Clopper-Pearson Bounds from HEP Data Cuts

Efficiencies:

P_s (to tag signals), P_b (to tag background),

N^Y tagged data.

Assume, N_s signals and N_b background data. We have

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

where N is the total number of data. The two equations solve for

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

What are the confidence limits, how does the signal distribution look like?

Clopper-Pearson Confidence Limits for Direct Signal Observation

(Biometrika 26 (1934) 404.)

Let p to be the (unkown) exact signal likelihood. The probability to observe k signals is given by the binomial coefficient

$$b_N(k,p) = \binom{N}{k} p^k (1-p)^{N-k} \ .$$

The probability to observe $k \geq N_s$ signals is

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

and to observe $k \leq N_s$ signals

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

For $N = 26$ and $N_s = 10$: Figure 1.

N^Y Tagged Data

(Berg and Riedler, Comp. Phys. Commun. 107 (1997) 39.)

Signal and background data can no longer be unambiguously distinguished. N_s fixed, the probability density of N^Y is given by the convolution

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

Summing over N_s, with N, p fixed, the probability to tag k events becomes

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

Fourier transformation of the convolution allows for an efficient calculation of the $P(k|N_s)$ coefficients. In analogy to the Clopper-Pearson case of deterministic signal detection, we find the probabilities to tag $k \geq N^Y$ and $k \leq N^Y$ events:

$$P_{k \geq N^Y}^Y(p) = \sum_{k=N^Y}^{N} b_N^Y(k,p)$$

and

$$P^Y_{k\leq N^Y}(p) = \sum_{k=0}^{N^Y} b^Y_N(k,p) \ .$$

For $N = 35$, $N^Y = 12$, $P_s = 0.8$ and $P_b = 0.05$: Figure 2.

Data Sets with Few Signals

$N^Y = 12 \to N^Y = 3$: Figure 3.

Signal Probability Distributions

The cumulative distribution function is

$$F(p) = \int_{-\infty}^{p} dp'\, f(p')$$

where $f(p)$ is the probability density. Bounds on $F(p)$ are

$$F_1(p) = 1 - P^Y_{k\leq N^Y}(p) = \sum_{k=N^Y+1}^{N} b^Y_N(k,p) \ \leq \ F(p)$$

and

$$F(p) \;\le\; F_2(p) = P^Y_{k\ge N^Y}(p) = \sum_{k=N^Y}^{N} b^Y(k,p)\,.$$

Figure 4 shows the result for the same data which were used in figure 2. It is instructive to define peaked distribution functions (Berg, to be published) by

$$F_i^{\text{peaked}}(p) = \begin{cases} F_i(p) & \text{for} \quad F_i(p) \le 0.5 \quad \text{and} \\ 1 - F_i(p) & \text{for} \quad F_i(p) \ge 0.5. \end{cases}$$

1. The ordinate becomes enlarged by a factor of two.
2. The estimated medians are located at the peaks and the probability content of the distribution is instructively displayed.
3. Probability bounds are easily read off.

Figure 5 for the same data as used in figure 4.

Probability densities follow from $f_i(p) = \frac{dF_i(p)}{dp}$: Figure 6. Their numerical calculation uses analytical expressions for the derivatives of the binomial coefficients.

At $p = 0$ the probability densities have δ-function contributions

$$P_i(p) \;=\; F_i(0)\,\delta(p) + \dots\,.$$

For figure 6 the $F_i(0)$ coefficient are practically zero, whereas for the probability densities corresponding to figure 3 there would be a substantial contribution: $F_1(0) = 0.096$ and $F_2(0) = 0.254$.

Conclusions

I have presented a method to calculate confidence limits, and limits of the entire cumulative signal distribution function, from the tagging efficiencies and the number N^Y of tagged data. In particular, the method allows to deal with few signals and yields then finite probability bounds for zero signal likelihood.