Quantitative Genomics and Genetics - Spring 2019
BTRY 4830/6830; PBSB 5201.01

Homework 3 (version 1)

Assigned February 12; Due 11:59PM February 16

Problem 1 (Easy)

Consider a coin (system), a 'one flip' experiment, a random variable X = 'number of Heads', a bernoulli probability model $X \sim bern(p)$ (where the true parameter value p is unknown), an iid sample $\mathbf{x} = [x_1, ..., x_{10}]$ produced by $n = 10$ flips of the coin, and an estimator $T(\mathbf{x}) = \hat{p}$:

a. Consider $\hat{p} = 0.5$, is this a legitimate estimator of the parameter p (explain your answer)? In what case will $\hat{p} = 0.5$ produce the correct result?

b. Given that it is possible for $\hat{p} = 0.5$ to be correct, why might you prefer a different estimator like $\hat{p} = \text{mean}(\mathbf{x})$ (explain your answer)?

Problem 2 (Medium)

Some of the answers to questions below will require coding in R (and some do not!). For questions requiring both, a complete answer will require BOTH your answers to the questions and R code used to generate your answers in an easy to run format (!!). For full credit, please make sure that what you submit clearly demarcates your answers to each question and code that is easy to follow and run (!!). You may use Rmarkdown and if you do, please submit your .Rmd script and a pdf for this problem (you may also use Rmarkdown for the entire assignment!).

For the questions a-e below, consider a coin (system), a 'one flip' experiment, a random variable X = 'number of Heads', a bernoulli probability model $X \sim bern(p)$, and assume that you know that the TRUE parameter value is $p = 0.3$.

a. Code a function to simulate M different iid samples of size n (i.e., M vectors of length n where the elements of each vector are 1's and 0's) assuming a parameter value p (hint: make use of 'rbinom( )' in your function), where your function also calculates the estimator $T(\mathbf{x}) = \text{mean}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{p} = MLE(\hat{p})$ for each sample (i.e., your function should calculate the mean of each sample) and plots a histogram of the M values of the estimator produced.

The inputs to your function should include the number of samples to simulate M, as well as sample size n and parameter p, where your function should output a vector that contains the estimator value for each of the samples. Use the output of your function to produce two histograms of the values taken by the estimator: one with $M = 1000, n = 10, p = 0.3$ and another with $M = 1000, n = 1000, p = 0.3$.

b. What is the difference between the two histograms you obtained in part [a] (i.e., describe the difference in your own words)? Which of the two sample sizes produced the right answer $p = 0.3$ more frequently?

c. Say you obtained the following (single!) iid sample: $\mathbf{x} = [1, 0, 1, 0, 0, 0, 0, 1, 0, 1]$. Given the likelihood function $L(p|x_1, ..., x_{10}) = \prod_{i=1}^{10} p^{x_i}(1-p)^{1-x_i}$, plot the likelihood of $p \in [0, 1]$ given this sample (you may construct this plot using 100 evenly spaced values of p between 0 and 1 or by plotting the continuous function). Also plot the log-likelihood for this same sample.

d. How do the graphs for the likelihood and log-likelihood in part [c] differ and in what way are they the same (i.e., describe in your own words)?

e. What is the value of p with the highest log-likelihood given your sample in part [c]? What is the log-likelihood of the correct parameter value $p = 0.3$ given your sample in part [c]? Is this the value of the parameter that produces the highest log-likelihood given your sample? If not, what value of p produces the highest log-likelihood given your sample and why? Why in general do we expect the value of a parameter with the highest log-likelihood given a sample NOT to be equal to the true parameter value (i.e., explain in your own words)?

For the questions f-j below, consider heights (system), a 'measure a person' experiment, a random variable X to model measured heights of individual people (in meters), a normal probability model $X \sim N(\mu, \sigma^2)$, and assume that you know that the TRUE parameter values are $\mu = 1.6, \sigma^2 = 1$.

f. Code a function to simulate M different iid samples of size n (i.e., M vectors of length n where the elements of each vector are measured heights) assuming parameter values $\mu = 1.6, \sigma^2 = 1$ (hint: make use of 'rnorm( )' in your function), where your function also calculates the estimators $T(\mathbf{x}) = \text{mean}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu} = MLE(\hat{\mu})$ and $T(\mathbf{x}) = \text{var}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - mean(\mathbf{x}))^2 = \hat{\sigma}^2 = MLE(\hat{\sigma}^2)$ for each sample (i.e., your function should calculate the mean and variance of each sample). The inputs to your function should include the number of samples to simulate M , the sample size n, as well as the parameters μ and σ^2, where your function should output 2 vectors (or an M by 2 matrix) that contains the values of the $\hat{\mu}$ and $\hat{\sigma}^2$ estimators for each of the samples. Use the output of your function to produce four histograms for the values taken by each of the two estimators for each of the following two sample sizes: one with $M = 1000, n = 10, \mu = 1.6, \sigma^2 = 1$ and another with $M = 1000, n = 1000, \mu = 1.6, \sigma^2 = 1$.

g. What is the difference between the $\hat{\mu}$ histograms for the two different sample sizes ($n = 10$ and $n = 1000$) you obtained in part [f] (i.e., describe the difference in your own words)? Which of the two sample sizes produced the right answer $\mu = 1.6$ more frequently? What is the difference between the $\hat{\sigma}^2$ histograms for the two different sample sizes ($n = 10$ and $n = 1000$) you obtained in part [f] (i.e., again describe the difference in your own words)? Which of the two sample sizes produced the right answer $\sigma^2 = 1$ more frequently?

h. Say you obtained the following (single!) iid sample:

$$\mathbf{x} = [2.22, 0.98, 2.63, 3.33, 1.86, 3.25, 2.25, 2.92, 1.78, 1.01] \quad (1)$$

Use the likelihood function $L(p|x_1, ..., x_{10}) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$, to plot the likelihood of $\mu \in [0, 3.2]$ given this sample after setting $\sigma^2 = 1$ and do the same for the likelihood of $\sigma^2 \in [0, 3]$ given this sample after setting $\mu = 1.6$ (you may construct these plots using 100 evenly spaced values for the range of μ and similarly for σ^2 or by plotting the continuous functions). Also plot the log-likelihoods for μ (setting $\sigma^2 = 1$) and for σ^2 (setting $\mu = 1.6$).

i. What values of μ and σ^2 produce the highest log-likelihood given your sample in part [h] and why? Why in general do we expect the values of the parameters μ and σ^2 with the highest log-likelihood given a sample NOT to be equal to the true parameter value (i.e., explain in your own words)?

j. Note that instead of using the $MLE(\hat{\sigma}^2)$ to estimate the true parameter value of σ^2 we could have used the estimator $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - mean(\mathbf{x}))^2$. In a real case where we do not know the true parameter values, do we know whether the $MLE(\hat{\sigma}^2)$ or this alternative estimator will produce a value closer to the true σ^2 for a specific iid sample (i.e., explain why or why not in your own words)? Why might we favor one of these estimator over the other if our sample size is small?

Problem 3 (Difficult)

Assume a single coin flip experiment / r.v. that is number of tails / sample n where the sample is iid. In this case, the possible samples are ten element vectors where each element takes the value zero or one according to a Bernoulli distribution, parameterized by the parameter $\theta = p$. Prove that the $MLE(\hat{p})$ of this sampling distribution produces the same $MLE(\hat{p})$ that we obtain for $X \sim Bin(n, p)$ when $n = 10$ by directly deriving the MLE's for each (note that you need to start with a different equation for the likelihood for each of these!).