MTH 3240 Lab 7

Due Thu., Mar. 12

1 Part A: Paired t Test and Confidence Interval

1.1 Clear-Cutting and Water Quality Data Set

In an impact assessment study of the effects of forest clear-cutting on water quality in an adjacent stream, several hydrological variables were measured on each of $n = 11$ days **upstream** and **downstream** of a clear-cutting operation.

The table below shows the **nitrate** concentrations (mg/L) and their ***differences***.

Nitrate Concentration

Date	Upstream	Downstream	Difference
08/15/97	1147.4	995.3	152.1
08/18/97	1412.2	1303.6	108.6
08/31/97	1613.9	1923.3	-309.4
09/18/97	763.3	747.8	15.5
11/04/97	1031.4	1082.9	-51.5
11/07/97	1093.2	1938.7	-845.5
02/27/98	390.8	338.8	52.0
07/14/98	909.8	776.8	133.0
08/25/98	1033.0	676.8	356.2
09/30/98	897.5	1291.0	-393.5
10/29/98	2314.0	1232.9	1081.1

We consider these samples (**upstream** and **downstream** nitrate measurements) to be ***paired by day***. Here are the data in a more convenient form:

Upstream	1147.4,	1412.2,	1613.9,	763.3,	1031.4,	1093.2,	390.8,	909.8,	1033.0,	897.5,	2314.0
Downstream	995.3,	1303.6,	1923.3,	747.8,	1082.9,	1938.7,	338.8,	776.8,	676.8,	1291.0,	1232.9

1. Use `c()` to create the two vectors `Upstream` and `Downstream`.
2. Type:

```
boxplot(Upstream, Downstream, col = "lightblue", names = c("Upstream", "Downstream"))
```

   to make side-by-side ***boxplots*** of the **upstream** and **downstream** nitrate measurements.
3. The `t.test()` function has an optional argument `paired` that allows us to specify that the samples are ***paired*** and that we want to carry out a ***paired t test*** on two data vectors.

   Carry out a ***paired*** t ***test*** of

$$H_0 : \mu_x - \mu_y = 0 \qquad \text{(or equivalently } H_0 : \mu_d = 0)$$
$$H_a : \mu_x - \mu_y \neq 0 \qquad \text{(or equivalently } H_a : \mu_d \neq 0)$$

where μ_x and μ_y are the true (unknown) mean *counts* using the **upstream** and **downstream** nitrate concentrations, respectively (and μ_d is the true unknown mean *difference*, **upstream** minus **downstream**), by typing:

```
t.test(Upstream, Downstream, mu = 0, paired = TRUE, alternative = "two.sided")
```

to decide if there's any **statistically significant difference** between the **upstream** and **downstream** nitrate concentrations.

4. The output from `t.test()` includes the ***95% paired t confidence interval***

$$\bar{D} \pm t_{0.025,n-1} \frac{S_d}{\sqrt{n}}$$

for the true difference $\mu_x - \mu_y$ in means (or equivalently for the true **mean difference** μ_d), with $\bar{D}$ and S_d being the sample mean and standard deviation of the **differences**. Look at the values of the endpoints of the confidence interval.

5. Create a vector containing the ***differences*** (**upstream** minus **downstream** nitrate) named `Diff` by typing:

```
Diff <- Upstream - Downstream
```

Then use `hist()` to make a **histogram** of the `Diff`erences. Does the histogram support the assumption that the **differences** are from a **normal** distribution (as required for the ***paired t test***)?

6. The ***paired t test*** is just a ***one-sample t test*** based on the **differences**.

Verify that you get the same results by using `t.test()` to carry out a ***one-sample t test*** on `Diff`:

```
t.test(Diff, mu = 0, alternative = "two.sided")
```

2 Part B: Signed Rank Test

2.1 Hospital Bacteria Failure Rates Data

The ***signed rank test*** is a **nonparametric** alternative to the ***paired t test*** that **doesn't** require the **normality** assumption for the differences.

A study was carried out to investigate the effectiveness of routine cleaning on reducing bacteria levels in a hospital.

For each of several surfaces in the hospital, the *failure rate* **before** and **after** cleaning was recorded. The **failure rate** is defined as the percentage of times that the bacteria level fails to meet standard specifications when tested.

The table below shows the failure rates (percent) before and after cleaning for ten surfaces in the bedroom and treatment room of one of the hospital's wards.

	Failure Rate (%)	
Surface	Before Cleaning	After Cleaning
Worktop	90	93
Treatment Room Tap Handle	60	64
Treatment Trolley	80	86
Door Handle	60	43
Fridge Handle	70	57
Treatment Room Bin Lid	80	50
Ward 4 Bed Tap Handle	90	86
Sink	90	78
Bed Rail	80	71
Ward 4 Bed Bin Lid	70	64

Here are the data in a more convenient form:

Before 90, 60, 80, 60, 70, 80, 90, 90, 80, 70

After 93, 64, 86, 43, 57, 50, 86, 78, 71, 64

1. Use `c()` to create the two vectors `Before` and `After`.

2. Use `boxplot()` to to make side-by-side ***boxplots*** of the **before** and **after** failure rates. (See step 2 of Part A if you forget the R command).

3. The function `wilcox.test()` has an optional argument `paired` that allows us to specify that the samples are ***paired*** and that we want R to carry out a ***signed rank test*** on two data vectors.

   Carry out a ***signed rank test*** of

   $$\begin{aligned} H_0 : \mu_x - \mu_y &= 0 \qquad \text{(or equivalently } H_0 : \mu_d = 0\text{)} \\ H_a : \mu_x - \mu_y &> 0 \qquad \text{(or equivalently } H_a : \mu_d \neq 0\text{)} \end{aligned}$$

   where μ_x and μ_y are the true (unknown) mean *failure rates* **before** and **after** cleaning, respectively (and μ_d is the true unknown mean *difference*, **before** minus **after**), by typing:

   ```
   wilcox.test(Before, After, paired = TRUE, alternative = "greater")
   ```

   to decide whether the routine cleaning reduces the failure rate. Use a level of significance level $\alpha = 0.05$.