Quantitative Genomics and Genetics - Spring 2016
BTRY 4830/6830; PBSB 5201.01

Homework 3 (version 2 - posted February 25)

Assigned February 19; Due 11:59PM February 26

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)?

b. In what case will $\hat{p} = 0.5$ produce the correct result?

c. 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)?

d. Assume that you are told that the coin is either a 'fair coin' OR a coin that produces only 'Heads' OR a coin that produces only 'Tails', i.e., no other cases are possible (you are provided no additional information!). Describe a $\hat{p}$ that you would use in this case (justify your choice!).

Problem 2 (Medium)

Many of the following questions a-i will require R code (!!) provide a separate text file with your R code used to generate your answers!

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. For an iid sample of size n, write the equation for the sampling distribution for cases of k 'Heads' and $n - k$ 'Tails', i.e. an equation that calculates $Pr([X_1, ..., X_n | k$ 'H', n-k 'T').

b. 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 method of moments estimator $T(\mathbf{x}) = \text{mean}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{p}$ 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 $M = 1000, n = 1000, p = 0.3$. Describe the difference between the two histograms and which of the two produced the right answer $p = 0.3$ more frequently.

c. The sampling (probability) distribution of the estimator $\hat{p}$ in part 'b' is the binomial distribution, i.e. $Pr(\hat{p}) \sim binom(n, p)$. Make use of rbinom( ) to directly simulate $M = 1000$ samples each of size of size $n = 10$ assuming $p = 0.3$ and for each sample calculate the estimator $\hat{p} = \text{mean}(\mathbf{x})$ (note that you will have to divide the outcomes of rbinom by n), then plot a histogram of your estimator values. Repeat this for $M = 1000$ samples each of size $n = 1000$ with $p = 0.3$ (i.e., you will produce two histograms total). Note that these histograms should look quite close to the histograms you produced in part 'b' (i.e., you have used two different approaches (!!) to simulate values of the estimator $\hat{p}$ obtained from $M = 1000$ samples of size $n = 10$ or $n = 1000$).

d. Say you obtained the following (single!) 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). What is the likelihood that $p = 0.3$? Is this the value of p with the highest likelihood? If not, what value of p has the highest likelihood (justify your answer)?

e. Plot the log-likelihood for the sample in part 'd'. Do the graphs in parts 'd' and 'e' look different (how so)? What value of p has the highest log-likelihood (justify your answer)?

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. For an iid sample of size n, write down the equation for the sampling distribution for $Pr([X_1, ..., X_n])$.

g. 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 μ and σ^2 (hint: make use tof ,'rnorm( )' in your function), where your function also calculates the method of moments estimators $T(\mathbf{x}) = \text{mean}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}$ and $T(\mathbf{x}) = \text{var}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - mean(\mathbf{x}))^2 = \hat{\sigma}^2$ 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 $M = 1000, n = 1000, \mu = 1.6, \sigma^2 = 1$.

Describe the difference between the $\hat{\mu}$ histograms and the $\hat{\sigma}^2$ histograms for the two different sample sizes, including a comment on which tended to produce estimator values closer to the true values of the parameters.

h. The sampling (probability) distribution of the estimator $\hat{\mu}$ in part 'g' is the normal distribution, i.e. $Pr(\hat{\mu}) \sim N(\mu, \frac{\sigma^2}{n})$. Make use of rnorm( ) to directly simulate $M = 1000$ samples each of size of size $n = 10$ assuming $\mu = 1.6, \sigma^2 = 1$, and for each sample calculate $\hat{\mu} = \text{mean}(\mathbf{x})$, then plot a histogram of your estimator values. Repeat this for $M = 1000$ samples each of size $n = 1000$ and $\mu = 1.6, \sigma^2 = 1$ (i.e., you will produce two histograms of the estimator $\hat{\mu}$). Note that these should look quite close to the histograms for $\hat{\mu}$ you produced in part 'g' (i.e., you have used two different approaches (!!) to simulate values of the estimator $\hat{\mu}$ obtained from $M = 1000$ samples of size $n = 10$ or $n = 1000$).

i. Say you obtained the following (single!) sample: $\mathbf{x} = [2.22, 0.98, 2.63, 3.33, 1.86, 3.25, 2.25, 2.92, 1.78, 1.01]$. 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 function). What is the exact likelihood that $\mu = 1.6$ and $\sigma^2 = 1$? Are these the value of the parameters that produce the highest likelihood? If not, what values of μ and σ^2 will produce the highest likelihood (justify your answer)?

j. Plot the log-likelihoods for μ (setting $\sigma^2 = 1$) and for σ^2 (setting $\mu = 1.6$) for the sample in part 'i'. Do the μ graphs in parts 'i' and 'j' look different (how so)? Do the σ^2 graphs in parts 'i' and 'j' look different (how so)? What value of μ has the highest log-likelihood (justify your answer)? What value of σ^2 has the highest log-likelihood (justify your answer)?

Problem 3 (Difficult)

a. For a 'one flip' experiment, a random variable X = 'number of Heads', a bernoulli probability model $X \sim bern(p)$, and an iid sample produced by n experimental trials, show that $\text{E}X = p$ (i.e., that the expected value of the random variable is equal to the value of the parameter p) and use this fact to demonstrate that the estimator $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is unbiased.

b. For a 'one flip' experiment, a random variable X = 'number of Heads', a bernoulli probability model $X \sim bern(p)$, consider two iid samples of size n_1 and n_2 where $n_1 < n_2$. For the estimator $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$, denote $\hat{p}_1$ to be this estimator for the sample of size n_1 and $\hat{p}_2$ to be this estimator for the sample of size n_2. Show that $\text{E}[(\hat{p}_1 - p)^2] > \text{E}[(\hat{p}_2 - p)^2]$ (i.e., show the expected squared difference between the estimator and the true parameter value gets smaller with increasing sample size or, stated another way, the expected difference between the estimator and the true parameter value gets smaller the bigger the sample size n).

HINT: There are several ways to show this, one relatively simple approach makes use of one of the formulas for variance of a random variable and the formulas for the algebras of expectations and variances (where this same approach would also work when considering $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ for iid samples when assuming a normally distributed random variable!).