Clopper-Pearson Bounds from HEP Data Cuts

B. A. Berg *

Department of Physics, The Florida State University, Tallahassee, FL 32306, USA.
E-mail: `berg@hep.fsu.edu`

Abstract. For the measurement of N_s signals in N events rigorous confidence bounds on the true signal probability $p_{\rm exact}$ were established in a classical paper by Clopper and Pearson [Biometrica 26, 404 (1934)]. Here, their bounds are generalized to the HEP situation where cuts on the data tag signals with probability P_s and background data with likelihood $P_b < P_s$. In particular, the method may be of interest in connection with the statistical analysis part of the ongoing Higgs search at the LEP experiments. The relevant Fortran code is available on the web.

INTRODUCTION

The general theory of confidence bounds (or fiducial intervals) was developed by Fisher (1), Neyman and Pearson (2). We consider a particular problem which is of interests when cuts are used to analyze high energy physics data. Typically, a neural network or some other method of performing the cuts results in probabilities (efficiencies) to tag signals more likely than background events. For instance by means of Monte Carlo (MC) simulations, these probabilities can normally be calculated. Let P_s be the probability to tag a signal and P_b be the likelihood to tag a background event, $0 < P_b < P_s < 1$. Out of a total number of N data one gets in this way

$$N^Y \ \text{ tagged data.} \tag{1}$$

It is easy to find from this the mean expectation for the signal probability. Assume that there are N_s signals and N_b background events in the data. Then we have

$$N = N_s + N_b \quad \text{and} \quad N^Y = P_s N_s + P_b N_b$$

and these two equations solve for

$$p_{\rm mean} = \frac{N_s}{N} = \frac{N^Y - P_b N}{N\,(P_s - P_b)}\ . \tag{2}$$

The question is, what are the implied confidence limits on the signal probability?

The special case $P_s = 1$ and $P_b = 0$ (sure signal detection) has been treated by Clopper and Pearson (3) in 1934. After briefly reviewing their approach in the next section, we derive and illustrate the general case in the subsequent section. This is, in part, based on Ref.(4). In particular, the method is valid when the number of tagged data is small and returns the probability P_0 for the case that there is no signal, i.e. that the exact signal probability is $p_{\rm exact} = 0$. This is of interest for the statistical analysis of the ongoing Higgs search at LEP (5). Discovery of the Higgs particle on the 5σ level would mean $P_0 \le 0.287\,10^{-6}$. Conclusions follow in the final section .

The Fortran code which produces the illustrations of this paper is available on the web. Start at the author's homepage (6) `www.hep.fsu.edu/~berg` and follow the `research` and from there the `Clopper-Pearson` hyperlink. Its use is explained in Ref.(7).

THE CLOPPER–PEARSON CONFIDENCE LIMITS

Let p be the likelihood that a data point is a signal. For N measurements the probability to observe k signals is given by the binomial coefficient

$$b_N(k,p) = \binom{N}{k} p^k\, q^{N-k} = \frac{N!}{k!\,(N-k)!}\, p^k\, q^{N-k} \tag{3}$$

with $q = p - 1$. The probability to observe $k \ge N_s$ signals is given by

$$P_{k\ge N_s}(p) = \sum_{k=N_S}^{N} b_N(k,p) \tag{4}$$

and the probability to observe $k \le N_s$ signals by

$$P_{k\le N_s}(p) = \sum_{k=0}^{N} b_N(k,p)\ . \tag{5}$$

For $N = 26$ and $N_s = 10$ the functional forms of $P_{k\ge N_s}(p)$ and $P_{k\le N_s}(p)$ are depicted in figure 1.

* In part supported by the DOE Grant DE-FG02-97ER41022.

Assume that N_s signals are found in N measurements and that a probability $Q^c < 0.5$ (typical values are $Q^c = 0.16$ or $Q^c = 0.025$) is given. We can solve equation (4) for $P_{k\geq N_s}(p_-) = Q^c$ and p_- is a lower bound on the true signal probability p_{exact}, such that the likelihood to find $k \geq N_s$ signals in N measurement is smaller than Q^c for every $p_{\text{exact}} < p_-$. Correspondingly, we can solve equation (5) for $P_{k\leq N_s}(p_+) = Q^c$ and p_+ is an upper bound on the true signal probability p_{exact}, such that the likelihood to find $k \leq N_s$ signals in N measurement is smaller than Q^c for every $p_{\text{exact}} > p_+$. Together, this combines into the Clopper–Pearson bounds: The probability to find the true signal probability in the range

$$p_- \leq p_{\text{exact}} \leq p_+ \quad \text{is larger or equal to} \quad P^c = 1 - 2Q^c \,. \tag{6}$$

In more details the meaning of the inequality is discussed in (4). For $P^c = 0.68$ ($Q^c = 0.16$) the $p^\pm$ values are indicated in figure 1. Approximately, this range corresponds to the confidence of a 1σ error bar. Similarly the confidence range corresponding to a 2σ error bar, etc., can be found.

CONFIDENCE LIMITS FROM HEP DATA CUTS

We are interested in the situation where signal and background data can no longer be distinguished unambiguously. Instead, a neural network or other device yields statistical information by tagging signals with efficiency P_s and background data with probability P_b, as discussed in the introduction.

Applying the cuts to all N data results in N^Y tagged data ($0 \leq N^Y \leq N$), composed of $N^Y = N^Y_s + N^Y_b$, where

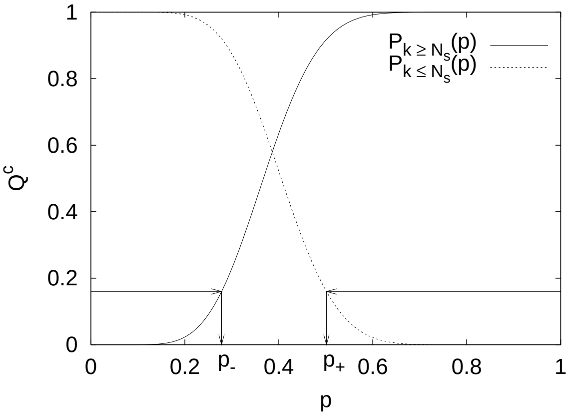

FIGURE 1. The probability functions $P_{k\geq N_s}(p)$ and $P_{k\geq N_s}(p)$ for $N = 26$ and $N_s = 10$ together with 68% confidence bounds.

N^Y_s is the number of tagged signals and N^Y_b is the number of tagged background data. Of course, the values for N^Y_s and N^Y_b are not known. Our task is to determine confidence limits for the signal probability p from the sole knowledge of N^Y. We proceed by writing down the probability density of N^Y for given p and, subsequently, generalizing the Clopper-Pearson method.

First, assume fixed N_s. The probability densities of N^Y_s and N^Y_b are binomial and thus the probability density for N^Y is given by the convolution

$$P^Y(N^Y|N_s) = \sum_{N^Y_s + N^Y_b = N^Y} b_{N_s}(N^Y_s, P_s)\, b_{N_b}(N^Y_b, P_b) \,. \tag{7}$$

Proof: For a signal event the probability to be tagged is P_s, so $b_{N_s}(N^Y_s, P_s)$ is the probability to tag N^Y_s out of the N_s signals. Similarly, the probability for a background event to become tagged is P_b and $b_{N_b}(N^Y_b, P_b)$ is the probability to tag N^Y_b of the $N_b = N - N_s$ background events. As these two probabilities are independent, the likelihood that precisely N^Y_s of the signals and N^Y_b of the background events are tagged becomes the product $b_{N_s}(N^Y_s, P_s)\, b_{N_b}(N^Y_b, P_b)$. Summing over all possibilities which add up to $N^Y_s + N^Y_b = N^Y$ gives the result.

Summing over N_s in (7) removes the constraint of fixed N_s and, with N, p fixed, the probability to tag k events becomes

$$b^Y_N(k, p) = \sum_{N_s=0}^{N} b_N(N_s, p)\, P^Y(k|N_s) \,. \tag{8}$$

Fourier transformation of the convolution (7) allows for an efficient numerical calculation of the $P(k|N_s)$ coefficients. In analogy with equations (4) and (5) we find the probabilities to tag $k \geq N^Y$ and $k \leq N^Y$ events to be

$$P^Y_{k\geq N^Y}(p) = \sum_{k=N^Y}^{N} b^Y_N(k, p) \quad \text{and} \quad P^Y_{k\leq N^Y}(p) = \sum_{k=0}^{N^Y} b^Y_N(k, p) \,. \tag{9}$$

For $N = 35$, $N^Y = 12$, $P_s = 0.8$ and $P_b = 0.05$ the functions $P^Y_{k\geq N^Y}(p)$ and $P^Y_{k\leq N^Y}(p)$ are depicted in figure 2. The 68% confidence range (6) is also indicated in figure 2, where the bound values $p_\pm$ are now defined as solutions of the equations

$$Q^c = P^Y_{k\geq N^Y}(p_-) \quad \text{and} \quad Q^c = P^Y_{k\leq N^Y}(p_+) \,. \tag{10}$$

The range $[p_-, p_+]$, obtained with $Q^c = 0.16$, guarantees the standard one error bar confidence probability of 68% for every true signal probability p_{exact}. For almost all values the actual confidence will be better. However, the bounds cannot be improved without violating the requested confidence probability for the case that p_{exact} happens to agree with either p_- or p_+. In the same way, bounds calculated with $Q^c = 0.023$ ensure the standard two error bar confidence level of 95.4% or better, and so on.

Data Sets with Few Signals

As outlined, data sets with few signals are of of particular interest in high energy physics. Let us replace $N^Y = 12$ of figure 2 by $N^Y = 3$. The resulting graph is depicted in figure 3. From the $P^Y_{k \geq N^Y}(p)$ curve we read off the finite probability $P_0 = 0.254$ for the likelihood that the true signal probability $p_{\text{exact}} = 0$ generates $k \geq N_y$ tags. Due to this probability the lower 68% confidence bound p_- disappears, whereas the upper p_+ bound does still exist. Let us note that for the data of figure 2 we have $P_0 = 0.69\,10^{-7}$, i.e. there $p_{\text{exact}} = 0$ is ruled out on the 5σ level.

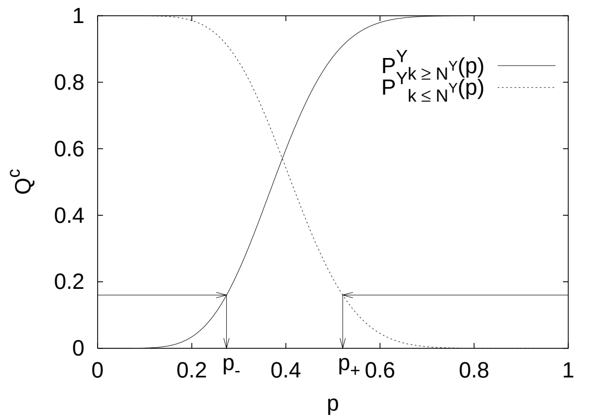

FIGURE 2. The probability functions $P^Y_{k \geq N_Y}(p)$ to find $k \geq N^Y$ tags in N events and $P^Y_{k \leq N^Y}(p)$ to find $k \leq N^Y$ tags in N events (9) are depicted for $N = 35$, $N^Y = 12$, $P_s = 0.8$ and $P_b = 0.05$. Symmetric 68% confidence bounds, found for $p_- = 0.274$ and $p_+ = 0.521$, are also indicated.

CONCLUSIONS

We have calculated confidence limits for an unknown true signal likelihood p_{exact}. The input used are the efficiency P_s for tagging signals, the probability P_b for tagging background events, the number N^Y of tagged data and the total number of data N. In particular, the method allows to deal with the situation where only few signals occur and yields then a finite probability for the likelihood that $k \geq N^Y$ tags are observed if the true signal probability is $p_{\text{exact}} = 0$. In real life the probabilities P_s and P_b are most likely estimators by themselves, i.e. quantities with error bars. This causes no major problem, one just has to apply our confidence calculations to an appropriate sample and to average over the results.

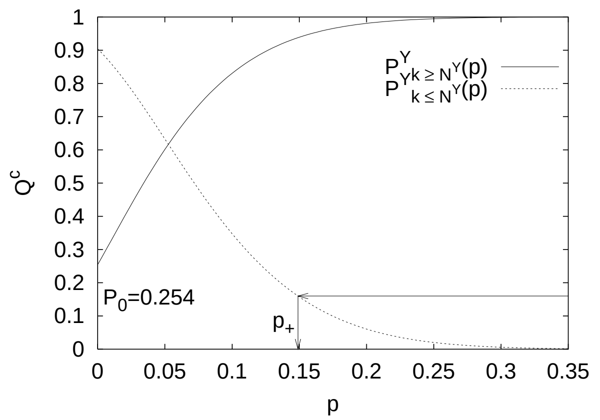

FIGURE 3. The same functions as in figure 2 are depicted, but for $N^Y = 3$ instead of $N^Y = 12$ tags. The lower 68% confidence bound does not exist anymore, instead the likelihood for $p_{\text{exact}} = 0$ has become $P_0 = 0.254$. The upper 68% confidence bound is found at $p_+ = 0.149$.

REFERENCES

1. R.A. Fisher, Proc. Camb. Phil. Soc. 26, (1930) 528; Proc. Roy. Soc. A139, (1933) 343.
2. J. Neyman and E.S. Pearson, Phil. Trans. Roy. Soc. A231, (1933) 289.
3. C.J. Clopper and E.S. Pearson, Biometrika 26, (1934) 404. For a textbook treatment see S. Brandt, *Statistical and Computational Methods in Data Analysis* (North-Holland, 1983).
4. B.A. Berg and J. Riedler, Comp. Phys. Commun. 107, (1997) 39.
5. `http://alephwww.cern.ch/ALPUB/seminar/wds/`
6. The address of the authors homepage and its tree structure are expected to be stable, whereas the absolute address where the programs are located is likely to change.
7. B.A. Berg, hep-ex/0010061.