Problems: Day 1

Kyle Cranmer
Center for Cosmology and Particle Physics, Physics Department, New York University, USA

Abstract
Homework problems to go along with lectures.

1 Modeling

Problem 1: The Poisson Model

Given the Poisson likelihood $\mathrm{Pois}(n|\nu) = e^{-\nu}\nu^n/n!$:

a) Make a sketch and describe the distribution n assuming the true value of ν is 3.5.

b) Make a sketch of the log likelihood function $-\ln L(\nu)$ for $n = 0, 1, 2, 3, 4$

c) Cary out the simple calculus to find $\hat{\nu}$ (the value of ν that maximizes the likelihood $L(\nu)$ or, equivalently, minimizes $-\ln L(\nu)$) in terms of n.

d) Make a sketch of the log likelihood ratio $-\ln[L(\nu)/L(\hat{\nu})]$ for $n = 0, 1, 2, 3, 4$

e) Make a sketch and describe the distribution $\hat{\nu}$ assuming the true value of ν is 3.5.

f) What is the value of $-\ln[L(\nu = 3.5)/L(\hat{\nu})]$ for $n = 0, 1, 2, 3, 4$

g) Make a sketch and describe the distribution of $-\ln[L(\nu = 3.5)/L(\hat{\nu})]$ assuming the true value of ν is 3.5.

Problem 2: A simple signal and background model

Using the same Poisson likelihood $\mathrm{Pois}(n|\nu) = e^{-\nu}\nu^n/n!$, but with the total rate $\nu = \epsilon\nu_S + \nu_B$, where ν_B is the background contribution and ν_S is the signal contribution with some efficiency ϵ

a) What is the maximum likelihood estimate (MLE) $\hat{\nu}_S$ in terms of n, ν_B, and ϵ.

Problem 3: A two-bin example

Consider two independent counting experiments $\mathrm{Pois}(n_1|\nu_1)$ and $\mathrm{Pois}(n_2|\nu_2)$ with different backgrounds and signal efficiencies, so that we $\nu_1 = \epsilon_1\nu_S + \nu_{B1}$ and $\nu_2 = \epsilon_2\nu_S + \nu_{B2}$ (note ν_S is common to both experiments).

a) Setup an equation for the maximum likelihood (MLE) $\hat{\nu}_S$ in terms of n_1, n_2, ν_{B1}, ν_{B2}, ϵ_1 and ϵ_2. (optional: solve the resulting equation for $\hat{\nu}_S$ – it is straightforward but messy and not very insightful).

b) Solve for $\hat{\nu}_S$ in the case there is no background ($\nu_{B1} = \nu_{B2} = 0$), where the algebra is simple and the answer is intuitive.

c) If the first experiment has a high purity, $\nu_S/\nu_{B1} \approx 1$ and the second experiment has a very low purity $\nu_S/\nu_{B2} \ll 1$, what do you expect for $\hat{\nu}_S$ dependence on n_2.

Problem 4: The "ABCD" method

Write the probability model for the "ABCD" method. Here one imagines four disjoint counting experiments measuring n_A, n_B, n_C, and n_D where the expected background in these regions is related by $\nu_A/\nu_B = \nu_C/\nu_D$. Furthermore, one expects the signal only to contribute to region A. Hint, you will need to introduce some proportionality coefficients.

Problem 5: Simple systematics

In a counting experiment the n selected events are required to have two electrons. The signal has two genuine electrons, while the background has one genuine electron and one jet faking an electron. The electron performance group has estimated that the electron efficiency to be $\epsilon_0 \pm \sigma_\epsilon$ and the jet fake rate to be $\rho_0 \pm \sigma_\rho$. Based on this estimate you expect some fraction ϕ_0 of your ν_S signal events to pass the cuts and a total of ν_{B0} background events to pass the cuts.

a) Write a probability model for the counting experiment alone.

b) Write a probability model including the auxiliary measurements made by the electron performance group. To simplify things, assume the auxiliary measurements can be described with Gaussian distributions and that the efficiency and fake rate are not correlated.

Problem 6: Neyman Pearson exercise

If your null hypothesis has the following multivariate Gaussian distribution for x, y:

$$f_{H_0}(x,y) = \frac{1}{2\pi}\exp(-\frac{x^2}{2})\exp(-\frac{y^2}{2})$$

and the alternate hypothesis H_1 has the shifted distribution

$$f_{H_1}(x,y) = \frac{1}{2\pi}\exp(-\frac{(x-1)^2}{2})\exp(-\frac{(y-1)^2}{2})$$

a) Make a sketch of the contours of f_{H_0} and f_{H_1} in the $x-y$ plane

b) What is the shape of the optimal acceptance region W for the null in this case? Hint, use the Neyman-Pearson lemma, and you might want to work with a log likelihood ratio.

c) Optional challenge: what is the exact boundary for a test with size $\alpha = 0.05$? Hint, you may need to use your computer to calculate erf(z) or erf$^{-1}(z)$ (google if you don't know the "erf" function) or, equivalently, use the cumulative of the Gaussian distribution. In ROOT, use
`ROOT::Math::gaussian_cdf(double x, double sigma = 1, double x0 = 0)`
and for the inverse use
`ROOT::Math::gaussian_quantile(double x, double sigma = 1).`

d) Optional challenge: what is the power $(1-\beta)$ of this test?

Problem 7: χ^2 cutoffs

Wilks's theorem states that the distribution of $-2\ln\lambda(\mu)$ follows a χ^2 distribution when μ evaluated at the true point. The number of degrees of freedom of this χ^2 distribution is given by the dimensionality of μ. The critical cutoff is defined such that the integral of the χ^2 distribution from 0 to the critical cutoff should be the desired confidence level: $\int_0^{\text{cutoff}} \chi^2_N(x)dx = (1-\alpha) = CL$

a) What is the χ^2 cutoff for a 95% confidence interval ($\alpha = 0.05$) for 1 parameter of interest?

b) What is the χ^2 cutoff for a 95% confidence interval ($\alpha = 0.05$) for 2 parameters of interest?

c) What is the χ^2 cutoff for a 68% confidence interval ($\alpha = 0.32$) for 1 parameter of interest?

d) What is the χ^2 cutoff for a 68% confidence interval ($\alpha = 0.32$) for 2 parameters of interest?

Hint: You are looking for the inverse of this integral, often called the quantile. In ROOT, use:
`ROOT::Math::chisquared_quantile(CL,ndof)`

Problem 8: Nuisance Parameters and the Profile Likelihood Ratio

Let's use a simplified version of Problem 5, where the electron reconstruction efficiency is known, but the fake rate ρ has uncertainty. As before, our probability model will be:

$$P(n, \hat{\rho}|\nu_S, \rho) = \text{Pois}\left(n|\nu_S + \frac{\rho}{\rho_0}\nu_B\right) \; G(\hat{\rho}|\rho, \sigma_\rho)$$

Here I used $\hat{\rho}$ for the random estimate that would change from experiment to experiment, while ρ_0 is the specific value that was used to estimate the background rate ν_B. If write the same equation in relative terms you can simplify the formula a bit. In particular, by introducing $\delta = \rho/\rho_0$, $d = \hat{\rho}/\rho_0$ and $\sigma_\delta = \sigma_\rho/\rho_0$ and then we have

$$P(n, d|\nu_S, \delta) = \text{Pois}(n|\nu_S + \delta\nu_B) \; G(d|\delta, \sigma_\delta)$$

a) Write $-2\ln L(\nu_S, \delta)$.

b) Find the maximum likelihood estimates $\hat{\nu}_S$ and $\hat{\delta}$ by solving the two equations:

$$\frac{\partial}{\partial \nu_S} - 2\ln L(\nu_S, \delta) = 0$$

and

$$\frac{\partial}{\partial \delta} - 2\ln L(\nu_S, \delta) = 0$$

Hint: Think about what you expect for part b). The parameter ν_S is completely free, so for any value of δ, there is a value of ν_S that will give you $n = \nu_S + \delta\nu_B$. For that reason, the Poisson term will not "pull" on δ at all. In that case, what do you expect for the best fit value of δ?

c) Find the conditional maximum likelihood estimate (or "profiled") $\hat{\hat{\delta}}(\nu_S)$. Recall, this is the best fit value of δ with ν_S fixed, thus it is a function of ν_S.

d) challenge: Plot

$$-2\ln\lambda(\nu_S) = -2\ln\frac{L(\nu_S, \hat{\hat{\delta}}(\nu_S))}{L(\hat{\nu}_S, \hat{\delta})}$$

for $\nu_B = 100$, $\sigma_\delta = 0.1$, $d = 1$ and $n = 120$.

e) challenge: Plot also the simple likelihood ratio with $\delta = d$ fixed

$$-2\ln\frac{L(\nu_S, \delta = d)}{L(\hat{\nu}_S, \delta = d)}$$

(note: you can just ignore the Gaussian constraint term because it will cancel in the ratio)
Is this curve wider or narrower than the profile likelihood ratio?

f) challenge: What is your 68% confidence interval on ν_S? Is the interval wider or narrower than the one neglecting uncertainty on δ?

Problem 9: A confidence interval

Based on your answers to problems 7a) and 8d), what is the 95% confidence interval for ν_S using $-2\ln\lambda(\nu_S)$ as a test statistic and assuming the asymptotic distribution is a good approximation in this case ($\nu_B = 100$, $\sigma_\delta = 0.1$, $d = 1$ and $n = 120$.)?

a) Draw a sketch of the confidence belt in ν_S and $-2\ln\lambda(\nu_S)$. Clearly label the observed test statistic, the acceptance region, and the upper- and lower-bounds of the confidence interval.

Problem 10: The James Stein Estimator

Consider a standard multivariate Gaussian distribution for $\mathbf{x}$ in n dimensions centered around $\boldsymbol{\mu}$

$$f(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu_i)^2}{2}\right) .$$

Let us take $n = 10$ and $\mu = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$. Use the computer programing language of your choice to generate a > 1000 samples of $\mathbf{x}$. Based on this ensemble of $\mathbf{x}$, estimate

a) the magnitude of the bias of the sample mean $\overline{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij}$ (where i is the component of the vector and j is an index over the samples).

b) the mean squared error ($E[||\bar{\mathbf{x}} - \boldsymbol{\mu}||^2]$) of the sample mean?

c) the magnitude of the bias of the James-Stein estimator

$$x_{JS} = \left(1 - \frac{n-2}{||\bar{x}||^2}\right)\bar{x}$$

d) the mean squared error of the James-Stein estimator

e) Which estimator has smaller mean-squared error?

(Note: The sample mean the vector based on averaging each component, , the bias is also a vector. The question is what is the magnitude of this vector. Recall that your estimate based on the finite size of the ensemble you are generating, which will have some fluctuations.)