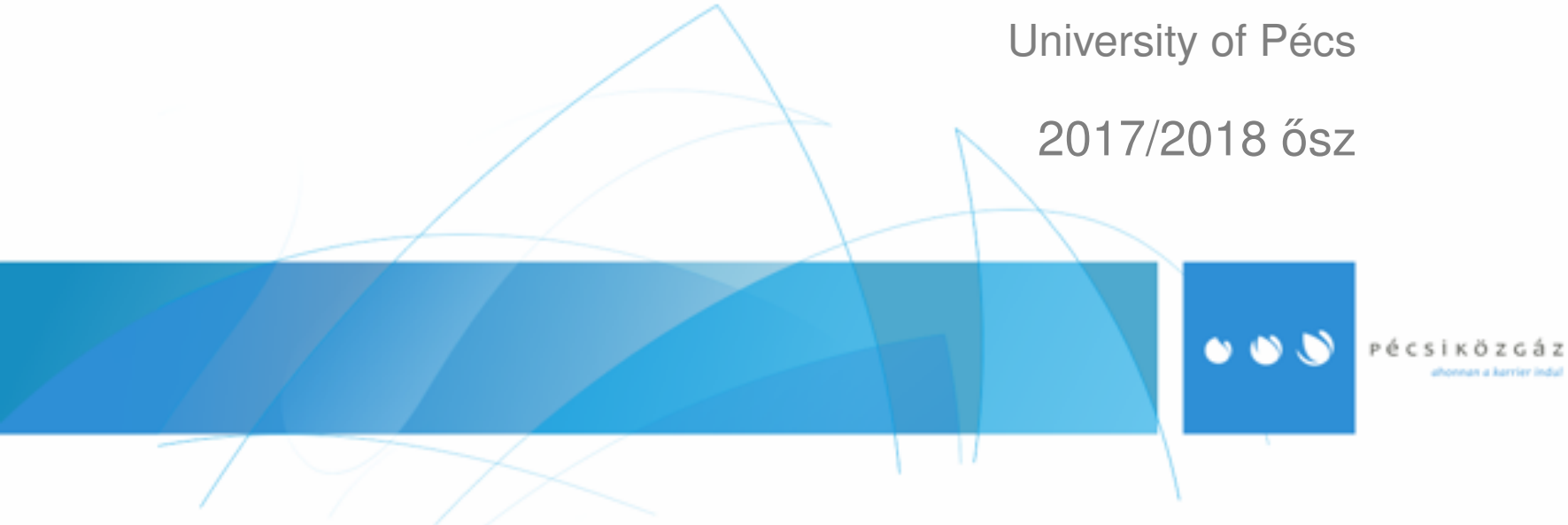
Introduction to R

Herczeg Róbert, Kehl Dániel

University of Pécs

2017/2018 ősz



Motivation

- R is used all over the world in many fields
  - social sciences
  - econometrics
  - bioinformatics: `http://www.bioconductor.org`
  - ...
- it is free, open source
- user contributions are possible, and great advantage
- one of the most enthusiastic user group (r-help mailing list)
- a lot of great books (R series of Wiley, UseR!)
- both in (reproducible) research and teaching
- and here is a long story (maybe a little biased):
  `http://r4stats.com/articles/popularity`

Table of contents

R basics

About the Project

- http://www.r-project.org
- developed from the S language
- free
- consists of so called packages
  - there are some basic packages (you have these after installing R)
  - you can find others (over 12350) on CRAN (The Comprehensive R Archive Network) – topic views
  - other user-provided functions and packages online
- also very popular in academics

R basics

Pros and cons

- you can basically find all (statistical) methods you might need
- produces high quality, customizable, publication ready graphs
- cooperates with other software packages (Excel, EViews, BUGS)
- freedom
- helps reproducible research
- script language (limited GUI and point-and-click functionality)
- running an analysis usually does not end with some tables and graphs (like in other software) but with objects containing information (and you can of course plot those or save them)

R basics

R Studio

- a popular IDE (integrated development environment)
- free, handy, convenient to use
- „Matlab like”
- you can download it at
  `https://www.rstudio.com/ide/download/`
- and here are some nice screenshots
  `https://www.rstudio.com/ide/screenshots/`
- we are going to use it on this short workshop

R basics

Basic functionality

- prompt: > - waiting for input
- try demo(), packages usually have demos, demo(persp), demo(graphics), demo(plotmath), demo(colors) etc.
- use it as a calculator, simple operators and functions
- define scalar variables: x = 1, x <- 1, 1 -> x (case sensitive!!!)
- access history by up and down arrows
- # indicates a comment in the code
- take a look at Appendix A (A sample session) in the R-intro.pdf!

R basics

Asking for help

- ?functionname or help(functionname)
  - first help pages might seem messy and too compicated, you have to get used to it, after some (a lot in fact) experience they are easy too use, informative and well structured
  - go for the examples if nothing else works (example(functionname))
- help("char") and ?"char" work in case of some special characters
- ?packagename
- help.start()
- help.search() or ??
- R-intro.pdf comes with R
- www, google, `http://stats.stackexchange.com`
- if no result, ask your question on the r-help mailing list
  - please read `http://www.r-project.org/posting-guide.html`

R basics

Installing packages

- install.packages(packagename)
  - choose a mirror, this downloads the files needed from CRAN
  - you only have to do this once
  - try install.packages(”googleVis”)
- library(packagename): activates the package (check library())
  - now you can use the functions in the package
  - you have to do this every time you need the package
  - try library(„googleVis”)
- take a look at your new package with ?googleVis
- try demo(WorldBank) and demo(AnimatedGeoMap), you probably have to wait a couple of seconds

Interactive graphs with Shiny

Webscraping with Shiny

- Collect semi-structured or unstructured data from websites
  - install shiny, shinydashboard package - app
  - install rvest package for webscraping
  - googleVis is already installed
- Develope a simple shiny app to visualize the data
- Use interactive graph to show the data

Interactive graphs with Shiny

Webscraping - download data

- Scraping data from web - ratebeer.com
  - html, body, table, tr, td
  - load rvest package
  - rvest::read_html - download webpage
  - rvest::html_table - get tables from webpage
  - find the right table
- data preprocessing

Interactive graphs with Shiny

Webscraping - Shiny, shinydashboard

- easily create webapps with Shiny
  - ui.R - user interface
  - server.R - server interface for the calculation
  - app.R - combine ui.R and server.R
- shinydashboard
  - increased UX - user experience
  - more options - sidebar, widgets, tabs etc.

Personalized exams

Statistics related courses – overview

- BA in Hungarian
  - Valószínűségszámítás és statisztika (Probability and Statistics)
    - 400/200 students, 1 lecture 7-8/3-4 computer lab sessions
  - Statisztikai modellezés (Statistical Modeling)
    - 100/250 students, 1 lecture 2/4-5 computer lab sessions
  - *Makrogazdasági adatok stat. el. (Applied Economic Statistics)*
    - 30-40 students, 1 lecture
  - *Statisztikai modellezés R-ben (Statistical Modeling in R)*
    - 20-30 students, 1 computer lab session
- BA in English
  - Probability and Statistics
    - 60-70 students, 1 lecture 2 computer lab sessions (autumn)
  - Business Statistics
    - 60-70 students, 1 lecture 2 computer lab sessions (spring)

Personalized exams

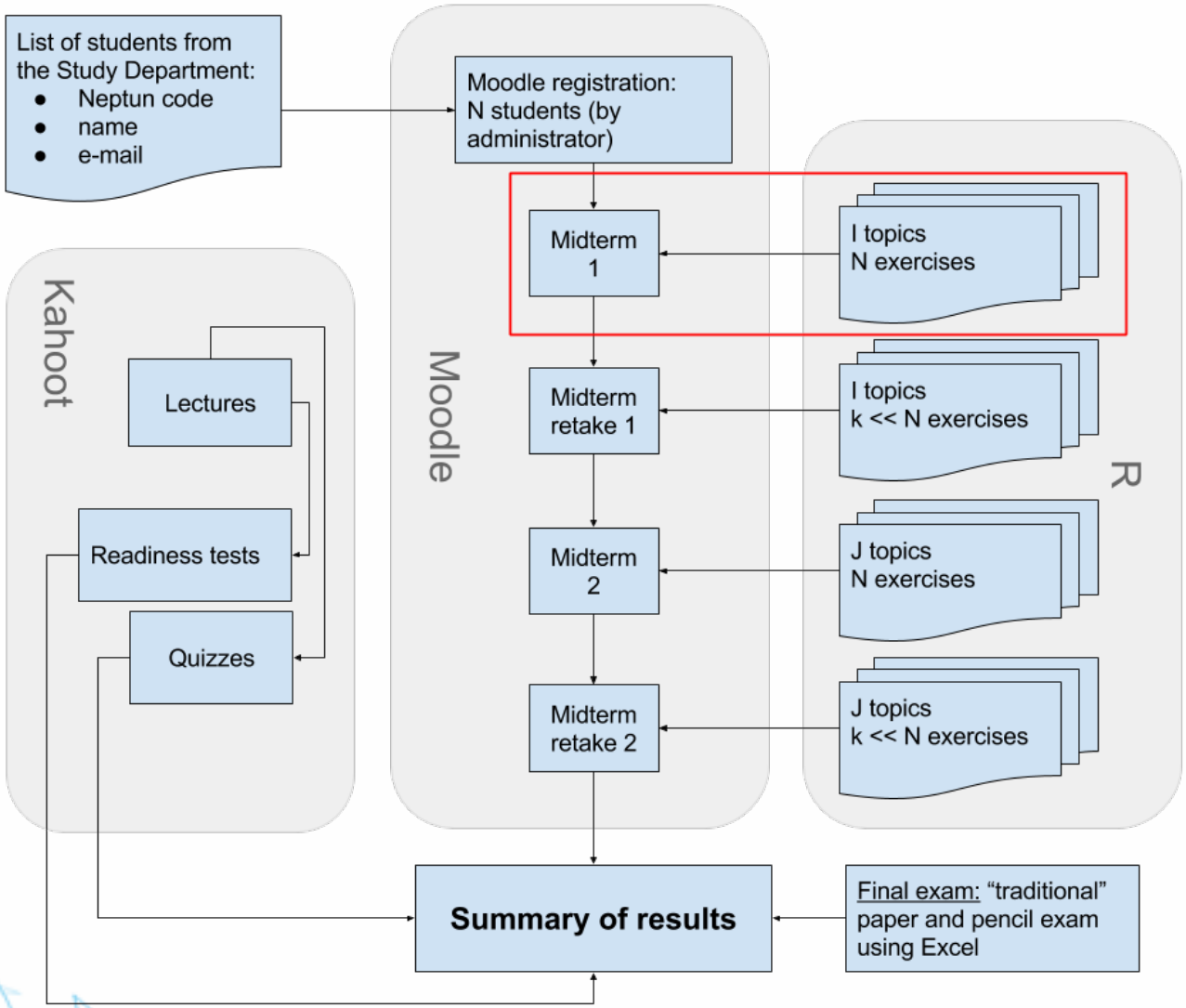
Main challenges in big classes

- high number of student especially on the Hungarian Programme
  - high number of exams, takes a lot of time grading, creating new problems, datasets etc.
  - after midterms students want to check their exams, solutions etc.
- desire to „force” students to continuously follow the course material throughout the semester
  - in-class short, 5 minute quizzes every lab session
  - two Excel-based midterms
- exams at the computer makes cheating easier

One possible solution is personalized exams with Moodle and R.

Personalized midterms

Assessment structure of the semester



Personalized midterms

Storing exercise text and data in spreadsheets

The example of estimation/hypothesis testing of the population mean

text	text2	mu	sigma	round
Egy üzleti döntés előkészítéséhez meg kell becsülnünk a termékünk ellenőrzéséhez szükséges átlagos időt (percben). Ehhez az alábbi nnnn elemű mintát választottuk.	Egy üzleti döntés előkészítéséhez meg kell vizsgálnunk a termékünk ellenőrzéséhez szükséges átlagos időt (percben). Ehhez az alábbi nnnn elemű mintát választottuk. A termelési igazgató állítása szerint az ellenőrzéshez szükséges átlagos idő mumumu	100	15	1
Egy nagyvállalatnál éves TeljesítményÉrtékelési Rendszer (TÉR) működik. A kitöltéshez szükséges átlagos időt kívánjuk megbecsülni. Véletlenszerűen kiválasztott nnnn munkavállaló esetén mértük ezt az időt.	Egy nagyvállalatnál éves TeljesítményÉrtékelési Rendszer (TÉR) működik. A kitöltéshez szükséges átlagos időt kívánjuk vizsgálni véletlenszerűen kiválasztott nnnn munkavállaló segítségével. A TÉR-t szállító külső vállalat állítása szerint az	25	5	1
A fogyasztói árindex számításához szükségünk van az egy kilogrammos fehér kenyér magyarországi átlagárára. Ennek érdekében az alábbi reprezentatív árösszeírásokat végeztük.	Az egyik politikai párt szerint az egy kilogrammos fehér kenyér magyarországi átlagára mumumu forint. Ennek ellenőrzésére az alábbi reprezentatív árösszeírásokat végeztük.	250	20	0
Egy vállalatnak készített marketing	Egy vállalatnak készített marketing			

Personalized midterms

The R-code – generating data and solution

```
<<echo=FALSE, results=hide>>=
## DATA GENERATION
szovegek <- read.csv2(file.path(mywd, "exercises", "prob_stat", "estimation", "mean_esthip.csv"), stringsAsFactors = FALSE)
szovegek <- szovegek[sample(1:nrow(szovegek), 1), ]
signLevel <- sample(c(.1, .05, .01), 1)

n <- sample(x = 6:8, size = 1)*5
smpl <- round(rnorm(n, szovegek$mu, szovegek$sigma), szovegek$round)

data <- matrix(smpl, ncol = 5)

dataDisp <- xtable(data, digits = szovegek$round)

## CALCULATIONS
m <- mean(smpl)
se <- sd(smpl) / sqrt(n)

mu0 <- round(m - runif(1, -4, 4) * se, szovegek$round)

temp <- (m-mu0)/se
tkrit <- -qt((signLevel) / 2, n - 1)

pertek <- 2*(1-pt(abs(temp), n - 1))
```

Personalized midterms

The R-code – generating questions, setting tolerances

```
25  ## QUESTION/ANSWER GENERATION
26  questions <- character(4)
27  solutions <- logical(4)
28  tolerances <- rep(0.0001, 4)
29
30  questions[1] <- "Mekkora a hipotézisellenőrzés során használt sztenderd hiba nagysága?"
31  solutions[1] <- se
32
33  questions[2] <- paste0("Adja meg a ", 100*signLevel, "\\%-os szignifikancia szinthez
    tartozó kétoldalú alternatív hipotézishez tartozó kritikus érték abszolút értékét!")
34  solutions[2] <- tkrit
35
36  questions[3] <- "Adja meg a próbafüggvény empirikus értékét!"
37  solutions[3] <- temp
38
39  if(runif(1) > 0.5) {
40    questions[4] <- "Adja meg a kétoldalú alternatív hipotézishez tartozó p-értéket!"
41    solutions[4] <- pertek
42  } else {
43    questions[4] <- "Adja meg az egyoldalú kisebb alternatív hipotézishez tartozó p
    -értéket!"
44    solutions[4] <- pt(temp, n - 1)
45  }
46  @
```

Personalized midterms

Output in Moodle

Egy nagyvállalatnál éves TeljesítményÉrtékelési Rendszer (TÉR) működik. A kitöltéshez szükséges átlagos időt kívánjuk megbecsülni. Véletlenszerűen kiválasztott 30 munkavállaló esetén mértük ezt az időt.

21,7 30,5 22,4 28,7 20,3
24,7 17,0 20,8 29,9 27,5
28,4 15,7 27,8 23,9 33,0
28,0 21,4 21,1 20,4 15,6
26,7 27,9 19,2 27,2 22,0
22,8 25,5 32,2 24,6 25,6

a. Adja meg a mintabeli szórást!

b. Mekkora a becslés hibahatára 90%-os megbízhatóság mellett?

c. Adja meg a 90%-os megbízhatósági szintű becslés konfidencia intervallumának alsó határát!

Egy nagyvállalatnál éves TeljesítményÉrtékelési Rendszer (TÉR) működik. A kitöltéshez szükséges átlagos időt kívánjuk megbecsülni. Véletlenszerűen kiválasztott 35 munkavállaló esetén mértük ezt az időt.

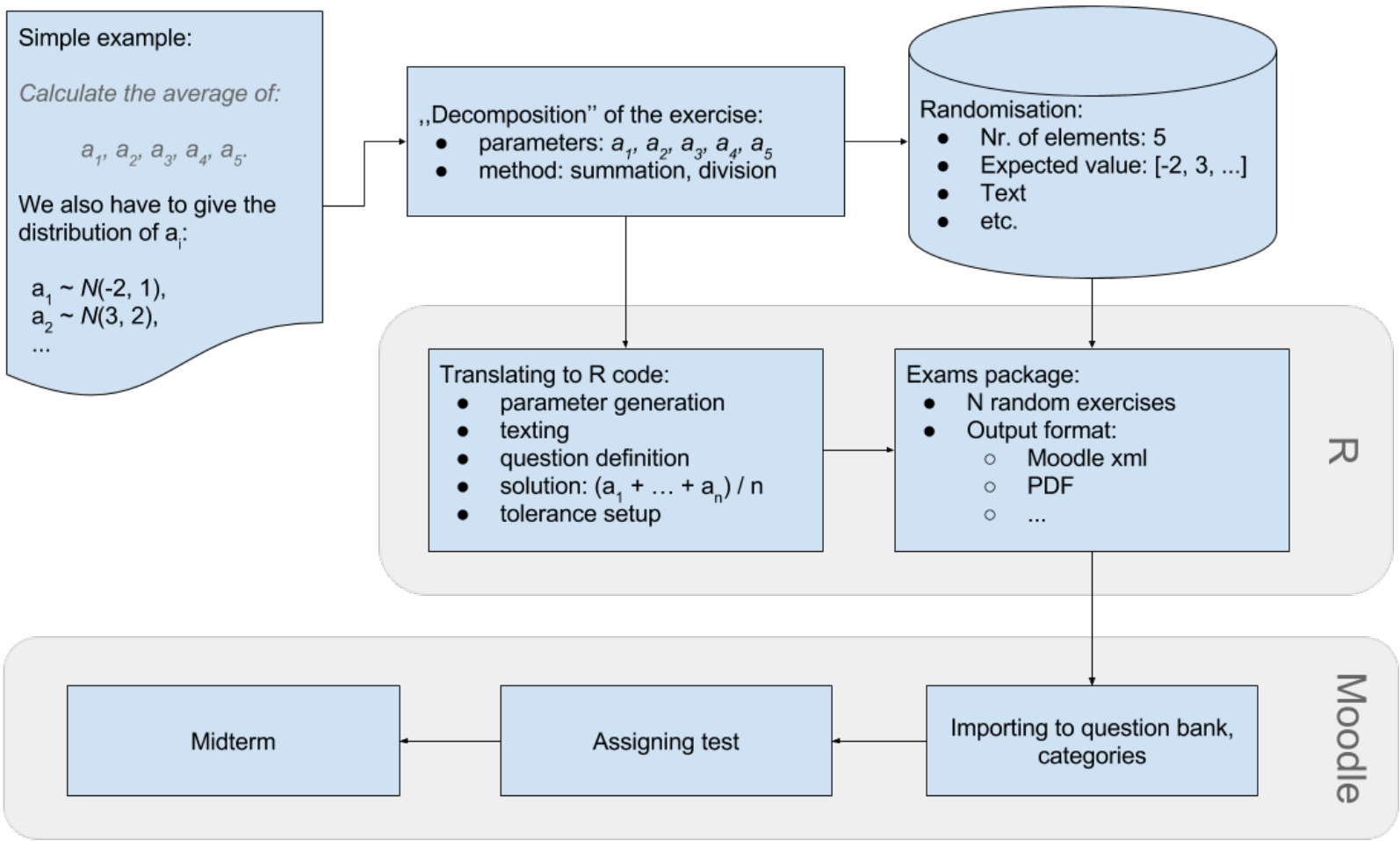
23,8 28,6 24,2 19,9 31,3
24,8 23,3 14,4 27,0 23,9
20,3 23,8 24,6 21,2 17,2
23,4 33,4 24,3 21,2 26,1
30,6 28,0 28,3 34,4 23,8
31,2 24,5 23,6 26,3 30,4
24,3 26,2 25,0 22,6 24,8

a. Adja meg a mintaátlagot!

b. Mekkora a becslés sztenderd hibájának nagysága?

Personalized midterms

The general idea – workflow



Personalized midterms

The R-code – generating a midterm

```
source("functions.R")

mywd <- getwd()
folder <- "exercises/prob_stat/estimation/"
exerc1617osz1 <- c("binom.Rnw", "hipgeom.Rnw", "poisson.Rnw", "norm.Rnw", "2valt.Rnw")
exerc1617osz2 <- c("mean_est.Rnw", "prop_est.Rnw", "prop_hip.Rnw", "f_egyez.Rnw")
exerc1617tavasz1 <- c("leiro_stat.Rnw", "binom.Rnw", "hipgeom.Rnw", "poisson.Rnw")
exerc1617tavasz1pot <- c("leiro_stat.Rnw", "binom.Rnw", "hipgeom.Rnw", "poisson.Rnw")
exerc1617tavasz2 <- c("mean_est.Rnw", "mean_hip.Rnw", "paros.Rnw", "prop_est.Rnw", "prop_h

exerc <- c("mean_est.Rnw", "prop_hip.Rnw", "f_egyez.Rnw")

myexam <- paste0(folder, exerc)

exams2pdf(myexam, n = 20, name = c(paste0(c(sub(".Rnw","",unique(exerc)),"exam"), collaps
                                   paste0(c(sub(".Rnw","",unique(exerc)),"solution"), colla
          encoding = "UTF-8",
          edir = "exercises",
          dir = "output",
          template = c("templates/exam.tex", "templates/solution.tex"),
          header = list(
            Date = "2017-09-09",
            ID = function(i) formatC(i, width = 5, flag = "0")
            )
          )

exams2moodle(myexam, n = 50, name = c("szeged"),
             encoding = "UTF-8",
             edir = "exercises",
             dir = "output")
```

As a result

our students face

- very similar question types (let's say one-sample t-test, ANOVA, estimation and linear regression)
  BUT
- different „stories"
- different questions (give the empirical value vs. give the critical value vs. give the p-value etc.)
- different datasets
- different numeric solutions
- immediate feedback and results
- possibility to „flag" questions in Moodle

Summary of our experiences

- results are fairly similar in comparison to previous years
- students do not complain about it, like the quick response
- preparing a midterm takes longer (see R code)
- setting up a question bank is an initial investment
- going through flagged questions is fairly quick
- saving a lot of time with automatic grading
- cheating seems to be harder

Useful links and materials

- https://moodle.org/
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Achim Zeileis, Nikolaus Umlauf, Friedrich Leisch (2014). Flexible Generation of E-Learning Exams in R: Moodle Quizzes, OLAT Assessments, and Beyond. Journal of Statistical Software 58(1), 1-36. doi:10.18637/jss.v058.i01
- https://cran.r-project.org/web/packages/exams/vignettes/exams.pdf
- https://cran.r-project.org/web/packages/exams/exams.pdf
- exams_skeleton function