STA130 Homework #2: Analysis of Marital Mortality

(Due in lecture at 2:10 pm sharp on Monday, February 27, 2017.)

NOTE: You may discuss this assignment with other students, but you must understand and type and create your solutions entirely by yourself. You can also ask your TA for assistance, either after a tutorial or by arranging to meet with them another time. And, don't forget the information at www.probability.ca/Rinfo.html about the "R" statistical software package.

A medical study in 2013 was announced in the media with headlines[1] like "Marriage Might Lengthen Life" and[2] "Stay single, die younger, say scientists", and advice[3] like "A new study might be evidence enough for single people to hightail it to Vegas and get hitched". The detailed study[4] performed many complicated statistical adjustments based on various covariates (i.e., characteristics of the patients studied), but the essence of the study was:

The study followed 4,802 people who were born in the 1940s and had attended the University of North Carolina (UNC) during 1964–1966. For each such person, it was determined if they had ever married (Yes/No), and also if they had died by August 2009 (Lived/Died). Out of the 4,802 people who were followed, the number in each of the four corresponding categories was as follows:

	Married?	
	Yes	No
Lived:	4,253	311
Died:	200	38

In particular, of the 4,802 people followed, a total of 238 of them died by August 2009, which as a fraction equals $p_{true} := 238/4,802 = 0.04956268$. For now, we will assume for simplicity that this p_{true} is the "true" precise probability that a random person in the study would die by August 2009. (This isn't strictly accurate, but it is approximately accurate, and it will simplify the remaining analysis.)

Based on the above information, consider the following questions. Answer them as best as you can, making use of "R" as necessary. You should then prepare answers to hand in at your tutorial. Your answers should include explanations in complete English sentences, together with precise numerical answers where appropriate. You should also include, with explanation, all of the "R" commands you used.

1. Of the 349 never-married people, what fraction of them died by August 2009? How does this fraction compare to the value p_{true} found above? What fact might this indicate, about the effect of marriage on mortality, i.e. on when people die?

---

[1] http://news.health.com/2013/01/11/marriage-might-lengthen-life/

[2] http://www.telegraph.co.uk/news/health/news/9796903/Stay-single-die-younger-say-scientists.html

[3] http://www.hlntv.com/article/2013/01/17/getting-married-40-leads-longer-life

[4] *Annals of Behavioral Medicine* **45(3)**, 338–347, 2013. Available at: www.probability.ca/sta130/maritalstudy.pdf

2. Suppose we wished to test whether the fact suggested in the previous question was real, or was "just luck". What null hypothesis should we assume? What alternative hypothesis?

3. In the previous question, under the null hypothesis, what would be the mean and variance and sd of the number of deaths in a random sample of 349 people from the study? Explain.

4. Specify (in words) what probability we would need to compute to obtain a P-value to test whether the fraction found in Question #1 above is "statistically significantly" different from p_{true}?

5. Using R, carry out the statistical test identified in the previous question. If possible, compute your P-value both exactly [using the binomial distribution], and approximately [using the normal distribution, both with and without the continuity correction], and compare the various different values that you obtain. For greater clarity, you may wish to include a graph diagram (similar to those in the lecture notes) showing which individual probabilities were counted when computing your P-value.

6. Based on the results of this test, what can you conclude? (Write your answer in complete English sentences, and be as clear and precise as possible.)

7. What are some limitations of the above study? That is, in what ways could the study's conclusions be flawed, or incorrect, or have an alternative explanation? List as many as you can, as clearly as you can. (If possible, consider such issues as "correlation does not imply causation" and "sampling bias", each of which you may need to look up.)

Hand in your answers and R commands and printed output to your TA, at the start of lecture at 2:10 pm on Monday February 27 (with penalties if it is even one minute late!).