LAST NAME (Please Print): KEY

FIRST NAME (Please Print): _____

HONOR PLEDGE (Please Sign): _____

Statistics 110

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: IN CLASS ON WEDNESDAY, SEPTEMBER 5.

In the context of driving safety, people often maintain that teenaged drivers are unsafe because they have more accidents than any other age group. But this does not account for the possibility that teenagers drive more miles than some other age groups; it could be that their accident rate, controlling for milage, 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`. Search on "age gender 2010", then go to the "NY State DMV Summary of Motor Vehicle Accidents" and use Table 8 (as a check, the total number of drivers with accidents is 518,750). To get information on the number of miles driven by age, the most recent 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 get the 2010 census counts of people in different age/gender groups in the state of New York, go to the U.S. Census Bureau's website at `www.census.gov` and explore around.

Note: I find find the Census site to be a bit of a pain, so be patient; as a check, the total is 19,378,102. 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—feel free to mention my name. Then go to our class's homework website and use the link I have posted there.

To find the Census site, (1) Go to the Census website, and search on "New York Population". (2) Open the first search result, "NY Quick Facts from the US Census". (3) At the top of the page, click on "more NY data sets". (4) Click link to "Demographic Profile" under the heading "Demographic Profile from 2010 Census".

37737.33 1. In New York State, the total number of accidents in 2010 for people under 16 is 55. What is the total number of accidents for people between 15 and 19, inclusive? (Prorate the 18-20 group under the assumption that the accident counts are equal for 18, 19, and 20 year-olds; also, assume that all of the "Under 16" group are 15.)

55 + 6328 + 5143 + 2*(22308 + 17009)/3

1,366,278 2. From the U.S. Census Bureau's website, what is the total number of people between 15 and 19, inclusive?

Just look it up in the Census table.

16,668,591.6 3 Assuming the New York State drivers are similar to those studied in the in the National Household Travel Survey, and that 2010 is much like 2001, what is your estimate of the total number of miles driven in one day by teenagers in New York State?

From the NHTS table, multiply 12.2 by the number of teens: 12.2 * 1,366,278 = 16,668,591.6.

4. Find the accident rates per 100,000 miles for teenagers and people who are 80 to 84. (This controls for mileage.) Note that the NY DMV gives only the count for people aged 80 and older; to prorate the count, assume that with each additional year, 10% of the people in that group stop driving. (Hint: if $|p| < 1$, then $\sum_{i=0}^{\infty} p^i = 1/(1-p)$.) Assume that the accident rate in this 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 roughly consistent with the trend in National Household Travel Survey).

For teenagers, there are 37,737.33 accidents. And the 1,366,278 teenagers drive 12.2 miles per day for 365 days (this is for 2010, which was not a leap year). Thus their accident rate per 100,000 miles is 37,737.33/(12.2 * 365 * 1,366,278/100000) = 0.62 accidents per 100,000 miles.

We need to first find the number of driving seniors who are 80 to 84 years old. According to the Census, there are 391,660 people in NY aged 80 to 84. Under our assumptions, on average they drive 5 miles a day, 365 days per week, for a total of 714,779,500 person-miles per year.

The hard part is to estimate the number of accidents they have. From the DMV table, people 80 and over had 6,292 + 4,775 = 11,067 accidents in 2010. Let x be the number of accidents that 80 year-olds have. Then

$$11,067 = x + .9x + .9^2x + \cdots = x(1 + .9 + .9^2 + \cdots) = x/(1 - .9)$$

(this uses the hint, which is fun and easy to prove). Solving shows $x = 1,106.7$. So the number of accidents for people between 80 and 84 inclusive is $1,106.7 + .9 * 1106.7, + .9^2 * 1,106.7 + .9^3 * 1,106.7 + .9^4 * 1,106.7 = 4.0951 * 1,106.7 = 4,532.05$. Putting all this together, the accident rate per 100,000 miles is 4,532.05/(714,779,500/100,000) = 0.63.

0.63 Seniors

0.62 Teenagers

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

After adjusting for the number of miles driven, seniors are slightly worse than teenagers.

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 milage 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.

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 many are three times as dangerous? Defend or refute.

   According to the NHTS table, 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 30,884 accidents that we ignored in our analysis. Why do you think the gender is unknown?

Surely these were hit-and-run accidents; this is corroborated by the relatively high counts of property damage by the driver who was the victim—the driver who stayed and filed a claim.

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: Recall Rick Perry's uncharacteristic healthcare position as Texas governor.)

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. The Public Health Service studied the effects of smoking on health for a large and random sample of U.S. residents (*The Health Benefits of Smoking Cessation: A Report of the Surgeon General*, 1990). It found that never-smokers were slightly healthier than smokers, but that smokers who had recently quit were much less healthy that current smokers. How do you explain this result?

Ex-smokers are a self-selected group, and many people give up smoking beacuse they are sick. So recent ex-smokers include a lot of sick people.

14. The following table shows the relationship between 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 perpsective on this at `www.ithaca.edu/staff/jhende` ⊕ `rson/titanic.html`; it goes down to the level of the lifeboats and the order in which they were launched—each lifeboat was a stage for its own drama, and a wide range of moral choices were made.

If your last name begins with A-I, please compare first and second class 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. Ignore the class that does not pertain to your initial.

	Women			Men			Children	
	Survival	Death		Survival	Death		Survival	Death
First Class	113	4		55	104		6	1
Second Class	78	13		13	135		25	0
Steerage	88	91		59	381		25	55

_____ What is the overall survival rate for the higher class ticket-holders?

The survival rate for First Class is $(113 + 55 + 6)/(117 + 159 + 7) = 0.615$.
The survivial rate for Second Class is $(78 + 13 + 25)/(91 + 148 + 25) = 0.439$.
The survival rate for Steerage is $(88 + 59 + 25)/(179 + 440 + 80) = 0.246$.

_____ What is the overall survival rate for the lower class ticket-holders?

_____ Use a weighted average to appropriately adjust the survival rate for the higher class ticket-holders.

To mimic the reasoning in the Berkeley graduate admissions example, note that survival is like being accepted at Berkeley, ticket class is like gender, and man/woman/child is the confounding variable that is like major.
First, we need to find the proportions of women, men, and children on the *Titanic*. Since (depending on your name) certain ticket classes are irrelevant, you should not count the men, women, or children from that group. Thus, if your name begins in A-I, the steerage people should not be counted. In that case, from the table, the proportions of women, men, and children are 208/547, 307/547, and 32/547, respectively. Specifically, There are 117+91 = 208 women in first and second class, and 283 + 264 = 547 people in first and second class.

Similarly, for J-R, the ratios of women, men and children are 296/982, 599/982, and 87/982. And for S-Z, the ratios are 270/963, 588/963, and 105/963.

For the A-I weighted average survival in first class, the formula from the Berkeley example sums, over all gender/child categories, the proportion in that category times the survival rate in that category for the first class. Thus: (208/547) * (113/117) + (307/547) * (55/159) + (32/547) * (6/7) = 0.612.

For the other comparisons, the weights will change depending on which pairs of ticket classes are being examined. For J-R, (296/982) * (113/117) + (599/982) * (55/159) + (87/982) * (6/7) = 0.578. And for S-Z, (270/963) * (78/91) + (588/963) * (13/148) + (105/963) * (25/25) = 0.403.

_____ Use a weighted average to appropriately adjust the survival rate for the lower class ticket-holders.

For A-I, the weighted average survival rate in second-class is (208/547) * (78/91) + (307/547) * (13/148) + (32/547) * (25/25) = 0.434. For J-R, the weighted average survival in steerage is (296/982) * (88/179) + (599/982) * (59/440) + (87/982) * (25/80) = 0.258. And for S-Z, the weighted average survival in steerage is (270/963) * (88/179) + (588/963) * (59/440) + (105/963) * (25/80) = 0.254.

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. 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.