Math 283, Fall 2023, Prof. Tesler
Homework #7, Due Wednesday November 22, 2023

Introduction

In this assignment, you'll examine a dataset with differential expression data from a microarray experiment and compare two hypothesis tests: paired t vs. Wilcoxon.

An overview of the dataset is given below. The problems for you to do are on page 2, and useful Matlab and R commands are on pages 3–4. References and optional extra details for those interested are on page 5.

Formatting your submission

Please submit your assignment to Gradescope as a single, legible, human-readable document in PDF format. Turn in a program listing as well, either at the end of the assignment or included with each problem. On the listing, we are not looking for programming style, just for correctness.

Please make sure that the pages of your submission are close to regular size, such as US Letter, A4, or your screen size. If your software makes extremely oversized pages, you will need to figure out how to paginate properly, or they may become illegible to the grader.

To create plots to embed in your document, see "Saving high-quality plots to a file" at the bottom of the Homework page on Canvas. In Matlab and R, you can output plots in PDF format (and other formats); this will generally be higher-quality than screenshots or using a camera to take photos of your screen.

Data

The dataset is from Perou et al (2000); details are at the end of this assignment for those interested, and some photos are in the class slides. The microarrays have 9216 probes per slide. A total of 84 slides were made with various types of breast cancer and controls. The data in this assignment is from 18 of those 84 microarray slides: 9 patients with "luminal-like ER+ tumors" before and after treatment. You'll test for which genes may have been differentially expressed due to the treatment.

The file `hw7data.txt` posted with this homework has 4526 rows (lines) and 19 columns (separated by tabs), which are a subset of the rows and columns in the published dataset:

- Each row corresponds to a spot on the microarray. The published dataset included rows for all 9216 spots, but for this assignment, rows with any bad readings in the selected microarrays were deleted.
- Column 1 is the spot ID number on the microarray slide, between 1 and 9216.
- Columns 2 through 10 are 9 patients before treatment.
- Columns 11 through 19 are those same patients, in the same order, after treatment.
- The numbers in columns 2–19 are normalized log intensity ratios between the red and green channels.
- The hypothesis tests compare means/medians of "before" vs. "after" samples to determine if differential expression occurred.
- If your program is slow on thousands of rows, then make a smaller file (say, the first 100 rows) to test on, and then run it on the full file once it's working.

Problems

These are the problems to do and turn in, H-701 through H-705.

Preparation: Your program should read the file and compute the following, which will be used to answer the subsequent questions. (There is no answer to submit for "Preparation" besides including it in the program listing and making use of it to answer the other problems.)

(a) For the paired t-test, compute the test statistic t and the P-value for all rows. Store these in arrays for the subsequent steps.

(b) Do the same but for the Wilcoxon signed rank test statistic, W (and its P-value), instead of the paired t-test.

Problem H-701.

(a) For the paired t-test, report the following for the top 10 most significant spots (most significant means smallest P-value): ID number, P-value, and test statistic.

(b) Do the same, but for W (and its P-value) instead of t.

The problems below should use all rows of the data set, not just the top 10 spots reported in H-701.

Problem H-702. The data set is not annotated with what spots truly are or are not differentially expressed, so we don't actually know which genes were correctly classified by these tests. For the purposes of this question, assume the paired t-test at $\alpha = 5\%$ gives the "true state of nature" for each spot. (We are pretending it is a gold standard test.) **Based on that assumption,** compute the following for the Wilcoxon test at $\alpha = 5\%$: (i) Confusion matrix, (ii) Type I error rate, (iii) Type II error rate, and (iv) False discovery rate (FDR). Do not use the Bonferroni or Šidák corrections.

Problem H-703.

(a) Make a histogram of the test statistic t. Comment on whether it shows H_0 is consistent or inconsistent with most of the data.

(b) Do the same, but for W instead of t.

Problem H-704.

(a) Make a histogram of the P-values for the paired t-test. Comment on whether it shows H_0 is consistent or inconsistent with most of the data.

(b) Do the same but for P-values of the Wilcoxon signed rank test instead of the paired t-test.

Problem H-705. Make a scatter plot comparing the P-values for the two tests: for each spot, plot (x, y), where x is the P-value for the paired t-test and y is the P-value for the Wilcoxon test. A scatter plot is just the points; do not connect the points by lines. (There is no natural order to put the points into, so connecting lines don't make sense and would just clutter the plot.) Comment on the consistency of the two tests.

Matlab and R tips

Matlab and R documentation links are on the class website Software page. If you're using something else like BOOST or Python, you'll need to find similar functions. Although these problems may seem amenable to using a spreadsheet, EXCEL and Google Sheets don't have all the statistical functions.

In the GUI versions of Matlab and R, you need to set the current directory (shown at the top) to where your files are located (use your directory name instead of the example shown):

Matlab: `cd('/home/user123/homework')`
R: `setwd('/home/user123/homework')`

Loading the file into a matrix variable named `data`:

Matlab:
```
data = load('hw7data.txt');
```

R:
```
data = read.delim('hw7data.txt',header=FALSE);
```
or:
```
data = read.table('hw7data.txt',header=FALSE,sep='');
```

Dimensions of matrix: Number of rows is $nr = 4526$. Number of columns is $nc = 19$.

Matlab:
```
[nr,nc] = dim(data);
```

R:
```
nr = nrow(data); nc = ncol(data);
```
or: `dim(data)` to get both

Using row r of the matrix:

	Matlab	**R**
Row r	`data(r,:)`	`data[r,]`
Spot ID	`spotID = data(r,1);`	`spotID = data[r,1];`
Vector of all patients before	`before = data(r,[2:10]);`	`before = as.double(data[r,2:10]);`
Vector of all patients after	`after = data(r,[11:19]);`	`after = as.double(data[r,11:19]);`
Patient $j = 1, \ldots, 9$ before	`before_j = data(r,j+1);`	`before_j = data[r,j+1];`
Patient $j = 1, \ldots, 9$ after	`after_j = data(r,j+10);`	`after_j = data[r,j+10];`

Paired t-test: H_0: $\mu_X = \mu_Y$ vs. H_1: $\mu_X \neq \mu_Y$.

For hypothesis test P-value p and test statistic t:

Matlab:
```
[h, p, ci, stats] = ttest(before, after); % ttest for paired t, ttest2 for two-sample
t = stats.tstat;
```

R:
```
test = t.test(before, after, paired=TRUE);
p = test$p.value;
t = test$statistic;
```

Wilcoxon signed rank test: H_0: Medians equal vs. H_1: Medians not equal

For hypothesis test P-value p and test statistic W:

Matlab:
```
[p,h,stats] = signrank(after-before); % Note p,h order is different than in ttest
W = stats.signedrank;
```

R:
```
test = wilcox.test(before,after,paired=TRUE);
p = test$p.value;
W = test$statistic;
```

SciPy: See note on class website

Listing the most significant *P*-values: Suppose your results are in a 2-dimensional array `results`, where row r has marker ID, P-value, test statistic, etc. To get a new array `sorted` where the rows are put in order smallest to largest by P-value (column 2), use

Matlab:
```
sorted = sortrows(result,2);
```

R:
```
perm = order(results[,2]); # permutation that puts column 2 into order
sorted = results[perm,];   # re-order all rows into that order
```

Confusion matrix and error rates: A straightforward way to get the counts is with a loop. Here are vectorized shortcuts. If `x` and `y` are equal length vectors,

Matlab:	`x <= .8`	is a vector in which entry i is 1 if $x(i) \leq .8$, or 0 if not
	`find(x <= .8)`	is a list of the indicies i for which $x(i) \leq .8$
	`length(find(x <= .8))`	counts the number of elements in `x` that are $\leq .8$
	`length(find(x <= .8 & x >= .4))`	counts the number of elements in `x` between .4 and .8
	`length(find(x <= .8 & y >= .4))`	counts the number of i's with $x(i) \leq .8$ and $y(i) \geq .4$.
R:	`x <= .8`	is a vector with entry i TRUE if $x(i) \leq .8$, or FALSE if not
	`which(x <= .8)`	is a list of the indicies i for which $x(i) \leq .8$
	`length(which(x <= .8))`	counts the number of elements in `x` that are $\leq .8$
	`length(which(x <= .8 & x >= .4))`	counts the number of elements in `x` between .4 and .8
	`length(which(x <= .8 & y >= .4))`	counts the number of i's with $x(i) \leq .8$ and $y(i) \geq .4$.

Printing results:

Matlab: Some results can be output by omitting the semicolon after the command. For C-style formatting, use `fprintf(1,'%d %d %f\n',i,j,x);` (also see `sprintf`). Variables can be printed with default formatting using `disp(x)`.

R: At the command line, variables can be output by typing just the variable name, but this does not work in a program. In a program, use `print`, `cat`, or `write`. For C-style formatting, use `cat(sprintf('%d %d %f\n',i,j,x));` (also see `format`).

Plots:

Matlab: Histograms can be made with `hist`, `histc`, or `bar`.

Scatter plots can be made with `plot(x,y,'.')` ('.' gives a dot marker shape; see the `plot` documentation for other marker shapes). `scatterplot` is another way to make scatter plots.

Plots can be saved/printed in various formats through the GUI, and also through the `print` command:

```
print('-depsc','file.eps');
print('-dpdf','file.pdf');
```

R: Histograms can be made with `hist` and `barplot`.

Scatter plots can be made with `plot(x,y)`. To shrink the dots, use `plot(x,y,cex=.1)` (cex is a scaling factor for the plot symbol).

Plots can be saved/printed in various formats through the GUI, and also through the `dev.print` command:

```
dev.print(postscript,'file.eps',horizontal=FALSE,onefile=FALSE,paper='special');
dev.print(pdf,'file.pdf');
```

Further details on dataset

This page has references for this dataset. You can skip this page if you want.

This data is for this article:

Perou et al., *Molecular portraits of human breast tumours,* *Nature* **406**, 747–752 (2000)
`https://dx.doi.org/10.1038/35021093`

This data was downloaded from GEO ("Gene Expression Omnibus").

`https://www.ncbi.nlm.nih.gov/geo/`

Search for "GSE61". Select "Molecular portraits of human breast tumors":

`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61`

Under "Download family", get the "Series Matrix File(s)" for the data set, and "SOFT formatted family file(s)" for a description of the microarray platform (description of the probes at each of the 9216 spots). Another description of them is also available on the "GPL180" platform link:

`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL180`

The "Matrix" file is called `GSE61_series_matrix.txt`. (There are also detailed files on each tissue sample available separately on the page, with the details of microarray image normalization; the "Matrix" file consists of the "UNF_VALUE" column in those files.) The last section of `GSE61_series_matrix.txt` is a large table. The file `hw7data.txt` was made by extracting the following columns from that large table. "BC107B-**BE**" / "BC107A-**AF**" are before/after treatment samples for "breast cancer patient 107."

Column # in `hw7data.txt`	Description in `GSE61_series_matrix.txt`	Column # in `hw7data.txt`	Description in `GSE61_series_matrix.txt`
1	ID_REF		
2	BC107B-BE	11	BC107A-AF
3	BC110B-BE	12	BC110A-AF
4	BC112B-BE	13	BC112A-AF
5	BC115B-BE	13	BC115A-AF
6	BC118B-BE	14	BC118A-AF
7	BC124A-BE	15	BC124B-AF
8	BC206A-BE	16	BC206B-AF
9	BC708B-BE	17	BC708A-AF
10	BC710A-BE	18	BC710B-AF

The full dataset record has links to more raw data on those, including before normalization.

Images of these microarray slides are available at a different website:

- `https://puma.princeton.edu/cgi-bin/publication/viewPublication.pl?pub_no=38`
- Hit the button "Display Data".
- Locate experiment "BC107B-BE" and follow the links to get microarray slide images, numeric data, etc. The 4×4 microarray icon with a black background (not the one with the white background) will give you a high resolution picture of the microarray.

Photos of the tissue samples: `https://tinyurl.com/4mvstm43`