LAST NAME (Please Print): KEY

FIRST NAME (Please Print): ____________________________________________

HONOR PLEDGE (Please Sign): ____________________________________________

Statistics 111

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.

DUE DATE: START of class on WEDNESDAY, SEPTEMBER 6.
In the context of driving safety, people often maintain that young drivers (19 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 – 2002, and examine Table 8 (as a check, the total number of drivers with accidents is 421,899). To get information on the number of miles driven by age, the best available information is at www.bts.gov: search for “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 2002, 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.

1. In New York State, the total number of accidents in 2002 for people under 16 is 299. What is the total number of accidents for people between 16 and 19, inclusive? (Prorate the counts under the assumption that the accident rates are equal within the age group.)

\[7910 + 6128 + \frac{2}{3} (20,189 + 13,416) = 36,441.33.\]

2. From the U.S. Census Bureau’s website, what is the total number of people between 16 and 19?

\[\frac{4}{5} (1,287,544) = 1,030,035.2\]

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 teen drivers (16 to 19) in New York State?

\[12.2 \times 1,030,035.2 = 12,566,429.44\]
4. Find the accident rates per 100,000 miles for people aged 16 to 19 and for people who are 82 to 85. (This controls for mileage.) Note that the Census and the NY DMV do not have corresponding age intervals.

To prorate the census, assume that from age 75 on, a proportion $p$ of people live to the next year. Solve for $p$, so that there are the appropriate number of people aged 85 and older, and use this to estimate the number of people aged 82 to 85.

To prorate the accidents, assume that after age 80, a proportion $2(1 - p)$ of elderly drivers stop driving. (Implicitly, this assumes that the accident rate in the 80+ group 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).

0.88 What is $p$?

The total number of people aged 75 and over is $860,818 + 311,488 = 1,172,306$. We need to find $p$ such that it produces exactly 311,488 people aged 85 and over. Let $x$ be the number of people who are exactly 75. Then this gives us two equations:

$$1,172,306 = \sum_{i=0}^{\infty} xp^i = x/(1 - p)$$

$$311,488 = \sum_{i=10}^{\infty} xp^i = xp^{10}/(1 - p).$$

Solving gives $x = (1 - p) * 1,172,306$. Plugging this into the second equation gives $311,488 = 1,172,306 * p^{10}$ or $p^{10} = 0.2657$. Taking the tenth root gives $p = 0.87587$. For future use, note that $x = 145,518.3438$.

0.79 Accident rate per 100,000 miles for drivers aged 16 to 19

The number of miles driven in a year is $365 * 12,566,429.44$. The number of accidents is $36,441.33$. So the number of accidents per 100,000 miles is $36,441.33/[ 365 * 12,566,429.44/100,000] = 0.7945$.

0.80 Accident rate per 100,000 miles for drivers aged 82 to 85

For seniors, there are $4,344 + 2,780 = 7,114$ accidents in the over-eighty group. Under the prorating assumption, the number of drivers drops by $2 * (1 - 0.87587) = 0.24826$ for each year, so if there are $x$ accidents for 80-year-olds, there are $0.75174x$ accidents for 81-year-olds, $0.75174^2x$ accidents for 82-year-olds, and so forth. Thus $7,114 = \ldots$
\[
\sum_{i=0}^{\infty} x \cdot 0.75174^i = x \cdot \frac{1}{1-0.75174} = 4.02804x, \text{ so } x = 1766.1195. \text{ Thus the number of accidents for people aged 82 to 85 is (}0.75714^2 + 0.75174^3 + 0.75174^4 + 0.75174^5) \cdot 1768.6021 = 2,798.8500.\]

To estimate the number of people aged 82-85, we use the fact that the number who are exactly 75 is \(x = 145,518.3438\). So we find \(145,518.3438 \cdot (p^7 + p^8 + p^9 + p^{10}) = 190,752.8463\).

So the number of accidents per 100,000 miles is \(2,798.8500/[(190,752.8463 \cdot 5 \cdot 365)/100,000] = 0.80398\).

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

Teen drivers are slightly safer than the 82-85 year-old drivers.

6. Why did it make sense to prorate the accidents with a factor of \(2(1 - p)\) rather than \(1 - p\)?

The death rate is \(1 - p\), but many elderly people stop driving before they die. So we need a larger drop out rate than death rate, and doubling the death rate seems like a good guess.

7. 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 10% of the seniors stop driving each year—that may be wrong. Similarly, we assumed constant accident rates in extreme old age.

5. The \(2(1 - p)\) factor is a guess.
6. There are others. I leave it to TA judgment.

8. 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.

9. 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.)

10. 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.

11. At the bottom of the DMV table, there are 310 accidents for which the gender is unknown. Why?

    Mostly hit-and-run drivers, of course.

12. 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'être is to provide exactly such information?

    Also, it is a bit annoying that the data are not available at a more disaggregated level. It would be nice to know how many men aged 65 live in New York State, and there are actually policy issues related to Medicare and workforce planning which should be informed by these data.
13. 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?

14. 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.

15. 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](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, compare first and steerage class survival rates. If your last name begins with S-Z, 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></th>
<th>Men</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 \((113 + 55)/(113 + 55 + 4 + 104)\) = 0.6087.
The unadjusted survival rate for second-class is \((78 + 13)/(78 + 13 + 13 + 135)\) = 0.3808.

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

The unadjusted survival rate for steerage is \((88 + 59)/(88 + 59 + 91 + 381)\) = 0.2375.

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

First, we need to find the proportions of women and men on the Titanic. There were 409 women and 1641 men, so those proportions are 0.1995 and 0.8005, respectively.
For each group, the Berkeley example formula sums, over both gender categories, the proportion in that category times the survival rate in that category for that class. Thus, for the crew, \((0.1995)[20/(20+2)] + (0.8005)[193/(193+701)]\) = 0.3542.
For the first class, the weighted average survival rate is \((0.1995)[113/(113+4)] + (0.8005)[55/(55+104)]\) = 0.4696. For grading purposes, this rounds to 0.47.
Similarly, for second class the weighted-average survival rate is \((0.1995)[78/(78+13)] + (0.8005)[13/(13+135)]\) = 0.2413. For grading purposes, this rounds to 0.24.

Use a weighted average to adjust the survival rate for the steerage group.
For steerage, the weighted-average survival rate is \((0.1995)(88/(88+91)) + (0.8005)(59/(59+381))\) = 0.2054. 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. 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. 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.

16. *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.