample of a given size (integrated luminosity) the expected number of events in the ith bin; suppose this is s_i for signal and b_i for background, where $i = 1, \ldots, N$.

For a given value of the strength parameter μ we can model the observed number n_i of events in bin i observed from the entire data set as following a Poisson distribution with a mean value $E[n_i] = \mu s_i + b_i$. As the bins are independent, the joint probability to observe the entire histogram is simply the product of the Poisson probabilities, and thus we find the likelihood function

$$L(\mu) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)} \; . \tag{21}$$

The likelihood ratio $L(\mu)/L(0)$ is therefore

$$\frac{L(\mu)}{L(0)} = \prod_{i=1}^{N} \left(1 + \frac{\mu s_i}{b_i} \right)^{n_i} e^{-\mu s_i} \; . \tag{22}$$

As our test statistic Q we can use as before its logarithm,

$$\ln \frac{L(\mu)}{L(0)} = -\mu s + \sum_{i=1}^{N} n_i \ln \left(1 + \frac{\mu s_i}{b_i} \right) \; , \tag{23}$$

where above we used the fact that the total expected number of signal event is obtained by summing over the bins, i.e., $s = \sum_{i=1}^{N} s_i$. As before we an drop the constant term μs in the definition of the test statistic

$$Q = \sum_{i=1}^{N} n_i \ln \left(1 + \frac{\mu s_i}{b_i} \right) \; . \tag{24}$$

We can now show that the statistic Q obtained here from the binned analysis of Eq. (24) is equivalent to what was found above in Sec. 2, Eq. (11) in the limit that the bin size Δy goes to zero. In this case the expected numbers of signal and backgrounds in the ith bin can be approximated as

$$s_i \approx s \, p(y_i|s) \, \Delta y \; , \tag{25}$$

$$b_i \approx b \, p(y_i|b) \, \Delta y \; , \tag{26}$$

where y_i is the value of y in the ith bin. Furthermore as the number of bins N becomes large the number of events that one observes in any given bin is either 0 or 1, so the sum in Eq. (24) gives no contribution if $n_i = 0$ and it enters once if $n_i = 1$. Out of the N bins, only $n = \sum_{i=1}^{N} n_i$ of them make a nonzero contribution, and so we find

$$Q = \sum_{i=1}^{n} \ln \left(1 + \frac{\mu s}{b} \frac{p(y_i|s)}{p(y_i|b)} \right) \; . \tag{27}$$

Equation (27) for Q is thus the same as what was found in the unbinned case of Eq. (11).

References

[1] Kyle Cranmer, Juan Pavez, Gilles Louppe, *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*, eprint: arXiv:1506.02169 [stat.AP] (2015).

[2] Jason Nielsen and Hongbo Hu, *Analytic Confidence Level Calculations using the Likelihood Ratio and Fourier Transform*, eprint: arXiv:physics/9906010 [physics.data-an] (1999).

[3] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-62.