Introduction to Statistics and R

Working with data sets

Eric Stemmler

24.02.2021

Contents

1 Recap: Exercise Solutions

1.0.1 Circles

1. Calculate the area of a circle $A = 2\pi r^2$ with $r = 2$ (Qian, 2016)
2. Write the circle area formula as a function with named as `circle` with parameter `r` and and calculate the area for `r <- seq(0, 3, 0.1)`
3. **Extension:** Extend your function to return a named `list` that contains the area and the circumference for a given parameter vector `r`

Өнгөрсөн долоо хоногт бид хоёр дасгал ажилласан. Эхнийх нь өгөгдсөн радиустай тойргийн талбайг тооцоолох тухай байв. Мөн тухайн дасгалд талбай ба тойрог гэсэн хоёр утга бүхий list-ийг гаргаж ирэх функцийг бичих ёстой байсан.

1.0.2 Normal Probability Density Function

1. Calculate the density of the normal probability distribution function $x \sim \mathcal{N}(2, 1.25)$ (mean $\mu = 2$ and standard deviation $\sigma = 1.25$) at `x <- seq(0, 4, 0.5)` by using the normal probability density formula $\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$, and verify your result by using the function `dnorm` (Qian, 2016).

Хоёр дахь дасгал нь дундаж μ болон стандарт хазайлт нь өгөгдсөн хэвийн магадлалын нягтын функцийн утгыг гаргаж ирэх функцийг бичих ёстой байсан. Томъёо нь өгөгдсөн байгаа бөгөөд хэвийн магадлалын нягтын функцийн x параметрүүдийн дарааллын утгыг тооцоолж R -д суулгасан dnorm() нэртэй функцтэй харьцуулах ёстой байсан.

1.1 Solutions

1.1.1 Circles

```
circle <- function(r) {
    result <- list(area = 2 * pi * r^2,
                circumference = 2 * pi * r)
    return(result)
}

circle(r = seq(0, 3, 0.5))
```

```
## $area
## [1]  0.000000  1.570796  6.283185 14.137167 25.132741
## [6] 39.269908 56.548668
##
## $circumference
## [1]  0.000000  3.141593  6.283185  9.424778 12.566371
## [6] 15.707963 18.849556
```

Тойргийн талаархи эхний дасгалын шийдэл энд харагдаж байна. Бид түлхүүр үгийн функцийг бичих замаар R-д фунцийг тодорхойлж байгаа ба энд r гэж нэрлэгдсэн нэг параметрийг тодорхойлно. Дараа нь бид энэ функцийг circle гэж нэрлэгдэх хувьсагчид хуваарилах бөгөөд үүнийг дараа нь функцийг дуудахад ашиглаж болно. Функц дотор бид result гэсэн хувьсагчийг тооцоолж, үүнийгээ талбай ба тойрог гэсэн хоёр утгыг агуулсан list гэж тодорхойлж, тооцоолно. Бид return/буцаах мэдэгдлийг ашиглан list-ээ гаргаж ирэх боломжтой

1.1.2 Normal Probability Density Function

```
npdf <- function(x, avg, stdev) {
    result <- 1.0 / (sqrt(2*pi*stdev^2)) *
        exp(-(x - avg)^2 / (2.0 * stdev^2))
    return(result)
}

npdf(x = seq(0, 4, 0.5), avg = 2, stdev = 1.25)
```

```
## [1] 0.08873667 0.15534884 0.23175324 0.29461611
## [5] 0.31915382 0.29461611 0.23175324 0.15534884
## [9] 0.08873667
```

```
    dnorm(x = seq(0, 4, 0.5), mean = 2, sd = 1.25)
```

```
## [1] 0.08873667 0.15534884 0.23175324 0.29461611
## [5] 0.31915382 0.29461611 0.23175324 0.15534884
## [9] 0.08873667
```

Энд харагдаж байгаа зүйл бол хоёр дахь дасгалын хариу юм. Бид дахин функцийг тодорхойлж npdf нэртэй хувьсагч руу хуваарилна. Энэ функцын хувьд бид x, average/дундаж/ гэсэн үгний товчлол avg, standard deviation/стандарт хазайлт/ гэсэн үгний товчлол stdev гэсэн гурван параметрийг тодорхойлно. Одоо бид функцээ x-ийн утгын болон дундаж утга ба стандарт хазайлт тус бүрийн хоёр дан утгын векторыг бэлтгэж дуудаж болно. Бидний функцийг R-д суурилуулсан функцтэй харьцуулахын тулд бид зүгээр л dnorm функцийг дуудаад л параметрүүдтэй ижил утгуудыг дамжуулна. Консол дээр хэвлэсэн үр дүнгүүд ижил байгааг бид харж байна.

2 R Tutorial

- Every Thursday, 02:00-03:00pm, Room: 314 (here)
- Practising R and working on exercises
- No lecture, just programming practise and answering questions
- Also possible: analysis of your data sets

Би та бүхэнд R-ийн хэрэглээний дадлага хийх тусдаа хичээлийн цагийг санал болгох гэсэн юм. Би үүнийг tutorial буюу давтлагын хичээл гэж нэрлэж байгаа ба бид энэ хичээлээр бие даан R программын талаар ярилцаж зөвлөлдөх боломжтой юм. Хэрэв та R-ийн талаар илүү ихийг мэдэхийг хүсч байвал бид энэ хичээл дээр ярилцах боломжтой бөгөөд R-ийг ойлгоход болон өгсөн дасгалуудыг ажиллахад бэрхшээлтэй тулгарч байгаа бол тус хичээлийн цагийг ашиглан шийдэх боломжтой юм . Мөн та бүхэн анализ дүн шинжилгээ хийх өөрийн өгөгдлийн багцыг чөлөөтэй авчрах боломжтой. Энэхүү давтлагын хичээл дээр ямар нэгэн лекц орохгүй болно.

3 Formatting a data set

3.0.1 CSV files

- CSV - **c**omma **s**eparated **v**alues
- easiest data format to work with
- can also be used by Microsoft Excel, etc.
- can easily be created from Microsoft Excel file (.xlsx)
- *is not a Microsoft Excel Spreadsheet!*

R хэл дээр бид R дотор үүсгэснээсээ илүүтэй бодит ертөнцөд цуглуулсан өгөгдлийн багцын статистик мэдээллийн анализ хийхийг ихэвчлэн хүсдэг. Ихэвчлэн бид өгөгдлийг тодорхой форматтай өгөгдлийн файлд цуглуулж хадгалдаг болно. R хэл дээр олон янзын форматтай өгөгдлийн файлуудыг унших боломжтой. Мэдээллийн санг ашиглах эсвэл интернетээс шууд татаж авах замаар илүү том өгөгдлийн багцыг ашиглах боломжтой. Гэсэн хэдий ч бид энэ талаархи хэт олон техникийн дэлгэрэнгүй ойлголт руу орохгүйгээр дан ганц өгөдлийн файлуудыг уншиж ажиллах болно. Хамгийн нийтлэг, амархан гарын доорх өгөгдлийн форматуудын нэг бол “CSV” гэж нэрлэгддэг файлын формат юм. CSV нь comma-separated values /таслалаар тусгаарлагдсан утга/-ийн товчлол бөгөөд энэ нь зүгээр л энгийн текстэн файл юм. Энэ нь та үндсэндээ CSV файл үүсгэхийн тулд дурын text editor /текст засварлагч-ийг ашиглаж болно гэсэн үг юм. Та нарын олонхи нь Microsoft Office програмын багц, тодруулбал Microsoft Excel програмыг сайн мэддэг байх. Excel-ийг CSV файл үүсгэхэд ашиглаж болно. Хэрэв танд Excel-ийн файл хэлбэрээр ихэвчлэн Xls файлын өргөтгөлтэй эсвэл илүү сүүлийн үеийн xlsx өргөтгөлтэй өгөгдөл байгаа бол энэ нь ялангуяа дадлага болж өгнө.

Example: `test.csv`

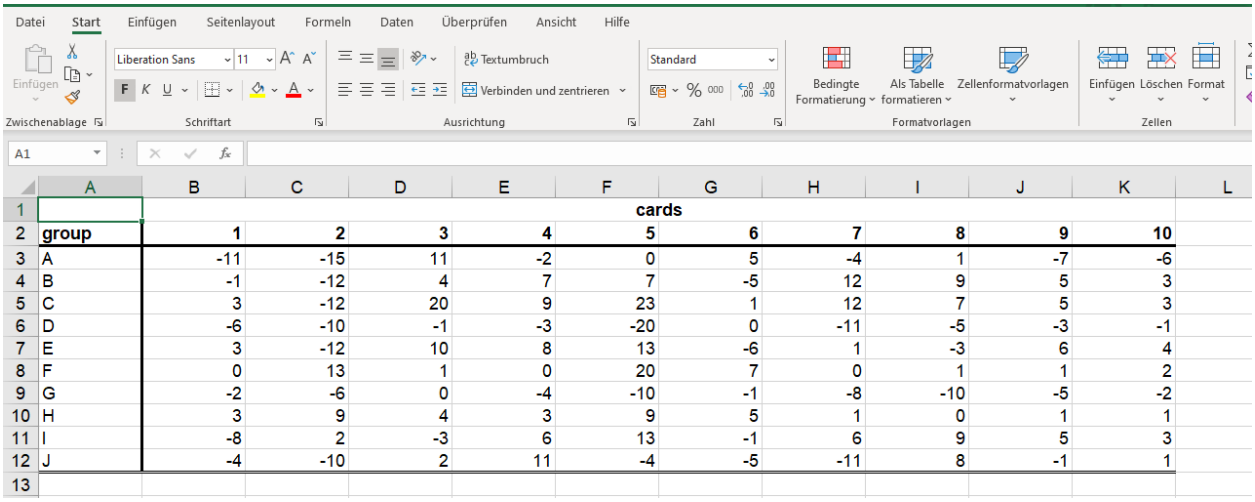
```
name;age;language;nationality;
eric;33;english;german;
stefanie;32;russian;german;
nino;2;mongolian;german;
```

- Open any text editor
- Enter example text
- File » Save As "test.csv"
- Open "test.csv" in Microsoft Excel

CSV файлын жишээг авч үзье. Та text editor/текст засварлагч-ийг нээж, дэлгэцэн дээр харуулсан текстийг бичнэ үү. Текст нь ердөө л таслалаар тусгаарласан үг, тооноос бүрдэнэ. Энэ файлыг өгөгдлийн хүснэгт хэмээж дараах байдлаар тайлбарлана: Багана, мөрийн хүснэгт хийх ба эхний багана болон

мөрийн нүдэнд "нэр" гэсэн үг байна, ижилхэн мөрөнд гэхдээ хоёр дахь баганын нүдэнд "нас" гэсэн үг байна. 1-р мөрийн 3, 4-р баганад "хэл", "үндэс угсаа" гэсэн үгс байна. Хэрэв бид CSV файлын хоёр дахь мөрөнд текст оруулбал энэ мөрний текстийг хүснэгтийн хоёр дахь мөрийн нүдэнд бичнэ. Таслалууд нь нүдэн дэх агуулгыг янз бүрийн баганад тусгаарлаж зааглаж өгнө. Хэрэв та энэ текстийг өөрийн сонгосон text editor /текст засварлагч- руу оруулбал файлыг жишээ нь test.csv байдлаар CSV файл болгон хадгалах замаар үр дүнг шалгах боломжтой ба File Open дээр дарж файлыг Excel дээр нээж болно. Үүний үр дүнд та текст файлын тухайн үгтэй тохирч буй агуулга бүхий хүснэгтийг харах болно.

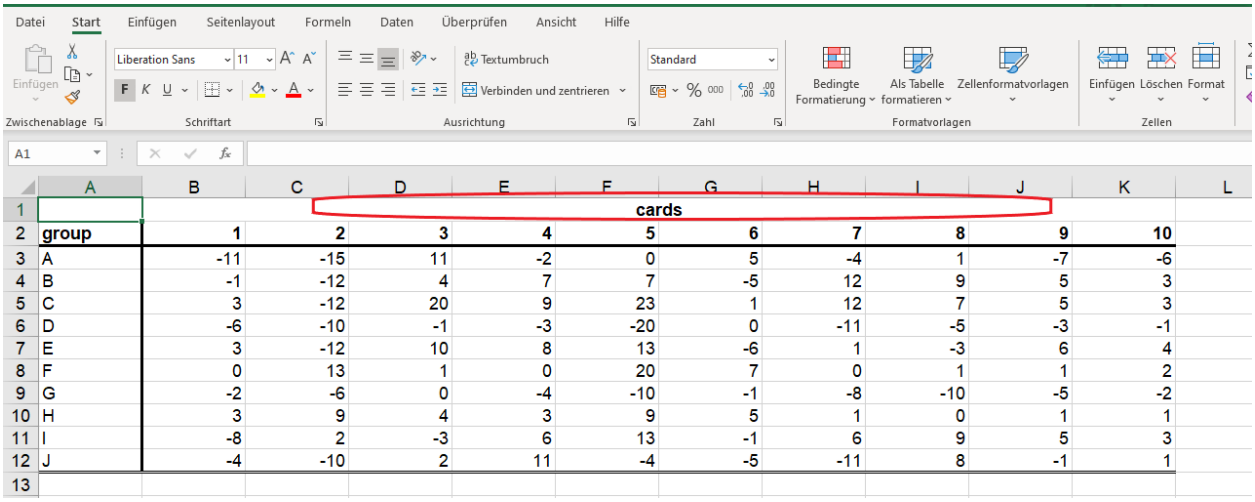
3.0.2 From Excel-file to CSV-file



	A	B	C	D	E	F	G	H	I	J	K	L
1						cards						
2	**group**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
3	A	-11	-15	11	-2	0	5	-4	1	-7	-6	
4	B	-1	-12	4	7	7	-5	12	9	5	3	
5	C	3	-12	20	9	23	1	12	7	5	3	
6	D	-6	-10	-1	-3	-20	0	-11	-5	-3	-1	
7	E	3	-12	10	8	13	-6	1	-3	6	4	
8	F	0	13	1	0	20	7	0	1	1	2	
9	G	-2	-6	0	-4	-10	-1	-8	-10	-5	-2	
10	H	3	9	4	3	9	5	1	0	1	1	
11	I	-8	2	-3	6	13	-1	6	9	5	3	
12	J	-4	-10	2	11	-4	-5	-11	8	-1	1	
13												

Typical Microsoft Excel table. Not suitable for using with R.

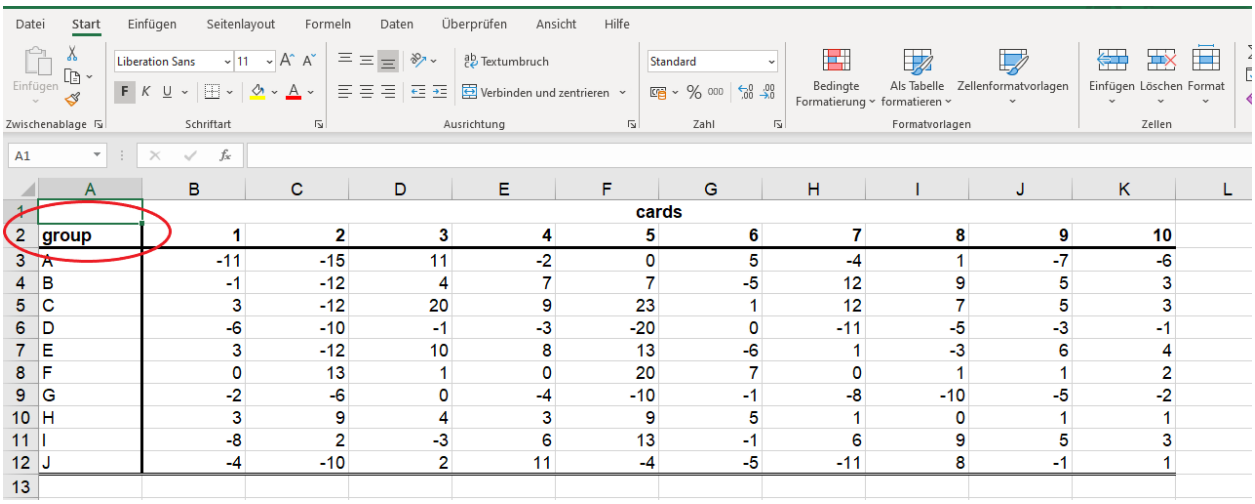
Одоо Microsoft Excel файлыг R дотор ашиглаж болох CSV файл болгон хэрхэн яаж өөрчлөхийг харуулья. Үүний тулд илүү практик дадлагад суралцах үүднээс өөрийн компьютер дээрээ харуулсан алхмуудыг хийгээд яваарай. Бидний одоо ажиллах гэж байгаа Excel файлыг энэ хичээлийн слайдуудтай хамт илгээснийг бүгд хүлээн авсан байгаа байх. Энэхүү Excel файл нь бидний нас таах туршилтын үеэр цуглуулсан өгөгдөл болох 10 баг тус бүрийн 10 гэрэл зураг тус бүрийн таамаглалын алдааг агуулсан байгаа. Мөн түүнчлэн хүн бүр компьютер дээрээ Microsoft Excel програм суулгасан байгаа байх гэж найдаж байна. Хэрэв суулгаагүй хүн байгаа бол өөр оролцогчийн компьютерийг хамт хэрэглэж, дагаад хийгээрэй. Файлыг нээхэд дэлгэц дээр та бүхэнд илгээсэн Excel файл харагдана. Энэ бол та бүхний ихэнх нь мэддэг ердийн Excel файл юм: Үүнд форматлагдсан текст, зарим нүд нь хоосон, бусад нүдэнд өнгө, ирмэг нэмсэн, заримдаа Excel файлууд бас томъёо агуулдаг. Гэсэн хэдий ч, энэ файлыг яг одоо байгаа байдлаар нь шууд R дахь өгөгдлийн файл болгон ашиглах боломжгүй юм. Яагаад гэдгийг харцгаая. Одоо Excel файлыг CSV болгон өөрчлөх боломжтой эсэхийг шалгах шаардлагатай зарим шалгуурыг танилцуулья



	A	B	C	D	E	F	G	H	I	J	K	L
1						cards						
2	group	1	2	3	4	5	6	7	8	9	10	
3	A	-11	-15	11	-2	0	5	-4	1	-7	-6	
4	B	-1	-12	4	7	7	-5	12	9	5	3	
5	C	3	-12	20	9	23	1	12	7	5	3	
6	D	-6	-10	-1	-3	-20	0	-11	-5	-3	-1	
7	E	3	-12	10	8	13	-6	1	-3	6	4	
8	F	0	13	1	0	20	7	0	1	1	2	
9	G	-2	-6	0	-4	-10	-1	-8	-10	-5	-2	
10	H	3	9	4	3	9	5	1	0	1	1	
11	I	-8	2	-3	6	13	-1	6	9	5	3	
12	J	-4	-10	2	11	-4	-5	-11	8	-1	1	
13												

Cells should not be merged.

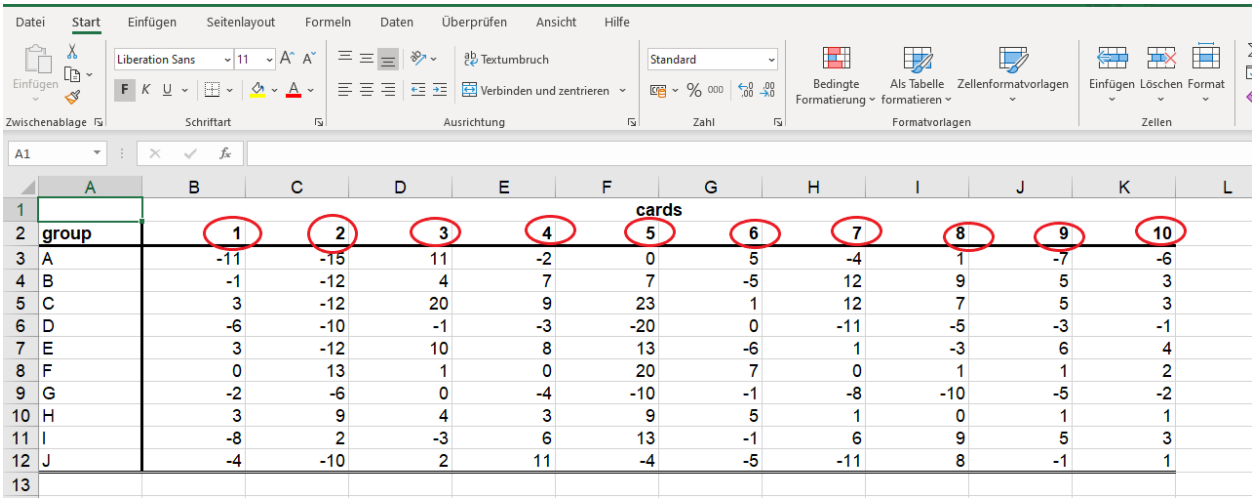
CSV файлын эхний жишээнээс харахад хүснэгтийн эхний мөрөнд хүснэгтийн баганын тоог зааж өгдөг. Гэхдээ энэ Excel файлд эхний эгнээний B-ээс K хүртэлх нүднүүд нэгтгэгдэж нэг нүдэнд орсон байна. Бид доорхи нүднүүдийн тоог тохируулахын тулд энэ нэгтгэсэн нүдийг хэд хэдэн нүдэнд дахин хуваах шаардлагатай. Мөн энэ хүснэгтэд эхнийх биш харин эхний хоёр мөр нь баганын нэрийг зааж өгч байна. Тиймээс бид тэдгээрийг нэг мөр болгон бичих шаардлагатай.



	A	B	C	D	E	F	G	H	I	J	K	L
1						cards						
2	group	1	2	3	4	5	6	7	8	9	10	
3	A	-11	-15	11	-2	0	5	-4	1	-7	-6	
4	B	-1	-12	4	7	7	-5	12	9	5	3	
5	C	3	-12	20	9	23	1	12	7	5	3	
6	D	-6	-10	-1	-3	-20	0	-11	-5	-3	-1	
7	E	3	-12	10	8	13	-6	1	-3	6	4	
8	F	0	13	1	0	20	7	0	1	1	2	
9	G	-2	-6	0	-4	-10	-1	-8	-10	-5	-2	
10	H	3	9	4	3	9	5	1	0	1	1	
11	I	-8	2	-3	6	13	-1	6	9	5	3	
12	J	-4	-10	2	11	-4	-5	-11	8	-1	1	
13												

The first row should be used for table header/ variable names

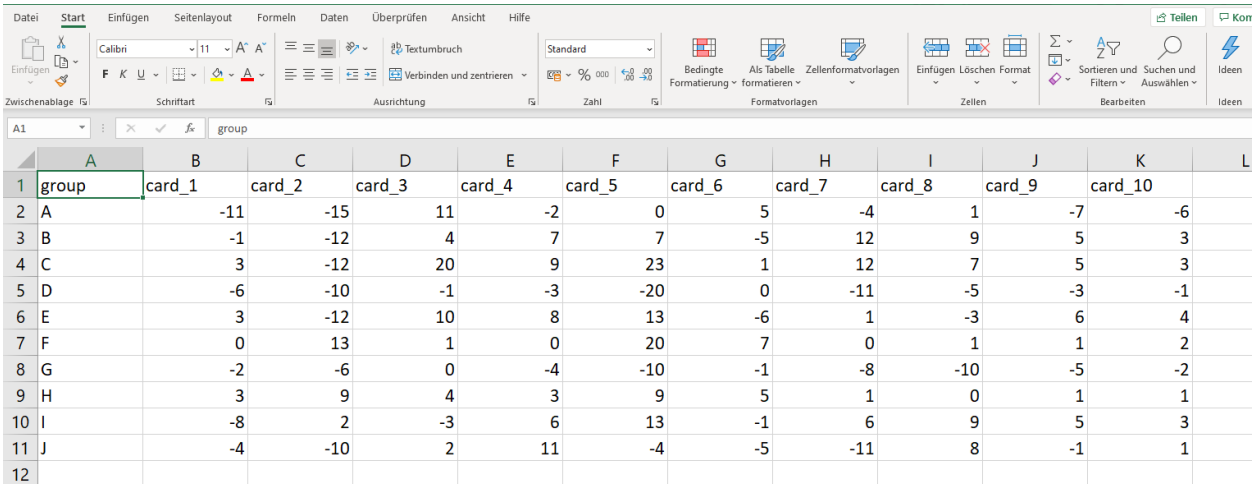
Нэмж дурдахад баганын нэр group-ийг эхний мөрөнд шилжүүлэх шаардлагатай болно, яагаад гэвэл энэ нь баганууудын нэрсийг зааж өгсөн мөр юм.



	A	B	C	D	E	F	G	H	I	J	K	L
1							cards					
2	group	1	2	3	4	5	6	7	8	9	10	
3	A	-11	-15	11	-2	0	5	-4	1	-7	-6	
4	B	-1	-12	4	7	7	-5	12	9	5	3	
5	C	3	-12	20	9	23	1	12	7	5	3	
6	D	-6	-10	-1	-3	-20	0	-11	-5	-3	-1	
7	E	3	-12	10	8	13	-6	1	-3	6	4	
8	F	0	13	1	0	20	7	0	1	1	2	
9	G	-2	-6	0	-4	-10	-1	-8	-10	-5	-2	
10	H	3	9	4	3	9	5	1	0	1	1	
11	I	-8	2	-3	6	13	-1	6	9	5	3	
12	J	-4	-10	2	11	-4	-5	-11	8	-1	1	
13												

- Variable names should be not start with a number or other non-readable characters.
- Variable names cannot contain characters like £, €, $, %, ^, *, +, -, (,), [ ], #, !, ?, <, > or anything that is a operator in R

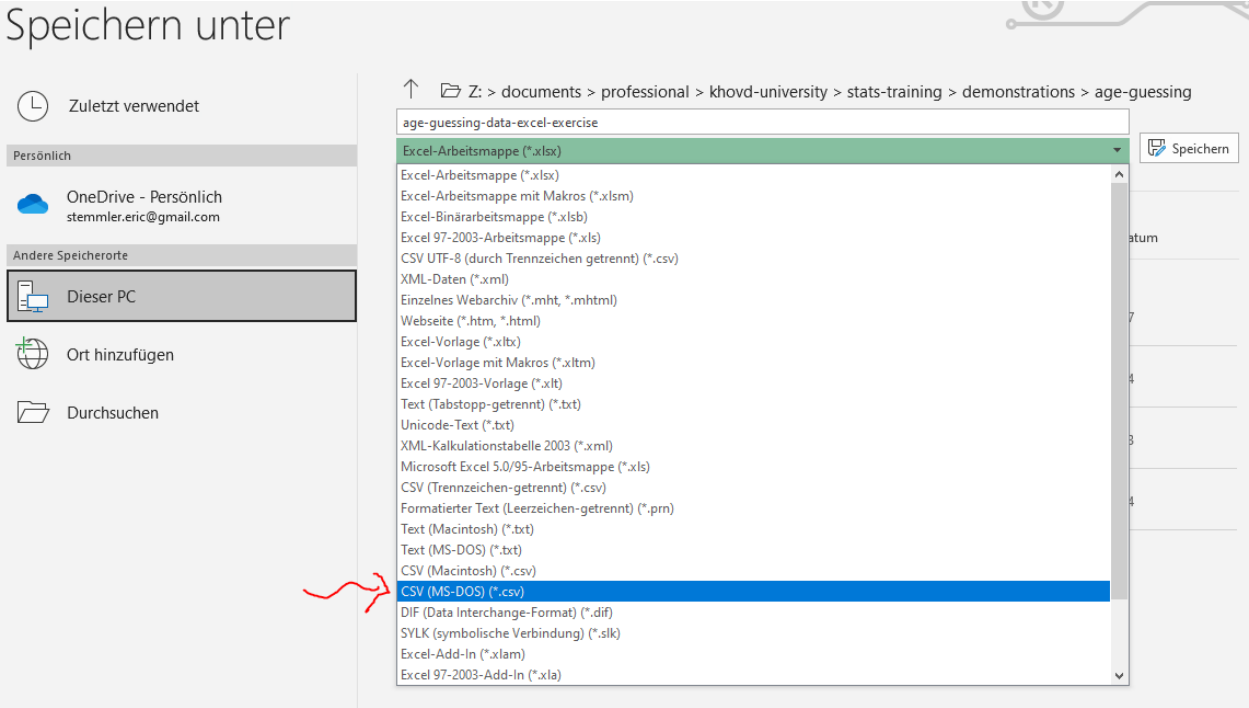
Энэ хүснэгтэд багануудын нэрсийг тоогоор дугаарлан нэрлэсэн байна. Excel-ийн хувьд энэ нь асуудал биш боловч бид энэ хүснэгтийг R-д уншихад эдгээр багануудыг тодорхойлоход хэцүү байх болно, учир нь R нь жишээ нь бидний "2" гэж нэрлэсэн баганы нэрийг 2 гэсэн тоог илэрхийлж байна гэж ойлгох болно. Тоогоор нэрлэгдсэн багануудтай тооцоолол хийх нь боломжгүй тул R нь "2" гэсэн нэртэй баганын бодит утгыг уншихын оронд үүнийг үргэлж тоо гэж тайлбарлах болно.



	A	B	C	D	E	F	G	H	I	J	K	L
1	group	card_1	card_2	card_3	card_4	card_5	card_6	card_7	card_8	card_9	card_10	
2	A	-11	-15	11	-2	0	5	-4	1	-7	-6	
3	B	-1	-12	4	7	7	-5	12	9	5	3	
4	C	3	-12	20	9	23	1	12	7	5	3	
5	D	-6	-10	-1	-3	-20	0	-11	-5	-3	-1	
6	E	3	-12	10	8	13	-6	1	-3	6	4	
7	F	0	13	1	0	20	7	0	1	1	2	
8	G	-2	-6	0	-4	-10	-1	-8	-10	-5	-2	
9	H	3	9	4	3	9	5	1	0	1	1	
10	I	-8	2	-3	6	13	-1	6	9	5	3	
11	J	-4	-10	2	11	-4	-5	-11	8	-1	1	
12												

Re-formated Microsoft Excel table suitable for using with `R`. Needs to be saved as `*.csv` file.

Одоо Excel файлыг R-д уншигдах боломжтой CSV файлуудтай нийцүүлэн өөрчлөцгөөе. Хэрэв та эхний мөрийн толгой хэсгүүд, хүчин төгөлдөр хувьсагчийн нэрээр, өөр форматгүйгээр өөрийн файлыг засч өөрчилж янзлахад ингэж харагдах болно.



Бидний хийх ёстой хамгийн сүүлийн алхам бол уг файлыг CSV форматаар хадгалах явдал юм. Учир нь xlsx файлын өргөтгөлтэй одоогийн өгөгдлийн форматыг унших боломжгүй тул үүнийг хийх шаардлагатай юм. Файлаа CSV файл хэлбэрээр хадгалахын тулд File цэсний Save As дээр дарж файлын нэрийн доор байрлах dropdown цэс дээр дарж "CSV" тохиргооны аль нэгийг сонгоно уу. CSV өгөгдлийн форматад зориулсан хэд хэдэн сонголтууд байж болно. Миний компьютер дээр MS-DOS, Macintosh гэх мэт хувилбарууд байна. Би MS-DOS хувилбарыг сонгохыг зөвлөж байна. Гэхдээ бусад хувилбарууд нь бас ажиллана. Мөн энэ файлыг хадгалахдаа эргээд санах folder/хавтас-ыг сонгохыг зөвлөж байна. Үүнээс гадна та энэ файлыг зүгээр л Desktop дээрээ хадгалах боломжтой.

4 R - reading in data

```
# On windows operating system
read.csv(file = "C:/Users/eric/Desktop/age-guessing-data-excel-exercise.csv")
```

```
##    group card_1 card_2 card_3 card_4 card_5 card_6 card_7 card_8 card_9 card_10
## 1      A    -11    -15     11     -2      0      5     -4      1     -7      -6
## 2      B     -1    -12      4      7      7     -5     12      9      5       3
## 3      C      3    -12     20      9     23      1     12      7      5       3
## 4      D     -6    -10     -1     -3    -20      0    -11     -5     -3      -1
## 5      E      3    -12     10      8     13     -6      1     -3      6       4
## 6      F      0     13      1      0     20      7      0      1      1       2
## 7      G     -2     -6      0     -4    -10     -1     -8    -10     -5      -2
## 8      H      3      9      4      3      9      5      1      0      1       1
## 9      I     -8      2     -3      6     13     -1      6      9      5       3
## 10     J     -4    -10      2     11     -4     -5    -11      8     -1       1
```

Ctrl + R (script editor)

Enter (console)

Бидний компьютерт хадгалагдсан CSV файлаар бид R-г эхлүүлж, нас таах өгөгдлийн багцаа уншуулахад бэлэн боллоо. Үүний тулд R програмыг эхлүүлж, шинэ script нээнэ үү. Дэлгэц дээр харуулсан командыг оруулна уу. Энэ команд нь read.csv нэртэй функцийг дууддаг бөгөөд бид file гэсэн ганц параметрийг бэлдэнэ. Энэ параметр нь бидний уншихыг хүсч буй файлын нэрийг агуулдаг. Хэрэв та өөрийн CSV файлыг өөр директорт хадгалсан бол шууд налуу (/) тэмдгийг ашиглан тухайн файл руу

чиглэсэн бүтэн замыг зааж өгөх шаардлагатай болж магадгүй юм. Хэрэв та энэ командыг R script editor-т бичсэн бол Ctrl + R товчийг дарж, хэрэв та консол дээр ажиллаж байгаа бол Enter товчийг дарж командыг ажиллуулна уу. Хэрэв та командыг зөв бичсэн бөгөөд зам нь мөн зөв бол слайд дээр харуулсны дагуу R нь өгөгдлийн файлын агуулгыг хүснэгт болгон консол дээр хэвлэнэ.

```
# install.packages("data.table")
library(data.table)
dt <- fread(input = "age-guessing-data-excel-exercise.csv")
dt
```

За, одоо нэг алхам урагшлъя. Хэрэв бид өгөгдлийн багц, график дээрээ тооцоо хийхийг хүсвэл хувьсагч дотор хадгалах шаардлагатай. Өгөгдлийн багцыг хадгалах, түүнтэй ажиллахад хялбар өгөгдлийн төрөл бол data.table юм. Энэ нь шууд R-д орж ирдэггүй өгөгдлийн төрөл юм. Тиймээс бид R багцын data.table-ийг суулгах хэрэгтэй. Хэрэв та энэ багцыг аль хэдийн суулгасан бол үүнийг зөвхөн нэг удаа суулгах шаардлагатай тул энэ алхам шаардлагагүй болно. Хэрэв бид R багцыг суулгасан бол үүнийг бас ачаалах шаардлагатай. Үүнийг library () функцийг багцын нэрээр параметр болгон дуудаж гүйцэтгэнэ. Хэрэв багц ачаалагдсан бол бид өөр функцийг ашиглан CSV файл дээрээ унших боломжтой. Энэ бол "файл унших" гэсэн утгатай fread нэртэй функц бөгөөд бид CSV файлын байршлыг өмнөх шигээ дамжуулдаг input/оролт нэртэй дан ганц параметрийг бэлтгэнэ. Шууд консол дээр ажилладаг хүмүүс дээшээ сумыг ашиглан өмнөх командуудыг гүйлгэн харах боломжтой гэдгийг санаарай. Манай CSV файлд унших fread функцаас гадна үр дүнг бид энэ тохиолдолд dt /data.table/ гэж нэрлэсэн шинэ хувьсагчид хадгална. Нэгэнт бид CSV файлыг уншиж, хувьсагч дотор хадгалснаас цааш хувьсагчийн нэрийг команд болгож оруулснаар хувьсагчийн агуулгыг дахин хэвлэх боломжтой юм.

```
##     group card_1 card_2 card_3 card_4 card_5 card_6 card_7 card_8 card_9
##  1:     A    -11    -15     11     -2      0      5     -4      1     -7
##  2:     B     -1    -12      4      7      7     -5     12      9      5
##  3:     C      3    -12     20      9     23      1     12      7      5
##  4:     D     -6    -10     -1     -3    -20      0    -11     -5     -3
##  5:     E      3    -12     10      8     13     -6      1     -3      6
##  6:     F      0     13      1      0     20      7      0      1      1
##  7:     G     -2     -6      0     -4    -10     -1     -8    -10     -5
##  8:     H      3      9      4      3      9      5      1      0      1
##  9:     I     -8      2     -3      6     13     -1      6      9      5
## 10:     J     -4    -10      2     11     -4     -5    -11      8     -1
##     card_10
##  1:      -6
##  2:       3
##  3:       3
##  4:      -1
##  5:       4
##  6:       2
##  7:      -2
##  8:       1
##  9:       3
## 10:       1
```

Слайд дээрээс харж байгаачлан R нь csv файлын агуулгыг дахин хэвлэж байна

```
dt[, total_error := abs(card_1) +
   abs(card_2) +
   abs(card_3) +
   abs(card_4) +
   abs(card_5) +
   abs(card_6) +
   abs(card_7) +
   abs(card_8) +
   abs(card_9) +
   abs(card_10)]
dt
```

- Function `abs(x)`: Compute absolute value of x
- Operator `:=`: Create new column in data table

Одоо сонирхолтой хэсэг рүүгээ орьё: Бид өгөгдлийн багц дээрээ зарим тооцоог хийдэг. Энэ слайд дээрх

код нь := операторыг хэрхэн ашиглаж бидний өгөгдлийн хүснэгтэд шинэ багана үүсгэхийг харуулж байна. Энэ баганыг бусад баганын утгыг үндэслэн тооцоолсон утгуудаар дүүргэх болно. Хэрэв та бидний нас таасан туршилтыг санаж байгаа бол бид баг бүрийн нийт таалтын алдааг гар аргаар тооцоолсоныг санаж байгаа байх. Бид R-д түүнтэй ижилхэн тооцоог хийж болох бөгөөд энэ нь илүү хурдан бөгөөд алдаа багатай байх болно. Бид data.table dt нэрний ард дөрвөлжин хаалт нээж бичээд л болоо. Дөрвөлжин хаалт нээх нь R-д бид data.table дотор ямар нэгэн зүйл хийх гэж байгаагаа илэрхийлж байгаа юм. Энэ кодонд битгий толгой эргээрэй, энэ нь юу болохыг дараа дэлгэрэнгүй тайлбарлах болно. Гол нь бид total_error нэртэй шинэ багана үүсгэж, түүний утгыг card_1, card_2 гэх мэтээр card 10 хүртэлх баганууудын абсолют утгуудын нийлбэрээс тооцох явдал юм. Бид dt/data.table/ хувьсагчийн агуулгыг хэвлэж, зөв хийсэн эсэхээ дахин шалгаж болно.

```
##     group card_1 card_2 card_3 card_4 card_5 card_6 card_7 card_8 card_9
##  1:     A    -11    -15     11     -2      0      5     -4      1     -7
##  2:     B     -1    -12      4      7      7     -5     12      9      5
##  3:     C      3    -12     20      9     23      1     12      7      5
##  4:     D     -6    -10     -1     -3    -20      0    -11     -5     -3
##  5:     E      3    -12     10      8     13     -6      1     -3      6
##  6:     F      0     13      1      0     20      7      0      1      1
##  7:     G     -2     -6      0     -4    -10     -1     -8    -10     -5
##  8:     H      3      9      4      3      9      5      1      0      1
##  9:     I     -8      2     -3      6     13     -1      6      9      5
## 10:     J     -4    -10      2     11     -4     -5    -11      8     -1
##     card_10 total_error
##  1:      -6          62
##  2:       3          65
##  3:       3          95
##  4:      -1          60
##  5:       4          66
##  6:       2          45
##  7:      -2          48
##  8:       1          36
##  9:       3          56
## 10:       1          57
```

Энэ слайд дээрээс харахад бид хүссэнээрээ шинэ багана үүсгэлээ. R нь мөр бүрийн үйлдлийг тус тусад нь автоматаар гүйцэтгэсэн байна.

```
dt[, mean_error := total_error / 10]
head(dt)
```

```
##    group card_1 card_2 card_3 card_4 card_5 card_6 card_7 card_8 card_9 card_10
## 1:     A    -11    -15     11     -2      0      5     -4      1     -7      -6
## 2:     B     -1    -12      4      7      7     -5     12      9      5       3
## 3:     C      3    -12     20      9     23      1     12      7      5       3
## 4:     D     -6    -10     -1     -3    -20      0    -11     -5     -3      -1
## 5:     E      3    -12     10      8     13     -6      1     -3      6       4
## 6:     F      0     13      1      0     20      7      0      1      1       2
##    total_error mean_error
## 1:          62        6.2
## 2:          65        6.5
## 3:          95        9.5
## 4:          60        6.0
## 5:          66        6.6
## 6:          45        4.5
```

- Function `head(x)`: Print the first 6 rows of x

Нас таах туршилтийн үеэр бид хамгийн сайн ажилласан баг байгаа эсэхийг сонирхож байсан. Бид дундаж алдааг тооцоолох замаар багуудыг харьцуулж болохыг мэдэж авсан. R дээр бас үүнийг адилхан хийж үзэцгээе. Бид дахин data.table дотор шинэ багана үүсгэе. Энэ удаад бид сая үүсгэсэн total_error баганы гэрэл зургийн тоо болох 10-д хувааж ашиглаж байна. Бүх зүйл зөв болсон эсэхийг дахин шалгана. Data.table-ийн агуулгыг хэвлэх өөр нэг практик арга бол head функцийг ашиглах явдал юм. Энэ функц нь ямар үүрэгтэйг нэрнээс нь аль хэдийн мэдэж болж байна: Энэ нь зөвхөн өгөгдлийн хүснэгтийн эхний мөрүүдийг хэвлэнэ. Хэрэв өгөгдлийн хүснэгт нь консол дээр хэвлэгдэх хэт олон мөр жишээ нь 1000 мөрийг агуулж байгаа бол энэ нь байх зүйл юм. Head нь анхдагч байдлаар эхний 6 мөрийг хэвлэнэ.

```
fwrite(x = dt, filename = "my-data-set.csv", sep = ";")
```

- Function `fwrite(x)`: Save data.table x as `*.csv` file

- ... as file named `filename`
- ... using separating character `sep`

Бид өгөгдлийн багцын нарийн төвөгтэй тооцооллыг хийсний дараа үр дүнгээ хадгалахыг хүсч байгаа. Бид R-д CSV файлуудыг уншиж чаддаг шигээ R-ээс CSV файлууд руу өгөгдөл бичих чадвартай байдаг. Энэ нь маш энгийн бөгөөд үүнийг fwrite функцийг ашиглан гүйцэтгэж болно. Энэ функц нь хэд хэдэн параметрийг авдаг. Энэ тохиолдолд бид бидний өгөгдлийн багцыг агуулсан x нэртэй хувьсагчийг, CSV файлаа хаана хадгалахаа зааж өгсөн файлын нэрийг, мөн ямар тэмдэгтийг нүдний тусгаарлагч болгон ашиглахыг зааж өгсөн sep параметрийг бэлддэг. Энд харуулсан команд нь тусгаарлагчийг таслалаар зааж өгөхийг харуулж байна, гэхдээ энэ нь анхдагч тусгаарлагч учраас шаардлагагүй юм. Хадгалагдсан файлыг Excel дээр нээж хадгалалт амжилттай болсон эсэхийг шалгаж болно. Энэ бол миний тус хичээлээр заахыг хүссэн зүйл маань юм. Одоо би R давтлагын хичээлийн үеэр болон гэртээ ажиллах өөр дасгалыг та бүхэнд өгье.

5 Exercises

5.1 Creating data sets

5.1.1 c()-function

Epidemiological data set: Observations of red deer and wild boar, tubercolosis (tb), parasite (ecervi) (Zuur et al., 2009)

farm	month	year	sex	length.class	length.ct	ecervi	tb
MO	11	2000	1	1	75.0	0	0
MO	7	2000	2	1	85.0	0	0
MO	7	2001	2	1	91.6	0	1
MO	NA	NA	2	1	95.0	NA	NA
LN	9	2003	1	1	NA	0	0
SE	9	2003	2	1	105.5	0	0
QM	11	2002	2	1	106.0	0	0

1. Use `c()` to create a vector of the `tb` values. Store the result in a variable named `tb`.

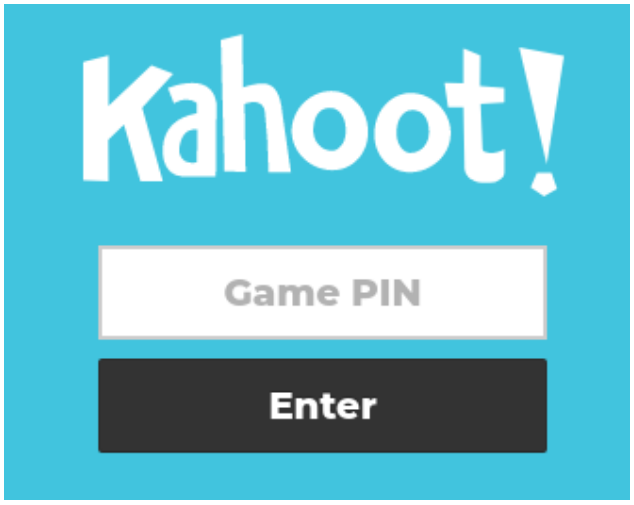
Энэ дасгал дээр та өгөгдлийн багц үүсгэх дадлага хийх болно. Энд харагдаж байгаа өгөгдлийн багц нь Испанид зэрлэг гахай, буга зэргийг хашаанд ажиглан хийсэн өвчний тархвар судлалын хэмжилтүүдийг агуулсан болно. Тус амьтдыг сүрьеэ болон тусгай төрлийн шимэгч хорхойгоор халдварласан эсэхийг шинжлэсэн байгаа. Нэмж дурдахад амьтдын бусад шинж чанарууд, мөн бас байршил, цаг хугацааны талаархи мэдээллийг бас бүртгэж оруулсан болно. Эхний даалгавар бол энэ хүснэгтийн "tb" /tuberculosis/ баганын утгыг агуулсан вектор үүсгэх R script-ийг бичих явдал юм.

5.1.2 list()-function

2. Use `c()` to create a second vector of the `length.ct` values. Store the result in a variable named `length.ct`.
3. Use `list()` to create a list using the two vectors `length.ct` and `tb`. Store the result as `deer`
4. Use the function `as.data.table()` to turn `deer` into a data.table. Print the result to the console.

Дасгал 2 болон 3-т дахин өөр нэг вектор үүсгээд түүнийгээ хувьсагч хэлбэрээр хадгалж, дараа нь үүсгэсэн хоёр векторыг хувьсагч хэлбэрээр хадгалагдах list-д нэгтгэнэ үү. Дасгал 4-т үүсгэсэн list-ийг ашиглаж, as.data.table() функцийг ашиглан list-ийг data.table болгон хөрвүүлж, гарсан объектыг хэвлэнэ үү.

6 Quiz



- Phone or Computer: https://www.kahoot.it
- Wifi: Platinum
- Password: H2SvsH2O

References

Song S Qian. *Environmental and ecological statistics with R*. CRC press, 2016.

Alain Zuur, Elena N Ieno, and Erik Meesters. *A Beginner's Guide to R*. Springer Science & Business Media, 2009.