

Tutorial 1

DSA1101

Introduction to Data Science

August 31, 2018

Exercise 1. Suppose we have two data vectors $x = c(x_1, x_2, \dots, x_n)$ and $y = c(y_1, y_2, \dots, y_n)$, both of length n . Suppose a and b are any two constants. Let $\overline{ax + b} = \frac{1}{n} \sum_{i=1}^n (ax_i + b)$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

(a) Show that $\overline{ax + b} = a\bar{x} + b$.

(b) Recall from lecture that the sample variance of x is given by $\text{var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. For the new data vector $ax = c(ax_1, ax_2, \dots, ax_n)$, show that $\text{var}(ax) = a^2 \text{var}(x)$.

Exercise 2. Data input and manipulation in R.

- (a) Read the data from the file `colleges.txt`

```
1 col=read.table("Colleges.txt",header=T,sep='\t')
2 View(col)
```

- (b) Draw a histogram of the `SAT` variable

```
1 hist(col$SAT,prob=T)
```

- (c) Draw a plot of the `DPerStudent` versus `GradPer` variables

```
1 plot(col$DPerStudent,col$GradPer)
```

- (d) Draw the histograms of `DPerStudent` separately for `LibArts` (liberal arts) and `Univ` (University) institutions

```
1 hist(col$DPerStudent[col$School_Type=="Lib Arts"])
2 hist(col$DPerStudent[col$School_Type!="Lib Arts"])
```

- (e) Draw the histograms of `GradPer` separately for `LibArts` (liberal arts) and `Univ` (University) institutions

```
1 hist(col$GradPer[col$School_Type=="Lib Arts"])
2 hist(col$GradPer[col$School_Type!="Lib Arts"])
```

- (f) Find out which institutions have more than 75% of faculty members with Ph.D. degrees

```
1 col$School[col$PerPhD>75]
```

- (e) Perform a linear regression of `Acceptance` on the variables `Top.10p`, `PerPhD` and `GradPer`

```
1 lm(Acceptance~ Top.10p+ PerPhD + GradPer, data=col)
```