Homework #2 - SOLUTIONS

Due: Tuesday, October 5 @ 6pm [39*points*]

Problem 1: [9*points*]

Suppose that a disease is inherited via a sex-linked mode of inheritance so that a male offspring has a 50% chance of inheriting the disease, but a female offspring has no chance of inheriting the disease. Further suppose that 51.3% of births are male.

a. Draw a probability tree and/or generate a contingency table representing the data above [3*points*]

Probability tree:

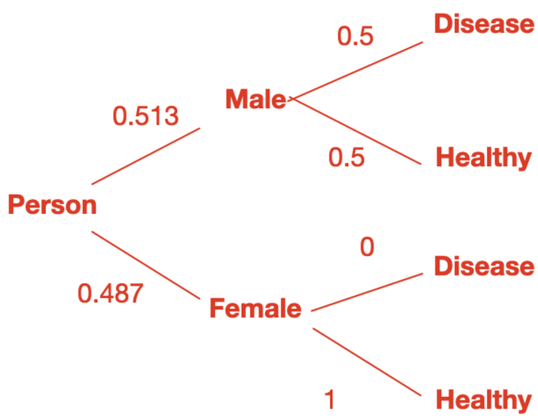

Table might look something like below. Note, the numbers in the table don't matter but the proportions have to be the same (i.e. in my table I had 100 people, but you could do the same with 200 or 500 people)

gender	disease	healthy	total
male	25.65	25.65	51.3
female	0.00	48.70	48.7
total	25.65	74.35	100.0

b. Are the two events (inheriting the disease and being male) disjoint or non-disjoint? Explain. [2*points*]

Non-disjoint because you can be male and have the disease at the same time

c. Are the two events (inheriting the disease and being male) independent or non-independent? Explain. [2*points*]

Dependent because being male changes the probability that you will have the disease. Likewise, having the disease changes your probability of being male.

d. What is the probability that a randomly chosen child will be affected by the disease? Be sure to show all work. [2*points*]

Pr(disease) = pr(male + disease) + pr(female + disease)

Pr(male + disease) = pr(male) * pr(disease | male)

Pr(male + disease) = 0.513 * 0.5 = 0.2565

Pr(female + disease) = pr(female) * pr(disease | female)

Pr(male + disease) = 0.487 * 0 = 0

Pr(disease) = 0.2565 + 0 = 0.2565

Problem 2: [$15 points$]

Suppose a test is 99% accurate: it gives a positive result 99% of the time if the patient is indeed infected (i.e. a 1% false-positive rate), and a negative result 99% of the time if the patient is indeed healthy (i.e. a 1% false-positive rate). For convenience, let pos/neg denote a positive/negative test result, and let I/H denote infected/healthy.

a. Using P(A | B) notation, write down the facts described above [$1 points$]

Grading note: giving just the first two or jus the second two is sufficient

P(pos|I) = 0.99; P(neg|H) = 0.99; P(neg|I) = 0.01; P(pos|I) = 0.01

b. Suppose I take the test and it comes up positive. I'd like to know what that means about the chances that I'm actually infected. Using P(A|B) notation, write down the quantity that I'm interested in (not the number, just the notation for the conditional probability that corresponds to this question). [$1 points$] $P(I|pos)$

c. Supposing my test came up positive, what can you tell me about the chances that I'm actually infected? Give numbers if you can; if you can't state what else you'd need to know. [*Hint: remember that the probability of a positive result is the sum of the probabilities of getting a true positive or of getting a false positive*] [$2 points$]

Per Bayes theorem:

$P(I|pos) = \frac{P(pos|I)*P(I)}{P(pos)}$

$P(I|pos) = \frac{P(pos|I)*P(I)}{P(pos|I)*P(I)+P(pos|H)*P(H)}$

$P(I|pos) = \frac{P(pos|I)*P(I)}{P(pos|I)*P(I)+P(pos|H)*(1-P(I))}$

$P(I|pos) = \frac{0.99*P(I)}{0.99*P(I)+0.01(1-P(I))}$

To solve this, we need to know $P(I)$, the probability that I'm infected regardless of the test result. This is not known.

d. Suppose we administered the test in a population where the prevalence of infection (i.e., the baseline probability that a given person is infected) is 1/1000. That is, P(I) = 0.001

- What fraction of all people would have a positive test result? [$2 points$]

P(pos) = P(pos|I)*P(I) + P(pos|H)*P(H) = 0.01098

- Of those people, what fraction of them would be truly infected? [$2 points$]

$P(I|pos) = \frac{P(pos|I)*P(I)}{P(pos)} = 0.00099/0.01098 = 0.09$

- What is the probability, then, that a person in this population with a positive result is truly infected? (Note: we call this the *posterior probability* or the *positive predictive value*) [1*points*]
  0.09
- Should people in this population believe that they're more likely infected than not if they get a positive result? Does this answer surprise you? Why or why not? [2*points*]

No; there is only a 9% chance they're infected. Surprises me!!!

e. Suppose now we do the same thing, but in a high-risk population where the prevalence is 1/3. How do your answers to (d) change? Between this result and your answer to part (d), what kind of recommendations would you make for administering this screening test? [2*points*]

The PPV goes up to 0.98! A positive result is much more reliable here

It probably only makes sense to administer this test in a high-risk population, since a positive result would be much more reliable

f. Suppose we go back to the low-risk group in (d) and re-administer the same test to those who tested positive the first time. What is the probability that someone who tests positive a second time (in addition to the first) is truly infected? [2*points*]

Of the people who get a positive result, only 0.09 are infected, so retesting in that population is like setting P(I) = 0.09. Testing them gives a PPV of 0.907. Testing positive **twice** is much more reliable than just testing positive once.

Problem 3: [9*points*]

The seeds of the garden pea (*Pisum sativum*) are either yellow or green. A certain cross between pea plants produced progeny in the ratio 3 yellow : 1 green. Imagine four randomly chosen progeny of such a cross are examined.

a. Does this variable fit the assumptions for a binomial random variable? Why or why not? [2*points*]

Yes! Binary (yellow or green), independent (each sample is randomly chosen), n (fixed n = 4), same p (always 3:1 ratio)

b. What is the probability that one is green and three are yellow? Be sure to show ALL work. [2*points*]

Pr(1 green) = pr(1 green)*(how many ways to get 1 green)

Pr(1 green) = binomial distribution = $(nCj)(p^j)(1-p)^(n-j)$

Pr(green) = 1/4 (1 green to 3 yellow)

Pr(1 green) = $(4C1)(1/4)^1(3/4)^3$

Pr(1 green) = 0.421875

Alt. in R: ‘dbinom(1, 4, 0.25)‘

c. Generate the probability distribution for every possible outcome given four randomly chosen progeny of such a cross are examined. [*Hint: first select which outcome will be viewed as "success" and create the probability table for that variable*] [2*points*]

green	yellow	probability
0	4	0.3164063
1	3	0.4218750
2	2	0.2109375
3	1	0.0468750
4	0	0.0039063

d. What is the probability that all four randomly chosen progeny are the same color? [1*points*]

Pr(all 4 same color) = Pr(4 green) + Pr(4 yellow)

Pr(all 4 same color) = Pr(4 green) + Pr(0 green)

Pr(all 4 same color) = 0.00390625 + 0.31640625 = 0.3203125

e. What is the expected value of green (or yellow) seeds? How does this compare to the known ratio of yellow:green seeds? [1*points*]

E(Y) = n * p = 4 * (1/4) = 1

E(Y) = 1 green seed, which makes sense because we were told there was a ratio of 3 yellow : 1 green seed!

Note: if you calculated the expected value of yellow seeds, you should get E(Y) = 3 (4 * (3/4) = 3)

f. What is the standard deviation of green (or yellow) seeds? [1*points*]

sd(Y) = sqrt(np(1-p)) = sqrt(4 * (1/4)(3/4)); SD(Y) = 0.866 (no matter if you chose yellow or green)

Problem 4: [6*points*]

When red blood cells are counted using a certain electronic counter, for a certain specimen, the true value is 5,000,000 cells/mm^3 and the standard deviation is 40,000. The distribution of repeated counts is approximately normal.

a. Suppose you get a reading of 4,900,000. What is the standardized z-score for this value? [2*points*]

mean = 5,000,000, sd = 40,000

$z = \frac{y-\mu}{\sigma}$

$z = \frac{4900000-5000000}{40000} = -2.5$

b. What is the probability that the counter would give a reading between 4,900,000 and 5,100,000? [2*points*]

Option 1 (manual): convert to z score

$z = \frac{5100000-5000000}{40000} = 2.5$

area between -2.5 and +2.5 = (p-value for z = 2.5) - (p-value for z = -2.5)

Can look up value in table, or use ‘pnorm()‘ in R:

‘pnorm(2.5) - pnorm(-2.5)‘ = 0.9876

Option 2: We can also use ‘pnorm‘ with non-standard mean and sd: ‘pnorm(5100000, 5000000,40000) - pnorm(4900000, 5000000, 40000)‘ = 0.9876

c. If the true value of the red blood count for a certain specimen is μ, what is the probability that the counter would give a reading between 0.98μ and 1.02μ? $[1 points]$

0.98*5000000 = 4900000, therefore, the answer is the same, 0.9876

d. A hospital lab performs counts of many specimens every day. For what percentage of these specimens does the reported blood count differ from the correct value by 2% or more? $[1 points]$

If 98.76% of the values lie within 2% of the true mean, then 1 - 0.9876 (1.24%) of the values will differ from the correct value by 2% or more.