Statistical Computing
Master Data Science
Winter Semester 2019/20

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 4
Measures of Dispersion*with solutions*

1 Workshop

Start R-Studio in the usual way and open a word document to include answers to any questions, graphics, etc.

1.1 Standard Deviation and Variance in R

Last week you worked through an exercise using the Prestige data. You will be using the same data this week.

(a) Load the package `car` and grab the data frame

```
> require(car)
> data(Prestige)
```

(b) Plot the variable income in a histogram, to get a feel for the data.

```
> hist(Prestige$income,breaks=15)
```

You learnt the commands to obtain the variance and standard deviation in Workshop 3. This week you will compute these statistics from scratch. Use the lecture notes to fill in the gaps

```
> xbar<-mean(Prestige$income)
> devs<-Prestige$income-xbar
> sqdevs<-???
> ssqdevs<-sum(sqdevs)
> varincome<-sqdevs/(????)
> sdincome<-????(varincome)
```

Check your values using the standard R functions.

(c) Obtain the range (`range()`) and interquartile range `(IQR())` of the income data. Notice that the IQR is a roughly similar value to the standard deviation. This is often the case.

1.2 Linear transformation

You will transform the data so that income is measured on a scale from 0 (smallest income) to 1 the largest income. The formula to do this is

$$y_i = \frac{x_i}{\text{xmax} - \text{xmin}} - \frac{\text{xmin}}{\text{xmax} - \text{xmin}}$$

This is a linear transformation of the form $y_i = ax_i + b$. What are the values of a and b in this case?

(a) Calculate the standard deviation of the income on this new scale using the formula in the Lecture and your result from Section 1.1.

(b) Check your by defining a new variable, checking that all the values lie between 0 and 1 and calculating its sd.

```
> Prestige$income01<-?????
> range(Prestige$income01)
> sd(Prestige$income01)
```

1.3 Coefficient of variation

There is another measure of dispersion not covered in the lecture called the coefficient of variation (CV). CV is the standard deviation divided by the mean.

$$CV_x = \frac{s_x}{\overline{x}}$$

This measure only makes sense if the data can only be positive.

The CV is "unit-less". It follows that if the the data is changed proportionately, for example from Canadian Dollars to US Dollars, the CV does not change. Formally If $y_i = ax_i$ for constant a, then $CV_y = CV_x$.

► Calculate the CV for the variable `income`

1.4 Rough interpretation of the standard deviation

(a) For the variable called `education` Calculate the interval: $[\overline{x} - 2s_x; \overline{x} + 2s_x]$.

(b) How many data points actually lie in this interval? Hint: see the lecture notes slide 9.

(c) What proportion of points lie in this interval?

(d) Does this proportion fit with the rule of thumb? *OK, 100 % are in the intervall*

(e) Repeat part the above steps for the variable `income`.

Tidying up At the end of the workshop: tidy up your script file, add comments to make the code readable and save your files.

Have a great weekend!

2 Exercises to do at home

Exercise 1: Variance

Three numeric variables a, b and c are given in the following table.

Variable	Data
a	1.0; 1.1; 1.2; 1.2; 1.3; 1.4; 1.5
b	20; 22; 25; 28; 30; 35; 40
c	-50; 20; 134; 219; 298; 504; 780; 1293

The three variances in random order are 50.62, 201900 and 0.02952.

Without using `R` assign the three variances to the three variables

Variance(a)= 0.02952,
Variance(b)= 50.62,
Variance(c)= 201900

Exercise 1 Calculating descriptive statistics

A small sample of lengths in millimetres is given in the following table. The values have been ordered to help you.

i	1	2	3	4	5	6	7	8	9	10
x_i	121	126	142	148	153	158	172	185	197	198

(a) Calculate the following without using R. Include the units in your answer.

(i) the sample size, *=10*

(ii) the mean, *=160 mm*

(iii) the variance, *sum of squared deviances=6780,* $s^2 = 753.3$ *mm*^{2}

(iv) the standard deviation, $\sqrt{753} = 27.4$ *mm*

(v) the median, *=155.5 mm*

(vi) the quartiles, *=142 und 185 mm* and

(vii) the coefficient of variation. 0.171

Exercise 2 Linear transformation

Let $y_i = ax_i + b$.
Use the definition of the mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ to prove that $\overline{y} = a\overline{x} + b$

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

Exercise 3 Coefficient of Variation

Let $y_i = ax_i$
Use the formulae for linear transformations in Lectures 3 and 4 to show that $CV_y = CV_x$.

$$\begin{aligned} CV_y &= \frac{\overline{y}}{s_y} \\ &= \frac{a\overline{x}}{as_x} \\ &= \frac{\overline{x}}{s_x} = CV_x \end{aligned}$$