

Bioinformatics Summer Institute - Probability/Statistics WorkshopInstructor: **Dr. Silvia Heubach****Workshop I – Probability, Counting Techniques and Bayes' Theorem****Assignments:****1. Sample Space and Basic Probability**

- 1.1** You are looking at a sliding frame of size five on the DNA strand. Define the sample space for this experiment. How many strings consist of one repeated letter?
- 1.2** Two dice are rolled. Let
 E = the event that the sum of the outcomes is odd
 F = the event that there is at least one 1.
- a) Interpret the events $E \cap F$, $E^c \cap F$ and $E^c \cap F^c$ both by describing the events in words, as well as by writing out the corresponding set of outcomes.
- b) Compute the probabilities for each of the five events $E, F, E \cap F, E^c \cap F$ and $E^c \cap F^c$.

2. Counting Methods

- 2.1** Given 10 different amino acids, how many pentapeptides (peptides of length 5) can you make if
- a) each amino acid can be used only once?
b) amino acids can be repeated?
- 2.2** Let A be the set of binary sequences (i.e., strings consisting of 0's and 1's) of length 12.
- a) How many elements are in A?
b) How many elements in A have exactly eight 0's (and four 1's)?
- 2.3** Let A be the set of DNA strings (i.e., sequences consisting of the four nucleotides A, T, C, and G) of length 8.
- a) How many elements are in A?
b) How many DNA strings have exactly 2 A's, 3 T's, 1 C and 2 G's?
- 2.4** Gametes carry 23 single chromosomes, each of which is equally likely to come from the father or mother. How likely is it that a certain gamete has 12 paternal and 11 maternal chromosomes?
- 2.5** There are 20^n different possible proteins of n residues (length n , 20 possible values per site). If the N- and C-terminus were indistinguishable and $R_1 R_2 \dots R_n$ did not have a given orientation, then the protein $R_1 R_2 \dots R_n$ would be indistinguishable from the protein $R_n R_{n-1} \dots R_1$. How many different proteins would exist in this case? (Hint: Count separately the proteins that are symmetric and those that are not, then adjust the count.)

2.6 This exercise demonstrates several different approaches to a particular counting problem, namely counting the number of paths (of length 4) from node 0 to node 8 in three different ways. In particular, the first method, which is easy enough in this example, would be very hard if the grid was larger, say consisting of 256 nodes, whereas the second and third methods can still be used for larger problems.

a) Method 1: Direct Count

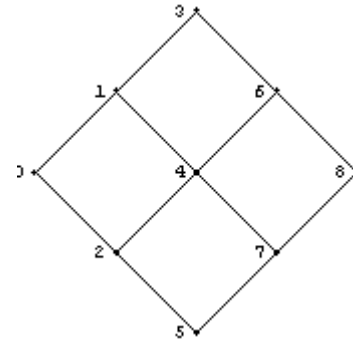
Write out all the different possible paths, then count.

Example: $0 \rightarrow 1 \rightarrow 3 \rightarrow 6 \rightarrow 8$

b) Method 2: Stitching together longer paths from shorter ones

Since each path from node 0 to node 8 has to pass through either node 3, node 4, or node 5, the problem can be broken down to counting the number of paths that pass through each of these nodes, then adding to get the total number of paths. (Which principle allows you to do that?)

To count the number of paths of length 4 that pass through node 3, we break down the problem further. Any such path consists of a path (length 2) from node 0 to node 3, “stitched together” with a path (of length 2) from node 3 to node 8. Count the number of these partial paths, then multiply their numbers to get all paths that pass through node 3. (Which principle have you used?) Repeat for paths (of length 4) that pass through nodes 4 and 5, respectively, and compute the total number of paths (which should agree with the answer in part a). Once you have understood the process, compute the number of paths of length 8 in a grid with 25 nodes. How many are there?



c) Method 3: Geometric approach – rephrasing the problem

Imagine we have placed this graph into a coordinate system, with nodes 0, 4 and 8 all positioned on the x-axis. Since node 0 and 8 are at the same level (same “y” coordinate) and the movements along the path consist of either (diagonally) “up” (↑) or (diagonally) “down” (↓), each “up” movement must be matched by a “down” movement and vice versa. Thus, each path can be described by a sequence of 4 instructions regarding the movement. For example, the path $0 \rightarrow 1 \rightarrow 4 \rightarrow 7 \rightarrow 8$ can be described as $\uparrow\downarrow\downarrow\uparrow$. Thus, the problem of counting paths has been rephrased as counting the number of permutations of ↑ and ↓, when there are two copies of each type. Use this method to check the answers for parts a) and b), then compute the number of paths of length 30 in a grid with 256 nodes.

3. Independence, Conditional Probability, Law of Total Probability & Bayes' Theorem

- 3.1** In data communication, a message transmitted is subject to various distortions and may be received with errors at its destination. Suppose that a message consisting of 64 bits (each bit is either a 0 or a 1) is transmitted and that each bit has a probability of 0.0001 of being received incorrectly, independently of the other bits. What is the probability that the message is received without errors?
- 3.2** Genes occur in different versions, called alleles. In human blood, there are three alleles, A, B, and O. A and B are dominant over O, and A and B are co-dominant. Therefore, there are three different blood types – A (either AA or AO), B (BB or BO) and AB. During fertilization, both parents randomly contribute one of their alleles. In a certain human population, the frequencies of the alleles A, B and O are 0.45, 0.20, and 0.35, respectively. If mother and son have blood type AB, and the father has blood type B, what is the probability that the father's genotype is BB?
- 3.3** For *Drosophila* (a kind of fruit fly), *B*, the gray body, is dominant over *b*, the black body, and *V*, the wild-type wing is dominant over *v*, *vestigal* (or very small wing). A geneticist, T.H. Morgan, when mating *Drosophila* of genotype *BbVv* with *Drosophila* of genotype *bbvv*, observed that 42% of the offspring were *BbVv*, 41% of the offspring were *bbvv*, 9% of the offspring were *Bbvv*, and 8% of the offspring were *bbVv*. Based on these results, should we expect that the body color and wing-type genes of *Drosophila* are on different chromosomes (i.e., are unlinked and hence independent)? Why or why not?
- 3.4** A laboratory blood test is 99% effective in detecting a certain disease when it is, in fact, present. However, the test also yields a “false positive” result for 1 % of the healthy persons tested. If 0.5% of the population actually has the disease, what is the probability that a person has the disease given that the test result is positive?
- 3.5** An insurance company classifies people into one of three classes – good risks, average risks, and bad risks. Their records indicate that the probabilities that a person classified as good, average and bad risk will be involved in an accident over a 1-year span are, respectively, 0.05, 0.15 and 0.30. Past data also indicates that 20% of the population are “good risks”, 50% are “average risks”, and 30% are “bad risks”.
- a)** What proportion of people have accidents (= probability of a random individual to have an accident) in a fixed year?
- b)** If a policyholder had no accidents in specific year, what is the probability that he or she is a good (average) risk?
- 3.6 (Bayesian Inference)** There are two coins on a table, one of which is a fair coin ($P(H) = P(T) = 0.5$), and one of which is biased ($P(H) = 2/3; P(T) = 1/3$). One of the coins is chosen at random, and tossed five times, resulting in the sequence **HHTHT**. You are to decide based on the data whether the coin that was tossed is the fair coin or the

biased coin. Let f stand for the event that the fair coin was chosen, and b for the event that the biased coin was chosen.

- a) In class, we computed the posterior probabilities for “fair coin” and “biased coin” based on seeing H on the first toss. Use these posterior probabilities, $P(f | H)$ and $P(b | H)$, as the prior probabilities for the second toss, and compute the new posterior probabilities after seeing the result of the second toss, another H . (This computation reflects that your view about the likelihood of the coin being fair or biased has changed as a result of seeing the first H .) The new posterior probabilities reflect an updated assessment on the likelihood of the coin being fair or biased after seeing the results of two coin tosses.
- b) Now compute the posterior probabilities after two coin tosses in a different way. Rather than updating the likelihoods after the first toss, we utilize knowledge of the first two coin tosses at once, i.e., you are to compute $P(f | HH)$ and $P(b | HH)$. (Recall that coin tosses are independent events, thus $P(HH) = P(H)P(H)$.) Your answers should be the same as in part a).
- c) The fact that the results in a) and b) are identical indicates that one can use the data from the five tosses at once to compute the posterior probabilities, rather than updating posterior probabilities one coin toss at a time and using the newly computed posterior probabilities as prior probabilities for the next step. Use this fact to compute the posterior probabilities after seeing the five coin tosses, $P(f | HHTHT)$ and $P(b | HHTHT)$.