Glen Cowan
RHUL Physics
8 July, 2014

Statistics Problem for Invisibles 2014

Consider a Poisson counting experiment in which the number of observed events n is modeled as following

$$P(n|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \;, \tag{1}$$

where s and b are the expected numbers of events from signal and background processes, respectively. Here s is the parameter of interest, and to establish discovery of the signal process we want to test the hypothesis $s = 0$. Suppose the expected background b is not known exactly but rather treated as a nuisance parameter. We regard our best estimate of b as a measured quantity, $\tilde{b}$, which follows a Gaussian distribution with mean b and standard deviation $\sigma_{\tilde{b}}$,

$$f(\tilde{b}|b,\sigma_{\tilde{b}}) = \frac{1}{\sqrt{2\pi}\sigma_{\tilde{b}}} e^{-(\tilde{b}-b)/2\sigma_{\tilde{b}}^2} \;. \tag{2}$$

Assuming n and $\tilde{b}$ are statistically independent, the likelihood function is given by the product of the Poisson and Gaussian distributions above.

1(a) Show that the log-likelihood function is (up to an additive constant)

$$\ln L(s,b) = n \ln(s+b) - (s+b) - \frac{1}{2}\frac{(b-\hat{b})^2}{\hat{\sigma}_{\hat{b}}^2} \;. \tag{3}$$

1(b) To test a hypothetical value of s we can use the profile likelihood ratio (see, e.g., [1]),

$$\lambda(s) = \frac{L(s,\hat{\hat{b}}(s))}{L(\hat{s},\hat{b})} \;, \tag{4}$$

where the double hat notation indicates the value of b that maximizes L for the given value of s, and single hats denote the (unconditional) maximum-likelihood estimators. In particular we are interested in testing $s = 0$, so we need $\lambda(0)$, and for this we require $\hat{\hat{b}}(0)$. Show that the ingredients are

$$\hat{s} = \frac{n-\tilde{b}}{s} \;, \tag{5}$$

$$\hat{b} = \tilde{b} \;, \tag{6}$$

$$\hat{\hat{b}}(0) = \frac{\tilde{b}-\sigma_{\tilde{b}}^2}{2} + \frac{1}{2}\sqrt{(\tilde{b}-\sigma_{\tilde{b}}^2)^2 + 4\sigma_{\tilde{b}}^2 n} \tag{7}$$

1(c) Using these quantities one can then evaluate the statistic

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{s} > 0 \ , \\ \\ 0 & \text{otherwise} \ . \end{cases} \tag{8}$$

One can show (see, e.g., Ref. [1]) that in the large-sample limit the discovery significance Z approaches

$$Z = \sqrt{q_0} \ . \tag{9}$$

This is equivalent to the p-value of the $s = 0$ hypothesis; the two quantities are related by

$$Z = \Phi^{-1}(1 - p) \ , \tag{10}$$

where Φ^{-1} is the quantile of standard Gaussian.

Suppose $n = 12$, $\tilde{b} = 6.0$ and $\sigma_{\tilde{b}} = 1.0$ Assuming the validity of the large-sample formulae given above, find the p-value of the $s = 0$ hypothesis and the corresponding discovery significance Z.

1(d) Suppose the nominal signal model predicts $s = 5.0$. Using the "Asimov approximation" (here, $n \to s + b \to s + \tilde{b}$) find the expected (median) discovery significance. Note that here one tests $s = 0$, but the median refers to $s = 5.0$.

1(e) Write a short Monte Carlo program to generate data sets (n, $\tilde{b}$) according to Eqs. (1) and (2) above using $s = 0$, $b = 6.0$, $\sigma_{\tilde{b}} = 1.0$. For each data set, compute the statistic q_0 using Eq. (8) and enter into a histogram. Compare the form of the histogram to the expected Asymptotic distribution from the lectures or Ref. [1]. Here this should be a delta function at $q_0 = 0$ with weight of one half plus a chi-squared distribution for one degree of freedom with weight one half.

1(f) Suppose as above one observes $n = 12$ and $\tilde{b} = 6.0$. Use your Monte Carlo program to compute the p-value of $s = 0$, assuming $b = 6.0$. (Note that the result should be insensitive to the value of b; test this.) Compare the result the one obtained in 1(c) from the asymptotic formula.

References

[1] Glen Cowan, Kyle Cranmer, Eilam Gross and Ofer Vitells, Eur. Phys. J. C 71 (2011) 1554; arXiv:1007.1727.