Zac Monroe

B365

HW2

	Diabetes	No Diabetes	Total
Kale	801	9192	9993
No Kale	9905	80122	90027
Total	10706	89314	100020

1. Problem 1

   (a) An exact value for p(Diabetes|Kale) cannot be computed because we can only compute an estimation for it based on these data. The true probability cannot be known based on a mere sample.

   (b) p̂(Diabetes|Kale) = 801 / 9993 ≈ 0.080156. This is the proportion of kale-eaters that have diabetes, which is in turn our estimated probability of a kale-eater having diabetes.

   (c) p(Diabetes|Kale) is 95% likely to be contained in the interval

   p̂(Diabetes|Kale) $\pm \frac{1.96}{\sqrt{4 * 9993}}$, or

   0.080156 ± 0.000098096, or

   (0.080058, 0.080254).

   p(Diabetes|No Kale) is 95% likely to be contained in the interval

   p̂(Diabetes|No Kale) $\pm \frac{1.96}{\sqrt{4 * 90027}}$, or

   0.11002255 ± 0.000010886, or

   (0.110011664, 0.110033436).

(d) It is definitely clear that the kale-eaters were much less likely to have diabetes than the kale-avoiders. The confidence intervals for the probability of kale-eaters to have diabetes and non-kale-eaters to have diabetes did not come close to overlapping, so I can conclude with a great deal of confidence that kale-eaters are less likely to have diabetes than those who do not eat kale.

(e) I cannot conclude that kale *causes* a lower rate of diabetes. I have only been given two points of data for each member in the population. There are probably many other differences between each population member other than just their kale consumption and whether or not they have diabetes. They are not identical in every other way, and these are just observed traits, so kale does not necessarily cause lower diabetes rates.

(f) It is possible that kale-eaters tend to just eat in a more healthy way, and live a more healthy lifestyle (frequent exercise, non-sedentary tendencies) than those who don't eat kale, and so because of all the other choices that they make surrounding their health, they are much less likely to get diabetes.

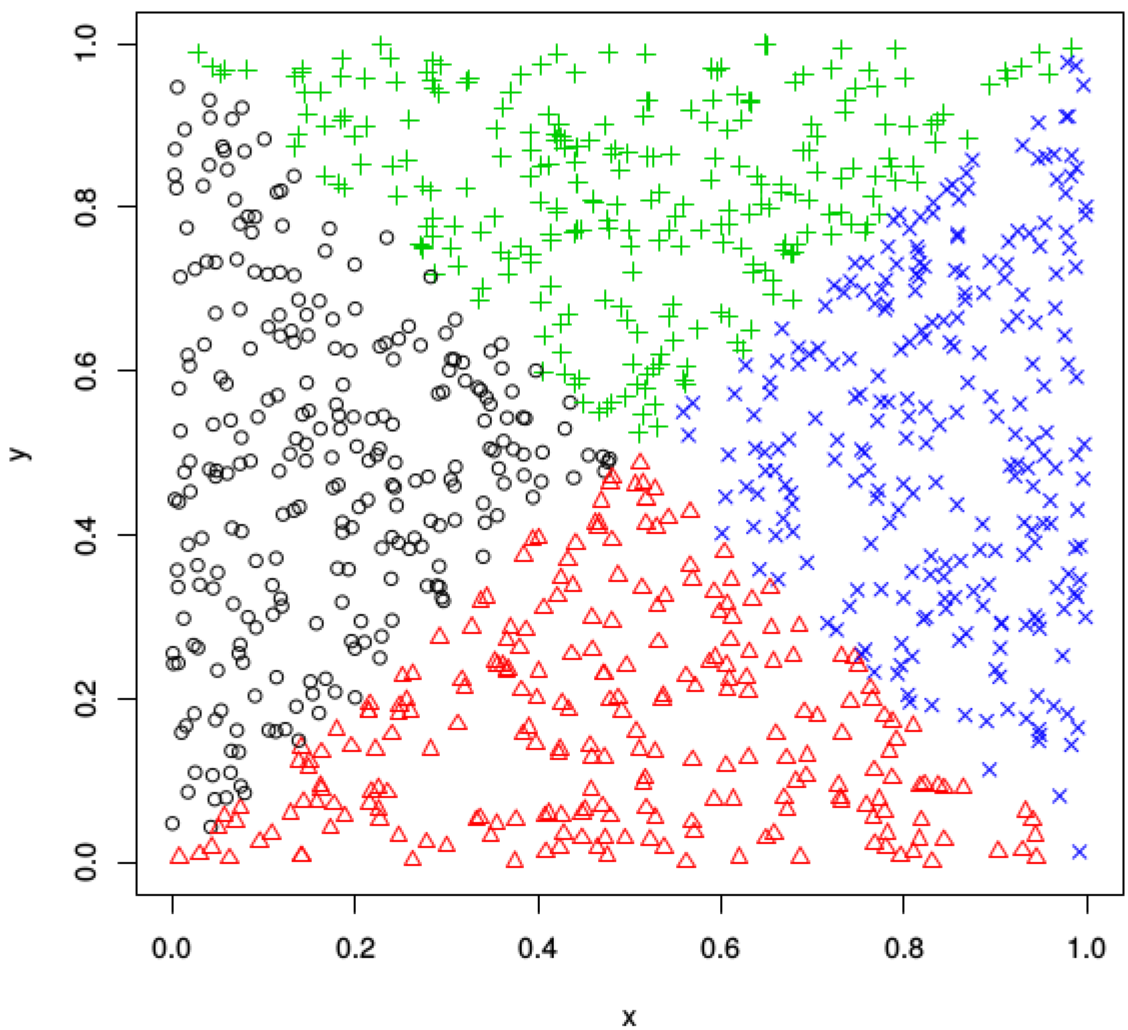
2. Problem 2

(a) It might seem at first glance that, in this situation, there would be evidence that kale causes a lower rate of diabetes. However, it is clear that there is a disproportionate number of kale-eaters to kale-avoiders, but the selection method sounded to give equal chance of kale consumption to each member of the population. It is likely that several even-numbered zip codes were selected that were contained inside a large city, thus skewing the opportunity for more people to eat kale and get more evidence. Because of this, there are 9-10x more kale-avoiders than kale-eaters. So it doesn't entirely make sense to conclude that kale can lead to a lower rate of diabetes.

(b) There is no evidence that kale causes a lower rate of diabetes in this situation because the criteria for selecting those who had to eat kale was completely subjective and based on the opinions of the members of the sample population. There are likely other substantive differences between people and so kale does not necessarily cause a lower diabetes rate.

(c) I believe that there is sufficient evidence to say that kale causes a lower rate of diabetes. The evidence is that the kale-eaters were literally randomly-assigned (or as close to random as we can get), and so there was no other criteria for eating kale, and everyone else probably continued with their life/diet as normal other than the kale requirements, so kale seems to have caused the lower rates of diabetes.

3. Problem 3 (prob3.r)



(a) Based on this plot (the black circles indicate that A and B both happened, red triangles indicate A but not B, green + signs indicate B but not A, and blue Xs indicate neither A nor B), I believe that A and B are independent events. There is a roughly equal number of data points that satisfy A and B, A but not B, B but not A, and neither A nor B, and so there seems to be no correlation between A and B being joint events.

(b) My confidence interval for P(A) is 0.508 +- .031, or [0.477, 0.539]. My confidence interval for P(A|B) is 0.51 +- .044, or [0.466, 0.554]. These intervals overlap by a large amount, so I believe that the probability of A happening is not related to whether or not B happens.

4. Problem 4

(a) The file prob4.r contains the work for this part. I got 0.500438 +- 0.00098 for my confidence interval for P(A), 0.500878 +- 0.00098 for P(B), and 0.250719 +- 0.00098 for P(A,B).

(b) Based on these data, A and B appear to be independent events. There is roughly a 50% chance for each of A and B to happen. If events are independent, P(Event 1, Event 2) = P(Event 1) * P(Event 2). P(A,B) is very close to P(A) * P(B), so I believe A and B to beNational Society of Black Engineers independent.

5. Problem 5

(a) These two events should be dependent because in most cases, the mayor appoints the chief of police to office. Politicians typically appoint like-minded people to themselves to offices, so if a person has one opinion about the mayor then they are likely to have a similar opinion about the chief of police.

(b) These two events should be independent because each event involves a different *independent* person whose political/personal views are not necessarily the same as those of another random citizen of the same city.

(c) These two events should be dependent because they cannot be true at the same time for any particular trial. The coin is either heads or tails, not both nor neither. If it is not one, it is the other.

(d) These two events should be dependent. I believe this because the movies that are of judgement are themselves related; one is a sequel of the other. They are similar enough in their overall existence that an opinion on one of them would likely be related to the opinion on the other one.