An experience using a programming language in statistics course

~~Imagine you~~ roll a die

Marcos Marvá, Fernando San Segundo
Departamento de Física y Matemáticas UAH

UTS Ingenieros Industriales
UNED 20.07.2016

Universidad de Alcalá

LibreTICs

UNED ETS de Ingenieros Industriales

1) The context and how we used to teach
2) What we did not like
3) What we are doing about
4) Conclusions – ideas – ongoing work

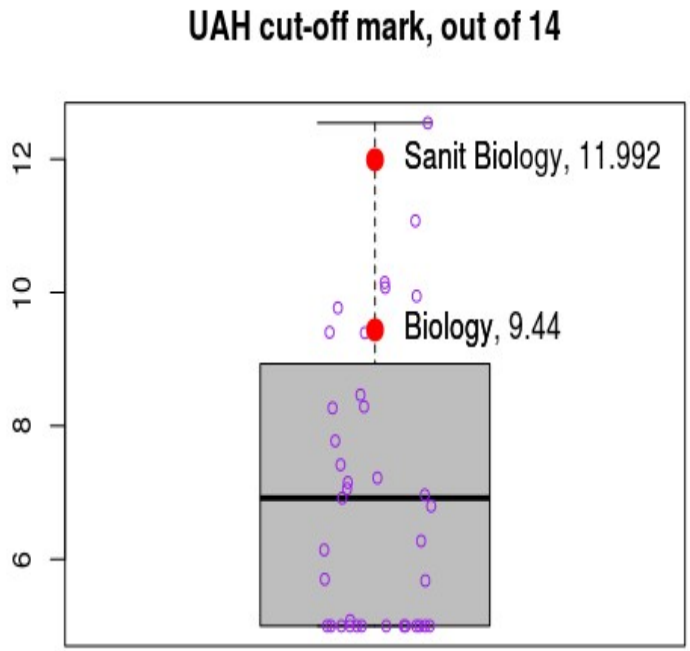
Context

Course on Statistics at:

- Sanitary biology degree, 1st term
- Biology degree, 3rd term

Organization/schedules

- 28h in large group (100 studentes)
- 22h in small group (25 students)

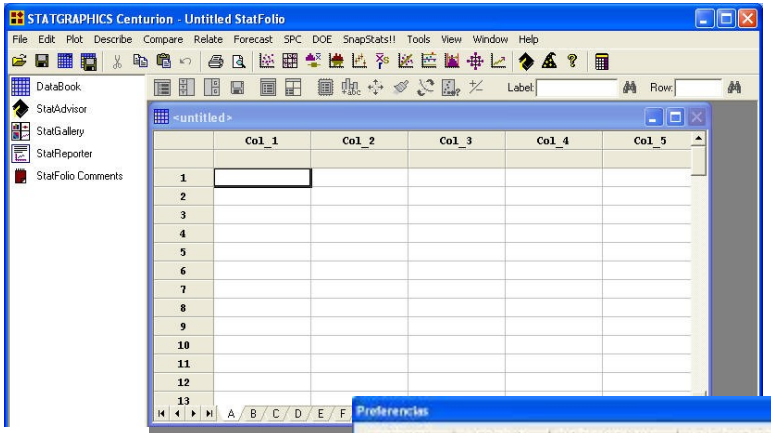


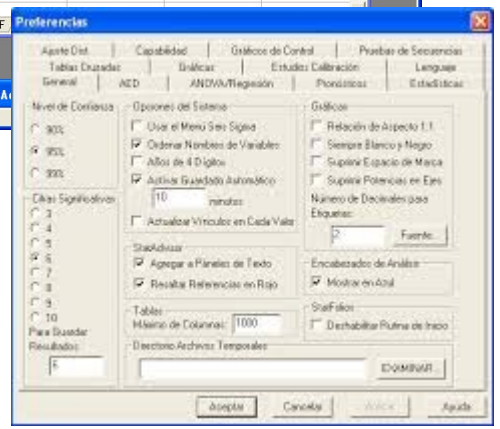
What we use to do? What is usually done*

Masterclasses (large group)

Computer labs (small group)

Seminars (small group)





$$\left[\hat{p}_1 - \hat{p}_2 \mp Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\left[\overline{X}_1 - \overline{X}_2 \mp Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

$$\frac{s_1^2}{s_2^2}$$

$$\left[\overline{X} \mp t_{\alpha/2,n-1}\ \frac{s}{\sqrt{n}}\right]$$

$$\frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

$$\left[\frac{s_1^2/s_2^2}{F_{\alpha/2,n_1-1,n_2-1}};\ \frac{s_1^2/s_2^2}{F_{1-\alpha/2,n_1-1,n_2-1}}\right]$$

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

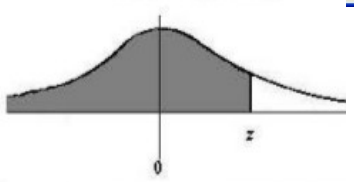
$$\left[\overline{X}_1 - \overline{X}_2 \mp Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

$$\frac{(n-1)\,s^2}{\sigma_0^2}$$

$$\left[\overline{X}_1 - \overline{X}_2 \mp t_{\alpha/2,n_1+n_2-2}\ s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$F_Z(z) = P[Z \le z]$$



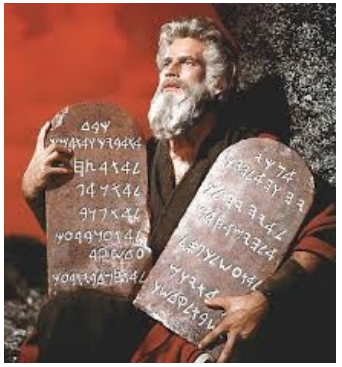
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131

* in our department

What we use to do?

Students (most of them):
- Proceeded mechanically – menu oriented thinking
- Lack of intuition on key concepts – no experimentation

We, teachers, (sometimes) felt
- Poor use of technology
- What about good students?
- Slightly unsatisfied, bored



Academic performance was good and not everything was wrong...
...BUT...there was room for improvement

Inspiration from

LibreTICs

TEDTalks EDUCATION

YouTube

Confessions of a converted lecturer

Eric Mazur (physician, Harvard University) – Peer instruction

MOOC

edX

coursera

El cerebro necesita emocionarse para aprender

Los nuevos experimentos en la enseñanza vislumbran el fin de las clases magistrales. Una de las tendencias es la neurodidáctica

45

ANA TORRES MENÁRGUEZ

Madrid - 18 JUL 2016 - 11:55 CEST

En el año 2010 un equipo de investigadores del Massachusetts Institute of Techonolgy (MIT), en Boston, colocaron a un universitario de 19 años un sensor electrodérmico en la muñeca para medir la actividad eléctrica de su cerebro las 24 horas durante siete días. El experimento arrojó un resultado inesperado: la actividad cerebral del estudiante cuando atendía en una clase magistral era la misma que cuando veía la televisión; prácticamente nula. Los científicos pudieron probar así que el modelo pedagógico basado en un alumno como receptor pasivo no funciona.

What to do instead?

Students learn, we do not teach
- But we can promove learning situtions

Learning enhanced by experimentation
- Reproducibility
- Interactivity

Programming must be a compulsory skill in science
- Compute the solution is part of the solution
- We prefer free software → 

….but we were a bit worried

What to do instead?

large groups

We

- Perform live simulations with R scripts
  (additional file central-limit-theorem-UNED-hospitalization-time.R)
- Intuition gained from interactive tools
  (Cap06-InterpretacionIntervalosConfianza_n.ggb)

Backed by

- Blog + lecture notes (→ blog + book)

What to do instead?

small groups

At the computers room

- Reproducible tutorial, docs...

1.4. Tablas de contingencia relativas en R.

Vamos a ver cómo utilizar R para obtener las tablas relativas que hemos discutido en la página 476 del libro. Concretamente, vamos a ver cómo reproducir los resultados del Ejemplo 12.1.6 en el que se analizaba la tabla de contingencia correpsondiente a una prueba diagnóstica, que hemos usado varias veces en el libro. Empezamos con la tabla de datos básica :

```
(tablaObservada = matrix( c(192, 4, 158, 9646),  nrow= 2))
```

```
##      [,1] [,2]
## [1,]  192  158
## [2,]    4 9646
```

Ponemos nombre a las filas y columnas:

```
colnames(tablaObservada) = c("Enfermos", "Sanos")
rownames(tablaObservada) = c("Positivo", "Negativo" )
tablaObservada
```

```
##          Enfermos Sanos
## Positivo      192   158
## Negativo        4  9646
```

y ya estamos listos para pasar a los valores marginales. Los añadimos a la tabla pero, además, calculamos la suma total:

```
(tablaObservadaMarg = addmargins(tablaObservada))
```

```
##          Enfermos Sanos   Sum
## Positivo      192   158   350
## Negativo        4  9646  9650
## Sum           196  9804 10000
```

```
(n = sum(tablaObservada) )
```

```
## [1] 10000
```

What to do instead?

small groups

At the computers room

- Reproducible tutorial, docs...
- Auxiliary templates

  instead of menus

```
1  ######################################################
2  # www.postdata-statistics.com
3  # POSTDATA. Introducción a la Estadísitica
4  # Tutorial-06.
5  #
6  # Fichero de instrucciones R para calcular un intervalo de confianza (1-alfa) para la
7  #
8  #    DESVIACION TIPICA de una  poblacion normal N(mu,sigma),
9  #
10 # a partir de una muestra de tamaño n. ste fichero usa los estadisticos  de una muestra,
11 # previamente calculados (numero de datos, media muestral, etc.)
12 ################################################################
13
14 rm(list=ls()) #limpieza inicial
15
16 # ATENCIÓN: Para usar este fichero
17 # la población debe ser (al menos aprox.) normal
18 # EN OTROS CASOS NO USES ESTE FICHERO!!
19 # ASEGURATE DE HABER ENTENDIDO ESTAS INSTRUCCIONES
20
21 # Introducimos el valor de la desviacion tipica muestral,
22 s =
23
24 # el tamaño de la muestra,
25 n =
26
27 # y el nivel de confianza deseado.
28 nc =
29
30 ################################################
31 #NO CAMBIES NADA DE AQUI PARA ABAJO
32 ################################################
33 # Calculamos alfa
34 alfa = 1 - nc
35
36 # y los grados de libertad:
37 (k= n - 1)
38
39 # Calculamos los valores criticos necesarios:
40 (chiAlfa2 = qchisq(1 - (alfa/2), df=k))
41 (chiUnoMenosAlfa2 = qchisq(alfa/2, df=k))
42
43 #Para la varianza, el intervalo de confianza sera
44 (intervaloVar = s^2 * k / c(chiAlfa2, chiUnoMenosAlfa2))
45
46 # Y para la desviacion tipica el intervalo de confianza es este:
47 (intervaloS = s * sqrt(k / c(chiAlfa2, chiUnoMenosAlfa2)))
48
```

What to do instead?

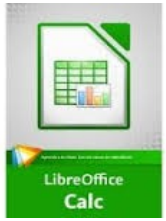
small groups

At the computers room

- Reproducible tutorial, docs...
- Auxiliary templates

  instead of menus
- A handful of free tools:

  R, Calc, GeoGebra,
  WolframAlpha, Wiris



ExamineR – test yourself

Quizz script

PostData **Problemas de Estadística**

www.postdata-statistics.com **Creados con el paquete exams de R**

1. **Problema**
   Calcula la media de este conjunto de números:

   8, 4, 7, 5, 3, 1, 2, 6, 0, 9.

   Redondea el resultado con 4 cifras significativas.

   Solución
   La respuesta es 4.5

2. **Problema**
   Calcula la **mediana** de este conjunto de números:

   9, 0, 9, 8, 5, 7, 6, 3, 9.

   Redondea el resultado con 4 cifras significativas.

   Solución
   La respuesta es 7

3. **Problema**
   Dada esta tabla de frecuencias de un conjunto de datos:

	6	13	14	22	25	29	43	46	62	63	70	89	91	95	100
Frecuencias	21	41	45	50	9	19	42	5	14	35	13	23	32	30	44

   calcula su **media**. Escribe tu respuesta con cuatro cifras significativas.

   Solución
   La solución es 50.57.

4. **Problema**
   Dada esta tabla de frecuencias de un conjunto de datos:

	7	9	11	16	17	24	29	30	54	61	65	71	79	89	93
Frecuencias	13	14	40	19	26	8	47	21	48	29	17	4	5	32	1

   calcula su **desviación típica (poblacional)** . Escribe tu respuesta con cuatro cifras significativas.

   Solución
   La solución es 26.02.

PostData **Problemas de Estadística**

www.postdata-statistics.com **Creados con el paquete exams de R**

1. **Problema**
   Calcula la **desviación típica (poblacional)** de este conjunto de números:

   9, 1, 7, 7, 5, 10, 6, 6, 7, 1.

   Redondea el resultado con 4 cifras significativas.

   Solución
   La respuesta es 2.809

2. **Problema**
   Dada esta tabla de frecuencias de un conjunto de datos:

	2	4	7	8	18	19	25	26	33	38	65	77	91	92	93
Frecuencias	44	1	28	19	31	32	5	7	30	34	48	4	37	26	23

   calcula su **varianza**. Escribe tu respuesta con cuatro cifras significativas.

   Solución
   La solución es 1119.

3. **Problema**
   Dada esta tabla de frecuencias de un conjunto de datos:

	7	12	20	25	29	34	35	44	50	58	59	62	63	74	77
Frecuencias	48	8	30	11	13	50	28	31	43	35	9	2	44	17	39

   calcula su **desviación típica (poblacional)** . Escribe tu respuesta con cuatro cifras significativas.

   Solución
   La solución es 21.57.

4. **Problema**
   Dada esta tabla de frecuencias de un conjunto de datos:

	2	5	7	11	12	21	24	50	56	63	74	84	89	95	99
Frecuencias	38	3	49	4	30	29	37	27	9	42	2	25	16	46	32

   calcula su **media**. Escribe tu respuesta con cuatro cifras significativas.

   Solución
   La solución es 46.39.

PostData **Problemas de Estadística**

www.postdata-statistics.com **Creados con el paquete exams de R**

1. **Problema**
   Calcula la **mediana** de este conjunto de números:

   12, 11, 6, 10, 6, 11, 1, 0.

   Redondea el resultado con 4 cifras significativas.

   Solución
   La respuesta es 8

2. **Problema**
   Dada esta tabla de frecuencias de un conjunto de datos:

	1	5	16	24	28	41	49	57	69	71	79	81	84	85	94
Frecuencias	8	14	6	46	2	16	32	7	24	18	20	12	27	3	25

   calcula su **media**. Escribe tu respuesta con cuatro cifras significativas.

   Solución
   La solución es 55.06.

3. **Problema**
   Dada esta tabla de frecuencias de un conjunto de datos:

	11	26	42	50	56	68	71	75	76	78	82	83	90	94
Frecuencias	49	26	31	15	39	29	19	18	23	14	17	34	20	16

   calcula su **varianza**. Escribe tu respuesta con cuatro cifras significativas.

   Solución
   La solución es 687.

4. **Problema**
   Calcula la **desviación típica (poblacional)** de este conjunto de números:

   3, 2, 12, 12, 2, 8, 2, 9, 9, 3, 5, 1.

   Redondea el resultado con 4 cifras significativas.

   Solución
   La respuesta es 3.923

What to do instead?

ExamineR – test yourself

Based on the `exams`[1] package

- Easy generation of quizzes
- 157 (and growing) randomly generated questions covering the program
- Github repo ExamineR

(1) A. Zeileis, B. Gruen, F. Leisch, N. Umlauf, D. Ernst
https://cran.r-project.org/web/packages/exams/index.html

ExamineR – test yourself

Question generation

```
<<echo=FALSE, results=hide>>=

################################################################################
# PostData Statistics:
# Fernando San Segundo, Marcos Marva
# Web: www.postdata-statistics.com
# Mail: postdatastatistics@gmail.com       (secondary: marcos.marva@gmail.es)
################################################################################
# calculate the mean value
## DATA GENERATION
signifDig <- 4
(n <- sample(8:13, 1))
(data <- c(sample(0:12, n)))
(dataString = paste(data, collapse=", "))
(sol <- signif(mean(data), digits = signifDig))

## QUESTION/ANSWER GENERATION
if(language=='en'){
  question=paste0(
"Find the mean of this set of numbers:\\begin{center}", dataString, ".\\end{center} Round the result  to ", signifDig, " significant digits."
  )
  answer=paste0("The answer is ", sol )
}else if(language=='es'){
  question=paste0(
    "Calcula la media de este conjunto de n\\'umeros:\\begin{center}", dataString, ".\\end{center}Redondea el resultado con ",
    signifDig, " cifras significativas."
    )
  answer=paste0("La respuesta es ", sol )
}
@

\begin{question}
<<echo=False, results=tex>>=
cat(paste(question,collapse=""))
@
\end{question}

\begin{solution}
<<echo=False, results=tex>>=
cat(paste(answer,collapse=""))
@
\end{solution}

%% META-INFORMATION
%% \extype{num}
%% \exsolution{\Sexpr{sol}}
%% \exname{Prediction}
%% \extol{0.00001}
```

ExamineR – test yourself

Question generation

```
<<echo=FALSE, results=hide>>=

################################################################################
# PostData Statistics:
# Fernando San Segundo, Marcos Marva
# Web: www.postdata-statistics.com
# Mail: postdatastatistics@gmail.com       (secondary: marcos.marva@gmail.es)
################################################################################
# calculate the mean value
## DATA GENERATION
signifDig <- 4
(n <- sample(8:13, 1))
(data <- c(sample(0:12, n)))
(dataString = paste(data, collapse=", "))
(sol <- signif(mean(data), digits = signifDig))

## QUESTION/ANSWER GENERATION
if(language=='en'){
  question=paste0(
"Find the mean of this set of numbers:\\begin{center}", dataString, ".\\end{center} Round the result  to ", signifDig, " significant digits."
  )
  answer=paste0("The answer is ", sol )
}else if(language=='es'){
  question=paste0(
    "Calcula la media de este conjunto de n\\'umeros:\\begin{center}", dataString, ".\\end{center}Redondea el resultado con ",
    signifDig, " cifras significativas."
    )
  answer=paste0("La respuesta es ", sol )
}else if(language=='ge'){
  question=paste0(
    "Finden Sie den Mittelwert dieser Satz von Zahlen:\\begin{center}", dataString, ".\\end{center} Rund um die Folge zu ",
    signifDig, " signifikanten Stellen."
    )
  answer=paste0("The Antwort ", sol )
}else if(language=='ga'){
  question=paste0(
    "Atope a media deste conxunto de n\\'umeros.:\\begin{center}", dataString, ".\\end{center} arredonda o resultado para",
    signifDig, " algarismos significativos."
    )
  answer=paste0("A resposta \\'e ", sol )
}
@

\begin{question}
```

ExamineR – test yourself

Quizz script

```
################################################################################
# PostData Statistics:
# Fernando San Segundo, Marcos Marva
# Web: www.postdata-statistics.com
# Mail: postdatastatistics@gmail.com       (secondary: marcos.marva@gmail.es)
################################################################################
################################################################################
# File to generate quizzes with the R Exams package (basic template)
# To run the code, press Ctrl+A (select all) Crtl+Intro (run selected code)
################################################################################

# clear the envirinment
rm(list=ls())

# uncomment the next line if you have not installed the package "exams".
# only needed once
# install.packages("exams")

# let know the code where it is to set the working directory
odir = getwd()
(tempDir=paste(odir,"/temp",sep=""))

# load the library "exams"
library("exams")

## Set the language code for your exam: "es" = spanish, "en"= english
language = "en"

# load templates for the exam and the answers

(templateExam=paste("PostData-Exam-",language,"-",sep="",collapse=""))
(templateAnswer=paste("PostData-Answer-",language,"-",sep="",collapse=""))

# use the desired number quizzes, for instance
nameFile=c("000001", "01010101", "020003", "020203", "020602", "03010202", "03020201", "040104", "04030201","050301", "06060302", "060803")
(UnExamen=as.character(paste(nameFile,".Rnw",sep="")))

# Exam in PDF format
 exams(UnExamen, dir=odir, template=c(templateExam,templateAnswer), n=3)

# Quiz in HTML format, uncomment the next line
# exams2html(UnExamen,dir=dirTrabajo,n=1,name=nameFile,mathjax=TRUE)

# Quiz in XML format ready for MOODLE, uncomment the next two lines
# library("tth")
# exams2moodle(file = UnExamen, n=7,name="moodle", nsamp = 5)

# Sometimes, it takes some time to compile the quizzes, the process will be finished when you see number pi
pi
```

Script StudentR

StudentR An script aided by the exams R package https://github.com/PostDataStatistics/ExamineR

N°: 2 Cox , Gertrude

1. Round the number

   3.11451492831111

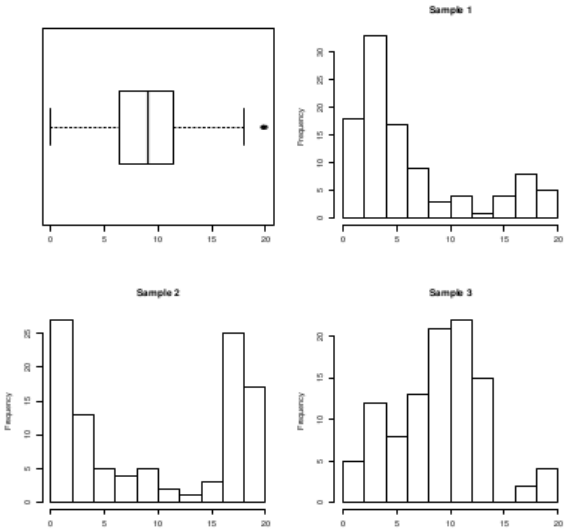
   to 3 significant figures.
2. Find the **median** of this set of numbers:

   6, 10, 2, 11, 3, 9, 5, 10, 7.

   Round the result to 4 significant digits.
3. Find the **(population) standard deviation** of this set of numbers:

   1, 0, 11, 9, 8, 4, 4, 2, 5.

   Round the result to 4 significant digits.
4. Which of the histograms corresponds to the boxplot?



StudentR An script aided by the exams R package https://github.com/PostDataStatistics/ExamineR

N°: 3 Fisher , Ronald

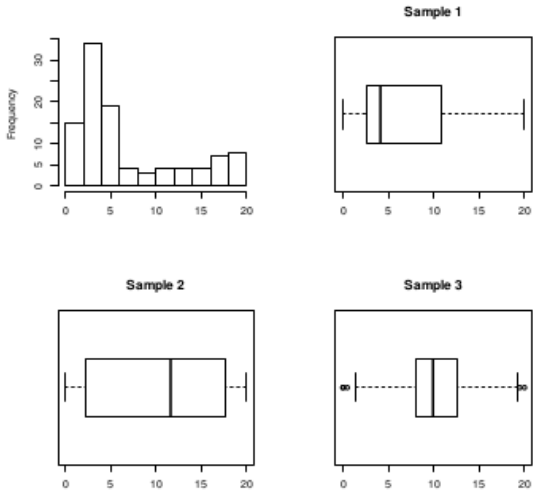
1. Round the number 770.438 to 3 significant digits.
2. Find the **median** of this set of numbers:

   0, 2, 2, 2, 5, 8, 4, 0, 7, 12, 4, 12, 2.

   Round the result to 4 significant digits.
3. Find the **(population) variance** of this set of numbers:

   5, 6, 0, 10, 9, 9, 7, 4.

   Round the result to 4 significant digits.
4. Which of the histrogram describes the data represented by one of the fo boxplots?



StudentR An script aided by the exams R package https://github.com/PostDataStatisti

N°: 9 Pearson , Karl

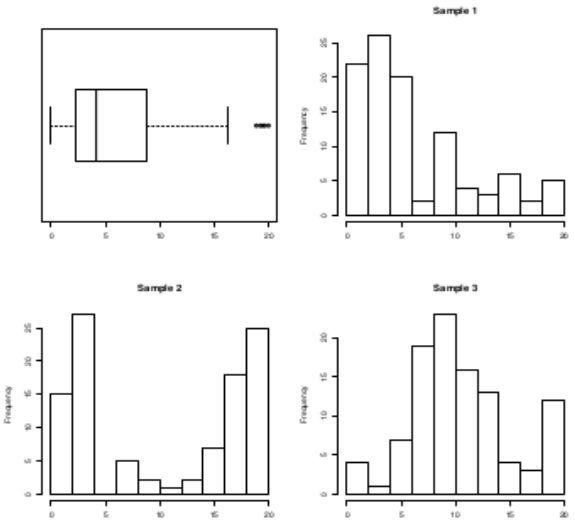
1. Round the number 691078 to 3 significant digits.
2. Find the **median** of this set of numbers:

   3, 6, 8, 10, 0, 5, 4, 1.

   Round the result to 4 significant digits.
3. Find the **(population) variance** of this set of numbers:

   0, 9, 0, 3, 1, 4, 6, 11, 7, 6, 3.

   Round the result to 4 significant digits.
4. Which of the histograms corresponds to the boxplot?



Results – conclusions – ideas

- “Recycle” through internet talks and open courses

- Rather natural use a programming language in masterclasses
  Simulations ↔ experiments enhance intuition and understanding

- No dificulties with scripts
  - Templates & reproducible documents
  - We share code chunk with students
  - Reproducible mistakes
  - Close to real work

- Many students use templates as menus
- Quite a lot students write/share their own scripts
- Some students search for “advanved” packages

Results – conclusions – ideas

- Students get used in open source software

Journal of Statistical Software

or Authors | JSS Style Guide | Editorial Board | Register

Home > Archives > Vol 22 (2007)

Vol 22 (2007)

Special Volume: Ecology and Ecological Modelling in R *(Editors: Thomas Kneib, Thomas Petzoldt)*

Bulgaria Got a Law Requiring Open Source

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Results – conclusions – ideas

- We have fun at lessons and feel satisfied

- Help those who want to "go further"
  do not harm those that just want to pass the course

Ongoing work

- Measure performance
- Assais coordinated with other courses
- R based degree capstone project