Instructions

- Submit your homework on Sakai before 4:40pm on the due date. Half credit for homework turned in after 4:40pm on the due date, no credit for homework turned in after the due date (because solutions will be posted).
- You may discuss these problems with each other verbally, but must write up the answers on your own, and may not share or show your answers to anyone else (this applies to code as well).
- Submit your R Script as a separate attachment. The script should be self-contained, so someone else should be able to run it as is and get your results. The script should also be well-commented, so it is clear which code goes with which question, and it is clear what each chunk of code does.
- Please be concise!

Data

In children at risk for malaria, research is conducted in stimulating the immune system with micronutrients supplementation, an important one being vitamin A[1]. Table 1 gives measurements of the malaria parasite Plasmodium falciparum (Pf), in $\log_e$ (counts / microL), for 6 children taking vitamin A supplementation and 6 children taking a control (placebo). The children were randomized in a completely randomized experiment. We wish to test Fisher's sharp null hypothesis of no treatment effect on these 12 children. Take the test statistic to be the difference in sample means.

Placebo	Vitamin A
8.62	7.53
1.48	0.06
8.93	2.19
9.57	7.32
2.65	1.72
7.30	7.62

Table 1: Parasite load [$\log_e$ (count/microL)]

[1]Shankar et al., (1999). "Effect of vitamin A supplementation on morbidity due to Plasmodium falciparum in young children in Papua New Guinea: a randomised trial," *Lancet* 354:203-9

Exercises

1. (1 point) Why did they use a placebo for the control group?
2. (1 point) Why is a sharp null hypothesis required to create a randomization distribution?
3. *Fisher's Exact p-value*
   (a) (3 points) Create the exact randomization distribution, and give the plot.
   (b) (2 points) Compute the exact p-value using this randomization distribution.
   (c) (2 points) Imagine you are a statistician consulting for these researchers. Explain to them what the p-value means for their study, and what conclusion they can make, without using any statistical terms.
4. *Approximate p-value*
   (a) (3 points) Create a randomization distribution using simulation.
   (b) (2 points) Calculate an approximate p-value based on this simulated randomization distribution.
   (c) (1 point) Run the code to compute this approximate p-value several times. Do you think your number of simulated randomizations is high enough? Explain.
5. *Neyman's Inference*
   (a) (2 points) Using Neyman's methods of inference, compute the p-value.
   (b) (2 points) Compare this to the p-value in 3b. Are they close? Which do you trust more, and why?
   (c) (2 points) Create a 95% confidence interval, and interpret in context.
   (d) (1 point) Show that, for a completely randomized experiment, $\hat{\tau} = \overline{\mathbf{Y}}_T^{obs} - \overline{\mathbf{Y}}_C^{obs}$ is an unbiased estimator for $\tau = \overline{Y}(1) - \overline{Y}(0)$.
6. *Bernoulli Randomized Experiment*
   Suppose that instead of a completely randomized experiment, this was a Bernoulli randomized experiment.
   (a) (3 points) Generate a randomization distribution (exact or simulated) and give the plot. (Hint: The function `rbinom` may be useful).
   (b) (1 point) If you look at the randomization distribution, you will notice that some of the values are NaN - R's value for "Not a number". What is going on in these cases? (Note: before computing the p-value, you will have to create a new randomization distribution that avoids these cases.)
   (c) (2 points) Compute the p-value (exact or approximate).
   (d) (2 points) Compare the standard error of the test statistic under the completely randomized experiment to the standard error under the Bernoulli experiment. What are the two standard errors, and why is one larger than the other?