Problem Set 1

Potential Outcomes Notation

1. Explain the notation $Y_i(1)$.
   The outcome of a given subject "i" when administered the treatment.
2. Explain the notation $Y_1(1)$.
   The outcome of subject 1 when administered the treatment.

3. Explain the notation $E[Y_i(1)|d_i = 0]$.
   The expectation of treatment group outcome when one village selected at random is not exposed to treatment.
4. Explain the difference between the notation $E[Y_i(1)]$ and $E[Y_i(1)|d_i = 1]$
   $E[Y_i(1)]$ is the average outcome of the treatment group whereas $E[Y_i(1)|d_i = 1]$ is the expected value of a member of the treatment group if selected at random.

Potential Outcomes and Treatment Effects

1. Use the values in the table below to illustrate that $E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - [Y_i(0)]$.

```
equal <- mean(table$y_1)-mean(table$y_0) == mean(table$tau)

print("Are the two statements equal?")
```

```
## [1] "Are the two statements equal?"
```

```
print(equal)
```

```
## [1] TRUE
```

2. Is it possible to collect all necessary values and construct a table like the one below in real life? Explain why or why not.

No - in real life it is impossible for the same subject to be in both the treatment and control group.

Visual Acuity

Suppose we are interested in the hypothesis that children playing outside leads them to have better eyesight.

Consider the following population of ten children whose visual acuity we can measure.

- Visual acuity is the decimal version of the fraction given as output in standard eye exams.
- Someone with 20/20 vision has acuity 1.0, while someone with 20/40 vision has acuity 0.5.
- Numbers greater than 1.0 are possible for people with better than "normal" visual acuity.

```
d <- data.table(
  child = 1:10,
  y_0 = c(1.2, 0.1, 0.5, 0.8, 1.5, 2.0, 1.3, 0.7, 1.1, 1.4),
  y_1 = c(1.2, 0.7, 0.5, 0.8, 0.6, 2.0, 1.3, 0.7, 1.1, 1.4)
)
```

In this table:

- `y_1` means means the measured *visual acuity* if the child were to play outside at least 10 hours per week from ages 3 to 6'

- `y_0` means the measured *visual acuity* if the child were to play outside fewer than 10 hours per week from age 3 to age 6;
- Both of these potential outcomes *at the child level* would be measured at the same time, when the child is 6.

1. Compute the individual treatment effect for each of the ten children.

```
tau = d$y_1-d$y_0
print(tau)
```

```
##  [1]  0.0  0.6  0.0  0.0 -0.9  0.0  0.0  0.0  0.0  0.0
```

2. Tell a "story" that could explain this distribution of treatment effects. In particular, discuss what might cause some children to have different treatment effects than others.

It could be that treatment effect is related to the initial visual acuity of the child in some way. It could be that for most children, the treatment effect is zero, but children at the high and low extremes of the acuity scale (less than 0.5 or greater than 1.5, for example) are more likely to experience a measurable change. There could be other aspects of a child's lifestyle that correlate with playing outside - better nutrition, rural environment, better weather - that have an effect on acuity as well and it could be for these individuals we are measuring those effects as well.

3. For this population, what is the true average treatment effect (ATE) of playing outside.

```
print(mean(tau))
```

```
## [1] -0.03
```

4. Suppose we are able to do an experiment in which we can control the amount of time that these children play outside for three years. We happen to randomly assign the odd-numbered children to treatment and the even-numbered children to control. What is the estimate of the ATE you would reach under this assignment? (Please describe your work.)

```
control = d[child%%2==0]
treatment = d[child%%2==1]

ate = mean(treatment$y_1)-mean(control$y_0)

print(ate)
```

```
## [1] -0.06
```

5. How different is the estimate from the truth? Intuitively, why is there a difference? The truth is an ATE of -0.03 whereas the estimate is an ATE of -0.06. The estimate is the same sign and double the magnitude of the actual. There is a difference because theory states that the expectation of the estimated ATE equals the true ATE over many tries, but individual estimates may differ from their true values because the two groups have different means purely as a result of random noise.

6. We just considered one way (odd-even) an experiment might split the children. How many different ways (every possible ways) are there to split the children into a treatment versus a control group (assuming at least one person is always in the treatment group and at least one person is always in the control group)?

```
N = 10
num_experiments = 0
```

```r
for (v in 1:9) {

  num_experiments = num_experiments + factorial(N)/(v*factorial(N-v))

}

print(num_experiments)
```

```
## [1] 749203
```

7. Suppose that we decide it is too hard to control the behavior of the children, so we do an observational study instead. Children 1-5 choose to play an average of more than 10 hours per week from age 3 to age 6, while Children 6-10 play less than 10 hours per week. Compute the difference in means from the resulting observational data.

```r
obv_control = d[child>5]
obv_treatment = d[child<=5]

obv_tau = mean(obv_treatment$y_1) - mean(obv_control$y_0)

print(obv_tau)
```

```
## [1] -0.54
```

8. Compare your answer in (7) to the true ATE. Intuitively, what causes the difference?

The ATE in part 7 is -0.54, which is far greater in magnitude than the true ATE of -0.03. This difference is caused by the fact that different groupings of subjects can have different means either by chance or because the grouping was not truly random.

Randomization and Experiments

1. Assume that researcher takes a random sample of elementary school children and compare the grades of those who were previously enrolled in an early childhood education program with the grades of those who were not enrolled in such a program. Is this an experiment, an observational study, or something in between? Explain!

This is an observational study, as the data was randomly selected from existing data rather than collected from individuals who were randomly assigned to treatment and control prior to treatment being administered. There may be commonalities between most or all children who enrolled in early childhood education (income, region, etc) that would weaken the study's causal claims.

2. Assume that the researcher works together with an organization that provides early childhood education and offer free programs to certain children. However, which children that received this offer was not randomly selected by the researcher but rather chosen by the local government. (Assume that the government did not use random assignment but instead gives the offer to students who are deemed to need it the most) The research follows up a couple of years later by comparing the elementary school grades of students offered free early childhood education to those who were not. Is this an experiment, an observational study, or something in between? Explain!

This can be considered either a poorly designed experiment or something in between an experiment and an observational study. A study can not be considered an RCT without random assignment of subjects to the treatment and control groups. In this case, it is explicitly stated that children are assigned to the control based on commonalities between them, which violates the principle that the control and treatment groups should be indistinguishable.

3. Does your answer to part (2) change if we instead assume that the government assigned students to treatment and control by "coin toss" for each student? Why or why not?

If the government uses the coin toss to randomly assign students to either attend free preschool or not attend free preschool, and these children took preschool at the same time and under the same conditions, than this would be an experiment.

Moral Panic

Suppose that a researcher finds that high school students who listen to death metal music at least once per week are more likely to perform badly on standardized test. :metal: As a consequence, the researcher writes an opinion piece in which she recommends parents to keep their kids away from "dangerous, satanic music".

- Let the potential outcomes to control, $Y_i(0)$, be each student's test score when listening to death metal at least one time per week.
- Let $Y_i(1)$ be the test score when listening to death metal less than one time per week.

1. Explain the statement $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ in words. First, state the rote english language translation – i.e. "The expected value of ... " – but then, second, tell us the *meaning* of this statement.

The expected value of the untreated potential outcome of the control group is equal to the expected value of the untreated potential outcome of the treatment group.

This means that, because we assign groups randomly, we would expect the outcomes for the control and treatment groups to be equal if, in fact, neither of them were given the treatment.

2. Do you expect that this circumstance actually matches with the meaning that you've just written down? Why or why not?

It's unclear in this situation if the researcher actually ran an experiment where she randomly divided students into treatment and control and assigned one group death metal to listen to, or if she simply used observational data. But assuming that random assignment did take place, you could reasonably expect this statement to hold true.