Prof. David Draper
Department of Statistics
University of California, Santa Cruz

STAT 206: Quiz 1 *[90 total points]*

Name: Kevin Guillen

Here is Your background information, translatable into $\mathcal{B}$, for this problem.

- *(Fact 1)* As a broad generalization (which you can verify empirically), statisticians tend to have shy personalities more often than economists do — let's quantify this observation by assuming (based on previous psychological studies) that 84% of statisticians are shy but the corresponding percentage among economists is only 10%.
- *(Fact 2)* Conferences on the topic of *econometrics* are almost exclusively attended by economists and statisticians, with the majority of participants being economists — let's approximately quantify this fact by assuming (based on data from previous conferences) that 95% of the attendees are economists (and the rest statisticians, except for a tiny proportion of people from other professions, which can be ignored).

Suppose that you (a physicist, say) go to an econometrics conference — you strike up a conversation with the first person you (haphazardly) meet, and find that this person is shy. The point of this problem is to show that the (conditional) probability p that you're talking to a statistician, given this data and the above background information, is only about 31%, which most people find surprisingly low, and to understand why this is the right answer. Let St = (person is statistician), E = (person is economist), and Sh = (person is shy).

(a) Identify (in the form of a proposition B_1, one of the elements of $\mathcal{B}$) the most important assumption needed in this problem to permit its solution to be probabilistic; explain briefly. *[5 points]*

Solution. Our most important assumption is that we haphazardly meet a person. Since translating this requirement to math language our B_1 is taking a random sample of 1 from the population, which is all conference attendees. $\square$

(b) Using the St, E and Sh notation, express the three numbers (84%, 10%, 95%) above, and the probability we're solving for, in conditional probability terms, remembering to condition appropriately on $\mathcal{B}$. *[5 points]*

Solution.

$$84\% = P(Sh \,|\, St, \mathcal{B}) \tag{1}$$
$$10\% = P(Sh \,|\, E, \mathcal{B}) \tag{2}$$
$$95\% = P(E \,|\, \mathcal{B}) \tag{3}$$

we also obtain the following,

$$5\% = P(St \,|\, \mathcal{B}) \tag{4}$$

□

(c) Briefly explain why calculating the desired probability is a good job for Bayes's Theorem. *[5 points]*

Solution. Well, we are trying to calculate $P(St \,|\, Sh, \mathcal{B})$ and we see that probability (1) from above is just the reverse ordering of conditioning of what we are trying to solve for, which is perfect for Bayes' Theorem.

$$P(St \,|\, Sh, \mathcal{B}) = \frac{P(St \,|\, \mathcal{B}) P(Sh \,|\, St, \mathcal{B})}{P(Sh \,|\, \mathcal{B})}$$

□

Table 1: *2 by 2 table cross-tabulating truth (statistician, economist) against data (shy, not shy) for the people at the conference, assuming a total number of attendees of 1,000.*

		Truth		
		Statistician	Economist	Total
Data	Shy	42	95	137
	Not Shy	8	855	863
	Total	50	950	1,000

The goal in the rest of the problem is for you to use all three of the methods developed in class — the 2 by 2 table cross-tabulating truth against data, Bayes's Theorem in odds ratio form, and calculating the denominator using the *Law of Total Probability*, by partitioning over the unknown truth — to compute $P(St \,|\, Sh, \mathcal{B})$, the posterior probability that the haphazard person is a statistician given that this person is shy (and given $\mathcal{B}$).

(d) Use the three numerical facts (84%, 10%, 95%) given at the beginning of the quiz to fill in all 8 of the entries marked '—' in Table 1, taking the total number of attendees at the conference to be 1,000 (*Hint:* All of these numbers are integers), thereby showing that $P(St \,|\, Sh, \mathcal{B}) = \frac{42}{137} \doteq 30.7\%$; show your work *[20 points]*.

Solution. First we can obtain the total number of economists and statisticians using (3) and (4) from part (b),

$$|E| = P(E \,|\, \mathcal{B}) \cdot 1000 = 0.95 \cdot 1000 = 950$$
$$|St| = P(St \,|\, \mathcal{B}) \cdot 1000 = 0.05 \cdot 1000 = 50.$$

Now with (1) and (2) from part (b) we can get how many of these statisticians and economists are shy,

$$|Sh, St| = P(Sh \,|\, St, \mathcal{B}) \cdot |St| = .84 \cdot 50 = 42$$
$$|Sh, E| = P(E \,|\, E, \mathcal{B}) \cdot |E| = .10 \cdot 950 = 95$$

now knowing the total number of economists and statisticians alongside of many of each are shy we can obtain how many of them are not shy,

$$|\neg Sh, St| = |St| - |Sh, St| = 50 - 42 = 8$$
$$|\neg Sh, E| = |E| - |Sh, E| = 950 - 95 = 855.$$

Which means given that the person we haphazardly meet at this convention is shy, we see from the table the probability that they are a statistician is,

$$P(St \,|\, Sh, \mathcal{B}) = \frac{42}{137} = 30.7\%$$

□

(e) Briefly explain why the following expression is a correct use of Bayes's Theorem on the odds ratio scale in this problem. *[5 points]*

$$\left[\frac{P(St \,|\, Sh, \mathcal{B})}{P(E \,|\, Sh, \mathcal{B})}\right] = \left[\frac{P(St \,|\, \mathcal{B})}{P(E \,|\, \mathcal{B})}\right] \cdot \left[\frac{P(Sh \,|\, St, \mathcal{B})}{P(Sh \,|\, E, \mathcal{B})}\right]$$
$$(1) \quad = \quad (2) \quad \cdot \quad (3)$$

Solution. Well we know the odds that the person we met is a statistician given they are shy and $\mathcal{B}$ is the ratio between the probability they are a statistician given they are shy and $\mathcal{B}$ versus the probability they are not a statistician given they are shy and $\mathcal{B}$, but to say not a statistician in this case is the same as saying they are an economist. In mathematical terms it is the following,

$$o_{St \,|\, Sh, \mathcal{B}} = \frac{P(St \,|\, Sh, \mathcal{B})}{P(E \,|\, Sh, \mathcal{B})}. \tag{5}$$

We know though $P(St \,|\, Sh, \mathcal{B})$ to be,

$$P(St \,|\, Sh, \mathcal{B}) = \frac{P(St \,|\, \mathcal{B}) P(Sh \,|\, St, \mathcal{B})}{P(Sh \,|\, \mathcal{B})}$$

and $P(E \,|\, Sh, \mathcal{B})$ to be,

$$P(E \,|\, Sh, \mathcal{B}) = \frac{P(E \,|\, \mathcal{B}) P(Sh \,|\, E, \mathcal{B})}{P(Sh \,|\, \mathcal{B})}.$$

Using this to expand equation (5) we get,

$$\begin{aligned} o_{St \,|\, Sh, \mathcal{B}} = \frac{P(St \,|\, Sh, \mathcal{B})}{P(E \,|\, Sh, \mathcal{B})} &= \frac{P(St \,|\, \mathcal{B}) P(Sh \,|\, St, \mathcal{B})}{P(Sh \,|\, \mathcal{B})} \cdot \frac{P(Sh \,|\, \mathcal{B})}{P(E \,|\, \mathcal{B}) P(Sh \,|\, E, \mathcal{B})} \\ &= \frac{P(St \,|\, \mathcal{B}) P(Sh \,|\, St, \mathcal{B})}{P(E \,|\, \mathcal{B}) P(Sh \,|\, E, \mathcal{B})} \\ &= \left[\frac{P(St \,|\, \mathcal{B})}{P(E \,|\, \mathcal{B})}\right] \cdot \left[\frac{P(Sh \,|\, St, \mathcal{B})}{P(Sh \,|\, E, \mathcal{B})}\right] \end{aligned}$$

which is the same as the expression give, which is why it is a correct use of Bayes' Theorem on the odds ratio scale in this problem. □

(f) Here are three terms that are relevant to the quantities in part (e) above:

- (Prior odds ratio in favor of St over E, given $\mathcal{B}$)
- (Bayes factor in favor of St over E, given $\mathcal{B}$)
- (Posterior odds ratio in favor of St over E, given $\mathcal{B}$)

Match these three terms with the numbers (1), (2), (3) in the second line of the equation in part (e). *[5 points]*

Solution.

(1) – (Posterior odds ratio in favor of St over E, given $\mathcal{B}$)
(2) – (Prior odds ratio in favor of St over E, given $\mathcal{B}$)
(3) – (Bayes factor in favor of St over E, given $\mathcal{B}$)

$\square$

(g) Compute the three ratios in part (e), briefly explaining your reasoning, thereby demonstrating that the posterior odds ratio o in favor of St over E (given $\mathcal{B}$) is $o = \frac{42}{95} \doteq 0.442$. *[15 points]*

Solution. Let us first solve for (2), this is simply the probability of haphazardly meeting a statistician given $\mathcal{B}$ over the probability of haphazardly meeting an economist given $\mathcal{B}$. We know 5 percent of the attendees are statisticians which means,

$$P(St \,|\, \mathcal{B}) = \frac{50}{1000} = \frac{1}{20}$$

then we know 95 percent of the attendees are economists,

$$P(E \,|\, \mathcal{B}) = \frac{950}{1000} = \frac{19}{20}.$$

Now solving for (2),

$$\frac{P(St \,|\, \mathcal{B})}{P(E \,|\, \mathcal{B})} = \frac{1}{20} \cdot \frac{20}{19} = \frac{1}{19}$$

Which means there is a 19 to 1 prior odds ratio against St given $\mathcal{B}$

Now solving for (3) this is the probability of the a person you haphazardly meet is shy given they are a statistician and $\mathcal{B}$ over the the same thing except they are an economist. We know though 84 percent of statisticians are shy and only 10 percent of economists are shy, which gives us,

$$\frac{P(Sh \,|\, St, \mathcal{B})}{P(Sh \,|\, E, \mathcal{B})} = \frac{0.84}{0.10} = \frac{42}{5}$$

Which means we have a $\frac{42}{5}$ Bayes factor in favor of St given $\mathcal{B}$

These two together will let us solve for (1),

$$\frac{P(St \,|\, Sh, \mathcal{B})}{P(E \,|\, Sh, \mathcal{B})} = \frac{1}{19} \cdot \frac{42}{5} = \frac{42}{95} = o$$

which means we have a 95 to 42 posterior odds against St given the data and the background statement, which matches the problem statement. □

(h) Use the expression $p = \frac{o}{1+o}$ to show that the desired probability in this problem — the conditional probability that you're talking to a statistician (given $\mathcal{B}$) — is $p = \frac{42}{137} \doteq 0.307$. *[5 points]*

Solution. We know $o = \frac{42}{95}$ so plugging in we get,

$$p = \frac{\frac{42}{95}}{1 + \frac{42}{95}} = \frac{42}{95} \cdot \frac{95}{137} = \frac{42}{137}$$

matching the problem statement. □

(i) Briefly explain why the following expression is a correct use of Bayes's Theorem on the probability scale in this problem. *[5 points]*

$$P(St \,|\, Sh,\, \mathcal{B}) = \frac{P(St \,|\, \mathcal{B})\, P(Sh \,|\, St,\, \mathcal{B})}{P(Sh \,|\, \mathcal{B})} \,. \tag{6}$$

Solution. Well we know from CS1 in class that,

$$P(\theta = 1 \,|\, y_1 = 1, \mathcal{B}) = \frac{P(\theta = 1 \,|\, \mathcal{B}) P(y_1 = 1 \,|\, \theta = 1, \mathcal{B}))}{P(y_1 = 1 \,|\, \mathcal{B})}$$

which is analogous to for the unknown we are trying to solve for now, which we can see through the table below,

(truth) unknown	data	problem
St or E	Sh or $\neg Sh$	Quiz 1
$\theta = 1$ or $\theta = 0$	$y_1 = 1$ or $y_1 = 0$	CS1

By analogy then, we have,

$$P(St \,|\, Sh,\, \mathcal{B}) = \frac{P(St \,|\, \mathcal{B})\, P(Sh \,|\, St,\, \mathcal{B})}{P(Sh \,|\, \mathcal{B})}$$

$\square$

(j) Notice as usual that you know both of the numerator probabilities in equation (7) but you don't (yet) know the denominator $P(Sh \,|\, \mathcal{B})$. Use the *Law of Total Probability*, partitioning over the unknown truth, to show that

$$P(Sh \,|\, \mathcal{B}) = \frac{137}{1000} = 0.137\,, \tag{7}$$

and use this to show that

$$P(St \,|\, Sh,\, \mathcal{B}) = \frac{(0.05)(0.84)}{0.137} = \frac{42}{137} \doteq 0.307\,. \tag{8}$$

[15 points]

Solution. Given that a person from the conference is shy we know they MUST either be and economist or a statistician but NOT both. So using the law of total probability (and rules for and/or when taking probability) we can express the denominator as,

$$\begin{aligned} P(Sh \,|\, \mathcal{B}) &= P((Sh \text{ and } St) \text{ or } (Sh \text{ and } E) \,|\, \mathcal{B}) \\ &= P((Sh \text{ and } St) \,|\, \mathcal{B}) + P((Sh \text{ and } E) \,|\, \mathcal{B}) \\ &= P(St \,|\, \mathcal{B}) \cdot P(Sh \,|\, St, \mathcal{B}) + P(E \,|\, \mathcal{B}) \cdot P(Sh \,|\, E, \mathcal{B}) \end{aligned}$$

in this form though we know all these probabilities, so plugging in we get the denominator to be,

$$\begin{aligned} P(Sh \,|\, \mathcal{B}) &= \frac{1}{20} \cdot \frac{21}{25} + \frac{19}{20} \cdot \frac{1}{10} \\ &= \frac{21}{500} + \frac{19}{200} \\ &= \frac{137}{1000} \\ &= 0.137 \end{aligned}$$

Now using this in Bayes' Theorem we get,

$$\begin{aligned} P(St \,|\, Sh, \mathcal{B}) = P(St \,|\, Sh,\, \mathcal{B}) &= \frac{P(St \,|\, \mathcal{B})\, P(Sh \,|\, St,\, \mathcal{B})}{P(Sh \,|\, \mathcal{B})} \\ &= \frac{(0.05)(0.84)}{0.137} \\ &= \frac{42}{137} = 0.307 \end{aligned}$$

which matches the problem statement as desired. $\square$

(k) Someone says, "That 30.7% probability can't be right: 84% of statisticians are shy, versus 10% for economists, so your probability p of talking to a statistician has to be over 50%." Briefly explain why this line of reasoning is wrong, and why p should indeed be less than 50%. *[5 points]*

Solution. While it is true a higher percent of statisticians are shy compared the percent of economists that are shy, but at this event there is a higher number of shy economists than there are shy statisticians. This is due to the large proportion of economists at this event versus the small proportion of statisticians (95 percent and 5 percent respectively). More specifically this person is focusing on the Bayes factor in favor of the person being a shy statistician and ignoring the prior odds ratio, which is 19:1 against the person being a statistician which is why they think it should be over 50 percent, but when one accounts for the prior odds ratio it should indeed be less than 50 percent. $\square$