Stat 185 - Review problems for the Midterm

In progress - still needs stuff on Hypothesis testing!!

1. 910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below as well as in figure 1.

		Political ideology			
		Conservative	Moderate	Liberal	Total
Response	(i) Apply for citizenship	57	120	101	278
	(ii) Guest worker	121	113	28	262
	(iii) Leave the country	179	126	45	350
	(iv) Not sure	15	4	1	20
	Total	372	363	175	910

   (a) What percent of these Tampa, FL voters identify themselves as conservatives?

   (b) What percent of these Tampa, FL voters are in favor of the citizenship option?

   (c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

   (d) What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

   (e) Do political ideology and views on immigration appear to be independent? Explain your reasoning.

2. According to `https://goo.gl/DanB3A`, Utah has the lowest percentage of adults who smoke at 9.1%.

   (a) Given a random sample of 120 adult Utahns, what is the probability that exactly half of them smoke?
   Assume that the 9.1% number is correct and write down you answer using a formula; a decimal approximation is not needed.

   (b) I don't actually believe the 9.1% so I went down to Woody's Tavern and surveyed 11 customers. As it turned out, 5 of them smoked! Identify any problems that you see with my "study".

3. Consider the data set $-1, 2, 2, 5$, which is a sample of a much larger population.

   (a) Write down a simple computation showing that the mean of this sample is 2.

   (b) Write down a simple computation showing that the standard deviation of this sample is $\sqrt{6}$.

   (c) What is the standard error of the data set?

4. I've got a six sided die with four sides labeled 0 and two sides labeled 1. Let X denote the result of one roll and let S denote the sum of the results of 99 rolls. Compute the following:

   (a) $P(X = 1)$

   (b) $P(S = 10)$

   (c) The mean of X

   (d) The mean of S

   (e) The standard deviation of X

   (f) The standard deviation of S

   (g) Use a normal distribution to estimate $P(30 < S < 40)$.

5. Suppose that a standardized tests' scores are normally distributed with a mean of 1500 and a standard deviation of 300.

   (a) According to the normal rules of thumb, what percentage of test takers score 2400 or above?

   (b) Use the normal table to compute $P(1400 < X < 1700s)$.

6. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter. The standard error for this estimate was 2.4%, and a normal distribution may be used to model the sample proportion.

   (a) *Quick*: Does the data provide a 95% level of confidence that more than half of U.S. adult Twitter users get some news through Twitter?

   (b) *Sorta quick*: Construct a rough 95% confidence interval for the fraction of U.S. adult Twitter users who get some news on Twitter.

   (c) If we construct a 90% confidence interval for the percentage of U.S. adults Twitter users who get some news through Twitter, this confidence interval will be wider than the corresponding 95% confidence interval.

   (d) What should we do if we would like to reduce the standard error of the estimate?

7. The Peachtree 10K is held every July 4th in Atlanta and regularly draws over 10,000 runners. In fact, more than 11,000 people ran the Peachtree in both 2010 and 2015. Let's try to use this data to address the following question: Do runners slow down between ages of 50 and 55?

   (a) Discuss the advantages and disadvantages of the following approaches to exploring the question. Be sure to use appropriate terminology.

   - Compare the times of all 1059 fifty year old runners in 2010 to the times of all 904 fifty-five year old runners in 2010.
   - Compare the times of all 293 fifty year old runners in 2010 who also ran in 2015 to their own time five years later.
   - Compare the times of all 461 fifty year old *female* runners in 2010 to the times of all 387 fifty-five year old *female* runners in 2010.
   - Compare the times of all 120 fifty year old female runners in 2010 who also ran in 2015 to their own time five years later.

   (b) Suppose I find twelve 50 year old women who raced the Peachtree in both 2010 and 2015. For each participant, I subtract their 2015 time from their 2010 time.

   Consider the set of all the men who ran the race in both 2010 and in 2015 an who were 50 years old in 2010. Let D_1 denote all the times of those men in 2010 and let D_2 denote all the times of those men in 2015. We form the pairwise difference

   $$D = D_1 - D_2.$$

   Of course, we might suppose that the mean of D_1 is less than the mean of D_2, indicating that the men have, on average, slowed down a bit. Suppose we actually compute that D has a mean of -5.2 minutes with a standard deviation of 9.3 minutes.

   - Write down an appropriate hypothesis test to explore the probability that runners are indeed slowing down between the ages of 50 and 55.
   - Use the information above to compute the T-score associated with your hypothesis.
   - Use the t-tables to decide if you should accept or reject the null hypothesis.

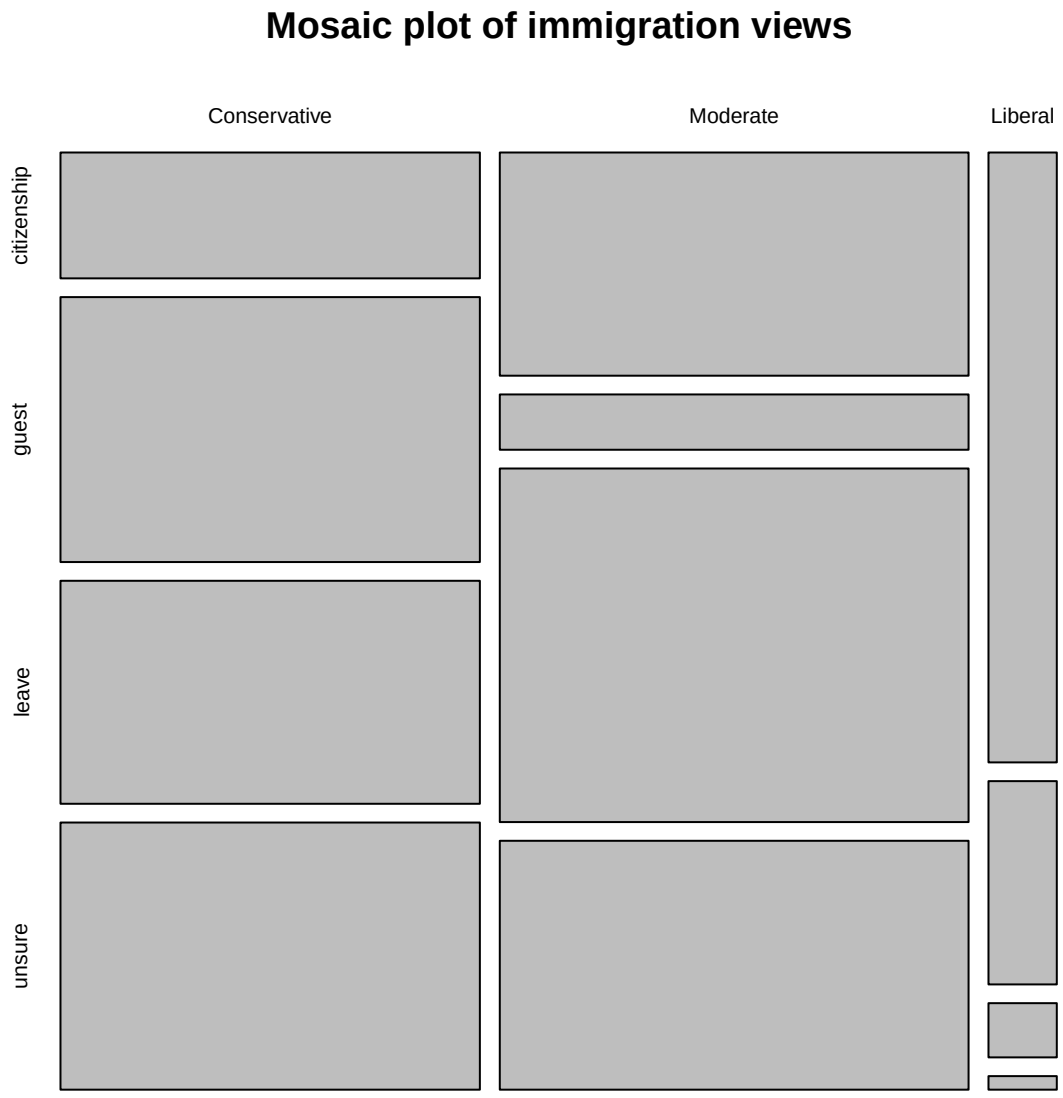

Figure 1: The Mosaic plot for problem 1