Biostatistical Methods I (BIOS 5710)
Breheny

Assignment 2
Due: Thursday, September 15

1. (a) Ten people in a room have an average annual income of $40,000. An 11th person, with an income of $1,000,000, enters the room. Find the average income of all 11 people.
   (b) How large would the 11th person's income have to be in order to raise the average income to $250,000?
   (c) Suppose that all of the original ten people had an annual income of $40,000. The millionaire from part (b) enters the room. Find the median income of all 11 people.
   (d) As we discussed in class, no one person can skew the median very much. What is the minimum number of wealthy individuals that would have to enter the room in order to raise the median income in the room to $100,000?

2. [vBFHL 3.15] This question deals with the data discussed in problem 3.15 in the text, which is also available on the course website (`hypercalcemia.txt`).
   (a) Calculate the mean, standard deviation, median, and IQR of plasma prostaglandin E for patients with hypercalcemia; do the same for patients without hypercalcemia.
   (b) Make box plots for plasma prostaglandin E levels for each group. What conclusions would you draw from the plots?
   (c) Create a histogram for iPGE levels, conditioning on hypercalcemia status.
   (d) Create a scatter plot of iPGE levels vs. calcium levels, and color the points according to hypercalcemia status. (i) What calcium level seems to be the threshold for determining whether a patient is "hypercalcemic" or not? (ii) Describe the relationship between calcium levels and iPGE levels.

3. About 51% of the population of the U.S. is female. Also, about 12% of the population is over age 65.
   (a) If sex and age are independent, what percent of the population is female and over age 65?
   (b) Still assuming independence, what percent of the population is either female or over age 65?
   (c) In parts (a) and (b), sex and age were assumed to be independent. Why is this probably not a valid assumption? If we took dependence between the two into account, how would your answers for (a) and (b) change? (i.e., go up/go down)

4. According to data collected by the Department of Highway Safety and Motor Vehicles in Florida on people involved in traffic accidents in 1998, 0.37% suffered a fatal accident, 28.4% did not wear a seat belt, and 0.28% did not wear a seat belt and suffered a fatal accident.
   (a) What is the probability that an accident involved a person not wearing a seat belt, given that it was a fatal accident?
   (b) What is the probability that an accident involved a fatal injury, given that the person was not wearing a seat belt?

(c) What is the probability that an accident involved a fatal injury, given that the person was wearing a seat belt? How does this compare to (b)?

5. Brief background on genetics: humans have two copies of each gene, one copy from their mother and the other from their father. There is an equal chance that each of the father's copies will be the one that is transmitted, and likewise for the mother. Some genetic diseases are *recessive*, meaning that a person who has a normal copy and a disease-causing copy will be unaffected by the disease – only individuals with two disease-causing copies will be affected. Cystic fibrosis is an example of a recessive disorder.

   (a) Suppose that a man and a woman, neither of whom are affected by the disease, have a son with cystic fibrosis. What is the probability that their second child will be unaffected by the disease?

   (b) Suppose that the man and woman from part (a) have an unaffected daughter. What is the probability that she carries a harmful copy of the cystic fibrosis gene, given that she is unaffected?

   (c) Suppose that the daughter from part (b) marries a man with cystic fibrosis. If the couple has a child together, what is the probability that their child will be unaffected?

   (d) (This is kind of a hard one) Suppose that the couple from part (c) have an unaffected child. What is the probability that their second child will be unaffected, given that their first child was unaffected?

6. In 1988, the CDC conducted a large study investigating the accuracy of HIV testing with the following results:

		Patient's HIV status	
		Positive	Negative
Test	+	6545	47
	−	21	3004

   (a) What is the observed sensitivity of HIV testing?

   (b) What is the observed specificity of HIV testing?

   (c) Suppose that population A consists of rural married men, while population B consists of single urban men. You decide to randomly screen individuals from each population. Without doing any calculations, will the probability that an individual has the disease, given that they test positive, be the same for each population? Why or why not?

   (d) Suppose that the prevalence of the disease in population A is 0.005%, and the prevalence in population B is 3%. For each population, find the probability that an individual who tests positive truly has the disease.

7. [vBFHL 4.5] For each of the following, how would you interpret them as a long-run probability? If a person said that the probability was, say, 20%, is there a repeatable random process that could objectively test the accuracy of that statement?

   (a) The probability of rain tomorrow.

   (b) The probability of you, the person reading this, dying of leukemia.

   (c) The probability of life on Mars.

8. [vBFHL 4.10] The following is a cross-classification table of race and blood type among the U.S. population (restricted to Caucasians and African Americans):

	A	B	AB	O	Total
White	0.352	0.064	0.024	0.360	0.800
Black	0.054	0.040	0.008	0.098	0.200
Total	0.406	0.104	0.032	0.458	1.000

So, for example, the probability that a person is both white and has blood group A is 35.2%.

(a) Are the events "blood group A" and "White" statistically independent?

(b) Are the events "blood group A" and "White" mutually exclusive?

(c) Assuming independence, what is the expected probability of "blood group A and white"?

(d) What is the conditional probability of blood group A given that the individual is white? What about if the individual is black?

9. [vBFHL 4.11] Among Caucasians, 35% of people have two copies of the Rh gene (+/+), 48% of people have one copy (+/-), and 17% of people have no copies (-/-). A person will express the Rh factor on the surface of their red blood cells if they have at least one copy of the gene that produces this factor. Expressing the Rh factor is an example of a *dominant* condition; possessing just a single copy of the gene will cause the individual to have the condition.

This is important because potential problems can occur when an Rh-positive male mates with an Rh-negative female, as this may result in *Rh incompatibility*. Rh incompatibility occurs when the fetus is Rh-positive and the mother is Rh-negative, meaning that the fetus's blood expresses the Rh factor and the mother's body recognizes that blood as a foreign antigen.

(a) Assuming random mating with respect to the Rh factor, what is the probability of an Rh-negative female mating with an Rh-positive male?

(b) Given such a mating – i.e., given that the father is Rh-positive and the mother is Rh-negative – what is the probability of an Rh incompatibility?

(c) Unconditionally, across all random matings, what is the probability of an Rh incompatibility?