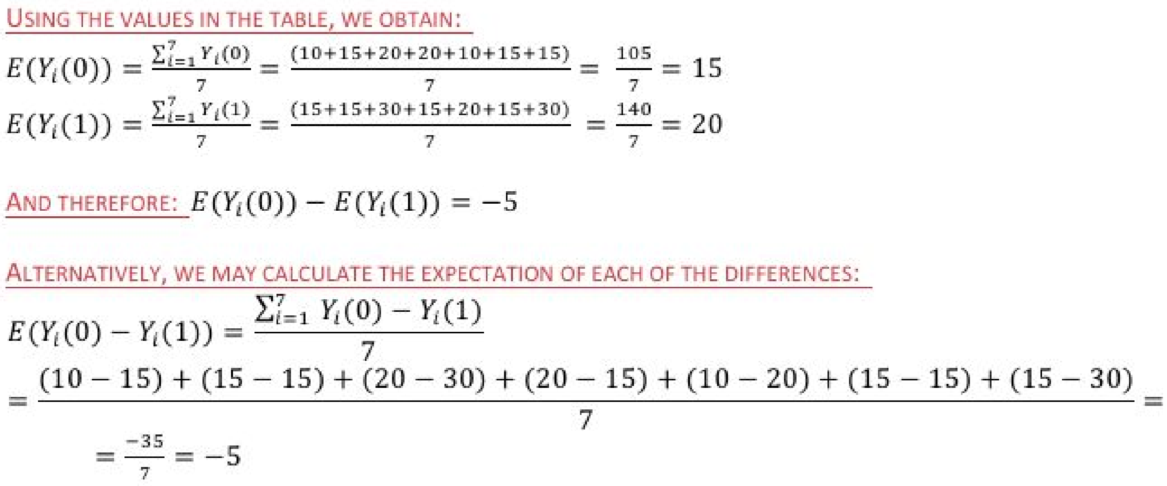
Solution set for Problem Set #1.

1. On the notation of potential outcomes:
   a. Explain the notation $Y_i(1)$.
      i. **The potential outcome for subject *i* when the treatment is present. If subject *i* were treated, this is what we would observe. Each Y for unit i is a function whose value depends upon whether that unit is in treatment or control.**
   b. Explain the notation $E[Y_i(1)|d_i=0]$.
      i. **The average outcome we would have observed for the subjects who actually did not receive treatment had they received treatment.**
   c. Explain the difference between the notation $E[Y_i(1)]$ and the notation $E[Y_i(1)|d_i=1]$.
      i. **The former is the mean of the treatment potential outcomes for all subjects, while the latter is the mean of the treatment potential outcomes only for subjects actually put in the treatment group in a particular experiment. The latter is what we would observe in a real experiment by calculating the mean of the treatment group outcome.**
   d. (Extra credit) Explain the difference between the notation $E[Y_i(1)|d_i=1]$ and the notation $E[Y_i(1)|D_i=1]$. Use exercise 2.7 from FE to give a concrete example of the difference.
      i. **d refers to the mean outcome for subjects in the treatment group for an actual realized randomization, such as the mean of Villages 3 and 7 from FE 2.7. The latter, with D, refers to the mean across all possible allocations of treatment assignment where two villages are in the treatment group. In this sense the letter could be written as $E[E[Y_i(1)|D_i=1]]$, because we are examining the "expected treatment group mean," averaging across two villages within patterns of treatment assignment and then across these averages to get the "average average."**
2. FE, exercise 2.2.



Using the values in the table, we obtain:

$$E(Y_i(0)) = \frac{\sum_{i=1}^{7} Y_i(0)}{7} = \frac{(10+15+20+20+10+15+15)}{7} = \frac{105}{7} = 15$$

$$E(Y_i(1)) = \frac{\sum_{i=1}^{7} Y_i(1)}{7} = \frac{(15+15+30+15+20+15+30)}{7} = \frac{140}{7} = 20$$

And therefore: $E(Y_i(0)) - E(Y_i(1)) = -5$

Alternatively, we may calculate the expectation of each of the differences:

$$E(Y_i(0) - Y_i(1)) = \frac{\sum_{i=1}^{7} Y_i(0) - Y_i(1)}{7}$$

$$= \frac{(10-15)+(15-15)+(20-30)+(20-15)+(10-20)+(15-15)+(15-30)}{7} =$$

$$= \frac{-35}{7} = -5$$

3. FE, exercise 2.3.

a-d)

		Y_i (1)		
Y_i(0)	15	20	30	p(Y_i(0))
10	1/7	1/7	0	2/7
15	2/7	0	1/7	3/7
20	1/7	0	1/7	2/7

p(Y_i(1))	**4/7**	**1/7**	**2/7**	**1.0**

e)

$$E(Y_i(0)|Y_i(1) > 15) = \sum_i Y_i(0)\frac{prob(Y(0)=Y_i(0);Y_i(1)>15)}{prob(Y_i(1)>15)} = 10 * \frac{\frac{1}{7}}{\frac{3}{7}} + 15 * \frac{\frac{1}{7}}{\frac{3}{7}} + 20 * \frac{\frac{1}{7}}{\frac{3}{7}} = 15.$$

f)

$$E(Y_i(1)|Y_i(0) > 15) = \sum_i Y_i(1)\frac{prob(Y(1)=Y_i(1);Y_i(0)>15)}{prob(Y_i(0)>15)} = 15 * \frac{\frac{1}{7}}{\frac{2}{7}} + 20 * \frac{0}{\frac{2}{7}} + 30 * \frac{\frac{1}{7}}{\frac{2}{7}} = 22.5.$$

4. More practice with potential outcomes. We are interested in the hypothesis that children playing outside leads them to have better eyesight.

   Consider the following population of ten representative 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.)

	$Y_i(0)$: Visual acuity after not playing outside	$Y_i(1)$: Visual acuity after playing outside.
Child 1	1.1	1.1
Child 2	0.1	0.6
Child 3	0.5	0.5
Child 4	0.9	0.9
Child 5	1.6	0.7
Child 6	2.0	2.0
Child 7	1.2	1.2
Child 8	0.7	0.7
Child 9	1.0	1.0
Child 10	1.1	1.1

In the table, state (1) means “playing outside an average of *at least* 10 hours per week from age 3 to age 6,” and state (0) means “playing outside an average of *less than* 10 hours per week from age 3 to age 6.” Y_i represents visual acuity measured at age 6.

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

Average of Yi(1) = 0.98, Average of Yi(0) = 1.02; 0.98 - 1.02 = -0.04

b. Compute the individual treatment effect for each of the ten children. Note that this is only possible because we are working with hypothetical potential outcomes; we could never have this much information with real-world data. (We encourage the use of computing tools on all problems, but please describe your work so that we can determine whether you are using the correct values.)

	$Y_i(0)$: Visual acuity after not playing outside	$Y_i(1)$: Visual acuity after playing outside.	Individual treatment effect
Child 1	1.1	1.1	0
Child 2	0.1	0.6	.5
Child 3	0.5	0.5	0
Child 4	0.9	0.9	0
Child 5	1.6	0.7	-.9
Child 6	2.0	2.0	0
Child 7	1.2	1.2	0
Child 8	0.7	0.7	0
Child 9	1.0	1.0	0
Child 10	1.1	1.1	0

c. In a single paragraph, tell a story that could explain this distribution of treatment effects. What might cause some children to have different treatment effects than others?

Most children’s eyesight is not impacted by playing outside. But, Child 2 typically spends most of her time inside playing computer games that ruin her eyes. Going outside means she avoids this activity. If Child 5 were to play outside, Child 5 would play a dangerous football game and be struck in the face, degrading his vision. Staying inside saves him from this fate.

d. 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 assign the odd-numbered children to treatment and the even-numbered children to control. What is the estimate of the ATE? (Again, please describe your work.)

Mean treatment potential outcome for odd-numbered children: (1.1 + 0.5 + 0.7 + 1.2 + 1.0)/5 = 0.9
Mean control potential outcome for even-numbered children: (0.1 + 0.9 + 2.0 + 0.7 + 1.1)/5 = 0.96

ATE estimate for this realization of treatment assignment = 0.9 - 0.96 = -0.06

e. How different is the estimate from the truth? Intuitively, why is there a difference?

This overstates the average deleterious effect by 0.02. This estimate differed from the truth by chance, because we happened to place students with worse vision in the treatment group.

f. We just considered one way (odd-even) to split the children into an experiment. How many different ways (*every* possible way) 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)?

2^10 - 2 = 1022

g. 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.

(1.1 + 0.6 + 0.5 + 0.9 + 0.7) / 5 – (2.0 + 1.2 + 0.7 + 1.0 + 1.1) / 5 = 3.8 / 5 – 6 / 5 = 0.76 – 1.2 = -0.44

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

The difference in part (g) is considerably larger in magnitude (more negative). This simply represents selection bias. Children 1-5 who chose to play outside a lot had considerably worse eyesight anyway. Why? One example explanation: perhaps children with worse eyesight choose to play outside because they have a harder time reading books.

5. FE, exercise 2.5.

a) All three physical methods of random assignment require that the person or persons in charge of implementing the randomization follow the intended

protocol: dice must be rolled once per subject, and cards or envelopes must be shuffled thoroughly. Assuming that the mechanics of each physical method of randomization are carried out, the limitation of the dice method is that possibility that the allocation of treatments could wind up being imbalanced; in principle, one could flip a coin 6 times and come up with 6 heads, in which case the treatments would not vary. The card method overcomes this problem and ensures that exactly half of the subjects will receive each treatment. The advantage of the sealed envelope method over the card method is the fact that envelopes help prevent the person who is allocating subjects from deliberately or unconsciously exercising discretion over who receives which treatment, thereby subverting the randomization.

b) **As the N increases, the dice method becomes more likely to produce a 50-50 division in treatments. For example, with 600 subjects, the probability of obtaining an assignment as imbalanced as 250-350 is less than 1-in-10,000.**

c) **The methods produce identical results, in expectation.**

6. FE, exercise 2.6.

This is an observational study. Subjects are not randomly assigned to the treatment, which in this case is taking the preparatory class. Instead, they self-select into the treatment for unknown reasons. The fact that the students were sampled randomly from the large population is immaterial; the key issue is whether students in the sample were randomly allocated to the treatment or control group. Note that this research method is prone to bias. If students with higher potential outcomes tend to take the prep class, this research design will tend to produce upwardly biased estimates of the ATE; if students with low potential outcomes tend to take the class in order to improve what they expect to be a sub-par score, this research design will tend to produce downwardly biased estimates of the ATE.

7. FE, exercise 2.8.

a) **Each of the treatments had a slight effect on the first outcome, the probability of residence verification. In the control group, this rate was 20/21 or approximately 95%. In the three treatment groups, the rate is 100%, implying an average treatment effect of approximately 100 – 95 = 5 percentage points. In terms of the median number of days until residence verification, the RTIA and NGO treatments were the same as the control group, implying an estimated ATE of 37 – 37 = 0. However, the Bribe group received their verification in only 17 days, which is 37 – 17 = 20 days faster than the control group.**

b) **In the control group, the rate was 5/21 or 24%. The NGO group fared slightly worse 3/18 = 17%. When a right to information request was filed, this rate jumped to 20/23 = 87%, which approaches the 24/24 = 100% success rate among those who paid a bribe.**

c) **Although the RTIA treatment does not appear to speed the process of residency verification, it does seem to increase the probability of receiving a card by 20/23 – 5/21 = 63 percentage points over the control group, which seems like a large effect, especially for a treatment that may be implemented inexpensively by applicants.**

8. FE, exercise 2.9. Please think of the outcome variable as an individual's answer to the survey question "Are you in favor of raising the estate tax rate in the United States?"

a) **This assumption may not be plausible in this application. Although lottery winners are chosen at random from the pool of players in a given lottery, this study does not compare (randomly assigned) winners and losers from a pool of lottery players. Instead, winners are compared to non-winners, where the latter group may include non-players. Winning is therefore not randomly assigned. If frequent players are more likely to win than non-players and the two groups have different potential outcomes, the comparison of the two groups may be prone to bias. For example, perhaps those that play the lottery are more conservative in the first place relative to those who do not; in this case, we may observe lottery winners also being more conservative than Americans who have not won the lottery.**

b) **The assumption is not rooted in a randomization procedure because frequent players are still more likely to be winners than infrequent players. Unfortunately, without detailed information about how many tickets were purchased for each lottery, we don't know the exact probability that each subject would win. If frequent and infrequent players have different potential outcomes, the comparison is prone to bias.**

9. FE, exercise 2.12(a). In your answer, give some intuitive explanation in English for what the mathematical expressions mean.

In plain(ish) language, the first assumption being made is "On average, the number of violent encounters those who do not read have when they actually do not read is the same as the number of violent encounters who do read would have had if they had not read." The second is "On average, the number of violent encounters those who do not read would have had if they did read is the same as the number of violent encounters those who do read have."

In this case, those who self-select into the treatment may have distinctive potential outcomes – bookish inmates may be less prone to violence. In that case, E[Y_i(0) | D_i=0] > E[Y_i(0) | D_i=1], meaning those who do read would have committed less violence even if they had not read. Thus, a comparison of readers and nonreaders will not tend to produce unbiased estimates of the ATE.