

**4. Building off of RQ#3 above, sort your observations so they are in ascending order based on the outcome under control. Randomize two at a time: one fair coin flip now assigns either the first or second observation to treatment (and the other to control). A second fair coin flip assigns either the third or the fourth observation to treatment (and the other to control). This continues so on and so forth. Use the appropriate permutation test to assess your data set using (i) Fisher's Sharp Null and (ii) the null hypothesis that the treatment adds five units to the outcome under control. Contrast the results here with the results from RQ#3.**

So what are we doing here? Specifically, why sort on  $r_C$ ?

| person | $r_c$ | $r_t$ |
|--------|-------|-------|
| 7      | -6.7  | -4.7  |
| 1      | -6.6  | -4.6  |
| 8      | -4.8  | -2.8  |
| 10     | -1.9  | 0.1   |
| 9      | -0.8  | 1.2   |
| 6      | 3.9   | 5.9   |
| 5      | 4.7   | 6.7   |
| 3      | 6.0   | 8.0   |
| 4      | 6.3   | 8.3   |
| 2      | 14.0  | 16.0  |

If you think about it, baseline covariates (i.e., the information we have about the observational units before the intervention) are useful primarily due to their prognostic ability<sup>1</sup>, their usefulness as predictive of the outcome under control. Taken together, the baseline covariates are what we have to form an approximation for the potential outcome of interest.

So what good does matching on  $r_C$  do us? It depends on how  $r_C$  and  $r_T$  are related. If you generated them completely independently from each other then matching won't help. But if there's any correlation between the potential outcomes then we should see a bit of a difference in the null distribution.

In RQ#4, there were  $2^{10} = 1024$  possible realizations of the randomization. Now we've constrained it quite a bit. Now there are only  $2^5 = 32$ . There are a few implications here. The most noticeable is that we can't get a tiny p-value any more. Even if we do a one-sided test, the smallest we can get is  $1/32$ . In RQ#4 we could get something more like  $1/(1024-2) = 0.0009...$

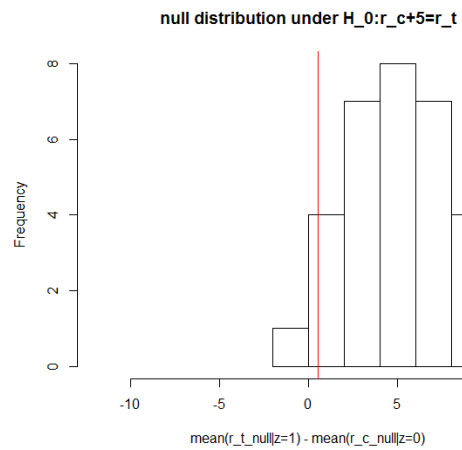
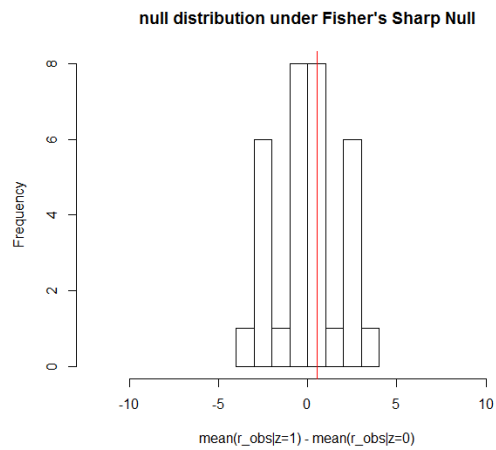
The table below shows a particular randomization using the matched pairs.

---

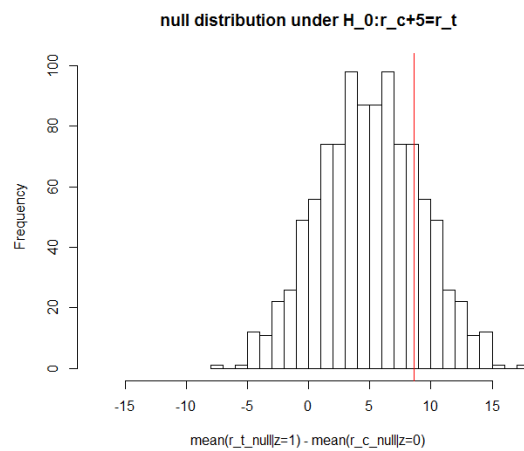
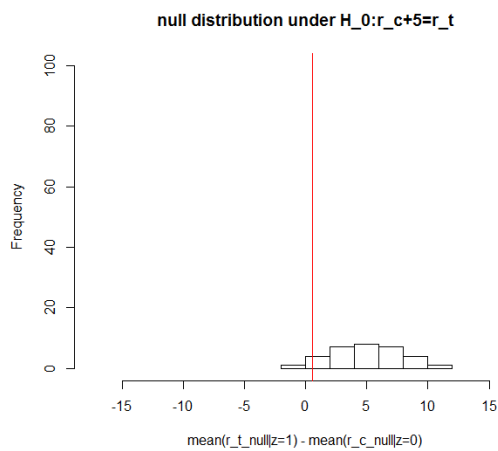
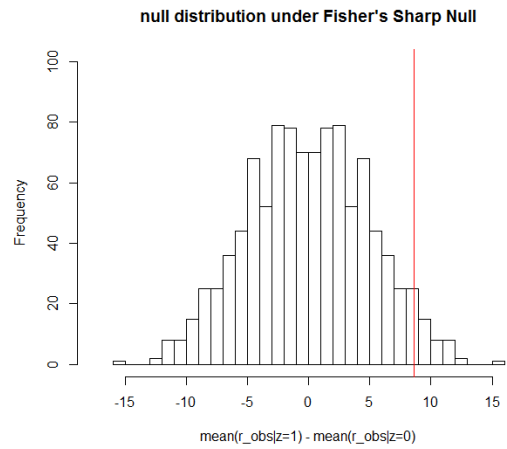
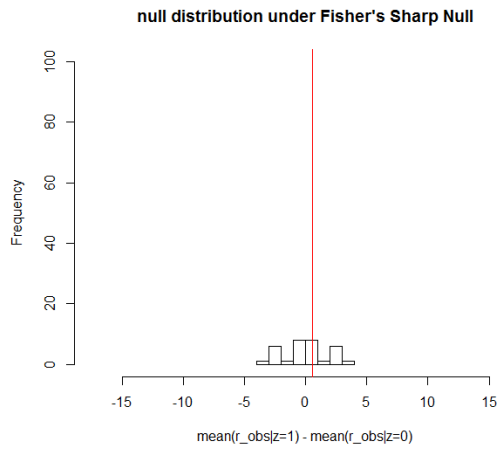
<sup>1</sup> Perhaps more practically, the covariates also help you can tell the difference between each participant in your trial.

| person | r_c  | r_t  | z | r_obs | delta |
|--------|------|------|---|-------|-------|
| 7      | -6.7 | -4.7 | 1 | -4.7  | 2     |
| 1      | -6.6 | -4.6 | 0 | -6.6  | 2     |
| 8      | -4.8 | -2.8 | 1 | -2.8  | 2     |
| 10     | -1.9 | 0.1  | 0 | -1.9  | 2     |
| 9      | -0.8 | 1.2  | 0 | -0.8  | 2     |
| 6      | 3.9  | 5.9  | 1 | 5.9   | 2     |
| 5      | 4.7  | 6.7  | 1 | 6.7   | 2     |
| 3      | 6.0  | 8.0  | 0 | 6.0   | 2     |
| 4      | 6.3  | 8.3  | 1 | 8.3   | 2     |
| 2      | 14.0 | 16.0 | 0 | 14.0  | 2     |

From this, we can build up the null distributions:



Using the same scaling as the null distributions in RQ#4:



The most obvious differences are the height and the width of the histograms. The height tells us a bit about how finely we'll be able to parse out our p-values. The width tells us how much variation we're seeing from permutation to permutation (under the null). Two questions: (1) Why would people want to trade off the histograms on the right for the ones on the left? (2) All of these histograms are symmetric. Is that necessarily true? You've got some code now that can help you build up enough experience to take a solid swing at #1 above. And #2 requires just a little bit of math.

The particular realizations of the randomizations (i.e., the  $r_{\text{obs}}$  from) are summarized by their average difference shown in red. They're different in the different plots because we had different randomizations.