MTH 3240 Lab 6

♣♣♣ Due Thu., Mar. 5 ♣♣♣

1 Part A: Two-Sample t Test (Cont'd from Lab 5)

1.1 Lakes' DOC Data Set

A ***BACI study*** was carried out to assess the impact of logging on the **dissolved organic carbon** (**DOC**) concentrations (μg/L) of nearby lakes.

DOC was measured on each of **9** lakes *before* and *after* their watersheds were **logged**, and on each of **13** *control* lakes whose watersheds remained **unlogged** over the same time period.

The table below shows the *before-* and *after-logging* **DOC** for each of the lakes along with the ***changes*** (after - before) in DOC.

DOC for Lakes in Unlogged Watersheds

Lake	Before	After	Change
AB220	4.1	4.4	0.3
AB35	5.4	5.6	0.2
CSL2	5.7	6.3	0.6
CSL5	10.5	10.6	0.1
DA4	8.5	9.0	0.5
DA9	9.7	10.4	0.7
DF4	9.2	9.4	0.2
K2	8.4	9.2	0.8
N35	4.4	4.1	-0.3
N43	9.5	9.4	-0.1
N55	5.0	5.2	0.2
N70	5.8	5.5	-0.3
N89	3.7	4.1	0.4

DOC for Lakes in Logged Watersheds

Lake	Before	After	Change
DF2	7.2	8.3	1.1
DF5	8.0	8.1	0.1
DF7	11.0	12.5	1.5
DF9	12.7	14.9	2.2
K1	6.5	6.7	0.2
K3	6.1	6.6	0.5
K4	7.3	8.2	0.9
K8	9.9	11.1	1.2
P109	6.1	7.0	0.9

One way to analyze the data is to compare the ***changes*** in **DOC** at the **logged** lakes to the ***changes*** at the **unlogged** lakes. *If logging had an effect*, we should see *larger changes* at the **logged** lakes than at the **unlogged** ones.

1. Here are the **DOC** ***changes*** for each of the **unlogged** and **logged** lakes (from the tables above):

UnloggedChange	0.3, 0.2, 0.6, 0.1, 0.5, 0.7, 0.2, 0.8, -0.3, -0.1, 0.2, -0.3, 0.4
LoggedChange	1.1, 0.1, 1.5, 2.2, 0.2, 0.5, 0.9, 1.2, 0.9

   Use `c()` to read each data set into its own vector.

2. Make side-by-side ***boxplots*** comparing the changes in DOC for lakes in **logged** and **unlogged** watersheds:

```
boxplot(UnloggedChange, LoggedChange,
        col = "lightblue",
        main = "Boxplots of DOC Changes",
        names = c("Unlogged", "Logged"),
        ylab = "DOC Change")
```

3. Make ***normal probability plots*** of the ***changes*** in **DOC** for lakes in **logged** and **unlogged** watersheds:

```
qqnorm(UnloggedChange, pch = 19)
qqline(UnloggedChange, col = "blue")
```

```
qqnorm(LoggedChange, pch = 19)
qqline(LoggedChange, col = "blue")
```

4. Recall that the `t.test()` function, when passed **two** vectors, will carry out a ***two-sample t test*** (and compute a ***95% two-sample t confidence interval***) for two population means $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$. Among its arguments are:

`x`	a data vector.
`y`	another data vector.
`alternative`	the direction for the alternative hypothesis, one of `"two.sided"`, `"less"`, or `"greater"`.
`mu`	the null hypothesized value for the unknown difference between population means, with default value 0.
`conf.level`	the confidence level for a confidence interval for the unknown difference between population means, with default value 0.95.

Use `t.test()` to carry out a ***two-sample t test*** of

$$\begin{aligned} H_0 : \mu_x - \mu_y &= 0 \\ H_a : \mu_x - \mu_y &< 0 \end{aligned}$$

(where $\boldsymbol{\mu_x}$ is the population mean **DOC** ***change*** for **unlogged** lakes and $\boldsymbol{\mu_x}$ is the population mean for **logged** lakes) to decide if the ***change*** in **DOC** was **greater** in **logged** lakes than in **unlogged** ones.

For example, you could type:

```
t.test(UnloggedChange, LoggedChange, alternative = "less")
```

2 Part B: Rank Sum Test

2.1 Suspended Solids Data Set

In a ***BACI study*** of the effects of a clear-cutting operation on a nearby stream's water quality, the **suspended solids** (mg/L) were measured **upstream** and **downstream** on each of **3** days *before* logging took place and on each of **17** days *after* it took place. The data are below.

Suspended Solids

Date	Period	Upstream	Downstream	Difference
04/03/98	Before	30.48	26.24	4.24
04/09/98	Before	20.20	17.84	2.36
04/15/98	Before	4.56	3.76	0.80
06/03/98	After	8.40	9.28	-0.88
06/09/98	After	3.24	0.40	2.84
06/15/98	After	4.30	3.04	1.26
06/22/98	After	1.96	1.92	0.04
06/29/98	After	3.72	3.18	0.54
07/09/98	After	1.34	0.32	1.02
07/15/98	After	1.72	1.86	-0.14
07/23/98	After	9.36	11.08	-1.72
07/31/98	After	21.06	2.50	18.56
08/17/98	After	0.88	1.12	-0.24
08/31/98	After	1703.50	484.80	1218.70
09/15/98	After	2.38	2.64	-0.26
10/08/98	After	1.08	1.36	-0.28
11/11/98	After	4.40	4.00	0.40
11/26/98	After	3.64	3.90	-0.26
04/12/98	After	23.26	20.42	2.84
04/29/98	After	1.00	1.40	-0.40

One way to analyze the data is to compare the **daily *differences*** (**upstream minus downstream**) from **before** the impact event to **daily *differences* after**. *If the impact event had an effect*, we should see a *change* in the *differences* from **before** to **after** the event.

1. Here are the **daily *differences*** (**upstream minus downstream**) for **before** and **after** the event (from the table above):

   BeforeDiff 4.24, 2.36, 0.80
   AfterDiff -0.88, 2.84, 1.26, 0.04, 0.54, 1.02, -0.14, -1.72, 18.56, -0.24, 1218.70, -0.26, -0.28, 0.40, -0.26, 2.84, -0.40

   Use `c()` to read each data set into its own vector.

2. Notice the extreme **outlier** in the difference data. Because the *two-sample t test* is based on sample means, which are affected by outliers, it wouldn't make sense to use a *two-sample t test* to compare the before and after differences.

   Instead, we'll use a ***rank sum test***. The **ranks** of a data set aren't affected by outliers.

The function `wilcox.test()`, when passed two vectors, will carry out a ***rank sum test*** for two population means μ_x and μ_y. Among its arguments are:

`x`	a data vector.
`y`	another data vector.
`alternative`	the direction for the alternative hypothesis, one of `"two.sided"`, `"less"`, or `"greater"`.
`mu`	the null hypothesized value for the unknown difference between population means, with default value 0.

Use `wilcox.test()` to carry out a ***rank sum test*** of

$$\begin{aligned} H_0 : \mu_x - \mu_y &= 0 \\ H_a : \mu_x - \mu_y &\neq 0 \end{aligned}$$

where $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$ are the true **population mean *differences*** in suspended solids concentrations **before** and **after** logging, respectively. For example, you could type:

```
wilcox.test(BeforeDiff, AfterDiff)
```

3 Part B: Deciding Which Test to Use

The ***two-sample t test*** requires that either the samples are from **normal** populations **or** the sample sizes $\boldsymbol{n_x}$ and $\boldsymbol{n_y}$ are both **large** (e.g. larger than about 30).

We check the normality assumption using a ***histogram*** or a ***normal probability plot***, i.e.

```
hist(x)
```

or

```
qqnorm(x)
qqline(x)
```

If the normality assumption **isn't** met (and n_x and n_y **aren't** large), the ***rank sum test*** is more appropriate.

3.1 Depleted Uranium Data Set

Nuclear reactors require uranium that is enriched beyond what occurs naturally. Depleted uranium (DU) is the remaining uranium after the enriched portion has been removed. It is an extremely dense heavy metal that is both toxic and radioactive.

DU particles left in the soil present a hazard to humans who inhale them after they are resuspended by the wind or by human activity. DU also presents a risk if it enters the drinking water or food chain.

Random samples of soil specimens were drawn from **two soil piles**, **six** specimens from the **first pile** and **nine** from the **second**, and the **DU** activity (pCi/g) measured in each specimen. The table below shows the data.

DU in Soil

Soil Pile B	Soil Pile D
57.3	26.4
19.9	42.0
20.5	17.1
27.6	21.7
15.3	11.0
21.4	17.0
	11.8
	10.8
	9.7

Here are the data in a more convenient format:

SoilPileB 57.3, 19.9, 20.5, 27.6, 15.3, 21.4

SoilPileD 26.4, 42.0, 17.1, 21.7, 11.0, 17.0, 11.8, 10.8, 9.7

We're interested in deciding whether there is **any significant difference** in the true mean DU levels for the two soil piles.

1. Use `c()` to create two data vectors, one containing the **SoilPileB** DU and the other **SoilPileD**.
2. Use `hist()` and `qqnorm()` followed by `qqline()` to check the **normality** assumption to decide which hypothesis test should be used (**two-sample *t*** or **rank sum**). You ***don't*** need to carry out the test, for example by typing:

```
hist(SoilPileB, col = "blue")
hist(SoilPileD, col = "blue")
```

```
qqnorm(SoilPileB, pch = 19)
qqline(SoilPileB, col = "blue")
```

```
qqnorm(SoilPileD, pch = 19)
qqline(SoilPileD, col = "blue")
```

3.2 Soil Mercury Data Set

The South Florida Ecosystem Assessment is a long-term research, monitoring and assessment project in the Florida Everglades conducted by the United States Environmental Protection Agency Region 4 in cooperation with various other government agencies and research laboratories.

At each of **25** sites in and around the Florida Everglades, numerous environmental and ecological variables were recorded. The table below contains a small subset of variables that were recorded. Below is a description of the variables:

`STA_ID` sampling station name
`SOIL_TYPE` description of soil type (Marl/Peat or Peat/Layers)
`THGSDE` Total Mercury in soil, μg/kg

Soil Mercury

STA_ID	SOIL_TYPE	THGSDE
M045	Marl/Peat	50
M112	Marl/Peat	18
M532	Marl/Peat	25
M580	Marl/Peat	40
M581	Marl/Peat	110
M662	Marl/Peat	64
M556	Marl/Peat	82
M704	Peat/Layers	130
M731	Peat/Layers	54
M051	Peat/Layers	120
M070	Peat/Layers	96
M071	Peat/Layers	150
M074	Peat/Layers	120
M079	Peat/Layers	100
M085	Peat/Layers	100
M087	Peat/Layers	120
M205	Peat/Layers	87
M219	Peat/Layers	57
M220	Peat/Layers	110
M228	Peat/Layers	110
M297	Peat/Layers	170
M303	Peat/Layers	190
M444	Peat/Layers	74
M462	Peat/Layers	110
M475	Peat/Layers	79

Here are the data in a more convenient format:

MarlPeat 50, 18, 25, 40, 110, 64, 82

PeatLayers 130, 54, 120, 96, 150, 120, 100, 100, 120, 87, 57, 110, 110, 170, 190, 74, 110, 79

We're interested in testing whether mean mercury (Hg) levels in the two types of soil, **Marl/Peat** and **Peat/Layers**, **differ** significantly.

1. Use `c()` to create two data vectors, one containing the **Marl/Peat** Hg and the other **Peat/Layers** Hg.
2. Use `hist()` and `qqnorm()` followed by `qqline()` to check the **normality** assumption to decide which hypothesis test should be used (**two-sample *t*** or **rank sum**). You ***don't*** need to carry out the test.

3.3 Sludge Copper Data Set

An experiment was carried out to decide if there's **any difference** between results obtained using **two methods** for measuring copper (Cu) in wastewater sludge.

Fourteen specimens of sludge were randomly allocated to two groups of **seven** specimens each. Cu leachate concentrations (μg/L) were measured using the Equilibrium Leach Test (**ELT**) in the first group, and the Toxicity Characteristics Leaching Procedure (**TCLP**) in the second. The data are below.

Cu in Sludge	
ELT	TCLP
126.1	156.0
148.5	206.0
208.8	239.0
87.2	175.0
82.6	310.0
94.2	359.6
115.4	199.5

Here are the data in a more convenient format:

ELT 126.1, 148.5, 208.8, 87.2, 82.6, 94.2, 115.4

TCLP 156.0, 206.0, 239.0, 175.0, 310.0, 359.6, 199.5

We want to carry out a test to decide whether there is **any difference** between the **two methods** for measuring Cu.

1. Use `c()` to create two data vectors, one containing the **ELT** Cu measurements and the other the **TCLP** ones.
2. Use `hist()` and `qqnorm()` followed by `qqline()` to check the **normality** assumption to decide which hypothesis test should be used (**two-sample *t*** or **rank sum**). You ***don't*** need to carry out the test.

3.4 Dissolved Organic Carbon Data Set

A researcher took water samples in a forest with a stream running through it. Water was collected from **streams** (surface water) on each of **31** days and **groundwater** was collected from organic soil on each of **44** (different) days. Each water specimen was analyzed for the dissolved organic carbon (**DOC**, mg/L). The data are below.

DOC in Water

Soil	Stream
22.74	10.83
29.80	8.74
27.10	9.20
16.51	8.12
6.51	7.60
8.81	6.30
5.29	6.68
20.46	7.34
14.90	9.52
14.86	11.94
15.91	12.40
15.35	10.30
9.72	10.48
19.80	12.88
14.86	19.01
8.09	19.19
17.90	13.14
18.30	12.51
5.20	15.76
11.90	10.96
14.00	19.78
7.40	20.56
17.50	17.93
10.30	14.28
11.40	13.11
5.30	12.27
15.72	16.47
20.46	15.03
16.87	14.53
15.42	12.04
22.49	16.82
	10.70
	16.00
	20.70
	13.70
	16.12
	15.77
	6.40
	14.19
	13.89
	8.69
	10.46
	16.23
	15.43

Here are the data in a more convenient format:

Ground 22.74, 29.80, 27.10, 16.51, 6.51, 8.81, 5.29, 20.46, 14.90, 14.86, 15.91, 15.35, 9.72, 19.80, 14.86, 8.09, 17.90, 18.30, 5.20, 11.90, 14.00, 7.40, 17.50, 10.30, 11.40, 5.30, 15.72, 20.46, 16.87, 15.42, 22.49

Stream 10.83, 8.74, 9.20, 8.12, 7.60, 6.30, 6.68, 7.34, 9.52, 11.94, 12.40, 10.30, 10.48, 12.88, 19.01, 19.19, 13.14, 12.51, 15.76, 10.96, 19.78, 20.56, 17.93, 14.28, 13.11, 12.27, 16.47, 15.03, 14.53, 12.04, 16.82, 10.70, 16.00, 20.70, 13.70, 16.12, 15.77, 6.40, 14.19, 13.89, 8.69, 10.46, 16.23, 15.43

We want to perform a test to decide if there's **any significant difference** between the mean **DOC** concentrations in the **two types** of **water**.

1. Use `c()` to create two data vectors, one containing the **Ground** DOC concentrations and the other the **Stream** ones.

2. Use `hist()` and `qqnorm()` followed by `qqline()` to check the **normality** assumption to decide which hypothesis test should be used (**two-sample *t*** or **rank sum**). You ***don't*** need to carry out the test.