

RMarkdown 1

Abari Kálmán

2021-09-16

Contents

Óravázlat	1
Ismétlő példa megoldása	2
Szöveg formázása	2
Táblázat beszúrása	2
Képek beszúrása	3
R csonkok	6
Képletek beszúrása	7
Hivatkozásra példák	8
Irodalomjegyzék	8

Óravázlat

1. Dokumentációk megbeszélése: <https://abarik.github.io/statisztika/>
2. A munka feltételeinek megbeszélése (az R a gyakorlatban videótutorial alapján):
 - Alap R
 - RStudio
 - RStudio alapbeállításai
 - a projektes használat
3. Az előző féléves anyag áttekintése: `r_parancsok.html`
4. Az `r_parancsok.html`-ből egy konkrét példa megoldása .R kiterjesztésű parancsállomány: *Géneexpressziós adatokból adott gének kiválasztása*
5. Az .Rmd kiterjesztésű RMarkdown állományok megbeszélése. Fontos segítség: RStudio Cheatsheets
6. Az RMarkdown fejléc szerkesztése
7. Az RMarkdown szövegek szerkesztése
 - Szöveg formázása (pl. címek, félkövér, dőlt, írógép, idézet, új sor, új bekezdés, felsorolás, számozás)
 - Latex képletek (bekezdés, inline)
 - Link és kép beszúrása
 - Táblázat beszúrása
 - Lábjegyzet beszúrása
 - Irodalomjegyzék és szövegek közötti hivatkozások beszúrása
8. Az RMarkdown R parancsok kezelése
 - Inline R parancsok
 - R csonkok

Ismétlő példa megoldása

Génexpressziós adatokból adott gének kiválasztása

Olvassuk be a `data_RNA_Seq_v2_mRNA_median_Zscores_IDC_lumA_v2.txt` állományt, amely Luminal A betegekre vonatkozó génexpressziós adatokat tartalmaz. A sorokban az egy génre vonatkozó adatok találhatók. Az első oszlopban a gén neve, a további 200 oszlopban 200 beteg adata szerepel. A megjelenő számok a génkifejeződés mértéke az egészséges normához képest. Olvassuk be a `gene_list.txt` állományt is, amely a számunkra érdekes gének nevét tartalmazza, és ez alapján szűrjük le a fenti adatbázist (használjuk az `%in%` relációs operátort).

Szöveg formázása

A mindennapi életben elének kerülő adatok a legalapvetőbb természetük alapján legalább 4 csoportba sorolhatók, *számszerű* (numerikus), *karakteres* (sztring), *logikai* (állítások igazságtartalma, igaz vagy hamis) és *dátum/idő* jellegű adatokra. A statisztika tudománya is elvégzi az adatok csoportosítását, e szerint egyrésztől megkülönböztetünk *nominális*, *ordinális*, *intervallum* és *arányskálájú* változókat (az alapján, hogy mit tudunk tenni a változó értékeivel), másrészt *diszkrét* és *folytonos* változókat (az alapján, hogy hány értéke van a változónak). Az adatok jellegének harmadik megközelítését adják az adott statisztikai programcsomag adattípusai, az R-ben például számunkra a legfontosabbak a **double**, **integer**, **karakteres**, **logikai**, **faktor**, **dátum** és **dátum-idő**. Az R nyelv adattípusait és kapcsolatukat a statisztikai változókategóriákkal és a mindennapi életben használt természetes felosztással a lenti táblázat tartalmazza.

Az R adattípusokat az 1. oszlop listázza. Az `str()` és `glimpse()` függvényben az adattípus jelölésére használt rövidítések a 2. oszlopban találhatók. A statisztikai skálák és az R adattípusok lehetséges összerendelésit a 3. oszlop mutatja. E szerint **egy adatbázisban többnyire double, integer vagy faktor oszlopokat várunk**. A statisztikai változók számossága és az R adattípus közötti összefüggés a 4. oszlopban olvasható. Az utolsó oszlopban az adott adattípus mindennapi életben használatos természetét emeltük ki¹.

A táblázat a 4.1.1 R verzió alapján készült.

Táblázat beszúrása

Table 1: Az R adattípusainak jellemzése.

R adattípus	<code>str()</code> , <code>glimpse()</code>	Skála	Számosság	Természet
double	<code>num</code> , <code>dbl</code>	intervallum/arány	folytonos/diszkrét	Numerikus
integer	<code>int</code> , <code>int</code>	ordinális/intervallum/arány	diszkrét	Numerikus
karakteres	<code>chr</code> , <code>chr</code>	-	-	Karakteres
logikai	<code>logi</code> , <code>lgl</code>	-	-	Logikai
faktor	<code>Factor</code> , <code>fct</code>	nominális/ordinális	diszkrét (kategorikus)	Karakteres
dátum	<code>Date</code> , <code>date</code>	-	-	Dátum/idő
dátum és idő	<code>POSIXct</code> , <code>dtm</code>	-	-	Dátum/idő

```
knitr::kable(  
  iris[1:10, ], longtable = TRUE, booktabs = TRUE,  
  caption = 'A table generated by the longtable package.'  
)
```

¹Természetesen más felosztás is elképzelhető.

Table 2: A table generated by the longtable package.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

Képek beszúrása



Figure 1: Az R kényelmes használata



Figure 2: Az R kényelmes használata

```
smoc2_rawcounts <- read.csv("data/fibrosis_smoc2_rawcounts_unordered.csv")
rownames(smoc2_rawcounts) <- smoc2_rawcounts$X
smoc2_rawcounts$X <- NULL

# Explore the first six observations of smoc2_rawcounts
head(smoc2_rawcounts)
```



Figure 3: Az R kényelmes használata

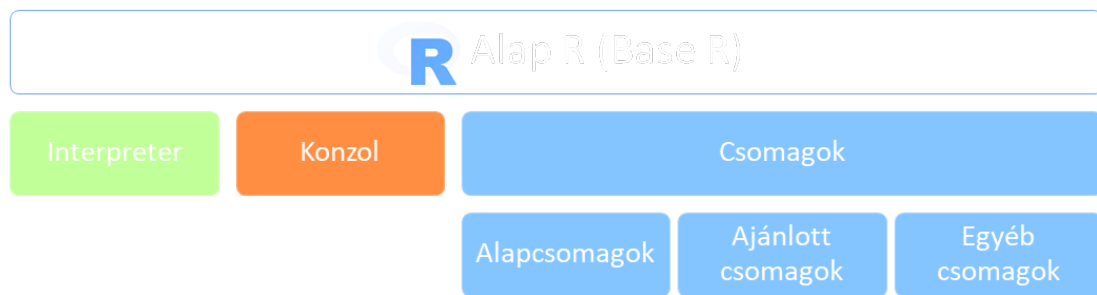


Figure 4: Az Alap R





	Konzol	Konzol + Grafikus felület
	R.exe	Rgui.exe
	R	
	R.app	

Figure 5: A konzolok

		Konzolos	Parancsállományos	Grafikus felületről
Alap R	Alap R Konzol	✓		
Alap R - Windows	RGui	✓	✓	
Alap R - Egyéb csomag	R Commander		✓	✓
rstudio.com	RStudio	✓	✓	
www.jamovi.org	jamovi			✓
jasp-stats.org	JASP			✓
www.blueskystatistics.com	BlueSky			✓

Figure 6: Az R használati módjai

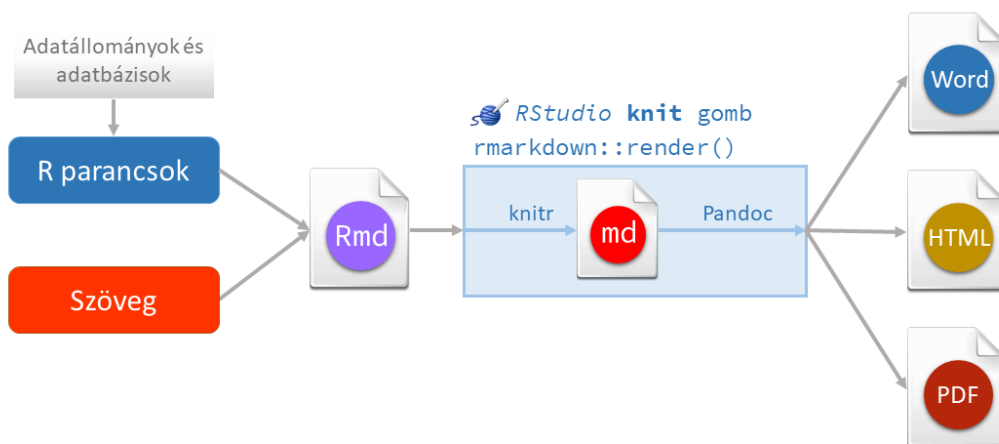


Figure 7: Az RMarkdown működése

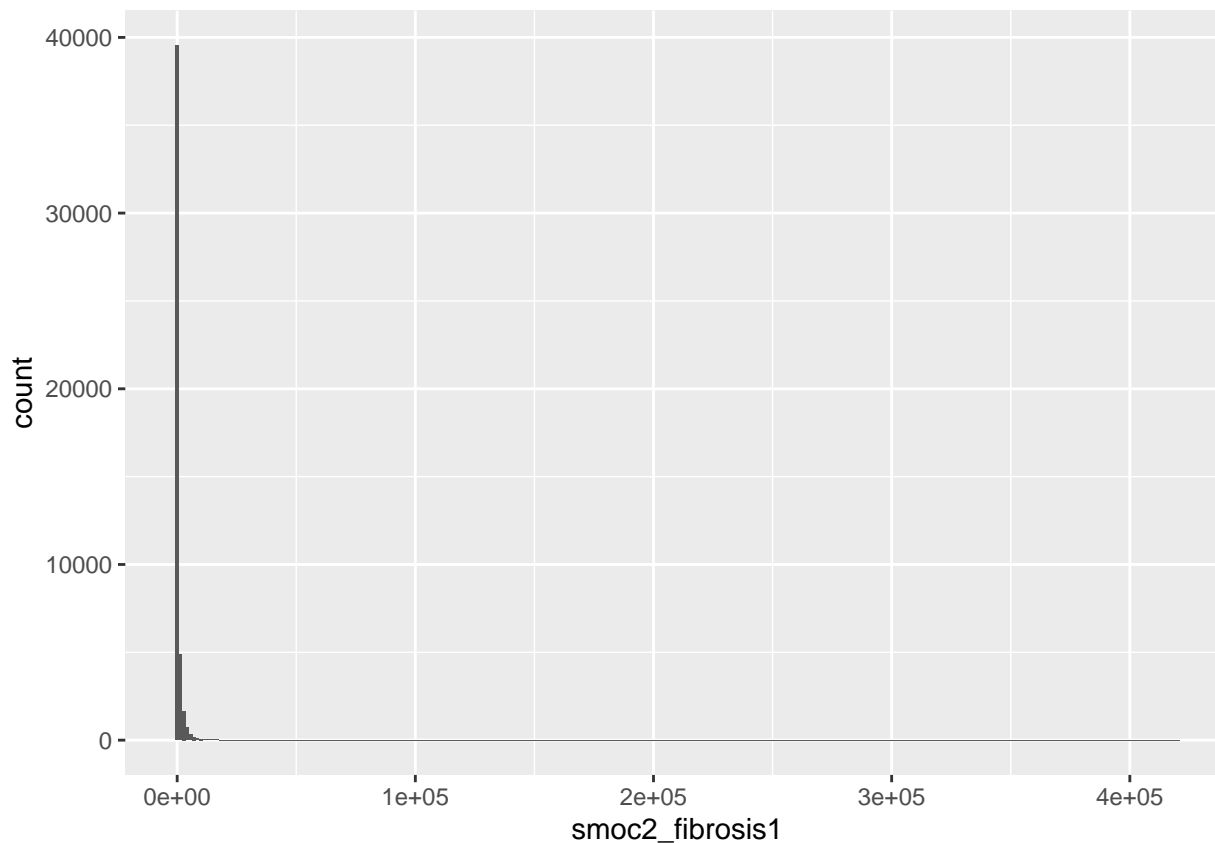
```
##          smoc2_fibrosis1 smoc2_fibrosis4 smoc2_normal1 smoc2_normal3 smoc2_fibrosis3
## ENSMUSG00000102693      0          0          0          0          0
## ENSMUSG00000064842      0          0          0          0          0
## ENSMUSG00000051951     72         30          0          3         36
## ENSMUSG00000102851      0          0          0          0          0
## ENSMUSG00000103377      0          0          1          0          0
## ENSMUSG00000104017      0          0          0          0          0
##          smoc2_normal4 smoc2_fibrosis2
## ENSMUSG00000102693      0          0
## ENSMUSG00000064842      0          0
## ENSMUSG00000051951      1         51
## ENSMUSG00000102851      0          0
## ENSMUSG00000103377      0          0
## ENSMUSG00000104017      0          0
```

```
# Explore the structure of smoc2_rawcounts
str(smoc2_rawcounts)
```

```
## 'data.frame':   47729 obs. of  7 variables:
## $ smoc2_fibrosis1: int  0 0 72 0 0 0 0 0 0 1 ...
## $ smoc2_fibrosis4: int  0 0 30 0 0 0 0 0 0 1 ...
## $ smoc2_normal1  : int  0 0 0 0 1 0 0 0 0 1 ...
## $ smoc2_normal3  : int  0 0 3 0 0 0 0 0 0 0 ...
## $ smoc2_fibrosis3: int  0 0 36 0 0 0 0 0 0 1 ...
## $ smoc2_normal4  : int  0 0 1 0 0 0 0 0 0 0 ...
## $ smoc2_fibrosis2: int  0 0 51 0 0 0 0 0 0 1 ...
```

R csonkok

```
library(ggplot2)
ggplot(smoc2_rawcounts) +
  geom_histogram(aes(x=smoc2_fibrosis1), bins = 300)
```



```
summary(smoc2_rawcounts$smoc2_fibrosis1)
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
##      0.0      0.0      1.0    579.7   184.0 420026.0
```

Képletek beszúrása

Diszperziós formula: $Var = \mu + \alpha \times \mu^2$

- Var : variancia
- μ : átlag
- α : diszperzió

A negatív binomiális modell:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = x_j \beta_j$$

- K_{ij} - nyers count az i. génben a j. mintában
- s_{ij} - size factor
- g_{ij} - normalizált count

A mintaátlag kiszámítása: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

A minta szórása:

$$s^* = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Hivatkozásra példák

A lenti workflow (Piper, 2020) alapján készült. Az elemzéshez a **DESeq2** csomagot használtuk (Love és mtsai., 2014) és figyelembe vettük Love és mtsai. (2020) leírását is.

A workflow-ban használt kutatási kérdések és adatbázisok a Gerarduzzi és mtsai. (2017) publikációján alapulnak, az RNS-szekvenálási adatok a Gene Expression Omnibus (GEO) adatbázisból letölthetők (GEO accession: GSE85209).

A szövegek közötti hivatkozás esetei:

- (Love és mtsai., 2014)
- Love és mtsai. (2014)
- (2014)
- (Gerarduzzi és mtsai., 2017; Love és mtsai., 2014)
- Love és mtsai. (2014p. 33)
- (Love és mtsai., 2014, pp. 33-35; Gerarduzzi és mtsai., 2017, ch. 1)

Irodalomjegyzék

- Gerarduzzi, C., Kumar, R. K., Trivedi, P., Ajay, A. K., Iyer, A., Boswell, S., Hutchinson, J. N., Waikar, S. S. és Vaidya, V. S. (2017). Silencing SMOC2 ameliorates kidney fibrosis by inhibiting fibroblast to myofibroblast transformation. *JCI Insight*, 2(8). <https://doi.org/10.1172/jci.insight.90299>
- Love, M. I., Anders, S. és Huber, W. (2020). *Analyzing RNA-seq data with DESeq2*. <https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>
- Love, M. I., Huber, W. és Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12). <https://doi.org/10.1186/s13059-014-0550-8>
- Piper, M. (2020). *RNA-Seq with Bioconductor in R*. <https://learn.datacamp.com/courses/rna-seq-with-bioconductor-in-r>