3 Hypothesis testing

Gregory Schott

Exercise 3.1: Significance

a) The p-value, calculated from the Poisson tail probability, is

$$p = \sum_{n=56}^{+\infty} \frac{40^n}{40!} e^{-40} = 1 - \sum_{n=0}^{55} \frac{40^n}{40!} e^{-40} = 0.00968\,.$$

The corresponding Z-value is $Z = \sqrt{2} \cdot \mathrm{erf}^{-1}(1-2p) = 2.34$. The approximate estimates based on the formulae mentioned in section 3.5.1 yield $S/\sqrt{S+B} = 2.14$, $2(\sqrt{S+B} - \sqrt{B}) = 2.32$ and $\sqrt{2(S+B)\log(1+S/B) - 2S} = 2.38$, where $B = \nu_b = 40$ and $S = N - \nu_b = 16$. The first of those estimates, which is also by far the most commonly used, is rather crude, being about 10% too low, while the other two yield a reasonably close result to the true value in that specific application.

b) With a Gaussian uncertainty considered, the p-value can be calculated from equation (3.10) as:

$$p = 1 - \int_{-40}^{+\infty} \mathrm{d}b' \sum_{n=0}^{55} \frac{(40+b')^n}{(40+b')!} e^{-40-b'} \frac{1}{\sqrt{2\pi}\Delta\nu_b} e^{-(b'/\Delta\nu_b)^2}\,,$$

where the integration is truncated in order to ensure a positive number of expected background events $\nu_b + \Delta\nu_b$. This can be approximated by a sum:

$$p = 1 - 0.03 \cdot \sum_{i=-500}^{+500} \sum_{n=0}^{55} \frac{(40+b')^n}{(40+b')!} e^{-40-b'} \frac{1}{\sqrt{2\pi}\Delta\nu_b} e^{-(b'/\Delta\nu_b)^2}\,,$$

where the continuous integration over variable b' has been replaced by an approximate sum over a discrete variable i with the relation $b' = \Delta\nu_b \cdot i/100$ (the sum is in a $\pm 5\,\sigma$ range with an increment every 0.03 in terms of deviations b'). One obtains $p = 0.0176$, i.e. $Z = 2.11$. Obviously the p-value is larger and the significance smaller than the figures above. An approximation is $Z = 2(\sqrt{S+B} - \sqrt{B}) \cdot B/(B+\Delta B^2) = 1.89$ which, as a result mentioned above, is also about 10% too low.

c) The 95% CL upper limit without systematics can be calculated by scanning values of ν_s in formula (3.5) until $p_1 = 0.95$ is reached. This occurs for $\nu_{s,UL} = 30.0$. The same can be done with systematic uncertainties based on formula (3.10) (note that this is just one possible approach) and the upper limit is obtained as $\nu_{s,UL} = 30.8$. Therefore, a value $\nu_s = 25$ is not excluded by the current data.

d) In a similar approach, one can scan Z values, varying a common luminosity-multiplying factor on ν_s and ν_b until $Z = 5$ is reached. A factor 1.5 of additional data (without systematics, or 3.5 when a systematic uncertainty is included) are needed to reach the discovery threshold. This is taking into account the already accumulated data which are included in this ratio. At this point, it would be good to work on trying to decrease the level of systematics in addition to collect more data in order to reach the discovery threshold sooner.

Exercise 3.2: Ω_c peak at ARGUS

1. a) Assuming the histogram is composed of only background events, the average number of background events per bin is $43/50 = 0.86$.

b) A rough $2\,\sigma$ region would be composed of the four bins lying in the range
$[2.7 - 2.748]$.

c) In that region, $N_{\text{cand},s} = 12$ events can be counted off the histogram.

d) An estimate of the number of background events in this region is $4 \cdot 0.86 = 3.44$.

e) The probabilty for the background to fluctuate from $\nu_b = 3.44$ expected events to $N_{\text{cand},s} = 12$ observed events or more is:

$$\sum_{n=N_{\text{cand}},s}^{+\infty} \frac{\nu_b^n}{n!} e^{-\nu_b} = 2.48 \cdot 10^{-4}\,,$$

i.e. a significance of about 3.5 standard deviations.

2. a) In the sideband region defined by the 46 bins outside the $2\,\sigma$ region one counts 31 events, i.e. an estimated average of 0.67 event per bin. Therefore a yield of 2.7 background events is estimated to contribute to the signal region.

b) The number of signal events is then $N_S = 12 - 2.7 = 9.3$ events, and with an uncertainty $\sigma_{N_s} = \sqrt{12} = 3.5$ events one estimates the signal significance to be about $N_s/\sigma_{N_s} = 2.7$.

Exercise 3.3: Goodness-of-fit tests

a) The table of measurements provided for this exercise should allow to reproduce the p-values of the Pearson χ^2 test and of the run test for both data sets following the procedure detailed in examples 3.3 and 3.4.

$$p_{\chi^2}(\text{sample 1}) = 0.328\,,$$

$$p_{\text{run}}(\text{sample 1}) = 0.773\,,$$

$$p_{\chi 2}(\text{sample 2}) = 0.00019\,,$$

$$p_{\text{run}}(\text{sample 2}) = 0.0274\,.$$

b) Both tests are usually mostly uncorrelated and the p-values can be combined with the Fisher method according to equation (3.15). One obtains:

$$p_{\text{combined}}(\text{sample 1}) = 0.328 \cdot 0.773 \cdot (1 - \log(0.328 \cdot 0.773)) = 0.60 \text{ and}$$

$$p_{\text{combined}}(\text{sample 2}) = 0.00019 \cdot 0.0274 \cdot (1 - \log(0.00019 \cdot 0.0274)) = 0.000069\,.$$

c) Assuming the errors on both sets of measurements to be independent and equal to $\sqrt{n_i}$ for each of the measuments n_i in each bin i, a two-sample test based on equation (3.33) gives $\chi^2 = 15.9$ (over the 20 bins); the data-sets seem to be compatible with one another.