Statistical Inference using R

Prepare data set. You can prepare the data by downloading it from `brick.txt`.

"Brick Weights"

1.099

1.103

1.154

...

...

...

1.206

1.066

1.132

The first line of the file, a part of which is shown above, is called a **header**, represents a **variable name** of the data set. To give variable names properly in the first line of the file, you should put it in the double-quotation marks (").

Read data file. The column data directly below the variable name are the actual data which should start from the second line. We now read the file "`brick.txt`" into **data frame** in the R programming by using `read.table()` function as follows:

```
> BrickData <- read.table("brick.txt", header=T)
```

NOTE: "`<-`" in R programming is supposed to play a role of "`=`" as in many other computer languages. Also, R programming environment is interactive, known as an "interpretor." For example, if you type "`x <- 3.14`", then you can find that "`x`" has the value 3.14 by simply typing "`x`" (then return).

Declare data frame in use. Before doing anything else, we have to declare the data frame *BrickData* in use by using `attach` function:

```
> attach(BrickData)
```

Now we can use the variable *Brick.Weights*. Here, the spaces (' ') in "`brick.txt`" are replaced with the periods ('.') inside the R programming. To see what variables are available, use `names` function as follows:

```
 > names(BrickData)
 [1] "Brick.Weights"
```

Use variable name in data frame. To calculate sample statistics, call the `summary()` com-

mand with variable name *Brick.Weights*.

```
> summary(Brick.Weights)
Min.  1st Qu.  Median Mean 3rd Qu.  Max.
0.874 1.083 1.110 1.111 1.140 1.257
```

Note that the sample size cannot be found from the `summary()` command. To find out the data size, use the `length()` command.

```
> length(Brick.Weights)
[1] 125
```

Use graphics for data exploration. To draw the histogram, we can use `hist` function as follows:

```
> hist(Brick.Weights)
```

We can change the "number of bands" by assigning the number to `breaks` in `hist` function as follows:

```
> hist(Brick.Weights, breaks=10)
```

NOTE: "`breaks=10`" is an optional argument. The argument is not necessarily specified, and it is set automatically if it is not given.

Use graphics, continued. We can also add a "color" and a "main title" as follows:

```
> hist(Brick.Weights, breaks=10, col="gray",
  main="Brick Weights in kg")
```

We can draw the boxplot by using `boxplot` function:

```
> boxplot(Brick.Weights, col="green", main="Brick
  Weights in kg")
```

QQ normal plot. The *quantile-quantile normal plot* (QQ normal plot) is one of the graphical methods to assess a fit of the data to a normal distribution.

```
> qqnorm(Brick.Weights, datax=T)
```

The values at the x-axis shows data (specified by `datax=T`), and the values at the y-axis correspond the quantiles from the standard normal distribution. For example, the values between -1.0 and 1.0 at the vertical axis consist of approximately 68% of the entire values, which corresponds to the sample data between $\mu - \sigma$ and $\mu + \sigma$ (at the x-axis) if the data are normally distributed. Thus, the straight line of plots indicate a fit to a normal distribution.

```
> qqline(Brick.Weights, datax=T,col='green')
```

Statistical inference. The hypothesis testing for population mean μ can be done by the `t.test` command. The `t.test` command calculate the p-value accordingly as

(a) $H_A: \mu \neq \mu_0$ (if `alternative="two.sided"` is specified);

(b) $H_A: \mu > \mu_0$ (if `alternative="greater"` is specified);

(c) $H_A: \mu < \mu_0$ (if `alternative="less"` is specified).

For example, if we construct the hypothesis testing problem

$$H_0: \mu = 1.1 \quad \text{versus} \quad H_A: \mu > 1.1$$

then the `t.test` command must include the options `mu=1.1` and `alternative="greater"`. The `t.test` command will return the following output on the display.

Statistical inference, output.

```
> t.test(Brick.Weights, mu=1.1, alternative="greater")
```

t = 2.2227, df = 124, p-value = 0.01402

alternative hypothesis: true mean is greater than 1.1

95 percent confidence interval:

1.102680 Inf

sample estimates:

mean of x

1.110536

The above result indicates that (i) t-statistic is 2.2227, (ii) p-value is 0.01402, and (iii) 95% one-sided confidence interval is $(1.102680, \infty)$. Then, we can reject H_0 with significance level 0.05, but we cannot reject H_0 with significance level 0.01. Thus, there is some evidence that the average brick weight is more than 1.1, but the evidence is modestly significant.

Statistical inference, continued. When you want the 99% two-sided confidence interval instead of the default 95% one-sided confidence interval, we can use the option `conf.level=0.99` together with `alternative="two.sided"` in the `t.test` command.

```
> t.test(Brick.Weights, mu=1.1, alternative="two.sided", conf.level=0.99)
```

.....

.....

99 percent confidence interval:

1.098135 1.122937

This gave the 99% two-sided confidence interval $(1.098135, 1.122937)$ for the population mean of brick weight.

Inference on Paired Data

Paired experiment. A researcher is interested in how a new class of drug treating a patient actually affects the patient's heart rate reduction. The pairs of heart rate reduction

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

of n participants under the standard drug and after taking the new drug are measured. The data file "`heart.csv`" of heart rate reductions is prepared in the "comma-separated values" (csv).

Patient, StdDrug, NewDrug

1, 28.5, 34.8

2, 26.6, 37.3

...

...

40, 40.1, 40.8

Read data set into R. We can read the csv file by using `raed.csv()`, and declare the use of data frame *HeartData* by `attach()`.

```
> HeartData <- read.csv("heart.csv")
> attach(HeartData)
```

The `summary` command will show you the variable names and their summary statistics. These variable names are *Patient*, *StdDrug* and *NewDrug* as indicated in the output below.

```
> summary(HeartData)
```

Patient StdDrug NewDrug

Min. : 1.00 Min. :21.60 Min. :22.40

1st Qu.:10.75 1st Qu.:27.45 1st Qu.:30.80

Median :20.50 Median :32.00 Median :34.25

Mean :20.50 Mean :31.18 Mean :33.84

...

Graphical presentations. Here we need the boxplot for each of *StdDrug* and *NewDrug* to compare the two samples graphically. The `boxplot` command will create the two boxplots in one figure.

```
> boxplot(StdDrug, NewDrug, names=c("Standard
  Drug", "New Drug"), col="gray", ylab="Heart
  rate reductions", main="Boxplots for Heart Rate
  Reductions")
```

Statistical inference. The paired sample test can be done by the `t.test()` command with the option `paired=T`. Suppose that we want to test

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_A : \mu_1 < \mu_2$$

where μ_1 and μ_2 are the true means of heart rate reductions with the standard drug and the new drug, respectively.

```
> t.test(StdDrug, NewDrug, alternative="less", paired=T)
```

t = -4.5016, df = 39, p-value = 2.974e-05

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -1.661287

sample estimates:

mean of the differences

-2.655

Statistical inference, continued. The output in the previous page shows that (i) t-statistic is -4.5016, (ii) p-value is 2.974×10^{-5}, and (ii) the 95% one-sided confidence interval $(-\infty, -1.661287)$ for the difference $(\mu_1 - \mu_2)$. Thus, we can reject H_0 with significance level 0.01. To obtain the 99% one-sided confidence interval, add the option "`conf.level=0.99`" as follows.

```
> t.test(StdDrug, NewDrug, alternative="less",
  paired=T, conf.level=0.99)
```

Inference on Two Independent Samples

Experimental studies. We often want to compare two independent samples. For example, a researcher tests the difference of nerve conductivity speed between healthy persons and patients with nerve disorder. The study considers a **control group** in which healthy subjects are examined, and an **experimental group** in which subjects with nerve disorder are participated. As a result of experiment, we obtain the measurements

$$X_1, \ldots, X_n$$

of the subjects from the control group, and the measurements

$$Y_1, \ldots, Y_m$$

of the subjects from the experimental group.

Experimental studies, continued. It is assumed that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent and normally distributed with (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively. Large sample sizes $(n, m \geq 30)$ ensure that the tests are appropriate even if they are not normally distributed. Then it becomes the hypothesis testing problem

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_A : \mu_1 \neq \mu_2.$$

where μ_1 and μ_2 are the respective population means of the control and the experimental groups

Pooled test procedure. Let S_x and S_y be the sample standard deviations constructed from $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively. When it is assumed that

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

we can estimate σ^2 by the **pooled sample variance**

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

Then we can construct $(1-\alpha)$-level confidence interval for the population mean difference $\mu_1 - \mu_2$ by

$$(\bar{X} - \bar{Y}) \pm t_{\alpha/2, n+m-2} S_p \sqrt{\tfrac{1}{n} + \tfrac{1}{m}}$$

Pooled test procedure, continued. The test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

has the t-distribution with $(n+m-2)$ degrees of freedom under the null hypothesis H_0. Thus, we reject the null hypothesis H_0 with significant level α when the observed value t of T satisfies $|t| > t_{\alpha/2, n+m-2}$. Or, equivalently we can compute the p-value

$$p^* = 2 \times P(Y \geq |t|)$$

with Y having a t-distribution with $(n+m-2)$ degrees of freedom, and reject H_0 when $p^* < \alpha$.

General procedure. Under the null hypothesis H_0, the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

has approximately the t-distribution with ν degree of freedom, where ν is the nearest integer to

$$\frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{S_x^4}{n^2(n-1)} + \frac{S_y^4}{m^2(m-1)}}.$$

Then we can construct $(1-\alpha)$-level confidence interval for the population mean difference $\mu_1 - \mu_2$ by

$$(\bar{X} - \bar{Y}) \pm t_{\alpha/2,\nu}\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$$

General procedure, continued. We reject the null hypothesis H_0 with significant level α when the observed value t of T satisfies $|t| > t_{\alpha/2,\nu}$. Or, equivalently we can compute the p-value

$$p^* = 2 \times P(Y \geq |t|)$$

with Y having a t-distribution with ν degrees of freedom, and reject H_0 when $p^* < \alpha$.

Read data set into R. The data set of nerve conductivity speeds is prepared in a 32-by-2 table. The first column of the table represents 32 healthy subject data, and the second column represents 27 disordered subject data. The asterisk (*) in the last 5 entries indicates that there is a difference in the column data sizes.

Healthy Disorder

52.20 50.68

...

...

55.90 53.98

52.23 *

54.90 *

55.64 *

54.48 *

52.89 *

Read data set into R, continued. To ignore the symbol (*) in reading the data file, the option "`na.strings="*"`" can be used in `read.table()` command. Now we read them into the data frame *NerveData* as follows.

```
> NerveData <- read.table("nerve.txt", header=T,
  na.strings="*")
```

Data exploration. Declare the use of data frame (`attach`), and then find out the variable names and their summary statistics (`summary`) as follows.

```
> attach(NerveData)

> summary(NerveData)

Healthy Disorder

Min. :52.20 Min. :44.86

...
```

The output from the `summary` command reveals *Healthy* and *Disorder* as the variable names. The `boxplot()` will be used to compare the two samples.

```
> boxplot(Healthy, Disorder, names=c("Healthy", "Nerve
  Disorder"), col="gray", ylab="Conductivity speeds",
  main="Boxplots for Nerve conductivity speeds")
```

Statistical inference. The `t.test()` can be used again for the two independent samples. Suppose that our hypothesis testing problem is

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_A : \mu_1 > \mu_2$$

where μ_1 and μ_2 are the true means of nerve conductivity speed for healthy subjects and nerve disorder subjects, respectively.

```
> t.test(Healthy, Disorder, alternative="greater")

data: Healthy and Disorder

t = 10.608, df = 32.684, p-value = 2.032e-12

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

4.538828 NA

sample estimates:

mean of x mean of y

53.99438 48.59370
```

Statistical inference, continued. The result shows that (i) t-statistic is 10.608, (ii) p-value is 2.032×10^{-12}, and (iii) the 95% one-sided confidence interval is $(4.538828, \infty)$ for the difference $(\mu_1 - \mu_2)$. Thus, we can reject H_0. If we wish to obtain the 99% two-sided confidence interval, then we use "`conf.level=0.99`" and "`alternative="two.sided"`" as follows.

```
> t.test(Healthy, Disorder, alternative="two.sided",
  conf.level=0.99)
```

The default call corresponds to general procedure. If equal variances are assumed, the pooled procedure can be used. In this case we need the option "`var.equal=T`".

```
> t.test(Healthy, Disorder, alternative="two.sided",
  conf.level=0.99, var.equal=T)
```

Manipulate two data frames. An engineer compares the sample of paint thicknesses (`line-a.txt`) from production line A with a sample of paint thicknesses (`line-b.txt`) from production line B. What conclusions should the engineer draw? Here we have two data sets in `line-a.txt` and `line-b.txt`. They should be read into two data frames *PA* and *PB* as follows.

```
> PA <- read.table("line-a.txt", header=T)
> PB <- read.table("line-b.txt", header=T)
```

Manipulate two data frames, continued. We can find out the variable name and summary statistics (`summary`) for each data frame as follows.

```
> summary(PA)
```

Paint.Thicknesses.in.mm

Min. :0.0760

...

```
> summary(PB)
```

Paint.Thicknesses

Min. :0.0230

...

It reveals that the data frame *PA* has the variable *Paint.Thicknesses.in.mm* and that the data frame *PB* has the variable *Paint.Thicknesses*.

Manipulate two data frames, continued. In order to manipulate the two data frames *PA* and *PB* simultaneously, we call the variables directly via

PA$*Paint.Thicknesses.in.mm*
PB$*Paint.Thicknesses*

without declaring `attach`. Then `boxplot` and `t.test` can be carried out as follows.

```
> boxplot(PA$Paint.Thicknesses.in.mm, PB$Paint.Thicknesses, names=c("Line A", "Line B"), ylab="Paint
thicknesses (mm)", main="Boxplots for Paint thickness")
```

Manipulate two data frames, continued.

```
> t.test(PA$Paint.Thicknesses.in.mm, PB$Paint.Thicknesses, alternative="two.sided")
```

data: PA$Paint.Thicknesses.in.mm and PB$Paint.Thicknesses

t = 2.5732, df = 154.713, p-value = 0.01102

alternative: true difference in means is not equal to 0

95 percent confidence interval:

0.007172363 0.054576254

sample estimates:

mean of x mean of y

0.2318133 0.2009390

The result shows that the p-value for the hypthesis testing problem

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_A : \mu_1 \neq \mu_2$$

is 0.01102. Thus, there is a fairly significant evidence that the paint thicknesses from production line A and those from production line B are different.

Simple Linear Regression

Linear regression model. Suppose that the researcher wants to find how the temperature of factory affects the labor efficiency to unload a truck. We conduct n independent experiments with different levels of temperature. The data set consists of the unloading time $Y_1, \ldots, Y_n$ paired with the respective temperature $x_1, \ldots, x_n$ of the factory.

Temperature	Unloading time
x_1	Y_1
$\vdots$	$\vdots$
x_n	Y_n

Linear regression model, continued. The relationship between the *explanatory variable* x_i and the *response variable* Y_i can be approximated by the **simple linear regression model**

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{8.1}$$

where ϵ_i is a "random error" due to other factors of condition. The standard assumption is that the random error terms $\epsilon_1, \ldots, \epsilon_n$ are iid normally distributed random variables with common variance σ^2.

Parameter estimates. The coefficients β_0 and β_1 of the linear regression model (8.1) are called the **intercept** and the **slope** parameters, respectively. The point estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the parameters β_0 and β_1 become

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

where the values $\bar{x}$, $\bar{Y}$, S_{xx}, and S_{xy} are computed as in the following table.

Variables	Mean	Sum of squares
Explanatory	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$
Response	$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$	$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$

Statistical properties. By constructing the **residual sum of squares** (**RSS**)

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

the point estimate $\hat{\sigma}^2$ of the variance σ^2 becomes

$$\hat{\sigma}^2 = \frac{RSS}{n-2}.$$

Then the statistics are summarized in the following table.

Coefficient	Estimate	Standard error	t-value
β_0	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$	$S_0 = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$	$T_0 = \frac{\hat{\beta}_0}{S_0}$
β_1	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	$S_1 = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$	$T_1 = \frac{\hat{\beta}_1}{S_1}$

Read and declare a data set. We use the data set of unloading time.

UnloadingTime Temperature

64 52

53 68

58 64

...

To read the data set, we use the `read.table`.

```
> TimeData <- read.table("time.txt", header=T)
```

Declare and use data set. We declare the use of data. To see sample statistics with variable names, we can use the `summary()`. Then, the first line of the output below displays the variable names *Time* and *Temperature*.

```
> attach(TimeData)
> summary(TimeData)
```

Time Temperature

Min. :38.00 Min. :52.00

...

The `plot` function can be used to show the scatter plot of temperature against time.

```
> plot(Temperature, Time, main="Scatter plot of
  temperature against time")
```

Statistical inference. To fit the data frame *TimeData* into a simple linear model, the `lm` function will be used and the result must be saved in a variable. Then, the `summary` function with the variable produced by the `lm` function can display the result.

```
> TimeLM <- lm(Time ~ Temperature)
> summary(TimeLM)
```

...

Coefficients:

```
 Estimate Std.Error t value Pr(>|t|)
```

(Intercept) 36.1935 16.9515 2.135 0.0585 .

Temperature 0.2659 0.2383 1.116 0.2905

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

...

Statistical inference, continued. In finding a trend, the result shows that (i) the estimate $\hat{\beta}_1$ of slope is 0.2659, and (ii) the p-value is 0.2905, which is insignificant. Thus, we cannot reject the null hypothesis $H_0 : \beta_1 = 0$, and therefore, there is not sufficient evidence to conclude that the unloading time depends on the temperature. And we should conclude that no relationship has been established between the two variables. As for the intercept, the result shows that (i) the estimate $\hat{\beta}_0$ of intercept is 36.1935, and (ii) the p-value is 0.0585, which is moderately significant. To see the fitted line graphically, we can use the `abline` function. It adds the fitted line on the scatter plot which was previously drawn.

```
> abline(TimeLM)
```

Residual analysis. To assess the fit graphically, we can use the following assortment of plots:

- **Residual-Fit spread plot** compares the spread of the fitted values with the spread of the residuals.
- **Normal Q-Q plot** provides a visual test of the assumption that the model's error terms are normally distributed.

We can create these diagnostic plots by using `plot` function with the model variable name *TimeLM*.

```
> plot(TimeLM, which=1)
> plot(TimeLM, which=2)
```

Assignment No.8

Computer Assignment. Investigate the following statistical studies, and write a short report on your own statistical analysis. Your report must include:

(a) sample statistics such as mean, median, and standard deviation;

(b) graphical presentations (histogram, boxplot, or scatter plot) of data;

(c) descriptions of hypothesis testing (null and alternative hypothesis);

(d) results of formal statistical inference (p-value), and your conclusions.

Study 1: Red blood cell adhesion.

Data set: `bloodcell.txt`

Researchers into the genetic disease sickle cell anemia are interested in how red blood cells adhere to endothelial cells, which form the innermost lining of blood vessels. A set of 14 blood samples are obtained, and each sample is split in half. One half of the blood sample is profuse over an endothelial monolayer of type A and the other half of the blood sample is profused over an endothelial monolayer of type B. The two types differs in respect to the stimulation conditions of the endothelial cells. The data represent the number of adherent red blood cells per mm^2. Is there any evidence that the different stimulation conditions affect the adhesion of red blood cells?

Study 2: Service times.

Data sets: `afternoon.txt` and `morning.txt`

The data set in `afternoon.txt` shows the service times (in second) of customers at a fast-food restaurant who were served between 2:00 and 3:00 on a Saturday afternoon. In addition, `morning.txt` shows the service times of customers at the fast-food restaurant who were served between 9:00 and 10:00 in the morning on the same day. What do these data sets tell us about the difference between the service times at these two times of day?

Study 3: Aerobic fitness.

Data set: `vo2max.txt`

The data concern the aerobic fitness of a sample of twenty male subjects collected at the Health and Performance Sciences Laboratory at Georgia Tech. An exercising individual breathes through an apparatus that measures the amount of oxygen in the inhaled air which is used by the individual. The maximum value per unit time of the utilized oxygen is then scaled by the person's body weight to come up with a variable VO2-max, which is a general indication of the aerobic fitness of the individual. Fit a linear regression model with VO2-max as the dependent variable (the response variable) and age as the explanatory variable. Is it clear that on average aerobic fitness decreases with age?