Tutorial 1

DSA1101
Introduction to Data Science

August 31, 2018

Exercise 1. Suppose we have two data vectors $x = c(x_1, x_2, ..., x_n)$ and $y = c(y_1, y_2, ..., y_n)$, 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 $\overline{ax+b} = a\bar{x} + b$.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(ax_i + b) &= \frac{1}{n}\sum_{i=1}^{n} ax_i + \frac{1}{n}\sum_{i=1}^{n} b \\
&= a\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{nb}{n} \\
&= a\bar{x} + b
\end{aligned}$$

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

$$
\begin{aligned}
\mathrm{var}(ax) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(ax_i - \overline{ax}\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(ax_i - a\bar{x}\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}a^2\left(x_i - \bar{x}\right)^2 \\
&= a^2\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \\
&= a^2\mathrm{var}(x)
\end{aligned}
$$

Exercise 2. Data input and manipulation in R.

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

```r
col=read.table("Colleges.txt",header=T,sep=’\t’)
View(col)
```

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

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

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

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

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

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

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

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

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

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

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

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