CS 109: Probability for Computer Scientists
Problem Set#2

Adonis Pugh

January 27, 2020

1. Let E and F be events defined on the same sample space S. Prove that:

$$P(EF) \geq P(E) + P(F) - 1$$

(This formula is known as Bonferroni's Inequality.)

Answer.

To begin, we know that $P(EF)$ is shorthand for $P(E \cap F)$ (1). While we may not know how to reverse engineer our result, we do know that if we compute in terms of complements it is easy to convert back to the originals. The use of complements here can allow us to take advantage of De Morgan's Laws. So, let us use $P(E \cap F)^C$. From De Morgan's Laws, we know that the complement of a union of events is the intersection of their complements (2). We also know that the probability of the intersection of two events must be at most equal to the sum of the probabilities of the individual events (3). The relationship we care about involves $P(E \cap F)$ (4). Now we convert from complements back to the original context (5). Simplifying and rearranging (6), we have our result (7).

$$P(EF) = P(E \cap F) \tag{1}$$
$$P(E \cap F)^C = P(E^C \cup F^C) \tag{2}$$
$$P(E^C \cup F^C) \leq P(E^C) + P(F^C) \tag{3}$$
$$P(E \cap F)^C \leq P(E^C) + P(F^C) \tag{4}$$
$$1 - P(E \cap F) \leq 1 - P(E) + 1 - P(F) \tag{5}$$
$$P(E \cap F) \geq P(E) + P(F) - 1 \tag{6}$$
$$\boxed{P(EF) \geq P(E) + P(F) - 1} \tag{7}$$

2. Say in Silicon Valley, 35% of engineers program in Java and 28% of the engineers who program in Java also program in C++. Furthermore, 40% of engineers program in C++.

    a. What is the probability that a randomly selected engineer programs in Java and C++?

    b. What is the conditional probability that a randomly selected engineer programs in Java given that they program in C++?

Answer.

a. Let $P(J) = 0.35$ be the probability that a randomly selected engineer programs in Java, and let $P(C) = 0.40$ be the probability that a randomly selected engineer programs in C++. We are looking for the probability of the intersection of Java and Python programmers, or $P(J \cap C)$. The conditional law of probability tells us that $P(J \cap C) = P(J)P(C|J)$. Since we know that $P(C|J) = 0.28$, we can compute $\boxed{P(J \cap C) = 0.35 \times 0.28 = 0.098 = 9.8\%}$.

b. We are looking for $P(J|C)$ this time. From the conditional law of probability, we know that $P(J|C) = \dfrac{P(J \cap C)}{P(C)}$. From part a, $P(J \cap C) = 0.098$, so we can compute our result $\boxed{P(J|C) = \dfrac{0.098}{0.40} = 0.245 = 24.5\%}$.

3. A website wants to detect if a visitor is a robot. They give the visitor three CAPTCHA tests that are hard for robots but easy for humans. If the visitor fails one of the tests, they are flagged as a robot. The probability that a human succeeds at a single test is 0.95, while a robot only succeeds with probability 0.15. Assume all tests are independent.

   a. If a visitor is actually a robot, what is the probability they get flagged (the probability they fail at least one test)?

   b. If a visitor is human, what is the probability they get flagged?

   c. The percentage of visitors on the site that are robots is 5%. Suppose a visitor gets flagged. Using your answers from part (a), what is the probability that the visitor is a robot?

Answer.

a. We are looking for the conditional probability that a visitor gets flagged given that they are a robot, $P(F \mid R)$. Let us enter the realm of the robot. Since the probability that a robot passes a single test is 0.15, and it needs to pass 3 successive times to succeed, the probability of success (is then 0.15^3. What we seek is the probability of failure, which is simply the complement of the probability of success, so the probability of failure given a visitor is a robot is $\boxed{P(F \mid R) = 1 - 0.15^3 = 0.996625 = 99.6625\%}$.

b. We are looking for the conditional probability that a visitor gets flagged given that they are a human, $P(F \mid H)$. Taking the same approach as part a, we arrive at the result $\boxed{P(F \mid H) = 1 - 0.95^3 = 0.142625 = 14.2625\%}$.

c. This is essentially the inverse of part a, rather than $P(F|R)$ we are looking for the $P(R|F)$. By the conditional law of probability, $P(R|F) = \dfrac{P(R \cap F)}{P(F)}$. Using the chain rule, we can expand the numerator $P(R \cap F) = P(R)P(F \mid R)$. This is now in terms that we know, so the numerator is $P(R \cap F) = P(R)P(F \mid R) = 0.05 \times 0.996625 = 0.04983125$. The denominator can be expanded by the law of total probability, which becomes $P(F) = P(R)P(F \mid R) + P(R^C)P(F \mid R^C) = P(H)P(F \mid H) = 0.05 \times 0.996625 + 0.95 \times 0.142625 = 0.185325$. Now, we can compute the probability that a visitor is a robot given that they are flagged, to be $\boxed{P(R \mid F) = \dfrac{P(R \cap F)}{P(F)} = \dfrac{0.04983125}{0.185325} \approx 0.268886 \approx 26.8886\%}$.

4. Say all computers either run operating system W or X. A computer running operating system W is twice as likely to get infected with a virus as a computer running operating system X. If 70% of all computers are running operating system W, what percentage of computers infected with a virus are running operating system W?

Answer.

Let $P(W) = 0.7$ be the percentage of computers running operating system W, and let $P(W^C) = P(X) = 1 - 0.7 = 0.3$ be the percentage of computers running operating system X. Here we are trying to find the conditional probability that a computer is running operating system W, given that it is infected. We can start off this problem using the law of conditional probability (1). The numerator can be expanded further expanded to reflect the context of the problem (2). We are unaware of the exact likelihood that a computer is infected given it is running operating system W, but we do know that it is twice as likely to be infected than if it were running operating system X, so let's reflect that in our equality (3). The denominator can be expanded by the law of total probability (4), and we can use the relation between W and X again (5). Revisiting the full equation (6), we can see that $P(I \mid X)$ can be factored out (7), a result of the relation between W and X we were given. We now know all these terms, and we can compute the percentage of infected computers running operating system W (8).

$$P(W \mid I) = \frac{P(W \cap I)}{P(I)} \tag{1}$$

$$P(W \cap I) = P(I \mid W)P(W) \tag{2}$$

$$P(W \cap I) = 2P(I \mid X)P(W) \tag{3}$$

$$P(I) = P(I \mid W)P(W) + P(I \mid X)P(X) \tag{4}$$

$$P(I) = 2P(I \mid X)P(W) + P(I \mid X)P(X) \tag{5}$$

$$P(W \mid I) = \frac{2P(I \mid X)P(W)}{2P(I \mid X)P(W) + P(I \mid X)P(X)} \tag{6}$$

$$P(W \mid I) = \frac{2P(W)}{2P(W) + P(X)} \tag{7}$$

$$\boxed{P(W \mid I) = \frac{2(0.7)}{2(0.7) + 0.3} \approx 0.8235 \approx 82.35\%} \tag{8}$$

5. Two cards are randomly chosen without replacement from an ordinary deck of 52 cards. Let E be the event that both cards are Aces. Let F be the event that the Ace of Spades is one of the chosen cards, and let G be the event that at least one Ace is chosen. Compute:

   a. $P(E \mid F)$

   b. $P(E \mid G)$

Answer.

a. We can calculate $P(E)$ (1) and $P(F)$ (2) directly using knowledge of combinatorics. $P(E \mid F)$ can be doubly expanded using the law of conditional probability (3). $P(F \mid E)$ is easier to calculate (4), so this form is more workable. We now know all the terms we need, and we can compute our result (5).

$$P(E) = \frac{\binom{4}{2}}{\binom{52}{2}} = \frac{1}{221} \tag{1}$$

$$P(F) = 1 - P(F^C) = 1 - \frac{\binom{51}{2}}{\binom{52}{2}} = 1 - \frac{25}{26} = \frac{1}{26} \tag{2}$$

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{P(F \mid E)P(E)}{P(F)} \tag{3}$$

$$P(F \mid E) = 1 - \frac{\binom{3}{2}}{\binom{4}{2}} = 1 - \frac{1}{2} = \frac{1}{2} \tag{4}$$

$$\boxed{P(E \mid F) = \frac{(\frac{1}{2})(\frac{1}{221})}{\frac{1}{26}} = \frac{1}{17}} \tag{5}$$

b. We can calculate $P(G)$ (6) in a similar manner to how we calculated $P(F)$. $P(E \mid G)$ can also be doubly expanded using the law of total probability (7). $P(G \mid E)$ is 1, because if both cards are Aces, there is one Ace chosen be definition. We know all the terms we need, and we can compute our result (8).

$$P(G) = 1 - P(G^C) = 1 - \frac{\binom{48}{2}}{\binom{52}{2}} = 1 - \frac{188}{221} = \frac{33}{221} \tag{6}$$

$$P(E \mid G) = \frac{P(E \cap G)}{P(G)} = \frac{P(G \mid E)P(E)}{P(G)} \tag{7}$$

$$\boxed{P(E \mid G) = \frac{\frac{1}{221}}{\frac{33}{221}} = \frac{1}{33}} \tag{8}$$

6. Two emails are received at a mail server. Suppose that each email is spam with probability 0.8 and that whether each email message is spam is an independent event from the other.

   a. Suppose that you are told that at least one of the two emails is spam. Compute the conditional probability that both emails are spam.

   b. Suppose now that one of the emails is randomly (accidentally) forwarded from the server to your account, and you see that this email is spam. What is the probability that both emails originally received by the server are spam in this case? Explain your answer.

Answer.

a. The sample space of possibilities is {(N,N), (S,N), (N,S), (S,S)}. Since we know that at least one of the emails is spam, we can eliminate (N,N) from the sample space. We can use the binomial distribution to compute the final probability.

$$\boxed{P(A \mid S) = \frac{P(A \cap S)}{P(S)} = \frac{\binom{1}{1}0.8^2(1-0.8)^{2-2}}{\binom{2}{1}0.8^1(1-0.8)^{2-1} + \binom{1}{1}0.8^2(1-0.8)^{2-2}} = \frac{2}{3}}$$

b. The sample space of possibilities is {(N,N), (S,N), (N,S), (S,S)}. Since we know that a specific email is spam, we can eliminate (N,N) *and* either (N,S) or (S,N). For simplicity, let's just say the spam email is the first one, so we can eliminate (N,S). We can use the binomial distribution to compute the final probability.

$$\boxed{P(B \mid S) = \frac{P(B \cap S)}{P(S)} = \frac{\binom{1}{1}0.8^2(1-0.8)^{2-2}}{\binom{1}{1}0.8^1(1-0.8)^{2-1} + \binom{1}{1}0.8^2(1-0.8)^{2-2}} = \frac{4}{5}}$$

7. After a long night of programming, you have built a powerful, but slightly buggy, email spam filter. When you don't encounter the bug, the filter works very well, always marking a spam email as SPAM and always marking a non-spam email as GOOD. Unfortunately, your code contains a bug that is encountered 10% of the time when the filter is run on an email. When the bug is encountered, the filter always marks the email as GOOD. As a result, emails that are actually spam will be erroneously marked as GOOD when the bug is encountered. Let p denote the probability that an email is actually non-spam, and let q denote the conditional probability that an email is non-spam given that it is marked as GOOD by the filter.

   a. Determine q in terms of p.
   b. Using your answer from part (a), explain mathematically whether q or p is greater. Also, provide an intuitive justification for your answer.

Answer.

a. Let $p = P(N)$ and $q = P(N \mid G)$, the probability that an email is not spam (N) and the probability that an email is not spam given it is marked as GOOD (G), respectively. By the conditional law of probability and the chain rule, we can expand $P(N \mid G)$ (1). Since we do not know $P(G)$ directly, we can use Bayes' theorem to expand the denominator (2). $P(G \mid N)$ has already been determined to be 1 since the bug in the program erroneously marks all emails as GOOD 10% of the time, so we can eliminate those terms from the expression as well define $P(G \mid N^C) = 0.10$ since there is a 10% chance a spam email is marked as GOOD (3). Finally, we can substitute $q = P(N \mid G)$, $p = P(N)$, and $1 - p = P(N^C)$ to yield our result (4).

$$P(N \mid G) = \frac{P(N \cap G)}{P(G)} = \frac{P(G \mid N)P(N)}{P(G)} \tag{1}$$

$$P(N \mid G) = \frac{P(G \mid N)P(N)}{P(G \mid N)P(N) + P(G \mid N^C)P(N^C)} \tag{2}$$

$$P(N \mid G) = \frac{P(N)}{P(N) + 0.10P(N^C)} \tag{3}$$

$$\boxed{q = \frac{p}{p + 0.1(1-p)} = \frac{p}{0.9p + 0.1}} \tag{4}$$

b. Since the expression for q contains p in the numerator and a term ≤ 1 in the denominator, $q \geq p$. This makes sense intuitively because the spam filter works in the majority of cases, so an email is more likely to be non-spam if the filter marked it as GOOD than it would be if it were chosen at random. At the extreme, if the filter worked perfectly ($q = 1$), q would always be greater than p unless the probability than an email was non-spam was 1, in which case $p = 1$ and $q = p$.

8. Consider a hash table with 15 buckets, of which 9 are empty (have no strings hashed to them) and the other 6 buckets are non-empty (have at least one string hashed to each of them already). Now, 2 new strings are independently hashed into the table, where each string is equally likely to be hashed into any bucket. Later, another 2 strings are hashed into the table (again, independently and equally likely to get hashed to any bucket). What is the probability that both of the final 2 strings are each hashed to empty buckets in the table?

Answer. This problem can be broken down into two phases: (1) the dispersion probabilities of the initial 2 strings being hashed, and (2) the probabilities of the final 2 strings going into empty buckets. For phase 1, there are 4 cases:

(a) Both strings go to used buckets
(b) One string goes to a used bucket, one goes to an empty bucket
(c) Both strings go to the same empty bucket
(d) Each string goes to a different empty bucket

These cases were chosen because these are all the possible ways the number of empty buckets can be altered. Each of these probabilities must be computed. To be sure these cases cover the entire sample space, we must make sure their sum is 1. In phase 2, the probability of case d must be computed once again for each probability. The sum of these four two-step probabilities is the probability that both of the final 2 strings are each hashed to empty buckets in the table.

$$\frac{(2+15-1)!}{2!(15-1)!} = \frac{16!}{2!5!} = 120\ (sample\ space) \tag{1}$$

$$\frac{\frac{(2+6-1)!}{2!(6-1)!}}{120} = \frac{7}{40} \rightarrow \frac{7}{40} \times \frac{\frac{(2+9-1)!}{2!(9-1)!} - 9}{120} = \frac{21}{400} \tag{2}$$

$$\frac{\frac{(1+6-1)!}{1!(6-1)!} \times \frac{(1+9-1)!}{1!(9-1)!}}{120} = \frac{18}{40} \rightarrow \frac{18}{40} \times \frac{\frac{(2+8-1)!}{2!(8-1)!} - 8}{120} = \frac{42}{400} \tag{3}$$

$$\frac{\frac{(1+9-1)!}{1!(9-1)!} \times 1}{120} = \frac{3}{40} \rightarrow \frac{3}{40} \times \frac{\frac{(2+8-1)!}{2!(8-1)!} - 8}{120} = \frac{7}{400} \tag{4}$$

$$\frac{\frac{(2+9-1)!}{2!(9-1)!} - 9}{120} = \frac{12}{40} \rightarrow \frac{12}{40} \times \frac{\frac{(2+7-1)!}{2!(7-1)!} - 7}{120} = \frac{21}{400} \tag{5}$$

$$\frac{21}{400} + \frac{42}{400} + \frac{7}{400} + \frac{21}{400} = \boxed{\frac{91}{400} = 0.2275 = 22.75\%} \tag{6}$$

9. Five servers are located in a computer cluster. After one year, each server independently is still working with probability p, and otherwise fails (with probability $1 - p$).

   a. What is the probability that *at least* 1 server is still working after one year?

   b. What is the probability that *exactly* 2 servers are still working after one year?

   c. What is the probability that *at least* 2 servers are still working after one year?

Answer.

a. The probability that at least 1 server is still working is the same after one year as 1 - the probability that no servers are working after one year. Using the binomial distribution, $\boxed{P(A) - 1 - P(A^C) = 1 - (1-p)^5}$.

b. The probability that exactly 2 servers are still working after one year is the probability of 2 success and 3 failures. Using the binomial distribution, $\boxed{P(B) = \binom{5}{2} p^2(1-p)^{5-2} = 10p^2(1-p)^3}$.

c. The probability that at least 2 servers are still working after one year is the same as 1 - the probability that no servers or 1 server is still working after one year. Using the binomial distribution, $\boxed{P(C) = 1 - (1-p)^5 - \binom{5}{1} p(1-p)^4 = 1 - (1-p)^5 - 5p(1-p)^4}$.

10. The Superbowl institutes a new way to determine which team receives the kickoff first. The referee chooses with equal probability one of three coins. Although the coins look identical, they have probability of heads 0.1, 0.5 and 0.9, respectively. Then the referee tosses the chosen coin 3 times. If more than half the tosses come up heads, one team will kick off; otherwise, the other team will kick off. If the tosses resulted in the sequence H, T, H, what is the probability that the fair coin was actually used?

Answer. We are trying to find the conditional probability that the fair coin was used $P(E_2)$, given that the sequence was H, T, H $P(S)$. This problem can be approached by using the conditional law of probability, the chain rule, and Bayes' theorem (which are all related) to expand $P(E_2 \mid S)$ into calculable terms.

$$
\begin{aligned}
&p_1 = 0.1,\ p_2 = 0.5,\ p_3 = 0.9 \\
&P(S \mid E_1) = p_1^2(1-p_1)^{3-2} = 0.1^2(1-0.1) = 0.009 \\
&P(S \mid E_2) = p_2^2(1-p_2)^{3-2} = 0.5^2(1-0.5) = 0.125 \\
&P(S \mid E_3) = p_3^2(1-p_3)^{3-2} = 0.9^2(1-0.9) = 0.081
\end{aligned}
$$

$$
\begin{aligned}
&P(E_2 \mid S) = \frac{P(E_2 \cap S)}{P(S)} = \frac{P(S \mid E_2)P(E_2)}{P(S)} \\
&P(E_2 \mid S) = \frac{P(S \mid E_2)P(E_2)}{P(S \mid E_1)P(E_1) + P(S \mid E_2)P(E_2) + P(S \mid E_3)P(E_3)} \\
&\boxed{P(E_2 \mid S) = \frac{0.125(\frac{1}{3})}{0.009(\frac{1}{3}) + 0.125(\frac{1}{3}) + 0.081(\frac{1}{3})} \approx 0.5814 \approx 58.14\%}
\end{aligned}
$$

11. A robot, which only has a camera as a sensor, can either be in one of two locations: L_1 (which does not have a window) or L_2 (which has a window). The robot doesn't know exactly where it is and it represents this uncertainty by keeping track of two probabilities: $P(L_1)$ and $P(L_2)$. Based on all past observations, the robot thinks that there is a 0.7 probability it is in L_1 and a 0.3 probability that it is in L_2.

    The robot then observes a window through its camera, and although there is only a window in L_2, it can't conclude with certainty that it is in fact in L_2, since its image recognition algorithm is not perfect. The probability of observing a window given there is no window at its location is 0.2, and the probability of observing a window given there is a window is 0.9. After incorporating the observation of a window, what are the robot's new probabilities for being in L_1 and L_2, respectively?

Answer. Let $P(L_1) = 0.7$ and $P(L_2) = 0.3$. In addition, let $P(O \mid L_1) = 0.2$ and let $P(O \mid L_2) = 0.9$. This problem can be solved using the conditional law of probability, the chain rule, Bayes' theorem (which are all related) to expand $P(L_1 \mid O)$ and $P(L_2 \mid O)$ into calculable terms.

$$P(L_1 \mid O) = \frac{P(O \mid L_1)P(L_1)}{P(O)} = \frac{P(O \mid L_1)P(L_1)}{P(O \mid L_1)P(L_1) + P(O \mid L_2)P(L_2)}$$

$$\boxed{P(L_1 \mid O) = \frac{0.2 \times 0.7}{0.2 \times 0.7 + 0.9 \times 0.3} \approx 0.3415}$$

$$P(L_1 \mid O) = \frac{P(O \mid L_2)P(L_2)}{P(O)} = \frac{P(O \mid L_1)P(L_1)}{P(O \mid L_1)P(L_1) + P(O \mid L_2)P(L_2)}$$

$$\boxed{P(L_2 \mid O) = \frac{0.9 \times 0.3}{0.2 \times 0.7 + 0.9 \times 0.3} \approx 0.6585}$$

12. The color of a person's eyes is determined by a pair of eye-color genes, as follows:

    - if both of the eye-color genes are blue-eyed genes, then the person will have blue eyes
    - if one or more of the genes is a brown-eyed gene, then the person will have brown eyes

    A newborn child independently receives one eye-color gene from each of its parents, and the gene it receives from a parent is equally likely to be either of the two eye-color genes of that parent. Suppose William and both of his parents have brown eyes, but William's sister (Claire) has blue eyes. (We assume that blue and brown are the only eye-color genes.)

    a. What is the probability that William possesses a blue-eyed gene?

    b. Suppose that William's wife has blue eyes. What is the probability that their first child will have blue eyes?

Answer.

a. Let *B* denote brown-eyed gene and *b* denote the blue-eyed gene. Since both of William's parents have brown eyes and William's sister has blue eyes, both of William's parents must be of type *Bb*. Here we are looking for the conditional probability that William possess a blue-eyed gene, i.e. is of type *Bb*, give n that he has brown eyes. Without knowing that he has brown eyes, he has a 0.25 chance of being *BB*, a 0.50 chance of being *Bb*, and a 0.25 chance of being *bb*. Since we know he has brown eyes, *bb*, and its associated probability, is no longer considered in the probability that he carries a blue-eyed gene. Using the conditional law of probability,

$$\boxed{P(Bb \mid BB \text{ } or \text{ } Bb) = \frac{P(Bb \cap BB \text{ } or \text{ } bb)}{P(BB \text{ } or \text{ } bb)} = \frac{\frac{1}{2}}{\frac{1}{4}+\frac{1}{2}} = \frac{2}{3}}$$

b. If William's wife has blue eyes, she is of type *bb* and can only pass on the blue-eyed gene. We can solve this part of the problem using the law of total probability, since $P(bb)$ is composed of parts. The first part is the probability that the child is bb given that Wiliam is BB, and the second part is the probability that the child is bb given that William is Bb.

$$\boxed{P(bb) = P(bb \mid BB)P(BB) + P(bb \mid Bb)P(Bb) = 0(\frac{1}{3}) + \frac{1}{2}(\frac{2}{3}) = \frac{1}{3}}$$

13. Your colleagues in a comp-bio lab have sequenced DNA from a large population in order to understand how a gene (G) influences two particular traits (T_1 and T_2). They find that $P(G) = 0.6$, $P(T_1 \| G) = 0.7$, and $P(T_2 \| G) = 0.9$. They also observe that if a subject does not have the gene G, they express neither T_1 nor T_2. The probability of a patient having both T_1 and T_2 given that they have the gene G is 0.63.

    a. Are T_1 and T_2 conditionally independent given G?

    b. Are T_1 and T_2 conditionally independent given G^C?

    c. What is $P(T_1)$?

    d. What is $P(T_2)$?

    e. Are T_1 and T_2 independent?

Answer.

a. By the law of conditional independence, if T_1 and T_2 were conditionally independent given G, then $P(T_1T_2 \mid G) = P(T_1 \mid G)P(T_2 \mid G)$. We can see from the data that $0.7 \times 0.6 = 0.63$, which means that trait T_1 and trait T_2 are conditionally independent given gene G.

b. T_1 and T_2 are conditionally independent given G^C. This is true because $P(T_1T_2 \mid G^C) = 0$, $P(T_1 \mid G^C) = 0$, and $P(T_2 \mid G^C) = 0$, so the law of conditional independence holds.

c.

$$\boxed{P(T_1 \mid G) = \frac{P(T_1T_2 \mid G)}{P(T_2 \mid G)} \to P(T_1) = \frac{P(T_1 \mid G)P(G)}{P(G \mid T_1)} = \frac{0.7 \times 0.6}{1} = 0.42}$$

d.

$$\boxed{P(T_2 \mid G) = \frac{P(T_1T_2 \mid G)}{P(T_1 \mid G)} \to P(T_2) = \frac{P(T_2 \mid G)P(G)}{P(G \mid T_1)} = \frac{0.9 \times 0.6}{1} = 0.54}$$

e. It is possible to compute $P(T_1T_2)$ directly and compare the result with the law of independence.

$$P(T_1T_2 \mid G) = \frac{P(G \mid T_1T_2)}{P(G)} \to P(T_1T_2) = \frac{P(T_1T_2 \mid G)P(G)}{P(G \mid T_1T_2)} = \frac{0.63 \times 0.6}{1} = 0.378$$

If T_1 and T_2 are independent, then $P(T_1T_2) = P(T_1)P(T_2)$, but $0.378 \neq 0.42 \times 0.54 \neq 0.2268$, so T_1 and T_2 are not independent.

14. See Problem Set #2 for a full write-up of this problem. You will submit parts a and b as code in Gradescope and write up answers to c and d (and optionally e) in this document.

Answer.

a. Python code

b. Python code

c. For a gene G_i to be independent of T, $P(G_i \mid T) = P(G_i)$. Based on the results of part a and b, it seems reasonable to conclude that G_1 and G_2 are independent of T, and G_3, G_4, and G_5 are dependent on T. Since the probability of T does not change significantly when G_1 or G_2 are present, this suggests that G_1 and G_2 are independent of T. Alternatively, I wrote a function to compute the conditional probability of G_i given T, and the results of the computation support that G_1 and G_2 are independent of T because their conditional probabilities given T are approximately the same as before. Technically, the probabilities would have to be exactly the same for these conclusions to be completely accurate, but the slight differences are assumed to be due to the sample size only being 100,000 bats, and the probabilities in theory should stabilize as the sample size approaches infinity.

d. The probability of T increases with the presence of G_3, G_4, and G_5. G_5 seems to correlate to T the most, followed by G_3 and G_4, respectively. Furthermore, the probabilities of G_3, G_4, and G_5 given T are all extremely high, all exceeding 97%. It might then be reasonable to indicate to pathologists that G_5, G_3, and G_4 seem to most responsible for T, in that order.