RMarkdown 1

Abari Kálmán

2021-09-25

Table of Contents

# 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 előzó féléves anyag áttekintése: r\_parancsok.html
4. Az r\_parancsok.html-ből egy konkrét példa megoldása .R kiterjesztésű parancsállomány: *Génexpressziós adatokból adott gének kiválasztása*
5. Az .Rmd kiterjesztésű RMarkdown állományok megbeszélése. Fontos segítség: [RStudio Cheatsheets](https://rstudio.com/resources/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 szövegközi 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 alaján szűrjük le a fenti adatbázist (használjuk az %in% relációs operátort).

# Szöveg formázása

A mindennapi életben elénk 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 egyrészről 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 ki[[1]](#footnote-24).

A táblázat a 4.1.1 R verzió alapján készült.

# Táblázat beszúrása

Az R adattípusainak jellemzése.

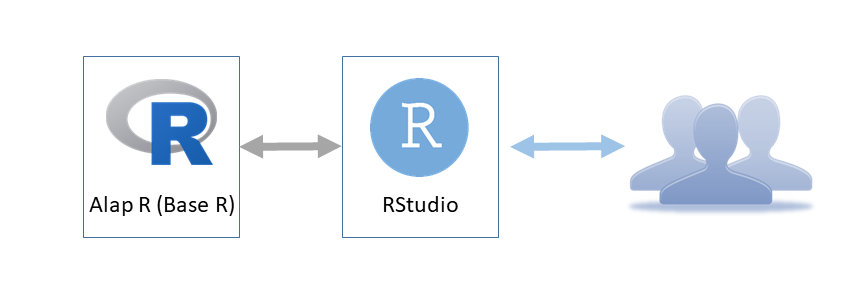
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| R adattípus | str(), glimpse() | Skála | Számosság | Természet |
| double | num, dbl | intervallum/arány | folytonos/diszkrét | Numerikus |
| integer | int, int | ordinális/intervallum/arány | diszkrét | Numerikus |
| karakteres | chr, chr | - | - |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

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

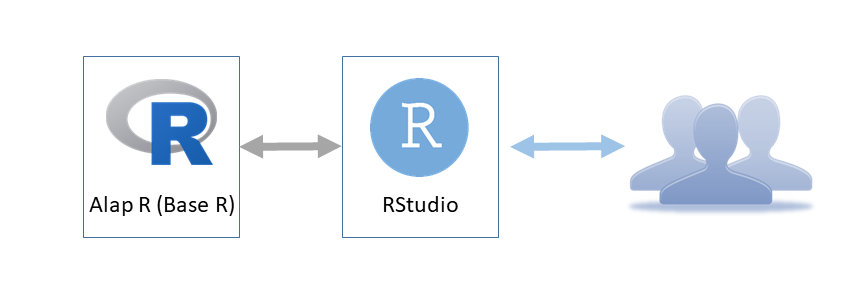
A table generated by the longtable package.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

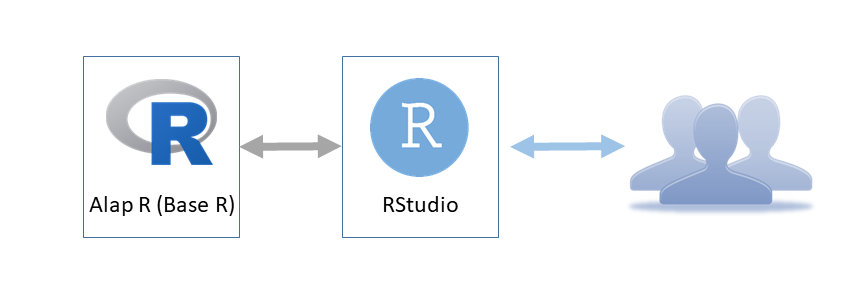
# Képek beszúrása



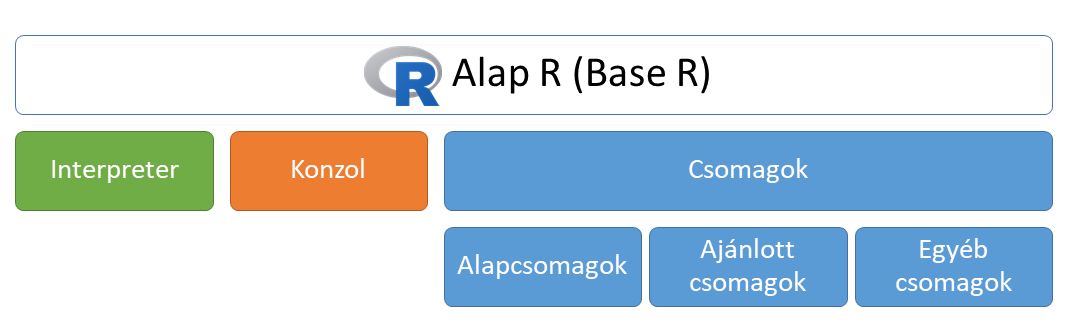
Az R kényelmes használata



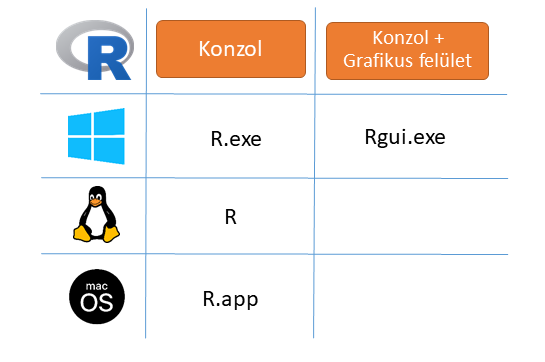
Az R kényelmes használata



Az R kényelmes használata



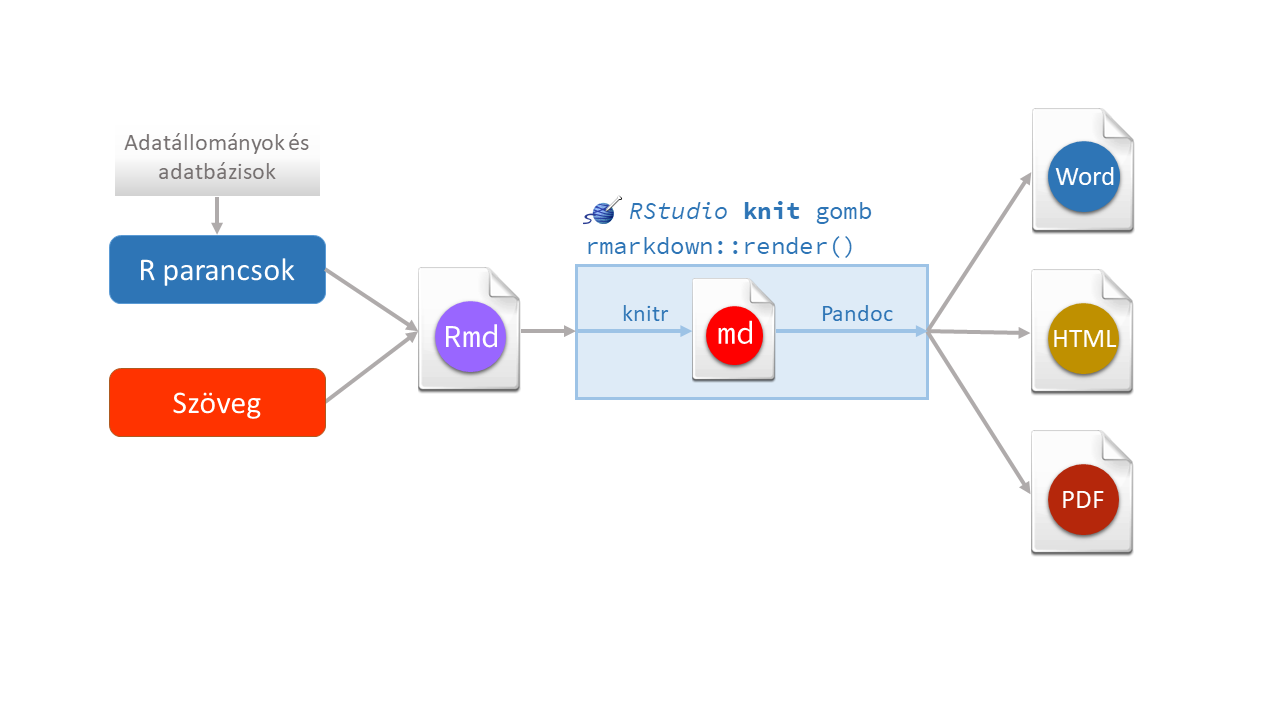
Az Alap R



A konzolok



Az R használati módjai



Az RMarkdown működése

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)

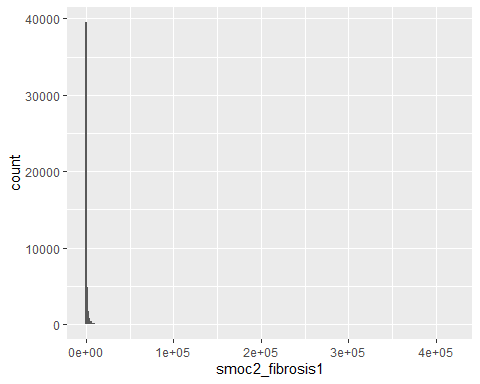
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | smoc2\_fibrosis1 | smoc2\_fibrosis4 | smoc2\_normal1 | smoc2\_normal3 | smoc2\_fibrosis3 | smoc2\_normal4 | smoc2\_fibrosis2 |
| ENSMUSG00000102693 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ENSMUSG00000064842 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ENSMUSG00000051951 | 72 | 30 | 0 | 3 | 36 | 1 | 51 |
| ENSMUSG00000102851 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ENSMUSG00000103377 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| ENSMUSG00000104017 | 0 | 0 | 0 | 0 | 0 | 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:

* : variancia
* : átlag
* : diszperzió

A negatív binomiális modell:

* - nyers count az i. génben a j. mintában
* - size factor
* - normalizált count

A mintaátlag kiszámítása:

A minta szórása:

# Hivatkozásra példák

A lenti workflow ([Piper, 2020](#ref-Pipperrana)) alapján készült. Az elemzéshez a **DESeq2** csomagot használtuk ([Love és mtsai., 2014](#ref-Love2014)) és figyelembe vettük [Love és mtsai.](#ref-anrlove2020) ([2020](#ref-anrlove2020)) [leírását](https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html) is.

A workflow-ban használt kutatási kérdések és adatbázisok a [Gerarduzzi és mtsai.](#ref-Gerarduzzi2017) ([2017](#ref-Gerarduzzi2017)) publikációján alapulnak, az RNS-szekvenálási adatok a Gene Expression Omnibus (GEO) adatbázisból letölthetők (GEO accession: [GSE85209](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85209)).

A szövegközi hivatkozás esetei:

* ([Love és mtsai., 2014](#ref-Love2014))
* [Love és mtsai.](#ref-Love2014) ([2014](#ref-Love2014))
* ([2014](#ref-Love2014))
* ([Gerarduzzi és mtsai., 2017](#ref-Gerarduzzi2017); [Love és mtsai., 2014](#ref-Love2014))
* [Love és mtsai.](#ref-Love2014) ([2014](#ref-Love2014)p. 33)
* ([Love és mtsai., 2014](#ref-Love2014), pp. 33-35; [Gerarduzzi és mtsai., 2017](#ref-Gerarduzzi2017), 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>

1. Természetesen más felosztás is elképezhető. [↑](#footnote-ref-24)