LAST NAME (Please Print): KEY

FIRST NAME (Please Print): __________________________________________

HONOR PLEDGE (Please Sign): ________________________________________

Statistics 101

Homework 1

You are allowed to discuss problems with other students, but the final answers must be your own work.

For all problems that require calculation, YOU MUST ATTACH SEPARATE PAGES, NEATLY WRITTEN, THAT SHOW YOUR WORK.

Please mark your answer in the space provided. As a general rule, each blank counts for one point. If necessary work is not shown, or if that work is substantially wrong, then you will not get credit even if the answer is correct. (The obvious purpose of this seemingly draconian policy is to prevent people from mindlessly copying each other’s answers.)

Report all numerical answers to at least two correct decimal places.

Upload to Gradescope by the START of class on TUESDAY, JANUARY 18.
In the context of driving safety, people often maintain that young drivers (17 and under) are unsafe because they have more accidents than any other age group. But this does not account for the possibility that they drive more miles than some other age groups; it could be that their accident rate, controlling for mileage, is pretty good.

Your job is to generalize the strategy for weighted averages so that one can make meaningful comparisons of accident rates across age groups. The raw information you need to do this can be found at three websites. To get the information on the numbers of driving accidents by age and gender in New York state, use the data at www.dmv.ny.gov/stats.htm. Go to the “Archives of Statistical Summaries,” select New York State Accident Data – 2001, and examine Table 8 (as a check, the total number of drivers with accidents is 548,530). To get information on the number of miles driven by age, the best available information is at www.bts.gov: search on “National Household Travel Survey,” click on “Highlights of the 2001 National Household Travel Survey” and go to table A-17 (as a check, “All persons 15 and older” drove an average of 29.1 miles per day). To approximate the age/gender mix in New York state in 2001, we shall use data from the 2000 decennial census. Go to the U.S. Census Bureau’s website at www.census.gov and explore around.

Note: I find the Census site to be a bit of a pain, so be persistent. As a check, the total is 18,976,457. If you cannot find this basic information on the Census website, click on the feedback link at the top right of the main Census webpage and express your dissatisfaction. And then look on our class’s FAQ site.

16,446 1. In New York State, the total number of accidents in 2001 for people under 16 is 664. What is the total number of accidents for people between 16 and 17, inclusive? (Prorate the counts under the assumption that the accident rates are equal within the age group; also, assume that all of the “Under 16” group are 15.)

This required no prorating—just sum the numbers for males and females.

515017.6 2. From the U.S. Census Bureau’s website, what is the total number of people between 16 and 17?

\[ \frac{2}{5} \times 1287544 = 515017.6. \]

6283214.72 3. Assuming the New York State drivers are similar to those studied in the in the National Household Travel Survey, what is your estimate of the total number of miles driven in one day by young drivers (16 to 17) in New York State?

12.2 \times 515017.6 = 6,283,214.72
4. Find the accident rates per 100,000 miles for people aged 16 to 17 and for people who are 85 to 88. (This controls for mileage.) Note that the Census and the NY DMV do not have corresponding age intervals. To prorate the counts, assume that with each additional year after 75, 12% of the people are removed from the driving population. (Hint: if \(|p| < 1\), then \(\sum_{i=0}^{\infty} p^i = 1/(1 - p)\).) Assume that the accident rate in those aged 85-88 does not change with age. Also assume that people who are 80 or older drive an average of 5 miles per day (which is consistent with the trend in National Household Travel Survey).

0.72 Young Drivers (16 to 17)

The number of miles driven in a year is 365 * 6,283,214.72. The number of accidents is 16,446. So the number of accidents per 100,000 miles is 16,446/[365 * 6,283,214.72/100,000] = 0.7171.

0.65 Seniors (85 to 88)

For seniors, there are 4535 + 2951 = 7486 accidents in the over-eighty group. Under the prorating assumption, the number drops by 12% for each year, so if there are \(x\) accidents for 80-year-olds, there are \(0.88x\) accidents for 81-year-olds, \(0.88^2x\) accidents for 82-year-olds, and so forth. Thus \(7486 = \sum_{i=0}^{\infty} x \cdot 0.88^i = x \cdot \frac{1}{1-0.88} = 8.3333x\), so \(x = 898.320004\). Thus the number of accidents for people aged 85 to 88 is \((0.88^5 + 0.88^6 + 0.88^7 + 0.88^8) \cdot 898.320004 = 1581.4440\).

To estimate the number of people aged 85-88, under our (somewhat loose) assumptions, we use the fact that 860,818 people are aged 75-84. If \(y\) is the number of 75-year-olds, then the previous logic says 860,818 = \(y + 0.88y + 0.88^2y + \cdots + 0.88^9y = 6.01249 \cdot y\), so \(y = 143,171.5866\). Then the number aged 85-88 is \(y(0.88^{10} + 0.88^{11} + 0.88^{12} + 0.88^{13}) = 143,171.5866 \cdot 0.92904 = 133,012.6474\).

So the number of accidents per 100,000 miles is 1,581.444/[(133,012.6474 * 5 * 365)/100,000] = 0.6515.

5. What conclusion do you draw from the above analysis?

16-17 year-old drives are more apt to have accidents than people aged 85-88.

6. Suppose you wanted to argue that the above analysis was not correct. List two assumptions you might reasonably question.
1. The NHTS national averages may not accurately represent New York behavior.

2. Teenage daily mileage seems low, and may be underreported (e.g., to reduce insurance costs).

3. Our assumption that seniors drive 5 miles per day on average may be wrong.

4. We have assumed that 12% of the seniors stop driving each year—that may be wrong. Similarly, we assumed constant accident rates in extreme old age.

7. In looking at the accident counts in New York broken out by severity, something is odd. What is it?

There are far too few property damage accidents.

8. How would you explain the oddity in the previous question.

People do not want to report accidents to insurance companies, and will pay out-of-pocket to keep their rates low. (Also, for collisions with parked cars there may be no victim present at the time to insist on reporting.)

9. In looking at the accident counts in New York broken out by gender, it is clear that men have about three times as many fatal accidents as women. Is it right to think men are three times as dangerous? Defend or refute.

According to NHTS table A-17, men drive about twice as many miles as women do. So men are only about 1.5 times more dangerous than women.

10. At the bottom of the DMV table, there are 17,355 accidents for which the gender is unknown. Why?

Mostly hit-and-run drivers, of course.
11. **Not Graded.** Could you find the NY State information on the Census website? If not, did you contact the Census Bureau about the poor design of their site, whose entire *raison d’etre* is to provide exactly such information?

12. Researchers at Kaiser Permanente found that women who used oral contraceptives had higher rates of cervical cancer than non-users, after controlling for age, education, and marital status. Their paper (“The incidence of cervical cancer and duration of oral contraceptive use,” by Peritz et al. in the *American Journal of Epidemiology*, 272, pp. 462–469) concluded that the birth control pill causes cervical cancer.

   Is this a designed experiment or an observational study? **Obs. Study**

   Women who used the pill probably differ from non-users in another way that was related to cervical cancer risk. Explain. (Hint: Rick Perry used to support a vaccination program in Texas—what was it?)

   Women on the pill tend to be more sexually active than those who are not. This exposes them to the risk of HPV, which is a cause of cervical cancer.

   **No** Were the conclusions of Peritz et al. justified?

13. Between 2000 and 2013, the median U.S. weekly wage, adjusted for inflation, grew by 0.9 per cent. But the median wage has actually fallen for high-school dropouts (7.9%), high-school graduates (4.7%), people with some college education (7.6%) and people with at least one college degree (1.2%). Explain this.

   Salary increases with education. Between 2000 and 2013, the proportion of the U.S. work force with some college or a college degree increased dramatically, leading to a net rise despite falling wages within each category.

14. The following table shows the relationship between crew, first, second, and steerage ticket classes and survival on the *Titanic* (it excludes servants in first class, but it is respectful to note that 0/24 female servants died, whereas 10/12 male servants died).
You can find a fascinating statistical perspective on this at http://www.icyousee.org/titanic.html; it goes down to the level of the lifeboats and the order in which they were launched—each lifeboat was the stage for its own drama, and a wide range of moral choices were made.

If your last name begins with A-I, please compare crew and steerage survival rates. If your last name begins with J-R, please compare first and steerage class survival rates. If your last name begins with S-Z, please compare second and steerage class survival rates. But use information from all four groups in calculating the adjusted survival rates.

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survival</td>
<td>Death</td>
<td>Survival</td>
</tr>
<tr>
<td>Crew</td>
<td>20</td>
<td>2</td>
<td>193</td>
</tr>
<tr>
<td>First Class</td>
<td>113</td>
<td>4</td>
<td>55</td>
</tr>
<tr>
<td>Second Class</td>
<td>78</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Steerage</td>
<td>88</td>
<td>91</td>
<td>59</td>
</tr>
</tbody>
</table>

What is the overall survival rate for the non-steerage group in your comparison?

The unadjusted survival rate for the crew is \((20 + 193)/(20 + 193 + 2 + 701)\) = 0.1385.

The unadjusted survival rate for first-class is \((113 + 55 + 6)/(113 + 55 + 6 + 4 + 104 + 1)\) = 0.6148.

The unadjusted survival rate for second-class is \((78 + 13 + 25)/(78 + 13 + 25 + 13 + 135 + 0)\) = 0.4394.

What is the overall survival rate for the steerage ticket-holders?

The unadjusted survival rate for steerage is \((88 + 59 + 25)/(88 + 59 + 25 + 91 + 381 + 55)\) = 0.2461.

Use a weighted average to appropriately adjust the survival rate for your non-steerage group.

First, we need to find the proportions of women, men, and children on the Titanic. There were 409 women, 1641 men, and 112 children, so those proportions are 0.1892, 0.7590, and 0.0518, respectively.
For each group, the Berkeley example formula sums, over all gender/child categories, the proportion in that category times the survival rate in that category for that class. Thus, for the crew, \((0.1892)[20/(20+2)] + (0.7590)[193/(193+701)]\) = 0.1639.

For the first class, the weighted average survival rate is \((0.1892)[113/(113+4)] + (0.7590)[55/(55+104)] + (0.0518)[6/(6+1)]\) = 0.4897. For grading purposes, this rounds to 0.49.

Similarly, for second class the weighted-average survival rate is \((0.1892)[78/(78+13)] + (0.7590)[13/(13+135)] + (0.0518)[25/(25+0)]\) = 0.2806. For grading purposes, this rounds to 0.28.

Use a weighted average to appropriately adjust the survival rate for the steerage group.

For steerage, the weighted-average survival rate is \((0.1892)[88/(88+91)] + (0.7590)[59/(59+381)] + (0.0518)[25/(25+55)]\) = 0.2110. For grading purposes, this rounds to 0.21.

Explain what is going on in this data set.

In a lifeboat situation, the rule is “women and children first”, and to a large degree this happened on the Titanic. For all ticket classes separately, it was better to be a woman or a child. The upper classes tended to have large proportions of women and children, but Steerage tended to include more men. So although the numbers make it clear that being in the first or second class was an advantage, some of this advantage was due to the relatively high proportions of women and children.

But, as a somber reflection on the value of human life in terms of social class, the adjusted survival rates are not very different. Class was more important than gender or age.

Strictly speaking, this is not a case of Simpson’s paradox—the survival rates do not reverse when controlling for gender/child status. But the class difference diminishes after such control, so there is a Simpsonesque effect at work.

15. Not Graded. How hard was this homework assignment? I expected it to require about five hours to complete—did you spend significantly more time than that? If you did, you should probably join a study group and work together.