B365 Homework 2

1. 100,020 Massachusetts adults were randomly sampled with two variables recorded: whether or not the individual had diabetes, and whether or not the person ate Kale. The following gives a table of the results.

	Diabetes	No Diabetes
Kale	801	9192
No Kale	9905	80122

   (a) We write $P(\text{Diabetes}|\text{Kale})$ for the probability that a Massachusetts adult who eats kale has diabetes. Either give a value for $P(\text{Diabetes}|\text{Kale})$ or explain why it cannot be computed.

   (b) We write $\hat{P}(\text{Diabetes}|\text{Kale})$ for the proportion of Kale-eating members of our sample above that had diabetes. Either give a value for $\hat{P}(\text{Diabetes}|\text{Kale})$ or explain why it cannot be computed.

   (c) Give 95% confidence intervals for the probability of having diabetes for both the kale-eating and non-kale-eating members of our Massachusetts adults.

   (d) Can you conclude that the kale-eaters are less likely to have diabetes? Explain your reasoning.

   (e) Can you conclude that kale consumption *causes* a lower diabetes rate in this population? Explain your reasoning.

   (f) Come up with a possible theory that explains why kale-eaters have a lower rate of diabetes, but does not assume that kale causes the lower rate.

2. Consider the same numerical data, but imagine that the people were assigned to eat kale for 10 years in the following ways. In each case say if you believe there is evidence that Kale *causes* a lower rate of diabetes and explain why.

   (a) The kale-eaters were chosen from 4 zip-codes, two with an odd last digit and two with an even last digit. The ones with an odd last were chosen to eat kale. The rest chosen to not eat kale.

   (b) People were asked if they considered themselves to be health-conscious. The health-conscious people were not allowed to eat kale while the non-health-conscious people were forced to eat kale.

   (c) Each person used the R program and was assigned to the kale-eating group if and only

   ```
   runif(1) < .2
   ```

   in their program.

3. Consider the following computer experiment. Generate two "uniformly distributed" numbers x, y in the interval $[0, 1]$ (this is what runif does). Let A be the event that $x + y < 1$ and B be the event that $x - y < 0$.

   (a) Generate 1000 (x, y) values and plot these in R. Create the 1000-length boolean vectors a and b according to whether or not the above events A and B are satisfied. There are 4 possible truth assignments of A and B: $\{TT, TF, FT, FF\}$. Use a different plot character for each possible truth assignment. Thus, for example, all of points where both A and B occur would have the same plot symbol, and similarly for the other possible truth assignments. From this picture argue that A and B either *are* or *are not* independent.

   (b) Using your samples from the first part, compute 95% confidence intervals for $P(A)$ and $P(A|B)$. Are these confidence intervals consistent with A and B being independent?

4. Consider the following experiment for generating two boolean variables corresponding to the events A, B. Here $x\%\%y$ is the remainder when x is divided by y, so $x\%\%1$ is the "decimal part" of x.

   ```
   x = runif(1);
   A = (x < .5)
   B = ((2*x) %% 1) < .5
   ```

   (a) Simulate this experient 1,000,000 times and generate confidence intervals for $P(A)$, $P(B)$, and $P(A \cap B)$ — the prob of A *and* B both occuring.

(b) Do A and B appear to be independent events?

5. Say if the following pairs of events should be modeled as independent or dependent. Explain your reasoning.

   (a) We choose a voter at random (all voters equally likely) from Bloomington and let A be the event that the voter thinks the mayor is doing a good job and B be the event that the voter thinks the police chief is doing a good job.

   (b) Two people are selected at random from Bloomington and let A be the event that the first person approves of the mayor, while B is the event that the 2nd person approves of the mayor.

   (c) Flip a coin and let A be the event that the coin is heads and B be the event that the coin is tails.

   (d) A person is selected at random from Bloomington. A is the event that the person likes the movie "The Incredibles" while B is the event that the person likes "The Incredibles 2."

6. 30% of a population have trait T. Of those that have trait T, 90% have trait S, while only 30% have S in the remaining population.

   (a) Are T and S indepedent? Explain your reasoning.

   (b) Suppose that a randomly-selected individual has trait S. What is the probability that the individual has trait T?

   (c) Using R, simulate 10,000 members of this population, assigning whether or not traits S and T occur according to the given model.

   (d) From your simulation above, estimate a 95% confidence interval for $P(T|S)$ using

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

   Is this consistent with your earlier answer to the 2nd part of this problem?