Homework 2

Statistics 140

Due February 18, 2026 at 9 pm

Problem 1.

a) The Fisher-exact p-value can be calculated exactly or can be approximated using a smaller number of randomized allocations. Discuss why, when possible, it is better to calculate it exactly.

b) Consider $N = 10$ experimental units (indexed by i) randomized to a binary treatment W, either active ($W_i = 1$) or control ($W_i = 0$). An outcome variable Y is measured after exposure to the treatment. Consider a completely randomized experiment with the number of treated units equal to $N_T = 4$. What is the minimum Fisher-exact p-value one can obtain from this experiment?

c) Load the `small_epi.RData` epigenetic dataset and calculate the minimum Fisher-exact p-value one can obtain from this completely randomized experiment randomizing participants to clean air vs ozone?

d) Consider $N = 20$ experimental units (indexed by i) randomized to a binary treatment W_i, either active ($W_i = 1$) or control ($W_i = 0$). An outcome variable Y is measured after exposure to the treatment. Consider a Bernoulli randomized experiment. What is the minimum Fisher-exact p-value one can obtain from this experiment?

e) Read the article from Zhong et al. and explain the limitations of Figure 2. Assume the assignment mechanism for this study was Bernoulli.

Problem 2.

A completely randomized experiment was done to evaluate the effect of honey treatment on nocturnal cough frequency in children. The outcome is measured on a scale from zero ("not at all frequent/severe") to six ("extremely frequent/severe"). Let us consider, for relative ease of exposition, a small dataset with six children. The following table provides the observed data on honey.

Unit i	W_i^{obs}	Y_i^{obs}
1	1	3
2	1	2
3	1	0
4	0	4
5	0	6
6	0	1

We are interested in assessing whether the honey treatment has an effect on nocturnal cough.

a) Provide the observed data table with the potential outcomes.

b) Specify the sharp null hypothesis.

Suppose we are testing the sharp null hypothesis specified in b). The goal is to calculate the Fisher-exact p-value using the mean difference as our test statistic T, i.e.,

$$T = \frac{1}{N_1} \sum_{i:W_i=1} Y_i^{obs} - \frac{1}{N_0} \sum_{i:W_i=0} Y_i^{obs},$$

where N_0 and N_1 are the number of children taking the control and active treatment, respectively. We expect honey to improve nocturnal cough, i.e., we expect T^{obs} to be negative .

c) First let us construct a matrix W with all the possible randomized allocations, but no duplicate vector.

d) Now, let us construct the null randomization distribution of T. For each possible randomized allocation, calculate the value of the test statistic under the sharp null hypothesis. Draw a histogram of these values.

e) Calculate the value of the observed test statistic and locate it on the histogram you constructed in d).

f) Calculate the Fisher-exact p-value and describe how you obtained it.

Problem 3.

Air pollution exposure has been shown to be associated with DNA methylation, which is an epigenetic process by which methyl groups are added to the DNA molecule. DNA methylation can change the activity of a DNA segment without changing the sequence. More data from the epigenetic randomized experiment are provided to examine whether ozone (vs. clean air) exposure changes DNA methylation measured at six CpG sites, which are regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide.

a) Import the `completelyrandomized.csv` dataset, which contains the DNA methylation measurements at six CpG sites after random exposure to ozone (`exp=2`) or clean air (`exp=0`) for seventeen participants.

b) Construct a assignment vector `W.obs` equal to `1` if participants were exposed to ozone and `0` if participants were exposed to clean air.

c) Using `W.obs`, construct a matrix W with all possible randomized allocations. What are the dimensions of your W matrix? Explain why.

d) Construct a vector `Y.obs` with the outcome measurement for CpG site `cg00000029`.

e) Now, we choose $T_{Welch} = \frac{\bar{Y}_1^{obs} - \bar{Y}_0^{obs}}{\sqrt{\frac{s_0^2}{N_0} + \frac{s_1^2}{N_1}}}$ as our test statistic, where s_0^2 and s_1^2 are the sample variance of the outcome Y after exposure to clean air and ozone, respectively. Calculate the observed value of T_{Welch} and call it `T.obs`.

f) Verify that the observed value of T_{Welch} you calculated in d) is the same as when you run the following R code `t.test(Y.obs[W.obs==1],Y.obs[W.obs==0])$statistic`, where `Y.obs[W.obs==0]` and `Y.obs[W.obs==1]` are the observed outcome vectors of the clean air and ozone groups, respectively.

g) Let us now construct the null randomization distribution of T_{Welch}. Draw a histogram of these values and locate `T.obs` on the histogram you constructed.

h) Calculate the Fisher-exact p-value assuming "extreme" corresponds to "greater than".

i) Compare the Fisher-exact p-value to the one you would obtain with the following R code `t.test(Y.obs[W.obs==1],Y.obs[W.obs==0],alternative="greater")$pvalue`. The approximating p-value returned by the `t.test` function assumes that T_{Welch} follows a Student's t distribution under the Neymanian null hypothesis.

j) Create a six-element vector `p.value`. Using two `for` loops, calculate the Fisher-exact p-values for the other five CpG sites in the dataset and fill in the following table, for which we have filled the first row. For the first three CpG sites, assume "extreme" corresponds to "greater than". For the last three CpG sites, assume "extreme" corresponds to "less than". Describe your results in light of problem 1.

CpG site	T^{obs}	Fisher-exact p-value	Approximating p-value
cg00000029	0.5724248	$\frac{5679}{19448} = 0.2920095$	0.2888397
cg09008103		$\frac{}{19448} =$	
cg14354270		$\frac{}{19448} =$	
cg21036194		$\frac{}{19448} =$	
cg00673208		$\frac{}{19448} =$	
cg20976708		$\frac{}{19448} =$	

k) Construct a figure with the six null randomization distributions and locate the six observed test statistics. Comment on the shape of the null randomization distributions.

Problem 4.

a) Using the same dataset, provide a point estimate (call it `tau.hat`) and a two-sided 95% confidence intervals (CI) for the mean difference in DNA methylation for each CpG site between ozone and clean air exposure (i.e., τ). Start with `cg00000029` and describe your results.

b) Verify your answers using the `t.test` function with the option `alternative`="two.sided".

c) Write a `for` loop to fill in the following table, for which we have filled in the first row. Round your estimates to three decimals.

CpG site	$\hat{\tau}$	Two-sided 95% confidence interval (CI)
cg00000029	0.009	[-0.026; 0.044]
cg09008103		
cg14354270		
cg21036194		
cg00673208		
cg20976708		