Stat 27850/30850: Problem set 1

1. Consider a multiple testing scenario where we are testing the global null. Suppose that you have n p-values $P_1, \ldots, P_n$. A proportion π of these p-values are true signals, distributed as

$$P_1, \ldots, P_{\pi n} \sim \text{Unif}[0, \tau]. \quad \texttt{tau smaller, true signal larger}$$

The rest of them are nulls,

$$P_{\pi \cdot n+1}, \ldots, P_n \sim \text{Unif}[0, 1].$$

And all the p-values are independent. For simplicity assume $\tau \geq \alpha/n$.

   (a) Write an exact expression for the probability of rejecting the global null, by using the Bonferroni correction at level α.

   (b) Calculate the probability of rejecting the global null using Fisher's test, using the normal approximation to the χ^2 distribution (note that χ^2_m has expected value m and variance $2m$, and the normal approximation is very accurate for large m due to the central limit theorem). You should show that this probability is (approximately) equal to $\Phi(\Phi^{-1}(\alpha) - \sqrt{n} \cdot \pi \log(\tau))$, where Φ is the CDF of a standard normal.

   (c) Now let $\pi = \frac{1}{n^{1/3}}$ and $\tau = \frac{1}{2}$. Show that the probability of rejecting the global null tends to 1 for Fisher's test, but tends to α for Bonferroni's (i.e. Bonferroni's test is no better than random guessing. It's trivial to see that the probability of rejection is *at least* α, i.e., at least as good as random guessing, so you only need to show an upper bound, i.e., the probability of rejection is upper bounded by a quantity that is $\approx \alpha$; this will be easiest to prove using a union bound).

   (d) Finally let $\pi = \frac{1}{n^{2/3}}$ and $\tau = \frac{1}{n}$. Show that the probability of rejecting the global null tends to 1 for Bonferroni's test, but tends to α for Fisher's (i.e. Fisher's test is no better than random guessing).

2. Next let's examine the difference between Bonferroni's and Fisher's methods numerically. We'll choose $\alpha = 0.05$, $n = 2^{30}$, $\pi = 2^{-i}$, and $\tau = 2^{-j}$, for $i, j \in \{1, \ldots, 30\}$. We are not running simulations, just doing calculations. For each value of i & j, based on your work in the previous problem, set
`Bonferroni[i,j]`
to be the exact probability that Bonferroni's method rejects the global null (with this choice of n, π, τ), and set
`Fisher[i,j]`
as the same calculation for Fisher (using the normal approximation as before). Now run this code to visualize your results:

```
par(mfrow=c(1,2))
image(1:30,1:30,Bonferroni,xlab="-log_2(pi)",ylab="-log_2(tau)",
    main="Bonferroni",col=gray((0:10)/10))
image(1:30,1:30,Fisher,xlab="-log_2(pi)",ylab="-log_2(tau)",
    main="Fisher",col=gray((0:10)/10))
```

These types of figures are known as "phase transition diagrams" where we see a sharp transition from a high chance of success to a high chance of failure. The grayscale corresponds to the chance of rejection: white = 100% chance of rejecting the global null, and black = 0% chance. You should see that higher values of π (i.e. lower i), and lower values of τ (i.e. higher j), improve the chances for both methods. However, the regions of π, τ values where the methods are successful, are different for the two methods. Describe and explain what you see.

3. Gene expression / COPD & statins. In class we saw a gene expression data set where for each patient, we observe

   - A label: whether the patient takes statins, or not
   - Covariates: disease/healthy, age, sex
   - Gene expression levels (log transformed) for each of $n = 12381$ genes (`X[k,i]` is the log-transformed gene expression level for person k and gene i)

Copy and paste the following code to download and organize the data:

```
download.file(
"ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE71nnn/GSE71220/matrix/GSE71220_series_matrix.txt.gz",
"GSE71220_series_matrix.txt.gz")

gene_expr = t(read.table("GSE71220_series_matrix.txt.gz",skip=66,nrows=12381)[,-1])
nsample = dim(gene_expr)[1]; n = dim(gene_expr)[2]

statin = read.table("GSE71220_series_matrix.txt.gz",skip=26,nrows=1)[-1]
statin = (strsplit(toString(unlist(statin)),"_")[[1]][2+3*(0:616)] == "statin")
# which patients take statins

disease = read.table("GSE71220_series_matrix.txt.gz",skip=37,nrows=1)[-1]
disease = (unlist(strsplit(strsplit(toString(unlist(disease)),": ")[[1]],","))[2*(1:nsample)] == "COPD")
# patients with COPD disease or healthy patients

age = read.table("GSE71220_series_matrix.txt.gz",skip=38,nrows=1)[-1]
age = as.numeric(unlist(strsplit(strsplit(toString(unlist(age)),": ")[[1]],","))[2*(1:nsample)])
# age of patient

sex = read.table("GSE71220_series_matrix.txt.gz",skip=39,nrows=1)[-1]
sex = (unlist(strsplit(strsplit(toString(unlist(sex)),": ")[[1]],","))[2*(1:nsample)])
# sex of patient (M or F)
```

Let's look only at patients with COPD disease, and only the first 200 genes to save computation time. We will ignore the age & sex covariates for now. We want to see if there is a true association between statin use and gene expression levels (overall, across the n genes—not necessarily making a claim for any specific gene). Let's test this with the statistic

$$T = \sum_{i=1}^{200} \big(\text{cor}(X_i, Y)\big)^2$$

where X_i is the vector of gene expression levels for gene #i, and Y is the binary vector indicating if the patient takes statins.

Next, we will try the following two variants of a permutation test, to compute a p-value based on T.

(a) For each $i = 1, \ldots, n$ — permute the vector X_i at random and compute $r_i^{\text{perm}} = \text{cor}(X_i^{\text{perm}}, Y)$. Then compute $T^{\text{perm}} = \sum_{i=1}^{200} (r_i^{\text{perm}})^2$.

(b) Permute the vector Y at random. Then for each $i = 1, \ldots, n$, compute $r_i^{\text{perm}} = \text{cor}(X_i, Y^{\text{perm}})$ (use the same vector Y^{perm} for all the i's). Again, compute $T^{\text{perm}} = \sum_{i=1}^{200} (r_i^{\text{perm}})^2$.

For each scheme, run it with 500 permutations and compute a p-value for the observed value of T.

1 Which scheme is a more valid way of testing the global null hypothesis (i.e., testing if there is any association between statin use and gene expression levels)? To explore this more, let's run the following simulation. First,

2 sample the vector Y (i.e., statins or no statins) randomly. Then using this simulated Y vector, run procedure (a) to produce a p-value via a permutation test, and do the same for procedure (b) (for each procedure, you can do a smaller number of permutations, like 50 or 100, to save time).

3 Now repeat this experiment a large number of times (at least 300 to see a clear trend — be aware this may take some time to run depending on how you implement it). Create a histogram of the p-values from scheme (a), and another for scheme (b). If the permutation test is valid, you should see (approximately) a uniform distribution, since we are simulating the procedure with a random vector Y.

Explain what you see, and which scheme is valid or invalid and why.