STATISTICS AND DATA ANALYSIS (2017/18)

PARAMETER ESTIMATION AND BAYESIAN STATISTICS

Exercise 1: Six independent observations from a Gaussian distribution $N(\mu, \sigma^2)$ are given by {1.017, 2.221, 1.416, 0.641, 0.124, 1.728}. If σ = *0.75* is known, find the symmetric confidence intervals for μ with confidence levels $1-\alpha$ = *0.68, 0.90,* and *0.95*, respectively. Discuss the case in which σ is unknown.

Exercise 2: Let's imagine that an LHC experiment measures the number of events produced in a certain decay channel of the Higgs particle. Let's call ν the expected number of events, assuming there is no background. Then, the probability to measure exactly N events in an experiment is given by the Poisson probability distribution:

$$P(N;\nu) = \frac{\nu^N}{N!}e^{-\nu}$$

I. Check that $P(N;\nu)$ is properly normalized. Given an observed value of N ($=N_{obs}$), find a frequentist unbiased estimator for ν. Compute its expected value and variance. *[Hint: compute* $\langle N \rangle$ *and* $\langle N(N-1) \rangle$ *first.]*

II. Assume now that the pdf for the estimator of ν is a Gaussian. Using the expected value and variance computed in part I, plot the Gaussian pdf corresponding to $N_{obs} = 150$.

III. Let's now construct the pdf of the estimator for ν using a Monte Carlo simulation:

   A. Let q be the estimator obtained for $N_{obs} = 150$. Generate 10^6 Monte Carlo experiments with $\nu = q$. Draw one value of $N(=N_i)$ from each one.

   B. For each experiment i, use N_i to compute the new estimator q_i.

   C. Plot the histogram of the values of q_i. Compare it with the result in part II. Is it a Gaussian?

IV. Assuming a prior for ν uniform between 0 and infinity, compute the Bayesian posterior pdf for ν. Note that it is not a Gaussian. Plot the pdf for $N_{obs} = 150$, and compare it with the results in II and III.

V. Let's now construct the Bayesian posterior pdf for ν using a Monte Carlo simulation:

   A. Generate 10^7 values for ν drawn from a uniform distribution between 0 and a large number (for instance, N_{obs} + *10*sqrt*(N_{obs})). This is our prior pdf.

   B. For each value of ν, generate one value of N according to a Poisson distribution with parameter ν.

   C. If $N = N_{obs}$, then keep the value of ν. Otherwise, discard.

   D. Plot the histogram of the values of ν that have been kept. Is it a Gaussian?

E. Compute the sample mean and sample variance from the histogram of $\mathbf{v}$. Plot a Gaussian with these mean and variance. Compare to V.D.

VI. Repeat for $N_{obs} = 10$.

VII. Repeat for $N_{obs} = 1$.

VIII. Repeat steps IV-VII using now a prior for $\mathbf{v}$ uniform in *log(*$\mathbf{v}$*)*. Compare to the previous results.

IX. What can you conclude about the similitudes and differences between the frequentist and Bayesian results? What can you conclude about the importance of choosing Jeffreys's prior? And about the validity of the Gaussian approximation for the pdfs?

Exercise 3: An experiment produces data that we believe can be described by a 2D correlated Gaussian. Let's use Bayesian inference to determine the means of the Gaussian.

I. Generate a simulated experimental data set **D** drawing 10^4 points from a 2D Gaussian with parameters $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\rho = 0.8$, using the transformation method based on the Cholesky decomposition.

II. Build the non-normalized likelihood for this data set as a function of the unknown parameters μ_1, μ_2: $L(D|\mu_1,\mu_2)$. Assume that $\sigma_1 = 1$, $\sigma_2 = 2$, $\rho = 0.8$ are known. Assuming a uniform prior pdf for μ_1, μ_2, build the non-normalized posterior pdf for μ_1, μ_2: $p(\mu_1,\mu_2|D)$.

III. Sample $p(\mu_1,\mu_2|D)$ using MCMC with a Metropolis-Hastings sampler. Choose an appropriate proposal distribution. Pay attention to terms that can be cancelled out in the pdf ratio. Pay attention to choose the proper scale for the steps in the proposal distribution. Apply an appropriate burn-in period. Compute the sample means and standard deviations of μ_1 and μ_2. Compare to the true values $\mu_1 = 2$, $\mu_2 = 1$.

IV. For extra credit, sample $p(\mu_1,\mu_2|D)$ using MCMC with a Gibbs sampler. Apply an appropriate burn-in period. Compute the sample means and standard deviations of μ_1 and μ_2. Compare to the true values $\mu_1 = 2$, $\mu_2 = 1$. Compare to the results in III.

V. Repeat steps I-IV with $\rho = 0.999$.