**ECO220Y: Homework 1 – SOLUTIONS**

**(For answers to the textbook exercises, see Appendix A of the textbook and/or the Readings page in Quercus.)**

**Required Problems:**

**(1) (a)** This is a stratified random sample. There are 6 strata, which are indicated in the graphic. When sampling with stratification, as opposed to collecting a simple random sample of any angler in Wisconsin, we will have to use sample weights in any subsequent analysis that wishes to combine strata. This is because we have deliberately over-sampled anglers in some strata (e.g. Stratum 1) and hence our random sample has many more anglers from these areas than we would get if we did a simple random sample of anglers in Wisconsin. Within a stratum of course we have a simple random sample.

**(b)** We would have to be very careful about making inferences about <u>all</u> Wisconsin anglers because not all areas of Wisconsin are included in the sample. Anglers living in the white area in the graphic above had no chance of inclusion in the sample. (An aside: it turns out that more than 80% of the population of Wisconsin lives in the sampled area.) Further some areas are over-represented in the sample (like Stratum 1) while other areas are under-represented in the sample (like Stratum 6). Hence to make inferences about the sampled area in general (all 6 strata), we would need to use weights.

**(c)** Anglers that live in Stratum 1. Anglers in this area are the most heavily sampled. In fact, the reason for this study was a natural resource damage assessment focused on PCBs in the Fox River (that goes right through Stratum 1) and in lower Green Bay. Given that anglers that live near these waters could be most adversely affected it could make sense to sample them most intensively. (Find Green Bay on the map above and then reconsider the sampling plan to see if it makes sense now that you have more information.)

**(d)** We may be concerned about non-response bias because many people contacted by telephone and asked to participate in a survey will choose not to participate: if these anglers are different than those who agreed to be in our sample then we will have non-response bias. They may well be different because people who are really enthusiastic about fishing are more likely to agree to be in a survey that talks about fishing (in this specific context this is also called an avidity bias). Another non-sampling error may be that some people do not have home telephone numbers and hence cannot be included in RDD sampling: i.e. the target population (anglers in certain areas of Wisconsin) is not the same as the sampled population (anglers in certain areas of Wisconsin who have home telephone lines).

**(2) (a)** There are two variables and each contains interval data: measures the number of children (real number measurement). There are 500 observations (records), which is the same thing as the sample size. These data are cross-sectional.

**(b)** Yes there is a positive association between kids_mom and kids_daughter. The cross-tabulation shows that those from larger families are more likely to have larger families.

**(c)** The cross-tabulation is appropriate for these data because there are not too many unique pairs of values (despite the fact that these are interval data). This is because the number of children is necessarily a small integer. If instead we had age there would be many unique pairs of values making a cross-tabulation unwieldy.

**(3) (a)**

| | sex | |
|---|---|---|
| age | Female | Male | Total |
| 15-24 | **892,000** | | |
| 25-54 | | | **1,794,000** |
| 55+ | | | |
| Total | | | **4,079,500** |

**(b)** Note: must use the columns of results in the original table labelled "Low wage employees as % of all employees."

| age | sex: Female | Male | Total |
|---|---|---|---|
| 15-24 | | | |
| 25-54 | | | |
| 55+ | | | |
| Total | **7,433,750** | **7,730,455** | **15,164,205** |

**(4) (a)** For infants between 1 and 12 months old born in the US to families that fall in the lowest categories of education or occupation (i.e. the lowest socioeconomic status), the death rate is a bit over 3 deaths per 1,000 infants. In contrast, for those families with the highest socioeconomic status the death rate is less than 1 death per 1,000 infants. The relationship is linear: as the socioeconomic status of a family rises in the U.S., the risk of infant death drops substantially with continued improvement as a family rises up the socioeconomic latter: it is not just the difference between being very poor and working to middle class. Overall, in the U.S. the death rates of infants is highly related to socioeconomic status of the family they are born into.

**(b)** For each socioeconomic class (families put into four categories), the United States has higher postneonatal death rates than Austria or Finland, which are very similar to each other. The gap is most pronounced for families outside the top socioeconomic status: the death rates are about three times higher in the US for these groups compared to families of similar status in Austria and Finland. Even for families in the highest socioeconomic class, the infant death rate is, perhaps surprisingly, still higher in the US than Austria. Hence the potential story is not simply one of widespread inequality in quality of healthcare for families in the US, but also one of generally worse outcomes even for those who can afford premium care. Most importantly, infant death rates in Austria and Finland vary substantially less by a family's socioeconomic status than in the US: in fact, it is really only the lowest of the four status categories that have notably worse outcomes in Austria and Finland. Hence, while there is some disadvantage to being in the lowest socioeconomic class in all three countries, the disadvantage is far worse in the US and there is a big disadvantage in the US even for middle status families.

Of course, we can also peek at the paper. Any well-written research paper not only presents its results in the form of figures and tables, but also explains them in words. Excerpt from p. 110 of the original journal article:

> We begin by investigating how postneonatal mortality rates vary by demographic group. Figure 6, panel A documents postneonatal death rates by education/socioeconomic status group, for which we observe four groups in the United States and Austria and three groups in Finland. In the United States, this is based on education: less than a high school degree, a high school degree, some college, and college degree or more. In Austria, we also use educational data: compulsory school, vocational school, high school with A-levels, and university or teaching college. In Finland, the groups are defined based on occupation: blue collar, lower white collar, and upper white collar or entrepreneur. The steeper socioeconomic gradient observed in postneonatal mortality within the United States is striking relative to the socioeconomic gradients observed in Austria or Finland. Notably, the within-US gradient is not simply due to high mortality rates in the least educated group; there is wide variation across the distribution; in contrast, to the extent that there is any inequality by socioeconomic status in Austria or Finland, it appears to be driven by the lowest education or occupation group.

**(5) (a)** The dashed blue line would continue to a dot of 0.70 (=77/110) for 2018. The solid red line would continue to a dot of 0.61(=47/77) for 2018.

**(b)** We know 72% of the papers are papers that use data, so that is 72 papers (72% of 100). Of those 72, we know that 46% are exempted, so that is about 33 papers (46% of 72). Hence, 33 papers would have been exempted from the data-sharing policy.

**(6) (a)** See completed table below. This *is* an example of Simpson's Paradox. We see that there is absolutely no difference in the admissions rate between applicants in Groups A and B in each department. However, overall it appears that Group B has an "unfair" admission advantage: 37.1% for Group B admitted versus 33.0% for Group A. However, this simply reflects the fact that Group B is more likely to apply to easier-to-get-into programs. For example, only 10% of the Group A applicants applied to the Department 4 program (with a 70% admission rate), whereas nearly 13% of the Group B applicants applied to Department 4.

| | **Group A** | | | **Group B** | | |
|---|---|---|---|---|---|---|
| | Admitted | Applied | $\hat{P}_A$ | Admitted | Applied | $\hat{P}_B$ |
| Department 1 | 30 | 100 | 0.300 | 660 | 2200 | 0.300 |
| Department 2 | 40 | 200 | 0.200 | 300 | 1500 | 0.200 |
| Department 3 | 60 | 150 | 0.400 | 1200 | 3000 | 0.400 |
| Department 4 | 35 | 50 | 0.700 | 700 | 1000 | 0.700 |
| Overall | 165 | 500 | 0.330 | 2860 | 7700 | 0.371 |

**(b)** See completed table below. This is *not* an example of Simpson's Paradox. There is nothing contradictory here. Group B has a lower admission rate overall and a lower admission rate in each department. It is very important to note that there are NO differences in between Group A and Group B in the likelihood of applying to each department. To help you notice this important fact, the table simply scaled the number of applicants to each department by 10 across Groups A and B.

| | **Group A** | | | **Group B** | | |
|---|---|---|---|---|---|---|
| | Admitted | Applied | $\hat{P}_A$ | Admitted | Applied | $\hat{P}_B$ |
| Department 1 | 40 | 100 | 0.400 | 300 | 1000 | 0.300 |
| Department 2 | 60 | 200 | 0.300 | 400 | 2000 | 0.200 |
| Department 3 | 75 | 150 | 0.500 | 600 | 1500 | 0.400 |
| Department 4 | 40 | 50 | 0.800 | 350 | 500 | 0.700 |
| Overall | 215 | 500 | 0.430 | 1650 | 5000 | 0.330 |

**(c)** See completed table below. This is *not* an example of Simpson's Paradox. There is nothing contradictory here. Group B has a higher admission rate overall and a higher admission rate in each department. Notice that there are no differences in the admissions rates across departments within each group.

| | **Group A** | | | **Group B** | | |
|---|---|---|---|---|---|---|
| | Admitted | Applied | $\hat{P}_A$ | Admitted | Applied | $\hat{P}_B$ |
| Department 1 | 350 | 1000 | 0.350 | 224 | 560 | 0.400 |
| Department 2 | 1050 | 3000 | 0.350 | 96 | 240 | 0.400 |
| Department 3 | 700 | 2000 | 0.350 | 312 | 780 | 0.400 |
| Department 4 | 1400 | 4000 | 0.350 | 120 | 300 | 0.400 |
| Overall | 3500 | 10000 | 0.350 | 752 | 1880 | 0.400 |

**(d)** In this specific case, a Simpson's Paradox requires BOTH that there are differences in the admissions rates across departments AND that there are differences across Groups A and B in the fraction applying to each department. Also, there will be an issue if both of these occur even if you do not get a dramatic result where the departmental admissions are starkly different from overall admissions. In other words, sometimes the bias caused by this situation is not so severe that it jumps out at you. Hence, watch out for cases where both requirements for a Simpson's Paradox are present.