STAT474/STAT574, Final Homework

Please get a printed copy of your solutions under my door by 5pm Friday May 8th. Turn in R code as an appendix only.

1. Consider Fisher's method of combining p-values versus using a combined sample. Suppose there are two samples, $x_1, \ldots, x_{20}$, and $y_1, \ldots, y_{30}$. Let

$$x_1, \ldots, x_{20}, y_1, \ldots, y_{30} \overset{iid}{\sim} N(\mu, 1)$$

(i.e., both samples have independent and identically distributed values from a $N(\mu, 1)$ distribution).

(a) Estimate the power for rejecting the null hypothesis

$$H_0 : \mu = 0$$

using Fisher's combined p-value when $\mu = 0.1, 0.2, \ldots, 1.0$. You may use the `combinedp()` function shown in class for the week 15 lectures to make the problem easier. To estimate the power, you should use several iterations for each value of μ (e.g., 100 or more iterations to make the power curve reasonably smooth).

(b) Repeat (a), but estimate the power using a t-test on the combined data for each value of μ. I.e., use samples of size 50 from a $N(\mu, 1)$ distribution and use single-sample t-tests.

(c) What do you conclude about the power of Fisher's combined p-value compared to getting a p-value directly from the combined data? Plot the power for both tests as a function of μ, making sure that the plot is well-labeled, and that you have separate curves on the same plot for the power using the t-test on the combined samples versus power for Fisher's combined p-value method.

2. Suppose you an evolutionary tree of HIV virus strains. The strains are labeled $a_1, \ldots, a_4$ and $b_1, \ldots, b_5$. (a) What is the probability that all of the $a_1, \ldots, a_4$ strains cluster together under the null hypothesis of no population structure? (b) What is the probability that all of the $b_1, \ldots, b_5$ strains cluster together? (c) What is the probability of reciprocal monophyly?

To do this problem, you can use probabilities from the file `w12-populationStructure.pdf` file from the week 12-13 notes.

3. Genetic association testing of just 100 candidate genes results in the following unsorted p-values:

```
6.769446e-06 8.970062e-01 4.560979e-04 3.847698e-02 6.748949e-03
1.940932e-03 2.506972e-01 7.299721e-01 3.520195e-02 1.602559e-03
5.729804e-01 6.004475e-01 2.605110e-02 2.405015e-01 1.526348e-04
2.854273e-04 1.040662e-01 1.123220e-02 1.562522e-03 1.155829e-02
9.236859e-03 4.466842e-04 1.915730e-06 3.114801e-03 2.991605e-04
3.434193e-02 1.331156e-01 8.334134e-02 2.351013e-01 1.480859e-01
8.026923e-04 8.163445e-03 7.490881e-01 4.259426e-09 8.610412e-02
3.346894e-01 9.606209e-02 4.326995e-01 1.399422e-02 2.135585e-02
1.598572e-02 1.202101e-03 2.224284e-01 3.958128e-02 3.671843e-02
1.753761e-04 2.399581e-01 1.975891e-02 1.468612e-01 1.023071e-01
1.375971e-03 9.100954e-01 4.435508e-01 6.456446e-01 5.839204e-01
9.219935e-01 2.270422e-01 4.867923e-02 6.338614e-04 5.460440e-05
1.089233e-02 3.144174e-03 2.323041e-01 6.485346e-01 3.905734e-01
4.667591e-01 1.239194e-06 1.519208e-01 2.895035e-01 5.852688e-07
1.197480e-04 5.830437e-04 4.763936e-02 1.452288e-01 2.393055e-01
6.872326e-02 1.158495e-01 6.008650e-01 4.434467e-05 5.272612e-02
2.228960e-01 9.851721e-03 3.583006e-02 3.360167e-02 4.062278e-01
1.966408e-03 6.093815e-03 5.349392e-02 1.048595e-07 8.445405e-03
8.044211e-01 4.378757e-01 3.926310e-01 2.656045e-02 5.365501e-01
3.128340e-01 2.148439e-04 6.335492e-02 3.366105e-04 5.660029e-03
```

Determine how many associations can be considered statistically signficiant using $\alpha = 0.05$ with (a) Bonferroni correction, (b) the Benjamini-Hochberg method controlling for FDR.

4. Consider the following family data. Within each family, the mother is the first row, the father is the second row, and the child is the third row. The variables are a family ID, allele 1, allele2, disease status (0=no disease, 1=disease). Here the genotype is the combination of the alleles. For example, allele 1 = 0 and allele 2 = 1 is a heterozygous individual, while allele 1 = 0 and allele 2 = 0 is a homozygous individual.

```
famID allele1 allele2 disease
001 0 1 0
001 0 1 1
001 0 0 1
002 0 0 0
002 0 1 1
002 0 1 1
003 0 0 1
003 0 1 1
003 0 0 1
004 0 1 0
004 0 1 1
004 1 1 1
005 0 1 0
005 0 0 1
005 0 0 1
```

Use this data to (a) set up a table of transmitted and nontransmitted alleles from the parents and perform a TDT test for whether the marker is associated with disease status. (b) Report the test statistic and p-value, and interpret the result. Do everything by hand, except getting the p-value, for which you can use the `pchisq()` function in R.