**Stat 321 - Day 8**
**Bayes' Theorem**

**Example: AIDS Testing**
The ELISA test for AIDS was used in the screening of blood donations in the early 1990's. As with most medical diagnostic tests, the ELISA test was not infallible. If a person actually carries the AIDS virus, experts estimate that this test gave a positive result 97.7% of the time. (This number is called the *sensitivity* of the test.) If a person does not carry the AIDS virus, ELISA gave a negative result 92.6% of the time (the *specificity* of the test). Estimates at the time were that 0.5% of the American public carried the AIDS virus (the *base rate* with the disease). We will determine the (conditional) probability that a person actually carried the AIDS virus given that he/she tested positive on the ELISA text.

(a) First, without doing any calculations, take a guess for the value of the (conditional) probability that a person who tests positive carries the virus:

Solution Through Two-Way Tables:
Before turning to a formal probability rule to address this issue, you can answer the question and develop some intuition for how the probabilities relate to each other by creating a 2x2 table. Imagine a hypothetical population of 1,000,000 people for whom these percentages hold exactly. (The population size is chosen to be so large in order to make the calculations all work out to be integers.) For questions (b)-(e) below, record your answers in the appropriate cells of the table.

(b) Assuming that 0.5% of the population of 1,000,000 people carries AIDS, how many such carriers are there in the population? How many non-carriers are there?

(c) Consider for now just the *carriers*. If 97.7% of them test positive, how many people test positive? How many carriers does that leave who test negative?

(d) Now consider only the *non-carriers*. If 92.6% of them test negative, how many test negative? How many non-carriers does that leave who test positive?

| | Positive test | Negative test | Total |
|---|---|---|---|
| Carries AIDS virus | (c) | (c) | (b) |
| Does not carry AIDS | (d) | (d) | (b) |
| Total | (e) | (e) | 1,000,000 |

(e) Determine the total number of positive test results and the total number of negative test results.

(f) *Of those who test positive*, what proportion actually carry the disease? How does this compare to your prediction above?

(g) Explain why this probability turns out to be small compared to the sensitivity and specificity. (Be sure to refer to calculations in the table.)

Derivation of Bayes' Theorem:

If the events $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive (i.e., they form the entire sample space with their union), then for any event *B*: $P(A_j \mid B) = \dfrac{P(A_j \cap B)}{P(B)} = \dfrac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{k} P(B \mid A_i)P(A_i)}$.

(h) What rule or definition justifies the first equality in this derivation?

(i) What rule justifies the numerator of the last step in this derivation? How about the denominator?

The probabilities $P(A_j)$ are sometimes called *prior probabilities*, and the conditional probabilities $P(A_j|B)$ are called updated, or *posterior probabilities*.

Bayes' Theorem provides a mechanism for updating uncertainty about a hypothesis (H) in light of new evidence (E). In its simplest form, it can be written as:

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E \mid H)P(H) + P(E \mid H')P(H')}.$$

You will now use Bayes' Theorem to re-create your ELISA analysis above. Let A denote the event that the person carries the disease, P denote a positive test result, and N denote a negative test result.

(j) Express the base rate of .005 as a probability in the notation of these events.

(k) Express the sensitivity .977 and specificity .926 of the test as conditional probabilities involving these events.

(l) Consider the conditional probability of having AIDS given a positive test result that you calculated in (f). Express this probability (only its notation) using these symbols.

(m) Use the simpler version of Bayes' Theorem to calculate this probability. Does the answer agree with that from your earlier analysis?

**Example: Programming Bugs**
Suppose that three programmers are designing computer code for a project: Alice has designed 60% of the code, Barb 30% and Chuck 10%. Suppose further that Alice has a bug in 3% of her work, Barb in 7% of her work, and Chuck in 5% of his.

(n) What percentage of the code written has a bug? In other words, what is the probability that a randomly selected piece of code has a bug? [*Hint*: Use the law of total probability.]

(o) Given that you find a bug in a line of code, who is most likely to have written it? Who is least likely? [*Hint*: Use Bayes' Theorem to find each person's conditional probability of having written the line given that it has a bug. Notice that you already calculated the denominator in (n).]

Amy:

Barb:

Chuck:

(p) Fill in the following probability table to represent this situation:

| | Alice | Barb | Chuck | Total |
|---|---|---|---|---|
| Bug | | | | |
| No bug | | | | |
| Total | | | | 1.000 |

(q) Construct a probability tree to represent this situation:

**Example: Multiple Choice Exams (cont.)**
Suppose that a student knows (with certainty) the answer to 50% of the questions on a multiple choice exam, while on the other 50% he is clueless and so guesses randomly among the four choices.

(r) Determine the probability that the student actually knew the answer, given that he answers correctly. [*Hints*: Define the relevant events carefully. Use the Law of Total Probability and Bayes' Theorem. Construct a probability table or tree if you wish.]

(s) Now suppose that there are $k$ choices on each question, where $k$ is some integer greater than one. Determine the probability in (r) as functions of $k$.

(t) Is your answer to (r) an increasing or decreasing function of $k$? What happens in the limit? Explain why these make sense.