BTRY 4830/6830: Quantitative Genomics and Genetics Fall 2014

Homework 2 (version 2)

Assigned September 10; Version 2 posted September 11 (changed due date); Due 11:59PM September 17

Problem 1 (Easy)

a. Using two sentences at most, provide an intuitive explanation as to why we will never know the 'true' probability model responsible for generating the sample that we observe from an experiment with absolute certainty.

b. For an experiment, an associated random variable, and an assumed family of probability distributions of your choice, provide an example of an estimator that is ALWAYS wrong.

Problem 2 (Medium)

Assume that we are conducting an experiment where we have assumed that the true probability distribution of our random variable is within the family of normal probability distributions, i.e. the correct distribution is $X \sim N(\mu, \sigma^2)$ where the μ and σ^2 are unknown to us.

Note that in R you can generate samples that are well-approximated by a normal probability model using the function:

rnorm(n, μ, $\sqrt{\sigma^2}$)

where n is the size of the sample you want to generate and μ, σ^2 are the true parameter values of the probability model.

FOR THIS PROBLEM (!!) provide a separate text file with your R code used to generate your answers!

a. For a sample produced by $n = 10$ experimental trials, we have a random vector $\mathbf{X} = [X_1, X_2, ..., X_{10}]$. Assume the sample is i.i.d. What is the marginal distribution of each X_i for this random vector? What is the correlation matrix of this random vector?

b. If we were to calculate the statistic $T(\mathbf{X}) = \frac{1}{10}\sum_{i=1}^{10} X_i$, what is the sampling (probability) distribution of this statistic $Pr(T(\mathbf{X}))$?

c. Assume that (unknown to us!) the true value of $\mu = 1$ and $\sigma^2 = 1$. Using R, simulate a SINGLE sample of size $n = 10$ and for this sample and calculate the likelihood of this sample setting $\mu = 1$ and $\sigma^2 = 1$, where you must directly calculate the likelihood using the likelihood equation:

$$L(\mu, \sigma^2|\mathbf{x}) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{1}$$

d. For the SAME sample you used in (c), calculate the likelihood for this sample setting $\mu = -5$ and $\sigma^2 = 0.5$. Explain why it makes sense that the likelihood of the sample under the parameter values set in (d) is less than the parameter values set in (c)?

e. For the SAME sample you used in (c) calculate the maximum likelihood estimators of μ and σ^2 using the following equations:

$$MLE(\hat{\mu}) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

$$MLE(\hat{\sigma^2}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - MLE(\hat{\mu}))^2 \tag{3}$$

and then, calculate the likelihood of the sample setting $\mu = MLE(\hat{\mu})$ and $\sigma^2 = MLE(\hat{\sigma^2})$. Why is this likelihood greater than the likelihood in (c) that set the parameters to the true values?

f. Continue to assume (still unknown to us!) the true value of $\mu = 1$ and $\sigma^2 = 1$, simulate a SINGLE sample of size $n = 10,000$ and calculate $MLE(\hat{\mu})$ and $MLE(\hat{\sigma^2})$. Are the estimates for this larger sample closer to the true values of the parameters than the MLE's calculated for the smaller sample in (e) and why?

g. What is the sample size required to guarantee that these MLE's will return the true values of the parameters?

h. Continue to assume (again, still unknown to us!) the true value of $\mu = 1$ and $\sigma^2 = 1$ and simulate a TOTAL of 1000 separate samples each of size $n = 100$ (e.g. imagine that you are looking at the results of $n = 100$ experimental trials from each of 1000 equivalent alternative universes...). For each of the 1000 samples, calculate $MLE(\hat{\mu})$ and produce a histogram of these values. What is the shape of this histogram and why does this make sense?

i. Almost none (or none!) of your estimates of μ and σ^2 in (h) resulted in the true value for these parameters. Explain why this is the case?

j. Continue to assume (and still unknown to us!) the true value of $\mu = 1$ and $\sigma^2 = 1$ and simulate 1000 samples each of size $n = 1000$. For each of the 1000 samples, calculate $MLE(\hat{\mu})$ and produce a histogram of these values. How does this histogram compare to the histogram in (h) and why does this make sense?

Problem 3 (Difficult)

Assume a single coin flip experiment, r.v. X that is number of tails, and we are assuming $X \sim Bern(p)$. For a sample of size n that is i.i.d. we have a random vector $\mathbf{X} = [X_1, X_2, ..., X_n]$, where

for $Pr(\mathbf{X})$ each marginal distribution is $X_i \sim Bern(p)$. Using the likelihood $L(p|\mathbf{X})$ for this case, prove that $MLE(\hat{p})$ produces the same $MLE(\hat{p})$ that we obtain for $X \sim Bin(p)$.