BTRY 4830/6830: Quantitative Genomics and Genetics Fall 2014

Homework 2 (version 2) - Key

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.

Many acceptable answers along the following lines: The true probability model is a function of unknown aspects of the system, the unknown conditions of the experiment, and our assumptions about the appropriate family of probability models. Since there are many unknowns and our sample (assuming non-infinite) is generated under these unknowns, where the result is one of many possible samples, any estimator we define on the sample will not return our true parameter value for every possible sample, such that we will never know if the estimated value of the parameter we obtained is correct.

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.

Many acceptable answers, one example: Assume a single coin flip example, random variable X = number of tails, Bernoulli probability model (e.g. with true parameter p), experiment of n flips, and the following estimator $\hat{\theta} = \hat{p} = T(\mathbf{x}) = 1.5$.

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?

$X_i \sim N(\mu, \sigma^2)$, the correlation is a diagonal matrix with one's in the diagonal entries (zero's for all other entries).

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 $Pr(T(\mathbf{X}))$?

$Pr(T(\mathbf{X})) \sim N(\mu, \frac{\sigma^2}{n})$

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:

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

Any code that answers the question correctly is acceptable, a compact example:
> x <- rnorm(10, 1, 1)
> prod((1/sqrt(2 * pi * 1)) * exp(−(x − 1)^2/2 * 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)?

> prod((1/sqrt(2 * pi * 1)) * exp(−(x − (−5))^2/2 * 0.05))
Since the sample was generated in an experiment where we assume the parameters of the distribution are $\mu = 1, \sigma^2 = 1$ for (almost all) samples the likelihood of parameters close to these values will be higher than the likelihood of parameters for a distribution that is not close to these parameters, as in the case $\mu = -5, \sigma^2 = 0.5$.

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?

$> 0.1 * sum(x)$
$> 0.1 * sum((x - sum(x))^2)$
OR
$> (9/10) * var(x)$
Since 'var' in R produces the unbiased estimator (not the MLE) of the variance!!

Since the MLE are functions of the observed sample that produce estimated values for the parameters that were 'most likely' to have generated the observed sample, if we calculate the likelihood of the sample given these estimates, it will always be higher than even the true parameter values (unless the estimated values happen to be the true parameter 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?

Any code that answers the question correctly is acceptable, a compact example:
> x2 <- rnorm(10000, 1, 1)
$> 0.1 * sum(x2)$
$> 0.1 * sum((x2 - sum(x2))^2)$

These estimates are closer to the true parameter values since the MLE are constructed such that the probability that they get closer to the true parameter values increases with sample size (that MLE are consistent estimators also accepted). Note there is a (very) small chance that the parameter estimates you obtained for the smaller sample were closer but you should then explain in your answer that this was unexpected.

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

$n \to \infty$ OR $n = \infty$

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?

As before, any code that answers the question correctly (e.g. including 'for' loops, using 'apply()' functions, etc.) is acceptable:
> x.umle <- NULL
> for(i in 1:1000){x.avg <-0.1∗sum(rnorm(100, 1, 1)); x.umle <-c(x.umle, x.avg)} > hist(x.umle)

The (histogram) plot looks like a normal distribution. This makes sense since we know the true sampling distribution of the statistic (the mean) under a normally distributed random variable will also be normally distributed (see part b.).

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?

Each of these is an estimate of the true parameter value based on a (non-infinite) sample so they will seldom be the exact values of the true parameters

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?

> x.umle2 <- NULL; for(i in 1:1000){x.avg <-0.1*sum(rnorm(1000, 1, 1)); x.umle2 <-c(x.umle2, x.avg)}; hist(x.umle2)

The histogram is less spread. Again (see above) we know that the $MLE(\hat{\mu}) \sim N(\mu, \frac{\sigma^2}{n})$ so with the larger samples size in part j. ($n = 1000$) compared to part i. ($n = 100$), it makes sense that samples from the model in j. will be less spread.

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 $X \sim Bin(p)$.

ANS: Note that the likelihood of an i.i.d. vector of $n = 10$ Bernoulli random variables $X_i \sim Bern(p)$ is:

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \tag{4}$$

and the log-likelihood is:

$$l(p\mathbf{x}) = \sum_{i=1}^{n} [x_i ln(p) + (1-x_i) ln(1-p)] \tag{5}$$

such that the first derivative is:

$$\frac{dl(p|\mathbf{x})}{dp} = \sum_{i=1}^{n} \left(\frac{x_i}{p} - \frac{1-x_i}{1-p} \right) \tag{6}$$

$$\frac{dl(p|\mathbf{x})}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n} (1-x_i)}{1-p} \tag{7}$$

$$\frac{dl(p|\mathbf{x})}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} \tag{8}$$

and by setting this equal to zero and solving for p we obtain:

$$0 = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} \tag{9}$$

$$\frac{\sum_{i=1}^{n} x_i}{p} = \frac{n - \sum_{i=1}^{n} x_i}{1-p} \tag{10}$$

$$(1-p)\sum_{i=1}^{n} x_i = p(n - \sum_{i=1}^{n} x_i) \tag{11}$$

$$\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i = pn - p\sum_{i=1}^{n} x_i \tag{12}$$

$$\sum_{i=1}^{n} x_i = pn \tag{13}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i = p \tag{14}$$

$$MLE(\hat{p}) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{15}$$

Also, note that each x_i can be either 1 or 0, such that if we define and x that is the number of 1's as:

$$x = \sum_{i=1}^{n} x_i \tag{16}$$

such that we have:

$$MLE(\hat{p}) = \frac{x}{n} \tag{17}$$

(note that you could have done this substitution earlier and the result would not change!!)

Next, note that the the likelihood of a Binomial random variable $X \sim Bin(n, p)$ with $n = 10$ is:

$$L(p|\mathbf{x}) = \binom{n}{x} p^x (1-p)^{n-x} \tag{18}$$

and the log-likelihood is:

$$l(p|\mathbf{x}) = ln\binom{n}{x} + xln(p) + (n-x)ln(1-p) \tag{19}$$

such that the first derivative is:

$$\frac{\partial l(p|\mathbf{x})}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p} \tag{20}$$

and by setting this equal to zero and solving for p we obtain:

$$MLE(\hat{p}) = \frac{x}{n} \tag{21}$$

which we can check by considering the second derivative:

$$\frac{\partial^2 l(p|\mathbf{x})}{\partial p^2} = -\frac{x}{p^2} + \frac{x-n}{(1-p)^2} \tag{22}$$

where this is always negative since $x < n$ (note that the denominators are squared and therefore always positive), indicating that the MLE is indeed a maximum. The MLE obtained is the same as above, proving the statement.