Sumarverkefni

Dæmi um skýrslu

1 Inngangur

Stutt sýnidæmi um hvernig skal skrifa skýrslu í RMarkdown með version-control.

1.1 Uppsetning

Fyrst skal:

1. Búa til nýtt repo á github (helst með README).
2. Í RStudio:
    - File -> New Project -> Version Control
    - Copy/Paste slóðina á repo (í mínu tilfelli (https://github.com/thorj/verkferli))

Þá ætti repo-ið að birtast á tölvunni ykkar. Það inniheldur bara README skrá en við bætum í það.

3. Stofna viðeigandi möppur (t.d. `scripts`, `data`, `img`, etc).
4. Prufa að commit og push á git í gegnum RStudio (líka hægt í terminal en þetta er "noob-vænna")

Getið notað ssh eða venjulegt log-in. Sjá nánar hér.

2 Sýnidæmi

Ef það er verið að vinna með stór gagnasett er tímasóun að vinna allt í einni Rmd skrá sem þarf að endurkeyra kóðann í hvert skipti sem hún er prjónuð. Betra að búta vinnsluna niður í nokkrar skriptur með vel skilgreind hlutverk. Til dæmis:

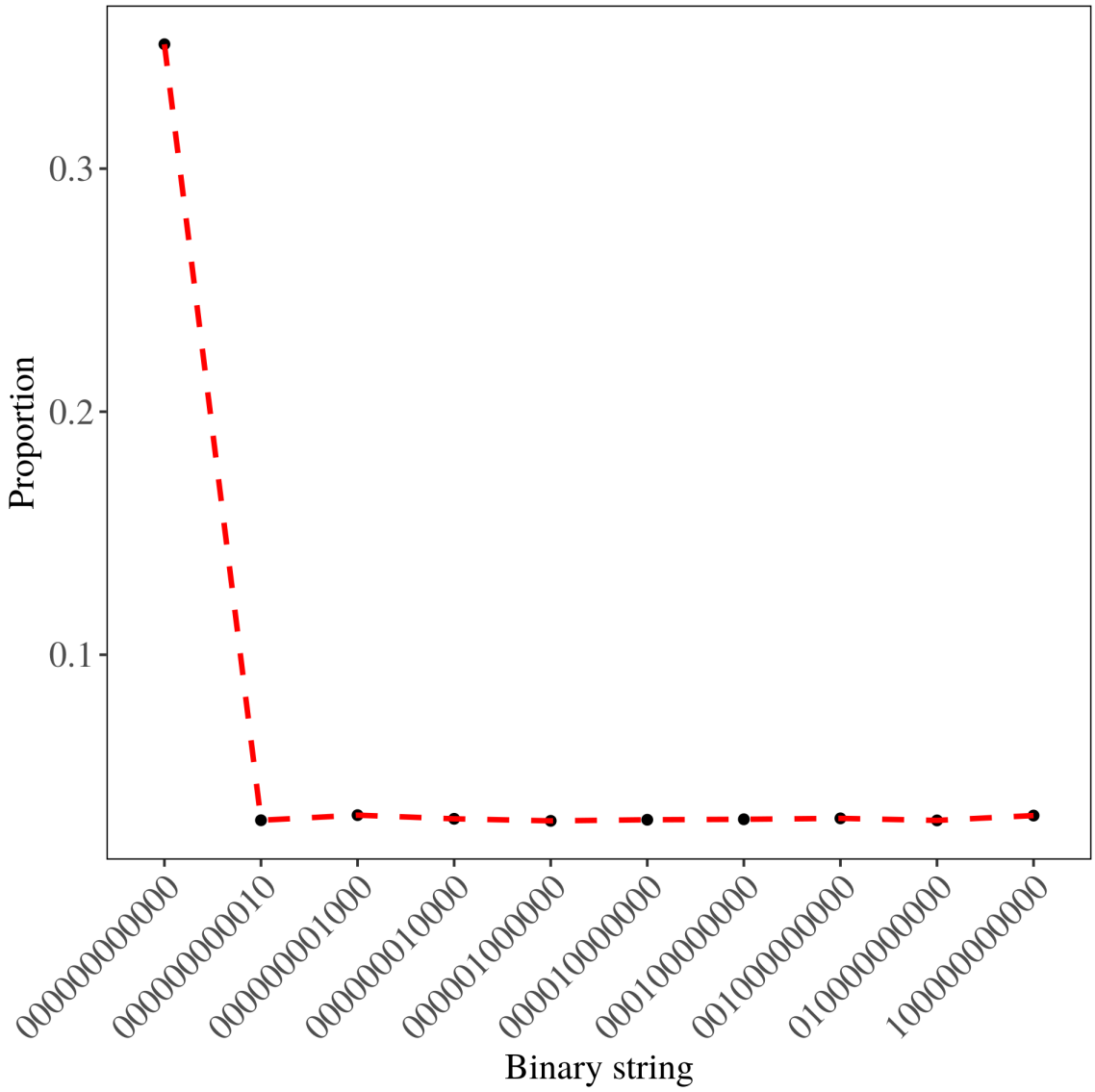
1. `settings.R`: Hleður inn öllum viðeigandi R pökkum og yfirskrifar default `ggplot2` stillingar.
2. `simulate_binomial.R`: Skripta sem býr til "stórt" gagnasett með óskilvirkum kóða. Ef hann væri í Rmd skránni tæki nokkrar mín að prjóna skjalið í hvert sinn.
3. `binomial_wrangl.R`: Eftir að það er búið að búa til gögnin með `simulate_binomial.R` þá eru allar myndir og töflur gerðar með þessari skriptu.

2.1 Lýsing

Til að búa til stutt sýnidæmi voru 50000 tvíkosta strengir hermdir og geymdir ásamt summu strengjanna. Algengustu 10 strengirnir voru fundnir ásamt hlutfalli þeirra af heildarfjölda allra strengja. Strengina ásamt hlutföllum þeirra má sjá í töflu 1 og á mynd 1.

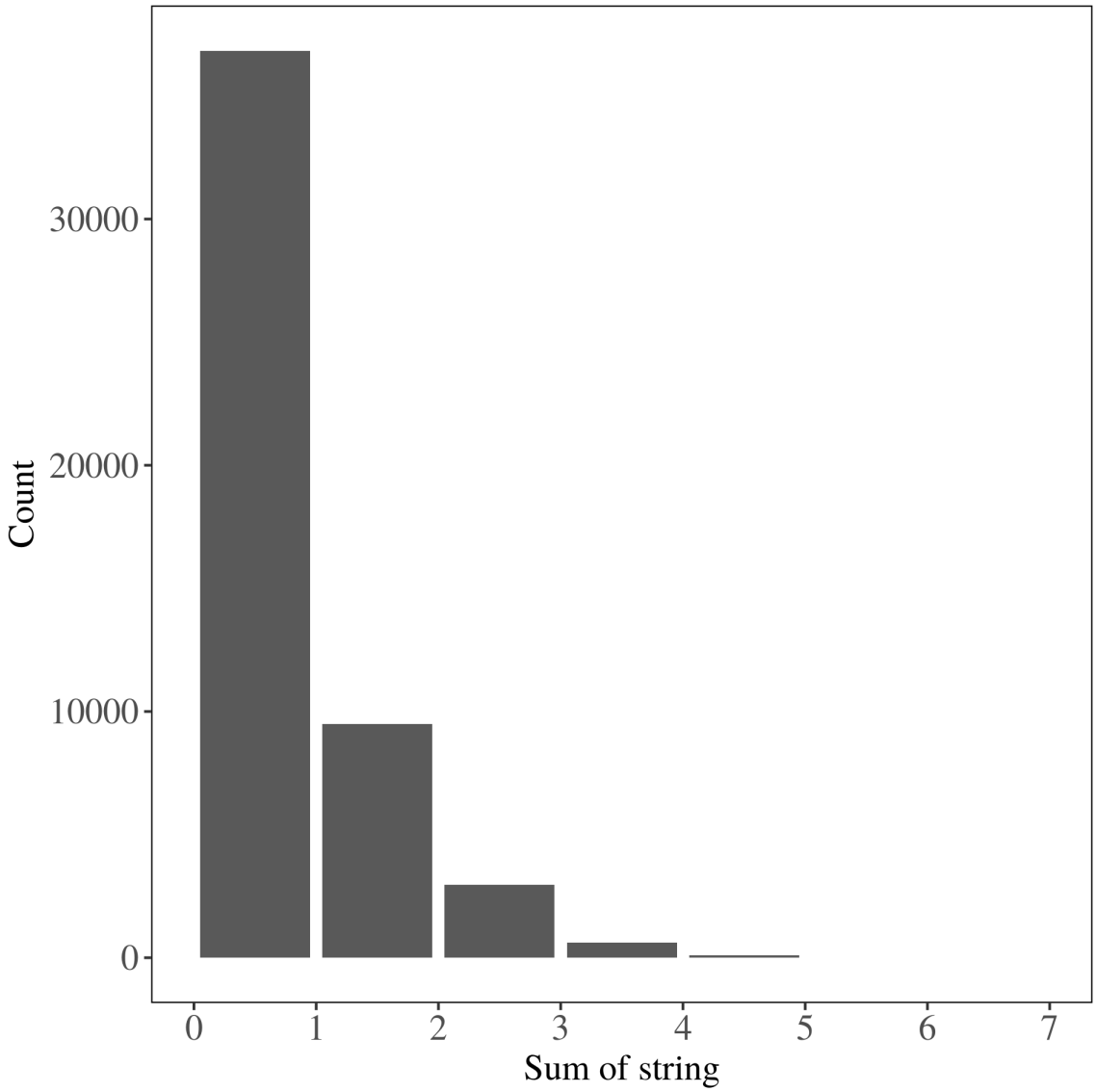
Tafla 1: Hér má sjá 10 algengustu strengina ásamt hlutfalli þeirra.

Strengur	**Hlutfall**
000000000000	0.35118
000000001000	0.03400
100000000000	0.03384
001000000000	0.03268
000000010000	0.03250
000100000000	0.03230
000010000000	0.03210
000000000010	0.03190
010000000000	0.03184
000001000000	0.03166



Mynd 1: Tíu algengustu strengirnir ásamt hlutfalli þeirra. Hér sést greinilega að 0-strengurinn er algengastur sem er afleiðing að því að strengirnir voru hermdir með litlu p.

Það var líka áhugi fyrir því að teikna stöplarit (súlurit?) af summu strengjanna. Hana má sjá á mynd 2.



Mynd 2: Mjög áhugaverð mynd.

2.2 Umræða

Tekur enga stund að keyra þessa skrá. Ef það þarf að breyta myndum eða töflum er það gert í `binomial_wrangl.R` og svo endurprjóna `index.Rmd`. Allar tölur eru birtar með vísum í minni svo það þarf ekki að endurskrifa neinar tölur. Hef notað þessa uppsetningu til að skila heimadæmum og skýrslum í áföngum eins og Hagnýt Bayesísk tölfræði og Stærðfræðigreining IV þar sem það þurfti að blanda saman kóða, töflum, myndum og stærðfræði. Dæmi um stærðfræði:

$$\mathrm{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t} dt.$$

3 Kóði

3.1 Stillingarskrá: settings.R

```
library(tidyverse)
library(gridExtra)
library(ggthemes)
library(ggpubr)
library(knitr)
library(kableExtra)
library(latex2exp)
library(bookdown)
library(scales)
theme_set(theme_tufte() +
            theme(panel.border = element_rect('black', fill = NA),
                  text = element_text(size = 14),
                  legend.text=element_text(size=14),
                  axis.text=element_text(size=14),
                  axis.title = element_text(size = 14),
                  plot.title = element_text(hjust = 0.5)))

ifelse(exists('d'), print('Loaded'), d <- read_csv('data/binom_sim.csv'))
```

3.2 Hermunarskrá: simulate_binomial.R

```
source('scripts/settings.R')
#' Simulate bernoulli strings of length n with success probability p
#' and store strings with their sums in a data frame.

#' Initialize
n <- 12
p <- 1/12
L <- 5e4
bi_sim <- data.frame(string = character(),
                     sum = numeric(),
                     stringsAsFactors = F)
set.seed(123)
#' Populate data frame
#' for loops are SLOW in R so this will run slowly (on purpose)
for(i in 1:L) {
  string <- rbinom(n, 1, p)
  bi_sim[i, 1] <- paste(as.character(string), collapse = '')
  bi_sim[i, 2] <- sum(string)
  if (i %% 1000 == 0) {
    print(i)
  }
}

# Export data
write_csv(x = bi_sim, path = 'data/binom_sim.csv')
```

3.3 Gagnavinnsluskrá: binomial_wrangl.R

```
source('scripts/settings.R')
d <- read_csv('data/binom_sim.csv')

# Plot barplot of sum
ggplot(data = d, aes(x = sum)) +
  geom_bar() +
  labs(x = 'Sum of string',
       y = 'Count') -> sum_bp
# Export plot
ggsave(filename = 'img/sum_bp.png', plot = sum_bp,
       width = 6, height = 6, dpi = 320)

# Get 10 most popular strings and their proportion
d %>%
  group_by(string) %>%
  summarize(p = n()/nrow(d)) %>%
  arrange(desc(p)) %>%
  ungroup() %>%
  filter(row_number() <= 10) -> pop_strings
# Export data
write_csv(x = pop_strings, 'tables/pop_strings.csv')

# Plot proprtions for pop_strings
ggplot(data = pop_strings, aes(x = string, y = p, group = 1)) +
  geom_point() +
  geom_line(size = 1, color = 'red', lty = 2) +
  labs(x = 'Binary string',
       y = 'Proportion') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) -> prop_plt
# Export
ggsave(filename = 'img/prop_plt.png', plot = prop_plt,
       width = 6, height = 6, dpi = 320)
```