Using the same rules as applied for Homework 1, Homework 2 will start with Part 2 of the lab we were working on in class. At least one additional questions will be added to this homework by Thursday of next week. You will have two weeks to complete this homework.

You will be following Li [1] to study the problem of SNP and genotype calling using Illumina-style next generation sequencing data. We assume whole genome or targeted resequencing has been undertaken of a random sample of individuals from a well-defined population. We assume the genome of the organism has been previously sequenced, and that the reads have been successfully aligned to a good reference genome. We assume sequencing errors at sites within the reads are independent, and we assume reads are independent. Following Li's notation, we assume a sample of size n individuals, and we focus on a particular site in the genome where there is read coverage from the sequencing data. For the ith individual, the ploidy level is m_i, the (unknown) genotype $G_i \in \{0, 1, 2, \ldots, m_i\}$ is the number of reference nucleotides, the sequence data are $\boldsymbol{D}_i = (D_{i1}, D_{i2}, \ldots, D_{ik_i})$, and the quality scores are $\boldsymbol{Q}_i = (Q_{i1}, Q_{i2}, \ldots, Q_{ik_i})$, where k_i is the coverage (number of sequences aligned) to the chosen genomic position. Throughout this problem, we will condition on the quality scores, assuming that they correctly represent the quality of the corresponding base. Though we condition on them, we will not write them in the conditions to avoid cluttered notation. When deriving equations, show your work and justify every step.

Part I: Genotyping a single individual.

1. There were a lot of assumptions listed above. Which are reasonable? For those that are unreasonable, explain how they may be violated.

   Solution: We will still discuss in class...

2. Illumina sequence data, distributed in FASTQ format, come as base calls and accompanying, digital quality scores, both encoded using ASCII symbols.

   (a) If $q \in \{0, 1, 2, \ldots, 41\}$ is the observed quality score, what is the probability of error?

   Solution: According to the latest version of Illumina, mentioned at the Wikipedia site linked above, the ASCII codes in the fastq files correspond to PHRED quality scores in the integer range 0 to 41. Furthermore, the quality score communicates the probability there has been an error in calling the corresponding base. Specifically, Wikedia provides the equation

   $$q = -10 \log_{10} p,$$

   where p is the probability of error. We can invert this equation to answer the question:

   $$\begin{array}{rcll} q & = & -10 \log_{10} p & \text{definition} \\ -q/10 & = & \log_{10} p & \text{divide by -10} \\ 10^{-q/10} & = & p & \text{inverse of } y = \log_{10}(x) \text{ is } x = 10^y \end{array}$$

   (b) Given the quality score q, what is the probability that true base b_1 is misread as base $b_2 \neq b_1$?

   Solution: Consider a random experiment where we read a single nucleotide B as base R in the read with accompanying quality score Q. Then, the probability requested is

   $$\begin{array}{rcll} \Pr(R = b_2 \mid B = b_1, Q = q) & = & \Pr(R = b_2 \mid R \neq B, B = b_1, Q = q) \Pr(R \neq B \mid B = b_1, Q = q) & \\ & & \text{CP with condition } \{R \neq B\} & \\ & = & \Pr(R = b_2 \mid R \neq B, B = b_1, Q = q) \Pr(R \neq B \mid Q = q) & \\ & & \{R \neq B\} \text{ is independent of } \{B = b_1\} \text{ conditional on } \{Q = q\} & \\ & & q \text{ provides a } \textit{marginal} \text{ error probability; it is the same } q \text{ regardless of } B \text{ (see a).} & \\ & = & \Pr(R = b_2 \mid R \neq B, B = b_1, Q = q) 10^{-q/10} & \\ & & \text{definition of quality score} & \\ & = & \frac{1}{3} 10^{-q/10} & \\ & & \textit{if} \text{ we assume all error substitutions equally likely} & \end{array}$$

3. In this question, we consider a single individual i, so we will drop the subscript i. WLOG, assume the first $l \leq k$ reads have the reference base r, and the remaining $k-l$ reads have another base. Throughout this question, also assume that there are only two nucleotides observed in the data $\boldsymbol{D}$ (Li [1] suggests dropping all but the two most common nucleotides). And we will assume the reads have been trimmed so all quality scores exceed 0.

   (a) Derive a formula for $\Pr(D_j = r \mid G = m)$ and $\Pr(D_j = r \mid G = 0)$ for $j = 1, 2, \ldots, l$.

Solution: If $G = m$, then the individual is homozygous reference. Therefore, the true nucleotide is always r, and there was no error. Let B be the true base that the read attempts to observe, then

$$\begin{aligned} \Pr(D_j = r \mid G = m) &= \Pr(D_j = r \mid B = r) && \text{equivalent events} \\ &= 1 - \Pr(D_j \neq r \mid B = r) && \text{addition rule applied to event \& complement} \\ &= 1 - 10^{-q/10} && \text{from 2a, with condition } Q = q \text{ implied} \\ & && \text{(see red text in intro).} \end{aligned}$$

On the other hand, when $G = 0$, the individual is homozygous non-reference, and the true nucleotide is always *not* r and observing $D_j = r$ implies an error. Now,

$$\begin{aligned} \Pr(D_j = r \mid G = 0) &= \Pr(D_j = r \mid B \neq r) && \text{equivalent events} \\ &= \tfrac{1}{3} 10^{-q/10} && \text{from 2b, with condition } Q = q \text{ implied.} \end{aligned}$$

   (b) Derive a formula for $\Pr(D_j = r \mid G = g)$ when $0 < g < m$.

Solution:
Each read selects one of the m chromosomes to read. Now, g of those chromosomes will have an r at the location, and $m - g$ will have the non-reference allele. If we assume chromosomes are selected without bias, then each chromosome is equally likely to be selected and the probability that the source chromosome has an r is g/m. Let Z_j indicate whether the source chromosome has an r. Then,

$$\begin{aligned} \Pr(D_j = r \mid G = g) &= \textstyle\sum_{z=0}^{1} \Pr(D_j = r \mid G = g, Z_j = z) P(Z_j = z \mid G = g) && \text{LTP} \\ &= \textstyle\sum_{z=0}^{1} \Pr(D_j = r \mid Z_j = z) P(Z_j = z \mid G = g) && \text{conditional independence} \\ & && G \text{ is irrelevant for determining } D_j \text{ when } Z_j \text{ known} \\ &= \Pr(D_j = r \mid Z_j = 0) \tfrac{m-g}{m} + \Pr(D_j = r \mid Z_j = 1) \tfrac{g}{m} && \text{unbiased read selection} \\ &= \tfrac{e_j(m-g)}{3m} + \tfrac{(1-e_j)g}{m} && \text{from 3(a)} \end{aligned}$$

where $e_j = 10^{-q_j/10}$ is the probability of error at the base in read j. Note, that the equations we derived in 3(a) are contained in this equation, so this equation works for all $G \in \{0, 1, \ldots, m\}$.

   (c) Also think about the same conditional probabilities when $D_j \neq r$, and combine the results into an equation for $\Pr(\boldsymbol{D} = \boldsymbol{d} \mid G = g)$. Compare the results to Li [1] Eq. (2). Do you completely agree?

Solution:

$$\begin{aligned} \Pr(\boldsymbol{D} = \boldsymbol{d} \mid G = g) &= \textstyle\prod_{j=1}^{k} \Pr(D_j = d_j \mid G = g) && \text{independence of reads} \\ &= \textstyle\prod_{j=1}^{l} \Pr(D_j = r \mid G = g) \\ & \quad \times \textstyle\prod_{j=l+1}^{k} \Pr(D_j \neq r \mid G = g) && \text{WLOG reorder} \\ &= \textstyle\prod_{j=1}^{l} \left[\tfrac{e_j(m-g)}{3m} + \tfrac{(1-e_j)g}{m} \right] \\ & \quad \times \textstyle\prod_{j=l+1}^{k} \left[\tfrac{(1-e_j)(m-g)}{m} + \tfrac{e_j g}{3m} \right] && \text{Part b} \end{aligned}$$

This is very similar to Li [1]'s Eq. (2), but for the $\frac{1}{3}$ factor. Interestingly, the same author has

the correct $\frac{1}{3}$ factor in an earlier 2008 publication [2].

4. How can you use the above model to decide the genotype of a single individual, given data $\boldsymbol{D}$?

Solution:

Maximum likelihood solution. We have been treating G as an unobserved random variable, but if we instead treated it as a model parameter, we could estimate it using the maximum likelihood estimator. G is a somewhat unusual parameter in that it is discrete $G \in \{0, 1, 2, \ldots, m\}$, which means that we cannot take derivatives and use calculus to find the maximum, but we can use brute force and try all $m + 1$ possible values and estimate $\hat{G}$ as the choice that maximizes the likelihood. Specifically, in this case, the likelihood with its unknown parameter G is

$$L(G \mid \boldsymbol{D}) = \Pr(\boldsymbol{D} = \boldsymbol{d}) = \Pr(\boldsymbol{D} = \boldsymbol{d} \mid G = g),$$

where we are just referring to the conditional probability to connect back to the work in previous parts. In this case, we estimate G as MLE

$$\hat{G} = \text{argmax}_{g \in \{0,1,\ldots,m\}} L(g \mid \boldsymbol{D}).$$

Bayesian solution. Another approach is to use Bayes' rule to obtain:

$$\Pr(G = g \mid \boldsymbol{D} = \boldsymbol{d}) = \frac{\Pr(\boldsymbol{D} = \boldsymbol{d} \mid G = g) \Pr(G = g)}{\sum_{h=0}^{m} \Pr(\boldsymbol{D} = \boldsymbol{d} \mid G = h) \Pr(G = h)}.$$

The difficulty is that we must specify the distribution $\Pr(G = g)$. For genotyping a known SNP locus, we may already have a very good estimate of this distribution for the population from which we have sampled this individual, say Ashkenazi Jews. We can simply plug in these values, then choose the G that maximizes $\Pr(G = g \mid \boldsymbol{D} = \boldsymbol{d})$. This estimator is called the maximum a posteriori estimate, the MAP, which is a point estimate commonly used in Bayesian statistics. In fact, we are being Bayesians when using this second approach. Since there are no other unknown parameters, it is a particularly simple Bayesian analysis. This is the method discussed in Li [1].

5. Suppose you are sequencing DNA from a tumor sample. How would you change your procedure to use the data to detect whether there is a novel mutation (not an existing SNP) at the locus in this individual?

Solution: Consider a locus that shows no variation in the population, so all normal individuals have genotype $G = m$. If this site is mutant in the tumor sample, then $G < m$. We can use a hypothesis test to detect mutation, specifically posing null hypothesis

$$H_0 : G = m$$

and rejecting it to conclude there is a mutation.

A relevant test statistic is $T(\boldsymbol{D})$, the number of reference alleles among the reads. Clearly, the smaller this test statistic, the more evidence that the tumor is mutant.

Under H_0, the probability distribution of $T(\boldsymbol{D})$ is determined by $\Pr(\boldsymbol{D} = \boldsymbol{d} \mid G = m)$ from above. Specifically, if Z_j indicates if the jth read matches the reference or not, then

$$\Pr\left[T(\boldsymbol{D}) = l\right] = \sum_{\substack{z_1=0}}^{1} \sum_{\substack{z_2=0}}^{1} \cdots \sum_{\substack{z_k=0 \\ z_1+z_2+\cdots+z_k=l}}^{1} \prod_{j=1}^{k} \left[\frac{e_j}{3}\right]^{\mathbb{1}\{z_j=0\}} \left[1 - e_j\right]^{1-\mathbb{1}\{z_j=1\}}.$$

where we condition throughout on the observed error probabilities $e_1, e_2, \ldots, e_k$. We sum over all possible read sets of size k with the same quality scores that have exactly l reference alleles.

Finally, if we observe $T(\boldsymbol{d}) = t$, the p-value is the probability of all outcomes less consistent with H_0, namely

$$p(\boldsymbol{D}) = \sum_{l=1}^{t} \Pr\left[T(\boldsymbol{D}) = l\right].$$

Part II: Combining multiple samples.

In this part, we will extend the analysis to n individuals. Suppose the individuals have been randomly sampled from a population where the reference nucleotide relative frequency is ψ (unknown).

6. What is Hardy-Weinberg equilibrium (HWE), and what does it imply about the genotype G_i of the ith individual?

7. If we observe data $\boldsymbol{D}_1 = \boldsymbol{d}_1, \boldsymbol{D}_2 = \boldsymbol{d}_2, \ldots, \boldsymbol{D}_n = \boldsymbol{d}_n$ and all the previous assumptions are true along with HWE, what is the likelihood function $L(\psi \mid \boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n)$?

8. Consider the data, `ngs_site_data.Rtxt`, available on Canvas, and shown partially below. The columns are individual i, read j, quality score q, and allele a, where allele 1 indicates the reference allele r and 0 the non-reference allele.

```
d <- read.table("ngs_site_data.Rtxt", header=T)
head(d)

##   i j  q a
## 1 1 1 23 1
## 2 1 2 38 1
## 3 1 3 31 1
## 4 1 4 23 1
## 5 1 5 37 1
## 6 1 6 39 1
```

   (a) Write an `R` function, e.g. `log.likelihood(psi, data)`, that computes the log likelihood (you may use another language if you are prepared to do part b in that language). Use it to plot the log likelihood as a function of ψ.

   (b) Use `R`'s `optim()` function to find the maximum likelihood estimate $\hat{\psi}$ for the population reference relative allele frequency. You can use `method="Brent"` as suggested by Li [1].

   (c) Do you know of any way to estimate the variance in the estimate, $\mathrm{Var}(\hat{\psi})$?

Part III: Handling uncertainty.

Please note that I was not expecting you to produce a variance in Question 8(c). This question will focus on dealing with uncertainty in the estimate $\hat{\psi}$.

9. The data from Part II are nested, reads (low) within individuals (high). It is not obvious (to anyone) how to perform bootstrap resampling in this case. Ren et al. [4] did some work to show that it is best to sample with replacement at the highest level (the usual bootstrap), but to sample *without* replacement at the lower levels. In other words, simply sample individuals with replacement, leaving their entire data intact. Plan and implement this approach to obtain bootstrap estimates $\hat{\psi}^{(b)}$ from these data.

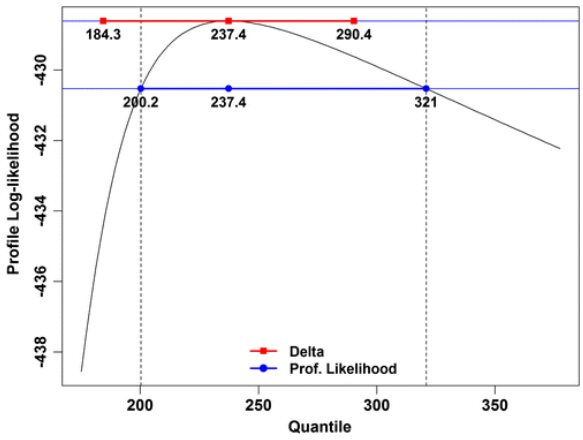

Figure 1: This figure from [3] demonstrates the profile log likelihood confidence interval estimation method, where the parameter of interest (x-axis) is a quantile for extreme flood events. In your case, the curve is your log likelihood, rather than a profile log likelihood, and your parameter of interest (x-axis) is ψ.

10. Make a histograph of the bootstrap estimates. Obtain a bootstrap confidence interval for ψ. Any concerns with the analysis?

11. An alternative to the above procedure for obtaining a confidence interval is to utilize the likelihood ratio test of null hypothesis

$$H_0 : \psi = \psi_0.$$

The likelihood ratio, with notation for this model and these data, is

$$\Lambda(\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n) = \frac{L(\psi_0 \mid \boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n)}{\sup\{L(\psi \mid \boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n) : \psi \in (0,1)\}},$$

and theory provides that the likelihood ratio test statistic $\lambda := -2 \ln \Lambda \dot{\sim} \chi^2_1$ has an asymptotic chi-squared distribution with one degree of freedom. Large values of λ indicate against H_0, and the approximation by the χ^2_1 distribution becomes better for large n. One can construct a confidence interval from those ψ_0 values for which H_0 is not rejected, which implies the confidence interval contains all ψ_0 where

$$-2 \left[l(\psi_0 \mid \boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n) - l(\hat{\psi} \mid \boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n) \right] \leq \xi_{1-\alpha},$$

ξ_p is the pth quantile of χ^2_1 and $1 - \alpha$ is the confidence level of the interval. The boundaries of this interval can be found by solving for the two ψ_0 for which the above inequality becomes an equality. This numerical problem is easily handled by the function `uniroot()` in `R`. A diagram to help you visualize what these equations and words are saying is given in Fig. 1. Find this likelihood ratio test-based confidence interval. Any concerns with the analysis?

References

[1] Heng Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." In: *Bioinformatics* 27.21 (2011), pp. 2987–2993.

[2] Heng Li, Jue Ruan, and Richard Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores." In: *Genome Res* 18.11 (2008), pp. 1851–1858.

[3] Jayantha Obeysekera and Jose D. Salas. "Quantifying the Uncertainty of Design Floods under Nonstationary Conditions". In: *Journal of Hydrologic Engineering* 19.7 (2014), pp. 1438–1446.

[4] Shiquan Ren et al. “Nonparametric bootstrapping for hierarchical data”. In: *Journal of Applied Statistics* 37.9 (2010), pp. 1487–1498.