CSE 3521 HOMEWORK 4

BRENDAN WHITAKER

1. The prior is the probability $P(B)$ we give to the random variable B before making any observations. We then make the observation B, and compute the probability of this observation given our random variable, as well as the probability of our observation $P(A)$, and thus, after normalizing by $P(A)$, we have the posterior probability $P(B|A)$, the probability of B given that we have this new information from our observation

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

2. (a) Let B represent having tuberculosis, and let A represent testing positive for tuberculosis. We know that $P(B) = 0.0001$, since Then we are given that the false positive rate is 1%, so $P(A|B) = 0.99$. Thus we have

$$P(B, A) = P(A|B)P(B) = 0.99 * 0.0001 = 0.000099,$$

which is the probability that you have tuberculosis, and the test comes back positive. Now we compute the probability that you do not have tuberculosis and the test comes back positive.

$$P(\neg B, A) = P(A|\neg B)P(\neg B) = 0.01 * 0.9999 = 0.009999.$$

Hence marginalizing, we have that $P(A) = P(B, A) + P(\neg B, A) = 0.000099 + 0.009999 = 0.010098$. Thus Bayes's rule gives us $P(B|A)$, the probability that you have tuberculosis given the positive test:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.99 * 0.0001}{0.010098} = 0.0098.$$

   (b) We let $P(B|A) = 0.5$, and solve for $P(A|B)$:

$$
\begin{aligned}
P(B|A) &= \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\neg B)P(\neg B)} \\
&= \frac{P(A|B)P(B)}{P(A|B)P(B) + (1 - P(A|B))(1 - P(B))} \\
&= \frac{P(A|B)P(B)}{P(A|B)P(B) + 1 - P(B) - P(A|B) + P(A|B)P(B)} \\
&= \frac{P(A|B)P(B)}{2(P(A|B)P(B)) + 1 - P(B) - P(A|B)} \\
\Rightarrow P(A|B)P(B) &= 0.5(2(P(A|B)P(B)) + 1 - P(B) - P(A|B)) \\
&= P(A|B)P(B) + 1/2 - \frac{1}{2}P(B) - \frac{1}{2}P(A|B) \\
\Rightarrow P(A|B) &= 1 - P(B) \\
&= 0.9999.
\end{aligned}
\tag{1}
$$

3. (a) No. We are able to calculate the values for $P(Bread, Relish, Cheese) = P(Bread, \neg Relish, Cheese) = 0.16$, and the values for $P(\neg Bread, \neg Relish, Cheese) = P(\neg Bread, Relish, Cheese) - 0.04$. However, the other 4 joint probabilities across all 3 variables cannot be computed individually.
   (b) We need to know $P(Bread, Relish, \neg Cheese)$ or something equivalent.

Date: AU17.

4. (a)
$$P(c) = \sum_i \left(P(c, b, a_i) + P(c, \neg b, a_i) \right)$$
$$= 0.012 + 0.024 + 0.084 + 0.09 + 0.072 + 0.018 \tag{2}$$
$$= 0.3.$$

(b) We compute the new probability table marginalized over c

	b	$\neg b$
a_1	0.061	0.153
a_2	0.122	0.156
a_3	0.427	0.081

(c) $P(\neg b|\neg c) = \frac{P(\neg b, \neg c)}{P(\neg c)} = \frac{\sum_i P(\neg b, \neg c, a_i)}{1 - P(c)} = \frac{0.063 + 0.084 + 0.063}{1 - 0.3} = \frac{0.21}{0.7} = 0.3.$

(d) C is not independent of A since

$$P(c|a_1) = \frac{P(c, a_1)}{P(a_1)} = \frac{0.102}{0.214} = 0.477 \neq 0.345 = \frac{0.096}{0.278} = \frac{P(c, a_2)}{P(a_2)} = P(c|a_2).$$

C is not independent of B since $P(c|b) = \frac{P(c,b)}{P(b)} = \frac{0.12}{0.73} = 0.164$,

$P(c|\neg b) = \frac{P(c, \neg b)}{P(\neg b)} = \frac{0.18}{0.27} = 0.667$, and these are clearly not equal.

(e) We show A is not conditionally independent of c given b. $P(a_1|b, c) = \frac{P(a_1, b, c)}{P(b, c)} = \frac{0.012}{0.12} = 0.1 \neq$
$0.836 = \frac{0.061}{0.73} = \frac{P(a_1, b)}{P(b)} = P(a_1|b).$

5. (a) No, they are not well formed. We could change the null hypothesis so it reads "The true proportion of seniors in S_1 is equal to the true proportion of seniors in S_2.", and the alternative could read "The proportion of seniors in S_2 is greater than that of S_1."

(b) Yes they are testable. We could perform a Chi-square goodness of fit test on each of them, but only if we knew the underlying distributions. Then, we could compare the results of these two tests and see if their difference is statistically significant.

6. (a) The null hypothesis will be $\mu_1 = \mu_2$ where these are the means of the delivery times in the old and new robot, respectively. The alternative hypothesis will be $\mu_1 < \mu_2$.

(b) We use student's t-test for difference of two means, since we are assuming independence between the two samples, and because time is quantitative data as opposed to categorical.

(c) We compute from a calculator the means, $\mu_1 = 30.9$, $\mu_2 = 26.6$, and the standard deviations $\sigma_1 = 24.6$, $\sigma_2 = 23.5$. Also, the number of data points for each is given by $N_1 = 10$, $N_2 = 12$. Thus we use 20 degrees of freedom, have a t-value of 0.420, and a P-value of 0.679. Since we are using a significance level of 0.05, and our P value is greater than this, we fail to reject the null hypothesis, and conclude that our alternative hypothesis is NOT supported.