Stat 502 HW1

Dongyang Wang

10/8/2021

Exercise 1

(a) The advantage of letting the children choose is to create an observational study, where their actions are genuine and not affected by the instructions or other factors. The disadvantage is also obvious: no control trial can be done in this case and the experiment itself becomes meaningless. Moreover, there is no randomization, possibility for replication, or any blocking. That means it is hard for us to draw causal conclusions even if there are any only correlational patterns.

(b) The advantage of giving the first 10 diet A and next 10 children diet B is to guarantee both diets are blocked so there we can compare across these two different diets and potentially see some difference. But a disadvantage here is the potential for systematic bias. It's possible, for example, people who are taller tend to stand back and shorter children stand near the front of line. So picking the first 10 may cause confounding factors to make an impact, in this case, height itself.

(c) The advantage of alternating is to ensure blocking, but for randomization, it does so only to a certain but not sufficient extent. Since A,B,A,B follows a clear pattern, the assignment of children to diets is not completely random. As discussed in (b), a clear pattern in the assignment introduces potential problems for confounding factors or systematic bias. For example, it is possible that after observing the pattern, some children switch the places with others to obtain the diet they want.

(d) The advantage of tossing a coin is the guarantee of randomization. However, it does not offer blocking. So it is possible that there are 15 children assigned to diet A but only 5 assigned to diet B. The validity of such an experiment is undermined because a conclusion is hard to draw when one group is more representative than the other. In other words, one group can be not represenatitive enough, in my example, B.

(e) This is the most appropriate method among the five choices. We have 10 children for each diet (blocking); we randomized the assignment entirely to get rid of confounding factors and we can replicate the experiment given the same methodology. One potential disadvantage is that children may not be very willing to eat what others want them to–but this disadvantage is avoidable only if we choose an observational study.

Exercise 2

(b)

```
vector1 <- rep(c("A", "B"), each = 10)
vector1
```

```
##  [1] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "B" "B" "B" "B"
## [20] "B"
```

(c)

```
vector2 <- rep(c("A", "B"), 10)
vector2
```

```
##  [1] "A" "B" "A" "B" "A" "B" "A" "B" "A" "B" "A" "B" "A" "B" "A" "B" "A" "B" "A"
## [20] "B"
```

(d)

```
g.binomial <- rbinom(20,1,0.5)
vector3 <- c()
for (i in g.binomial){
  if (i == 1){
    vector3 <- append(vector3, "A")}
  else{
    vector3 <- append(vector3, "B")}
}
vector3
```

```
##  [1] "B" "A" "A" "B" "B" "B" "A" "A" "B" "B" "B" "A" "B" "B" "A" "B" "B" "B" "B"
## [20] "A"
```

(e)

```
vector4 <- rep(0, 20)
index <- sample(20,10, replace = FALSE)
vector4[index] <- "A"
vector4[vector4 == 0] <- "B"
vector4
```

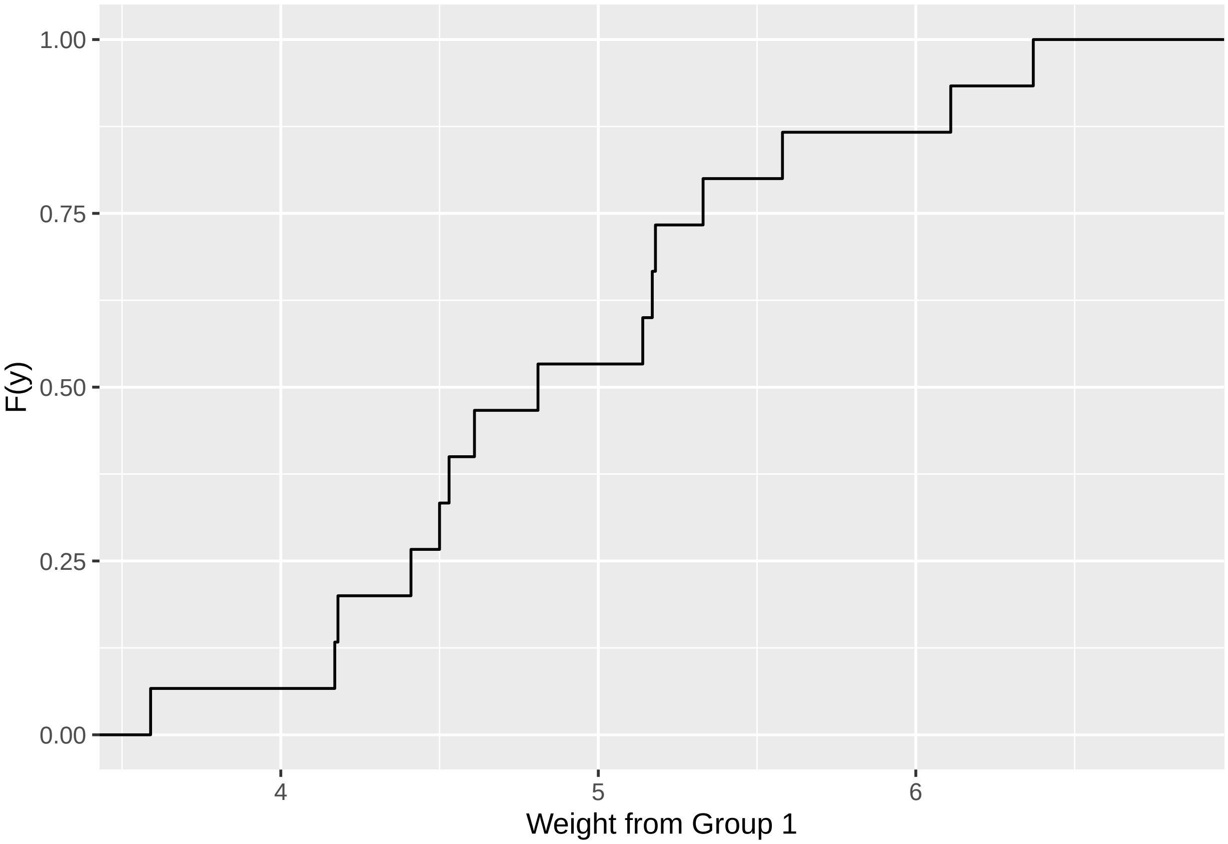
```
##  [1] "A" "A" "A" "A" "B" "B" "B" "A" "A" "B" "B" "A" "B" "A" "A" "A" "B" "B" "B"
## [20] "B"
```

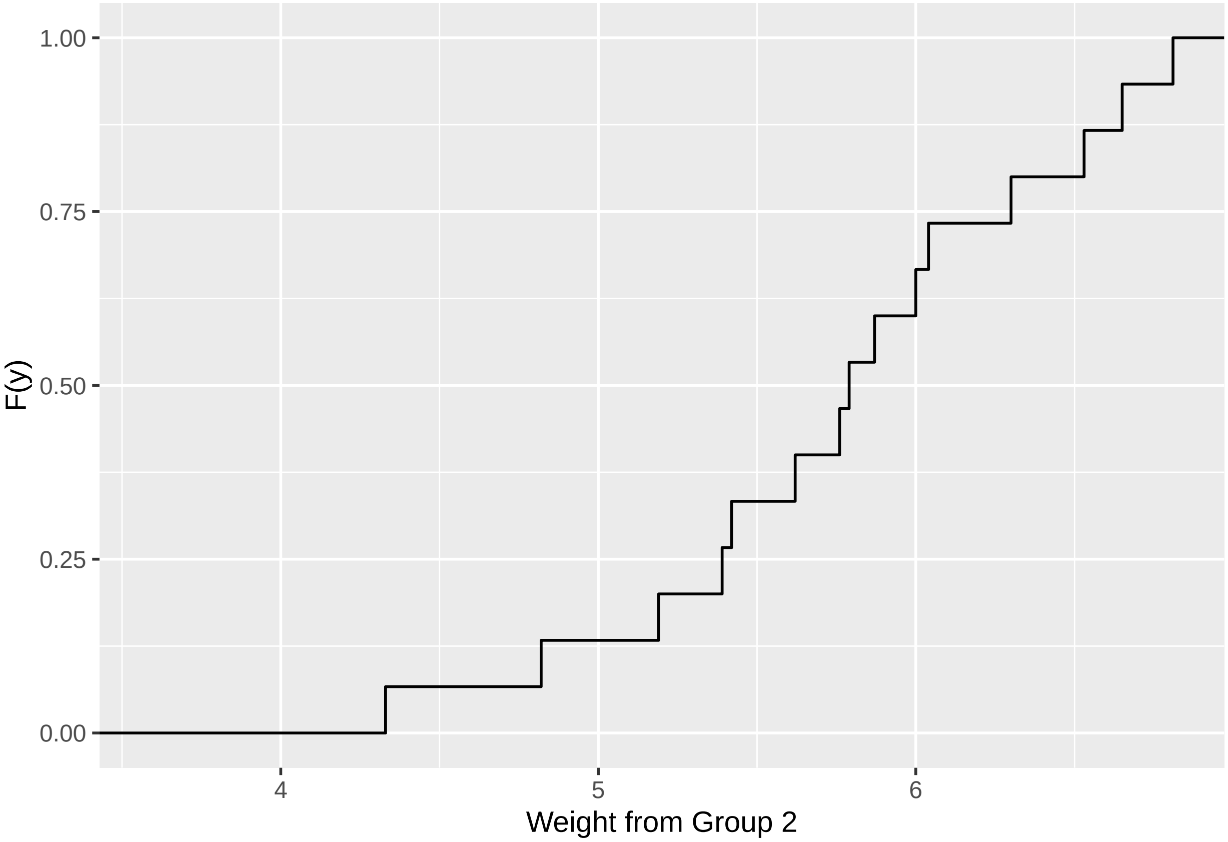
Exercise 3

(a)

For group 1, the mean weight is 4.912, median is 4.810, standard deviation is 0.7500305; For group 2, the mean weight is 5.768, median is 5.790, standard deviation is 0.6788983.

The following page contains the cdf for the two groups.

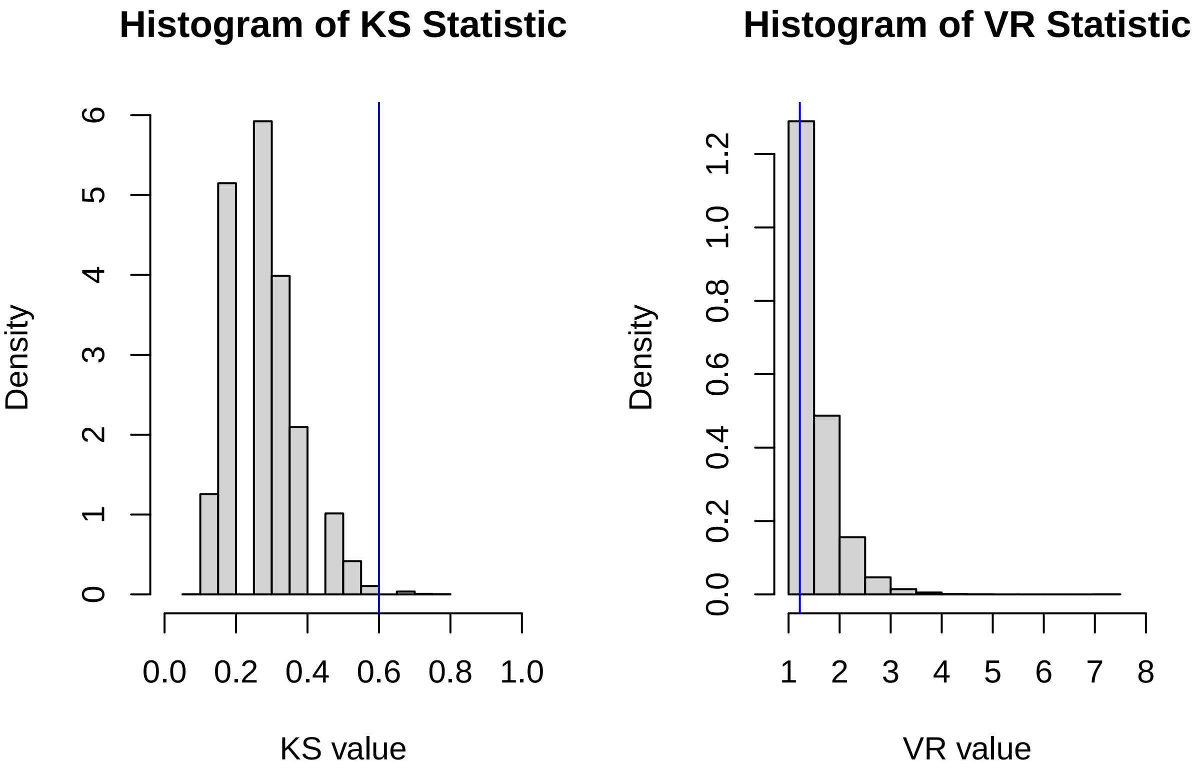

F(y)
1.00
0.75
0.50
0.25
0.00
4
5
6
Weight from Group 1


F(y)
1.00
0.75
0.50
0.25
0.00
4
5
6
Weight from Group 2

(b)

i. The KS-statistic is 0.6; the variance ratio statistic is 1.22053.

ii.



iii.

We have calculated that the p-value is 0.0077 for the KS statistic. and 0.6591 for the variance ratio statistic. Therefore, we can reject the null hypothesis given the KS statistic, but we fail to reject the null with the variance ratio statistic. As discussed in class, the KS statistic is sensitive to any difference between the two treatments. In comparison, the variance ratio statistic focuses more on the overall difference (or variation in general). Note that our null hypothesis is that there is no difference between the two treatments in terms of plant growth. Therefore, as the ks statistic is sensitive to difference, which is our topic of concern, we can reject the null hypothesis with the support of the KS statistic and conclude that there are differences between the two treatments.

(c) We know that for any hypothesis testing, the probability of rejecting the true null hypothesis is α and 0.05 in our case. We consider independence of the tests and if they are identical. If two tests are identical and independent, we would expect to see the probability of at least one test rejecting the true null hypothesis to be $\alpha = 0.05$. However, as the question prompt says, the tests in 3b are not identical. Then, we consider two cases. First, if the tests are independent, we know from the definition of significance level that each test has a probability of α in rejecting the true null hypothesis. So, with some calculation of probability, at least one of the tests rejects the true null hypothesis is $1 - (1 - \alpha) \times (1 - \alpha) = 2\alpha - \alpha^2 = 0.1 - 0.0025 = 0.0975$. On the other hand, if the tests are not independent, it is more complicated. In our example, the two tests are kind of related. Since if one test tends to fail to reject the null given the other fails too, and failing to reject happens 95% of the time for a given test, we know that the probability of at least one test is rejecting the true null is falling. So the upper bound is still as calculated above, 0.0975.

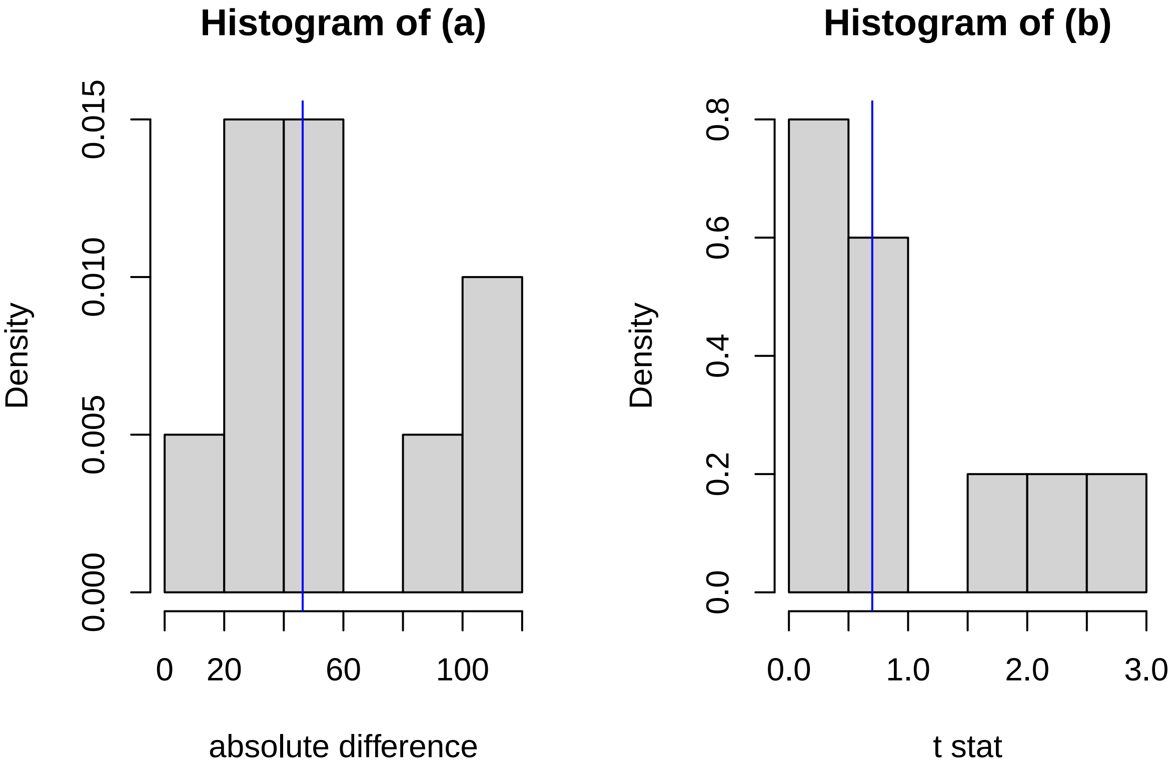
Exercise 4

(a)

A randomization test of the hypothesis returns a p-value of 0.4. Therefore, we reject the null hypothesis that the treatment makes any difference.

(b)

A randomization test of the hypothesis returns a p-value of 0.4. Therefore, we reject the null hypothesis that the treatment makes any difference. The two tests in (a) and (b) have the same p-value.



(c) See next page of handwritten notes.

Since $s_p^2 = \frac{1}{(n_A-1)+(n_B-1)} \left(h(y_A, y_B) - \frac{n_A n_B}{n_A+n_B} (\bar{y}_B - \bar{y}_A)^2 \right)$

and $s_p^2 = \frac{n_A - 1}{n_A - 1 + n_B - 1} s_A^2 + \frac{n_B - 1}{n_A - 1 + n_B - 1} s_B^2$

we have $(n_A - 1) s_A^2 + (n_B - 1) s_B^2 = h(y_A, y_B) - \frac{n_A n_B}{n_A + n_B} (\bar{y}_B - \bar{y}_A)^2$

So,

$$h(y_A, y_B) = (n_A - 1) s_A^2 + (n_B - 1) s_B^2 + \frac{n_A n_B}{n_A + n_B} (\bar{y}_B - \bar{y}_A)^2$$

$$= \sum_{i=1}^{n_A} (y_{Ai} - \bar{y}_A)^2 + \sum_{j=1}^{n_B} (y_{Bj} - \bar{y}_B)^2 + \frac{n_A n_B}{n_A + n_B} \left(\frac{\sum_{j=1}^{n_B} y_{Bj}}{n_B} - \frac{\sum_{i=1}^{n_A} y_{Ai}}{n_A} \right)^2$$

$$= \sum_{i=1}^{n_A} (y_{Ai}^2) \pm 2 \cdot \frac{\sum_{i=1}^{n_A} y_{Ai} \cdot \sum_{i=1}^{n_A} y_{Ai}}{n_A} + \frac{\sum_{i=1}^{n_A} y_{Ai} \cdot \sum_{i=1}^{n_A} y_{Ai}}{n_A} + \sum_{j=1}^{n_B} (y_{Bj}^2) - \frac{\sum_{j=1}^{n_B} y_{Bj} \sum_{j=1}^{n_B} y_{Bj}}{n_B}$$

$$+ \frac{n_A \sum_{j=1}^{n_B} y_{Bj} \sum_{j=1}^{n_B} y_{Bj}}{(n_A + n_B) n_B} + \frac{n_B \sum_{i=1}^{n_A} y_{Ai} \sum_{i=1}^{n_A} y_{Ai}}{(n_A + n_B) n_A} - 2 \frac{\sum_{j=1}^{n_B} y_{Bj} \cdot \sum_{i=1}^{n_A} y_{Ai}}{n_A + n_B}$$

$$= \sum_{i=1}^{n_A} (y_{Ai}^2) - \frac{n_A \left(\sum_{i=1}^{n_A} y_{Ai} \right)^2}{(n_A + n_B) n_A} - \frac{n_B \left(\sum_{j=1}^{n_B} y_{Bj} \right)^2}{(n_A + n_B) n_B} + \sum_{j=1}^{n_B} (y_{Bj}^2) - \frac{2 \sum_{j=1}^{n_B} y_{Bj} \sum_{j=1}^{n_A} y_{Ai}}{n_A + n_B}$$

$$= \sum_{i=1}^{n_A} (y_{Ai}^2) + \sum_{j=1}^{n_B} (y_{Bj}^2) - \frac{\left(\sum_{i=1}^{n_A} y_{Ai} \right)^2 + 2 \sum_{j=1}^{n_B} y_{Bj} \sum_{i=1}^{n_A} y_{Ai} + \left(\sum_{j=1}^{n_B} y_{Bj} \right)^2}{n_A + n_B}$$

$$= \sum_{i=1}^{n_A} (y_{Ai}^2) + \sum_{j=1}^{n_B} (y_{Bj}^2) - \frac{1}{n_A + n_B} \cdot \left(\sum_{i=1}^{n_A} y_{Ai} + \sum_{j=1}^{n_B} y_{Bj} \right)^2$$

Therefore, $h(y_A, y_B)$ can be written as a function of y_A and y_B regardless of the test. And it is only related to the total sample size $n_A + n_B$ which does not change across tests. And $h(y_A, y_B)$ doesn't depend on the labeling of data.

In the ~~ttest~~ t-statistic test, $\frac{n_A n_B}{n_A + n_B} (\bar{y}_B - \bar{y}_A)^2$ will be cancelled out and g_t is only related to $h(y_A, y_B)$.

$$g_t(y_A, y_B) = \frac{|\bar{y}_B - \bar{y}_A|}{\sqrt{\frac{1}{n_A - 1 + n_B - 1} \left[h(y_A, y_B) - \frac{n_A n_B}{n_A + n_B} (\bar{y}_B - \bar{y}_A)^2 \right]} \cdot \sqrt{\frac{n_A + N_B}{n_B n_A}}}$$

Regarding $h(y_A, y_B)$ as constant and we will be able to cancel $|\bar{y}_B - \bar{y}_A|$ and $\sqrt{(\bar{y}_B - \bar{y}_A)^2}$. Both g, g_t will be affected by y_A, y_B.