Basic Data Analysis Using R
Exercise sheet

Slide 5
Ex1.1: We will use the dataset survey in the package MASS.

```r
library(MASS) #Load the package
data(survey) #Make the dataset available

head(survey) #Check the dataset to see what it looks like
dim(survey)
str(survey)

complete.cases(survey) #Return Booleans - TRUE/FALSE:
# TRUE: complete case/no missing value(s)
# FALSE: incomplete case/missing value(s)
sum(complete.cases(survey)) #Total # of cases w/o no missing values
```

Ex1.2: Let’s create a new dataset called survey.complete, which contains no missing values. In other words, we want to remove all the rows with missing values.

```r
survey.complete <- na.omit(survey)
```

Slide 8
Ex2.1: We will use the newly created dataset survey.complete. Let’s check the mean and median of height, age, and pulse.

```r
mean(survey.complete$Height); median(survey.complete$Height)
mean(survey.complete$Age); median(survey.complete$Age)
mean(survey.complete$Pulse); median(survey.complete$Pulse)

#We can also use colMeans() if we want to compute the means for multiple variables/columns.
colMeans( survey.complete[, c("Height", "Age", "Pulse")]

#There is no equivalent function for medians. We have to use apply()
# apply(X, MARGIN, FUN)
apply(survey.complete[, c("Height", "Age", "Pulse")], 2, median)
```

Ex2.2: Let’s compute the standard deviation, variance, and range for Height, Age, and Pulse.

```r
apply(survey.complete[, c("Height", "Age", "Pulse")], 2, sd)
apply(survey.complete[, c("Height", "Age", "Pulse")], 2, var)
apply(survey.complete[, c("Height", "Age", "Pulse")], 2, range)
```

Slide 9
Ex3.1: Let’s use the function summary().

```r
summary(survey.complete)
```

Ex3.2: Let’s use the function describe() from the package `psych`.

```r
# if you don’t have psych already installed
install.packages("psych")
```
#load `psych`
load(psych)

describe(survey.complete)
#Those with an asterisk are categorical variables. Since the function  
#force-convert categorical variables into numerical variables, the  
#outputs do not make sense, and can be ignored.

### Slide 12

**Ex4.1:** Test whether the mean age is significantly different from 19.75.

```r
plot(density(survey.complete$Age))  #density plot
hist(survey.complete$Age, col="gray")  #histogram
boxplot(survey.complete$Age)  #boxplot

t.test(survey.complete$Age, mu = 19.75)
```

### Slide 13

**Ex4.2:** Test whether the mean height of female students is significantly different from the mean height of male students.

```r
height.m <- subset(survey.complete, subset=Sex=="Male", select=Height)
height.f <- subset(survey.complete, subset=Sex=="Female", select=Height)
boxplot(Height ~ Sex, data=survey.complete)

t.test(height.m, height.f)  #Welch’s test
t.test(height.m, height.f, var.equal=TRUE)  #Student’s t
```

### Slide 14

**Ex4.3:** Use `t.test(formula, data)` to test the hypotheses tested in Ex4.2.

```r
t.test(Height ~ Sex, data = survey.complete)
```

### Slide 15

**Ex4.4:** Test whether the mean Writing hand span (`Wr.Hnd`) and the mean Non-writing hand span (`NW.Hnd`) differ significantly.

```r
t.test(survey.complete$Wr.Hnd, survey.complete$NW.Hnd, paired=TRUE)
```

### Slide 17

**Ex5.1:** Test whether the mean pulse rates differ amongst the different exercise groups (None, Some, and Freq)

```r
boxplot(Pulse ~ Exer, data=survey.complete)

pulse.aov <- aov( Pulse ~ Exer, data = survey.complete )
summary( pulse.aov )  #Significance test (p vals, F statistics, etc)
```

### Slide 18

**#Import data**

```r
hsb2 <- read.table("http://www.ats.ucla.edu/stat/r/faq/hsb2.csv", sep=""," header=TRUE")
```
Ex5.2: Test the main effects of race & schtyp and the interaction race:schtyp.

```r
boxplot(read ~ race * schtyp, data=hsb2)
```

```r
read.aov<-aov(read~as.factor(race)*as.factor(schtyp), data=hsb2)
summary(read.aov)
```

---

# Import data

expenditure <- read.table("http://dornsife.usc.edu/assets/sites/210/docs/GC3/educationExpenditure.txt", sep=";", header=TRUE)

---

Ex5.2: education regresses on income.

```r
plot(expenditure$income, expenditure$education)
ed.lm1 <- lm(education ~ income, data=expenditure)
ed.lm1
summary(ed.lm1)
```

```r
abline(ed.lm1$coefficients)  # regression line
```

```r
ed.res1 <- residuals(ed.lm1)  # Extract residuals
ed.pred1 <- predict(ed.lm1)  # Extract predicted/fitted values
```

```r
plot(ed.pred1, ed.res1, main="Scatterplot of residuals and predicted values", xlab="predicted values", ylab="residuals")
```

# You can also use plot(ed.lm1) to obtain additional diagnostic plots.

---

Ex5.2: education regresses on income.

```r
ed.lm2 <- lm(education~income + young + urban, data=expenditure)
ed.lm2
summary(ed.lm2)
```

```r
par(mfrow=c(2, 2))  # create a plot window that holds 4 plots in a 2 by 2 grid
plot(ed.lm2)
```