Solutions to Assignment #8

1. *(Problem 4.2.3 on page 128 in Allman and Rhodes).* Consider the 20–base sequence

$$AGGGATACATGACCCATACA.$$

   (a) Use the first five bases to estimate the probabilities P_A, P_G, P_C and P_T.

   Solution: $P_A \approx \frac{2}{5}$, $P_G \approx \frac{3}{5}$, $P_C \approx \frac{0}{5} = 0$, and $P_T \approx \frac{0}{5} = 0$.

   (b) Repeat part (a) using the first 10 bases.

   Solution: $P_A \approx \frac{4}{10} = \frac{2}{5}$, $P_G \approx \frac{3}{10}$, $P_C \approx \frac{1}{10}$, and $P_T \approx \frac{2}{10} = \frac{1}{5}$.

   (c) Repeat part (a) using all the bases.

   Solution: $P_A \approx \frac{8}{20} = \frac{2}{5}$, $P_G \approx \frac{4}{20} = \frac{1}{5}$, $P_C \approx \frac{5}{20} = \frac{1}{4}$, and $P_T \approx \frac{3}{20}$.

   (d) Is there a pattern to the way the probabilities you computed in parts (a)–(c) changed? If so, what features of the original sequence does this pattern reflect?

   Solution: P_A did not change. This reflects that the A base is more evenly distributed than the other ones.

2. *(Problem 4.2.5 on page 128 in Allman and Rhodes).* A simple model for human offsprings is that each child is equally likely to be male or female. With this model, a three–child family can be thought of as an ordered triple of F's for females and M's for males, in which each of the triples is determined at random.

   (a) What are the 8 possible outcomes? What are their probabilities?

   Solution: The possible eight ordered triples are:

$$\begin{array}{ccc} F & F & F \\ F & F & M \\ F & M & F \\ F & M & M \\ M & F & F \\ M & F & M \\ M & M & F \\ M & M & M \end{array}$$

   and their corresponding probabilities are all 1/8 by the assumption of equal likelihood. □

(b) What outcomes make up the event "the oldest child is a daughter" and what is the probability of this event?

Solution: The event "the oldest child is a daughter" consists of the outcomes

$$\begin{matrix} F & F & F \\ F & F & M \\ F & M & F \\ F & M & M \end{matrix}$$

and the probability of this event is $1/2$. □

(c) What outcomes make up the event "the family has one daughter and two sons" and what is its probability?

Solution: The event 'the family has one daughter and two sons" is made up of the outcomes

$$\begin{matrix} F & M & M \\ M & F & M \\ M & M & F \end{matrix}$$

and the probability of that event is $3/8$. □

(d) What is the complement of the event in part (c)? List the outcomes in it and describe it in words. What is its probability?

Solution: The complement of the 'the family has one daughter and two sons" is made up of the outcomes

$$\begin{matrix} F & F & F \\ F & F & M \\ F & M & F \\ M & F & F \\ M & M & M \end{matrix}$$

that is, "the family has either no daughters or at least two daughters. The probability of this event is $5/8$. □

(e) What outcomes make up the event "the family has at least one daughter"? What is its probability?

Solution: The outcomes making up the event "the family has at least one

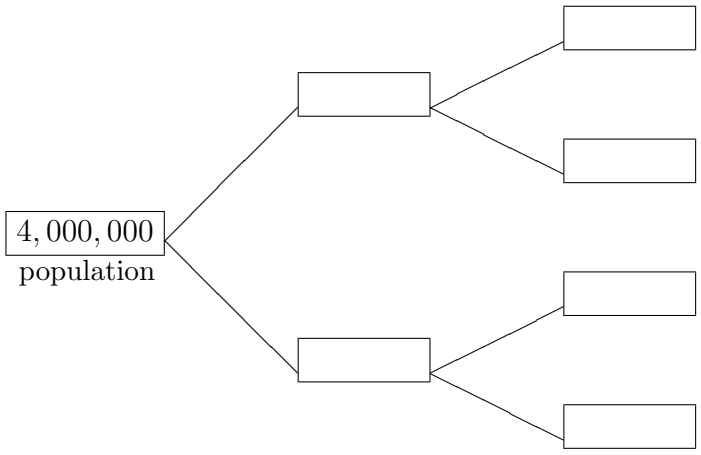
daughter" are

$$\begin{array}{ccc} F & F & F \\ F & F & M \\ F & M & F \\ F & M & M \\ M & F & F \\ M & F & M \\ M & M & F \end{array}$$

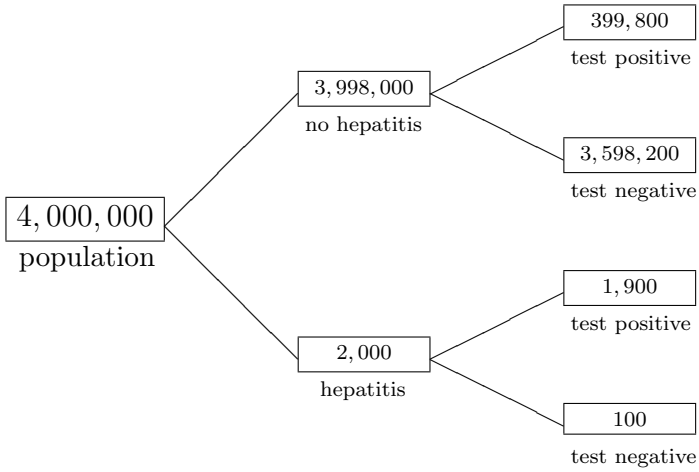
that is, the complement of the outcome "the family has no daughters." The probability of this event is 7/8. □

3. Suppose a clinic in a large city uses a new test to determine if a patient has hepatitis. If a person tests *positive*, then the test is telling us that he/she does have hepatitis (the test could be wrong!) Similarly, if a person tests negative, then the test's conclusion is that he/she does not have hepatitis. Assume that out of every 100 people who do have hepatitis, 95 test positive (that is, the test result is that they do have hepatitis) and 5 test negative. Out of every 100 people who don't have hepatitis, 90 test negative and 10 test positive. Suppose that among the 4 million people who live in the city, 0.05% do have hepatitis.

   (a) Complete the chart below. First decide on what the branches represent (there are two ways to do the branching but the information provided is suitable for only one of them). Then, for each of the boxes, find the corresponding number of people. Be careful not to make any assumptions other than the facts presented above.



   Solution: First, we figure out that 2000 people, out of the 4 million who live in the city, actually have the disease. This gives the first branching in the chart with $3,998,000$ people not having the disease:



The other branchings in the figure are those of people who do not have hepatitis but test positive and those who test negative, and those who do have hepatitis but test negative and those who test positive. The chart shows those figures.

(b) A person is selected at random and given the test. If the test is positive, what is the probability that the person actually has hepatitis.

Solution: Here we look at all those who test positive, namely $399,800 + 1,900 = 401,700$. Out of those, $1,900$ actually have hepatitis. Thus, the probability that a person who tests positive actually have the disease is

$$\frac{1,900}{401,700} \approx 0.0047 \quad \text{or} \quad 0.47\% \quad \square$$

(c) Is your answer to the previous part surprising? In what way?

Solution: Yes; the probability is smaller than one would expect. $\square$

4. *(Problem 4.3.3 on pages 134 and 135 in Allman and Rhodes).* Medical tests, such as those for diseases, are sometimes characterized by their *sensitivity* and *specificity.* The sensitivity of a test is the probability that a person with the disease will show a positive test result (a correct positive). The specificity of a test is the probability that a person who does not have the disease will show a negative test result (a correct negative).

   (a) Both sensitivity and specificity are conditional probabilities. Which of the following are they?

   $$P(-\text{result} \mid \text{disease}), \quad P(-\text{result} \mid \text{no disease})$$
   $$P(+\text{result} \mid \text{disease}), \quad P(+\text{result} \mid \text{no disease})$$

   Solution: sensitivity $= P(+\text{result} \mid \text{disease})$ and
   specificity $= P(-\text{result} \mid \text{no disease})$. □

   (b) The other conditional probabilities listed in (a) can be interpreted as probability of false positives and false negatives. Which is which?

   Solution: $P(\text{false positive}) = P(+\text{result} \mid \text{no disease})$ and
   $P(\text{false negative}) = P(-\text{result} \mid \text{disease})$. □

   (c) A study by Yerushalmy *et al.*[1] investigated the use of X–ray readings to diagnose tuberculosis. Diagnosis of $1,820$ individuals produces the data in Table 1. Compute both the sensitivity and specificity for this method of diagnosis.

Table 1: Data from Tuberculosis (TB) Diagnostic Study

	Persons without TB	Persons with TB
Negative X–ray	$1,739$	8
Positive X–ray	51	22

Solution: $$\text{sensitivity} = P(\text{positive X–ray} \mid \text{TB}) = \frac{22}{8+22} = \frac{11}{15} \approx 73.3\%$$

$$\text{specificity} = P(\text{negative X–ray} \mid \text{no TB}) = \frac{1,739}{1,739+51} \approx 97.2\%.$$

□

[1] Yerushalmy, J., Harkness, J. T., Cope, J. H., and Kennedy, B. R. (1950). *The role of dual readings in mass radiography.* Am. Rev. Tuber., **61**, 443–464

5. *(Problem 4.3.4 on page 135 in Allman and Rhodes).* Ideally, the specificity and sensitivity of medical tests should be high (close to 1). However, even with a highly sensitive and specific test, screening a large population for a disease that is rare can produce surprising results.

   (a) Suppose the specificity and sensitivity of a test for the disease are both 0.99. The test is applied to a population of $100,000$ individuals, only 100 of whom have the disease. Compute how many individual with or without the disease would be expected to test positive or negative.

   Solution: Organize the results in the following table

   Table 2: Solution of Problem 4.3.4 (a) in Allman & Rhodes

	Persons without disease	Persons with disease
Negative result	$98,901$	1
Positive result	999	99

   □

   (b) Compute the conditional probability that a person who tests positive actually has the disease.

   Solution: Use the numbers in Table 2 to get

   $$P(\text{disease} \mid \text{positive test}) = \frac{99}{999+99} = \frac{99}{1098} \approx 9\% \quad \square$$