Math 183 HW15, Name: _____, PID: _____

On this homework assignment, feel free to use R whenever you must run a chi-squared test. Whenever you use R to do anything, simply include the code you typed and weave this into your answer.

6.40 True or false, Part II. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) As the degrees of freedom increases, the mean of the chi-square distribution increases.

(b) If you found $\chi^2 = 10$ with $df = 5$ you would fail to reject H_0 at the 5% significance level.

(c) When finding the p-value of a chi-square test, we always shade the tail areas in both tails.

(d) As the degrees of freedom increases, the variability of the chi-square distribution decreases.

6.42 Evolution vs. creationism. A Gallup Poll released in December 2010 asked 1019 adults living in the Continental U.S. about their belief in the origin of humans. These results, along with results from a more comprehensive poll from 2001 (that we will assume to be exactly accurate), are summarized in the table below:[61]

	Year	
Response	2010	2001
Humans evolved, with God guiding (1)	38%	37%
Humans evolved, but God had no part in process (2)	16%	12%
God created humans in present form (3)	40%	45%
Other / No opinion (4)	6%	6%

(a) Calculate the actual number of respondents in 2010 that fall in each response category.

(b) State hypotheses for the following research question: have beliefs on the origin of human life changed since 2001?

(c) Calculate the expected number of respondents in each category under the condition that the null hypothesis from part (b) is true.

(d) Conduct a chi-square test and state your conclusion. (Reminder: Verify conditions.)

6.48 Coffee and Depression. Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.[63]

		Caffeinated coffee consumption					
		≤ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	≥ 4 cups/day	Total
Clinical depression	Yes	670	373	905	564	95	2,607
	No	11,545	6,244	16,329	11,726	2,288	48,132
	Total	12,215	6,617	17,234	12,290	2,383	50,739

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

(b) Write the hypotheses for the test you identified in part (a).

(c) Calculate the overall proportion of women who do and do not suffer from depression.

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2/Expected$.

(e) The test statistic is $\chi^2 = 20.93$. What is the p-value?

(f) What is the conclusion of the hypothesis test?

(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study.[64] Do you agree with this statement? Explain your reasoning.

6.50 How's it going? The American National Election Studies (ANES) collects data on voter attitudes and intentions as well as demographic information. In this question we will focus on two variables from the 2012 ANES dataset:[65]

- region (levels: Northeast, North Central, South, and West), and
- whether the respondent feels things in this country are generally going in the right direction or things have pretty seriously gotten off on the wrong track.

To keep calculations simple we will work with a random sample of 500 respondents from the ANES dataset. The distribution of responses are as follows:

	Right Direction	Wrong Track	Total
Northeast	29	54	83
North Central	44	77	121
South	62	131	193
West	36	67	103
Total	171	329	500

(a) Region: According to the 2010 Census, 18% of US residents live in the Northeast, 22% live in the North Central region, 37% live in the South, and 23% live in the West. Evaluate whether the ANES sample is representative of the population distribution of US residents. Make sure to clearly state the hypotheses, check conditions, calculate the appropriate test statistic and the p-value, and make your conclusion in context of the data. Also comment on what your conclusion says about whether or not this sample can be considered to be representative.

(b) Region and direction:

(i) We would like to evaluate the relationship between region and feeling about the country's direction. What is the response variable and what is the explanatory variable?

(ii) What are the hypotheses for evaluating this relationship?

(iii) Complete the hypothesis test and interpret your results in context of the data and the research question.

Do not write your answer on this page (but do scan it in your PDF document!). Write your answer on the next page, which has been left blank to give you plenty of room.

DO NOT WRITE IN THIS SPACE!

(Answer to 6.50 on this page)

R Questions

R1. Mars Candy announces a new holiday line of M&Ms that is supposed to contain candies in these percentages: 30% red, 30% green, 25% white, and 15% silver. You buy a jumbo bag of the new line and count how many of each color appear: 97 red, 84 green, 70 white, and 62 silver. You are curious if your bag provides strong evidence against the Mars published color percentages.

Do these things: (1) state your hypotheses clearly, (2) discuss what conditions are needed to test this scenario and if you meet them, (3) write R code that performs the test, and (4) give a clear conclusion based on the P-value assuming a significance level of 0.05.

R2. Is the sum of two independent chi-squared distributions again a chi-squared distribution? To find out, set $X = \chi^2_m$ and $Y = \chi^2_n$ (the chi-squared distributions with *m* and *n* degrees of freedom). Using R, draw a random sample of 5000 numbers from *X* (choose some value for *m*) and 5000 numbers from *Y* (choose some value for *n*). Add these vectors and make a histogram of the result. Does it look like a chi-squared distribution? If so, which one (that is, how many degrees of freedom)? Explore different values for *m* and *n* until you think you know the answer. Once you have a conjecture, use Google to determine if you are right. **[code to explore the question, distribution that describes $\boldsymbol{X + Y}$]**