1. The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other than English (foreign language) at home, and 4.2% fall into both categories.

   (a) Are living below the poverty line and speaking a foreign language at home disjoint?

   No, there are people who are both living below the poverty line and speak a language other than English at home.

   (b) Draw a Venn diagram summarizing the variables and their associated probabilities.

   The diagram should show two circles that overlap. The speak-FL circle has .165 in the non-overlap; the below PL circle has .104 in the non-overlap; the intersection has .042.

   (c) What percent of Americans live below the poverty line and only speak English at home?

   Each person living below the poverty line either speaks only English at home or doesn't. Since .146 of Americans live below the poverty line and .042 speak a language other than English at home, the other .104 only speak English at home.

   (d) What percent of Americans live below the poverty line or speak a foreign language at home?

   Using the General Addition Rule:

$$\begin{aligned} P(below\, PL\, or\, speak\, FL) &= P(below\, PL) + P(speak\, FL) - P(both) \\ &= 0.146 + 0.207 - 0.042 = 0.311 \end{aligned}$$

   (e) What percent of Americans live above the poverty line and only speak English at home?

   P(neither below PL nor speak FL) = 1 - P(below PL or speak FL) = 1 - 0.311 = 0.689

   (f) Is the event that someone lives below the poverty line independent of the event that the person speaks a foreign language at home?

   *Two approaches: 1) Using the multiplication rule: P(below PL) * P(speak FL) = 0.146*0.207 = 0.030, which does not equal P(below PL and speak FL) = 0.042, therefore the events are dependent. 2) Using Bayes' Theorem: If the two events are independent, then P(below PL | speak FL) = P(below PL).*

$$\begin{aligned} P(below\, PL \,|\, speak\, FL) &= \frac{P(below\, PL\, and\, speak\, FL)}{P(speak\, FL)} \\ &= \frac{0.042}{0.207} \approx 0.203 \end{aligned}$$

2. In parts (a) and (b), identify whether the events are disjoint, independent, or neither (events cannot be both disjoint and independent).

   (a) You and a randomly selected student from your class both earn A's in this course.

   (b) You and your class study partner both earn A's in this course.

   (c) If two events can occur at the same time, must they be dependent?

   (a) If the class is not graded on a curve, they are independent. If graded on a curve, then neither independent nor disjoint – unless the instructor will only give one A, which is a situation we will ignore in parts (b) and (c).

   (b) They are probably not independent: if you study together, your study habits would be related, which suggests your course performances are also related.

(c) *No. See the answer to part (a) when the course is not graded on a curve. More generally: if two things are unrelated (independent), then one occurring does not preclude the other from occurring.*

3. Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.

   (a) What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?

   (b) What is the probability that a student chosen at random misses no more than one day?

   (c) What is the probability that a student chosen at random misses at least one day?

   (d) If a parent has two kids at a DeKalb County elementary school, what is the probability that neither kid will miss any school? Note any assumption you must make to answer this question.

   (e) If a parent has two kids at a DeKalb County elementary school, what is the probability that both kids will miss some school, i.e. at least one day? Note any assumption you make.

   (f) If you made an assumption in part (d) or (e), do you think it was reasonable? If you didn't make any assumptions, double check your earlier answers.

   (a) *P(no misses) = 1 - (0.25 + 0.15 + 0.28) = 0.32*

   (b) *P(at most 1 miss) = P(no misses) + P(1 miss) = 0.32 + 0.25 = 0.57*

   (c) *P(at least 1 miss) = P(1 miss) + P(2 misses) + P(3+ misses) = 1 - P(no misses) = 1 - 0.32 = 0.68*

   (d) *For parts (d) and (e) assume that whether or not one kid misses school is independent of the other. P(neither miss any) = P(no miss) * P(no miss) = 0.32 * 0.32 = 0.1024*

   (e) *P(both miss some) = P(at least 1 miss) * P(at least 1 miss) = .68 * .68 = 0.4624*

   (f) *These kids are siblings, and if one gets sick it probably raises the chance that the other one will get sick as well. So whether or not one misses school due to sickness is probably not independent of the other.*

4. P(A) = 0.3, P(B) = 0.7

   (a) Can you compute P(A and B) if you only know P(A) and P(B)?

   (b) Assuming that events A and B arise from independent random processes,

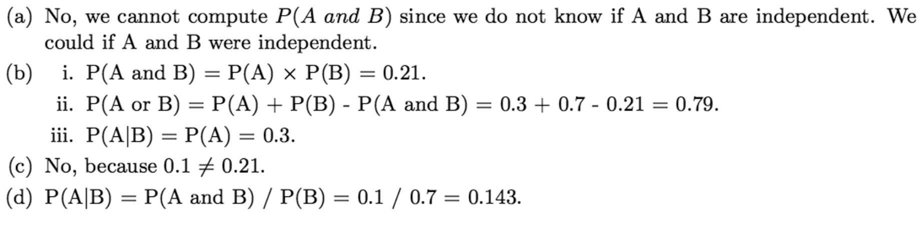
      (i) what is P(A and B)?

      (ii) what is P(A or B)?

      (iii) what is P(A|B)?

   (c) If we are given that P(A and B) = 0.1, are the random variables giving rise to events A and B independent?

   (d) If we are given that P(A and B) = 0.1, what is P(A|B)?



(a) No, we cannot compute $P(A \textit{ and } B)$ since we do not know if A and B are independent. We could if A and B were independent.

(b) i. P(A and B) = P(A) × P(B) = 0.21.

   ii. P(A or B) = P(A) + P(B) - P(A and B) = 0.3 + 0.7 - 0.21 = 0.79.

   iii. P(A|B) = P(A) = 0.3.

(c) No, because $0.1 \neq 0.21$.

(d) P(A|B) = P(A and B) / P(B) = 0.1 / 0.7 = 0.143.

5. A 2010 Pew Research poll asked 1,306 Americans "From what you've read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?". The table below shows the distribution of responses by party and ideology, where the counts have been replaced with relative frequencies.

		Response			
		Earth is warming	Not warming	Don't Know Refuse	Total
Party and Ideology	Conservative Republican	0.11	0.20	0.02	0.33
	Mod/Lib Republican	0.06	0.06	0.01	0.13
	Mod/Cons Democrat	0.25	0.07	0.02	0.34
	Liberal Democrat	0.18	0.01	0.01	0.20
	Total	0.60	0.34	0.06	1.00

(a) Are believing that the earth is warming and being a liberal Democrat mutually exclusive?

(b) What is the probability that a randomly chosen respondent believes the earth is warming or is a liberal Democrat?

(c) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a liberal Democrat?

(d) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a conservative Republican?

(e) Does it appear that whether or not a respondent believes the earth is warming is independent of their party and ideology? Explain your reasoning.

(f) What is the probability that a randomly chosen respondent is a moderate/liberal Republican given that he does not believe that the earth is warming?

(a) No, 0.18 of respondents fall into this combination.

(b) P(earth is warming or liberal Democrat) =
= P(earth is warming) + P(liberal Democrat) - P(earth is warming and liberal Democrat)
= 0.60 + 0.20 - 0.18 = 0.62

(c) P(earth is warming | liberal Democrat) $= \frac{.18}{.20} = .33$

(d) P(earth is warming | conservative Republican) $= \frac{.11}{.33} = .33$

(e) No, the two appear to be dependent. The percentages of conservative Republicans and liberal Democrats who believe that there is solid evidence that the average temperature on earth has been getting warmer over the past few decades are very different.

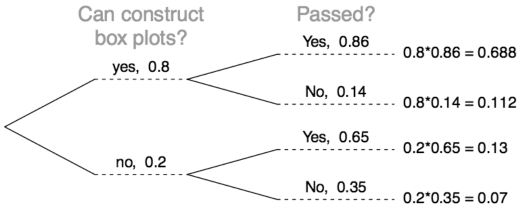
(f) P(moderate/liberal Republican | not warming) $= \frac{.06}{.34} = .18$

6. After an introductory statistics course, 80% of students can successfully construct box plots. Of those who can construct box plots, 86% passed, while only 65% of those students who could not construct box plots passed.

(a) Construct a tree diagram of this scenario.

(b) Calculate the probability that a student is able to construct a box plot if it is known that he passed.

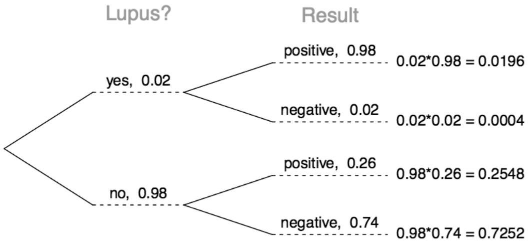
(a) A tree diagram of this scenario is below:



(b)

$$\begin{aligned} P(\textit{can construct} \mid \textit{pass}) &= \frac{P(\textit{pass and can construct})}{P(\textit{pass})} \\ &= \frac{0.80 \times 0.86}{0.80 \times 0.86 + 0.20 \times 0.65} \\ &= \frac{0.688}{0.818} \\ &\approx 0.84 \end{aligned}$$

7. Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease. The test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease. There is a line from the Fox television show *House* that is often used after a patient tests positive for lupus: "It's never lupus." Do you think there is truth to this statement? Use appropriate probabilities to support your answer.



$$\begin{aligned} P(\textit{lupus}|\textit{positive}) &= \frac{P(\textit{lupus and positive})}{P(\textit{positive})} \\ &= \frac{0.0196}{0.0196 + 0.2548} \\ &= 0.0714 \end{aligned}$$

Even when a patient tests positive for lupus, there is only a 7.14% chance that he actually has lupus. House may be right.

8. At a university, 13% of students smoke.

(a) Calculate the expected number of smokers in a random sample of 100 students from this university.

(b) The university gym opens at 9 am on Saturday mornings. One Saturday morning at 8:55 am there are 27 students outside the gym waiting for it to open. Should you use the same approach from part (a) to calculate the expected number of smokers among these 27 students?

(a) E(X) = 100 × 0.13 = 13

(b) No, these 27 students are not a random sample from the university's student population. For example, it might be argued that the proportion of smokers among students who go to the gym at 9am on a Saturday morning would be lower than the proportion of smokers in the university as a whole.

9. Consider the following card game with a well-shuffled deck of cards. If you draw a red card, you win nothing. If you get a spade, you win \$5. For any club, you win \$10 plus an extra \$20 for the ace of clubs.

   (a) Create a probability model for the amount you win at this game. Also, find the expected winnings for a single game and the standard deviation of the winnings.

   (b) What is the maximum amount you would be willing to pay to play this game? Explain your reasoning.

(a) The probability model and the calculation of the expected value and standard deviation are shown below:

Event	X	$P(X)$	$X \cdot P(X)$	$(X - E(X))^2$	$(X - E(X))^2 \cdot P(X)$
Red	0	$\frac{26}{52}$	$0 \times \frac{26}{52} = 0$	$(0 - 4.14)^2 = 17.1396$	$17.1396 \times \frac{26}{52} = 8.5698$
Spade	5	$\frac{13}{52}$	$5 \times \frac{13}{52} = 1.25$	$(5 - 4.14)^2 = 0.7396$	$0.7396 \times \frac{13}{52} = 0.1849$
Club	10	$\frac{12}{52}$	$10 \times \frac{12}{52} = 2.31$	$(10 - 4.14)^2 = 34.3396$	$34.3396 \times \frac{12}{52} = 7.9245$
Ace of clubs	30	$\frac{1}{52}$	$30 \times \frac{1}{52} = 0.58$	$(30 - 4.14)^2 = 668.7396$	$668.7396 \times \frac{1}{52} = 12.8604$
			$E(X) = 4.14$		$V(X) = 29.5396$ $SD(X) = \sqrt{V(X)} = 5.4350$

(b) If you are playing with the goal of making money, you should not play unless the cost of the game is less than \$4.14.

10. An airline charges the following baggage fees: \$25 for the first bag and \$35 for the second. Suppose 54% of passengers have no checked luggage, 34% have one piece of checked luggage and 12% have two pieces. We suppose a negligible portion of people check more than two bags.

    (a) Build a probability model, compute the average revenue per passenger, and compute the corresponding standard deviation.

    (b) About how much revenue should the airline expect for a flight of 120 passengers? With what standard deviation? Note any assumptions you make and if you think they are justified.

(a) The probability model and the calculation of average revenue per passenger (expected value) are as follows:

Event	X	P(X)	X · P(X)	$(X - E(X))^2$	$(X - E(X))^2 \cdot P(X)$
No baggage	0	0.54	0	$(0 - 15.70)^2 = 246.49$	$246.49 \times 0.54 = 133.10$
1 checked bag	25	0.34	8.5	$(25 - 15.70)^2 = 86.49$	$86.49 \times 0.34 = 29.41$
2 checked bags	60	0.12	7.2	$(60 - 15.70)^2 = 1962.49$	$1962.49 \times 0.12 = 235.50$
			E(X) = \$15.70		V(X) = \$398.01 $SD(X) = \sqrt{V(X)} = \$19.95$

(b) We assume independence between individual fliers. This probably is not exactly correct, but it would provide a helpful first approximation for the true revenue.

$$\begin{aligned} E(X_1 + \cdots + X_{120}) &= E(X_1) + \cdots + E(X_{120}) = 120 \times 15.70 = \$1,884 \\ V(X_1 + \cdots + X_{120}) &= V(X_1) + \cdots + V(X_{120}) = 120 \times 398.01 = \$47,761.20 \\ SD(X_1 + \cdots + X_{120}) &= \sqrt{47,761.20} = \$218.54 \end{aligned}$$

11. The relative frequency table below displays the distribution of annual total personal income (in 2009 inflation-adjusted dollars) for a representative sample of 96,420,486 Americans. These data come from the American Community Survey for 2005-2009. This sample is comprised of

59% males and 41% females.

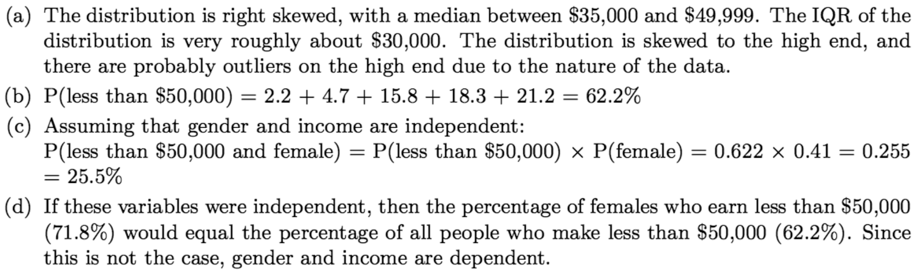
(a) Describe the distribution of total personal income.

(b) What is the probability that a randomly chosen US resident makes less than $50,000 per year?

(c) What is the probability that a randomly chosen US resident makes less than $50,000 per year and is female? Note any assumptions you make.

(d) The same data source indicates that 71.8% of females make less than $50,000 per year. Use this value to determine whether or not the assumption you made in part (c) is valid.

Income	*Total*
$1 to $9,999 or loss	2.2%
$10,000 to $14,999	4.7%
$15,000 to $24,999	15.8%
$25,000 to $34,999	18.3%
$35,000 to $49,999	21.2%
$50,000 to $64,999	13.9%
$65,000 to $74,999	5.8%
$75,000 to $99,999	8.4%
$100,000 or more	9.7%



(a) The distribution is right skewed, with a median between $35,000 and $49,999. The IQR of the distribution is very roughly about $30,000. The distribution is skewed to the high end, and there are probably outliers on the high end due to the nature of the data.

(b) P(less than $50,000) = 2.2 + 4.7 + 15.8 + 18.3 + 21.2 = 62.2%

(c) Assuming that gender and income are independent:
P(less than $50,000 and female) = P(less than $50,000) × P(female) = 0.622 × 0.41 = 0.255 = 25.5%

(d) If these variables were independent, then the percentage of females who earn less than $50,000 (71.8%) would equal the percentage of all people who make less than $50,000 (62.2%). Since this is not the case, gender and income are dependent.

Extra Challenge Problem: A *chord* of a circle is a straight line segment whose endpoints both lie on the circle. For a fixed circle, what is the probability that the length of a randomly drawn chord will exceed that circle's radius?