Seth Metcalf

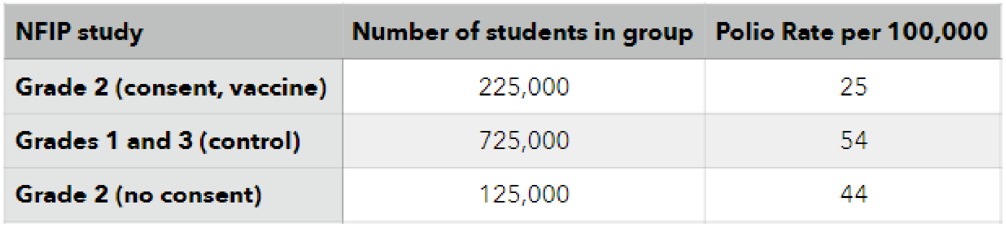
Prof. Hank Ibser

Homework 1

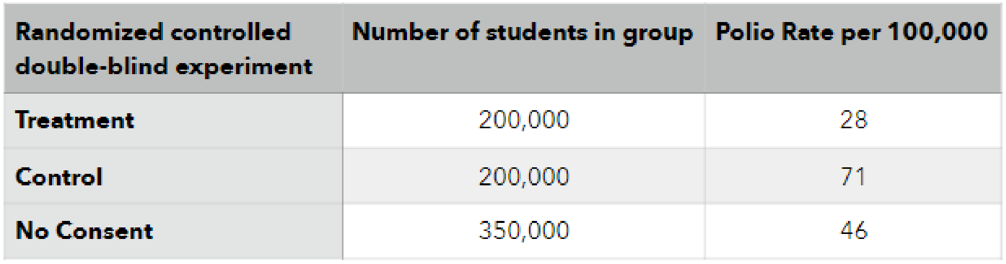
29 January 2021

Statistics HW 1

1. FPP Ch 2, based on A2-A4:



NFIP study	Number of students in group	Polio Rate per 100,000
Grade 2 (consent, vaccine)	225,000	25
Grades 1 and 3 (control)	725,000	54
Grade 2 (no consent)	125,000	44



Randomized controlled double-blind experiment	Number of students in group	Polio Rate per 100,000
Treatment	200,000	28
Control	200,000	71
No Consent	350,000	46

   a. These two findings are confirmed by the fact that the Treatment in the Randomized Controlled experiment and the Grade 2 (consent, vaccine) had similar rates to one another. Since we know that the rate of Polio affecting children differs based off of wealth (the wealthier you are the more likely you are to get Polio), we know that they must have similar background in these areas or the rates per 100,000 would differ more.
   b. Similarly, to how problem (a) was solved, the data in the no-consent groups have similar Polio rates to one another. Because of this we know that they come from similar background otherwise the data in the Polio rate would differ from one another.
   c. Between the control groups, there is a very large gap between the NFIP study with 54 Polio cases per 100,000 vs the Randomized controlled experiment with the Control Polio Rate being 71 per 100,000. This disparity shows that the two control groups most likely come from differing backgrounds from one another (with the Randomized controlled Control group being from a wealthier background).
   d. The (no consent) group in the NFIP study had a lower rate of Polio than the (control) group because it can be assumed that since the parents are cautious when it comes to vaccinating their child, that similarly they would be more cautious in day to day life than the average parent. This would cause the rate of Polio to be lower as they are more cautious in day to day life and would keep their child more protected.
   e. They should not compare the (consent, vaccine) group with the (no consent) treatment group to prove that the vaccine works; since that group has external factors that could affect how the rate of Polio in the children. The better group would be to compare the vaccinated group to the control group.

   f. There could be a slight bias in the results of the NFIP due to this. The idea behind doing a blind experiment is to make sure there are no external factors that could affect it, but since the NFIP design was not done blind there could be external differences in actions done that could lead to the children having differing results. For example, if it is known that they got the vaccine, maybe the parents/children would act differently (either more safe or more dangerous), which would lead to a biased Polio rate.
   g. No. Just because the areas selected had a 25% higher rate of Polio than the rest of the country it can not be concluded that this is due to the field trials. There could be a multitude of factors affecting this; maybe it is due to the Public Health Services selecting, on average, wealthier areas than the rest of the country, or maybe it is something else.
2. FPP Ch 2 Rev 10:
   a. An observational study. There as no treatment administered and was only looking at the correlation of “controlling” mothers and the child’s weight.
   b. They seem to have reached that conclusion based off the fact that the article was titled “Parents of Fat Kids Should Lighten Up” and earlier in the sentence it states “A study of young children found that those with more body fat tended to have more ‘controlling’ mothers.”
   c. Yes, if controlling behavior had a direct correlation with making your child eat more then it would explain the higher rate of obesity in the children of controlling mothers.
   d. No, that would not explain the association unless that same gene also caused mothers to become more cautious. If anything, all a gene that causes obesity would do would be to represent a confounding factor in the experiment.
   e. An alternative would be what I mentioned in the previous response. If a gene existed that increased the controlling aspect of the mother – and that same gene increased the chance of obesity in the child, then that would be an alternate way to explain the results.
   f. No. A controlling mother is not the reason for obesity in children, there just seems to be a correlation. A child may still be overweight regardless of the attitude of the mother.
3. FPP Ch 4 A4, 5, 6:
   a. The average of the group of people would rise as new data is being put into the data set. With the introduction of the new 6’5” person, the average would go from 5’6” to 5’7”. (66 inches is the average before, the new person being added is 77 inches tall. The accumulation of all the previous people is 660 inches + 77 inches is 737, divide this by the number of people, 11 – you get 67 inches, or 5’7” as the average).
   b. Similar to the problem before; the accumulation of all the people in the data set is 1386 inches, the addition of the new person is 77 inches, making the total 1463 inches. Divide this by the total number of people in the set, 22, and you get the average of the set – 66.5 inches or 5’6.5”.
   c. In order to increase the data set by 1 inch, the person has to be quite tall because there are a large number of people in the data set. In order to get an average height of 67 inches with 22 people, the accumulated total would have to equal 1474 inches. Since we know the average height of 21 people in the data set is 5’6” (or 66 inches), we can calculate that before adding the final person, the total height was 1386 inches. The

difference between the two numbers (1474 inches – 1386 inches) is 88 inches. That means that the final person added to the group to raise the average height by a full centimeter has to be 88 inches tall – or 7’4”.

4. FPP Ch 4 Rev 9:
    a. Yes, this does affect the average of the data set. Due to the miscalculation, it would raise the mean of the data set by about $887.4.
    b. No, it does not affect the median as the median is not affected by outliers.

R Questions

1. If you type in 1:3 + 1:4, what is the output in R? Explain briefly why R gives this as an answer.

When doing 1:3 + 1:4 it has the output 2 4 6 5. This is because 2 4 6 5 is what is given when adding the two vectors (1 2 3) and (1 2 3 4). Since there is a shorter vector between the two, when it goes to add to the last number of the larger vector (4), it cycles back to (1) in the shorter vector. So 2 4 6 5 is the output of the following mathematical sequence: (1) + (1), (2) + (2), (3) + (3), and (1) + (4).

2. Compute the mean and median age of everyone in the data set.

The mean and median age is given to us by doing:
mean(family$age), which gives us 48.14286 and
median(family$age), which gives us 47.5.

3. Compute the mean age of the males and the mean age of the females. Hint: the select () function will return a data frame, and the mean() function requires a vector as the input. You can use the $ notation to get/reference a vector in a data frame.

The mean age of the males and females is given to us by doing:
mean(filter(family,gender==”m”)$age), which gives the output 53.25 and
mean(filter(family,gender==”f”)$age), which gives the output 41.33333.

4. Who is the shortest person in the dataset? Write code to return the answer, don’t just look at the data.

The shortest person’s name in the dataset is given to us by:
filter(family,height==min(family$height)), in which we get Zoe as the shortest person in the dataset.

5. Make a vector of weights in order from the youngest person to the oldest.
In order to sort the weight from the youngest person to the oldest:
attach(family), use this so that it can be sorted
sortedfamily = family[order(age),], this sorts the family by age and attaches it to a new name
sortedfamily$weight gives us the vector of weights in the family dataset sorted by age