Statistical Computing: A Language or a Graphical User Interface?

Daniel T. Kaplan and Jason Wilson
kaplan@macalester.edu & jason.wilson@biola.edu

eCOTS, May 16, 2012
Electronic Conference on Teaching Statistics

A Discussion of Pros and Cons

There are many reasonable choices in statistical software: Minitab, SPSS, Fathom, Excel, JMP, ...

The choice of one over the others is shaped by many factors:

- Background of students
- Future trajectory of students
- Institutional support and facilities
- Instructor experience

We focus here on a "generic" issue: the advantages and disadvantages of a Graphical User Interface (GUI) versus a Command Line interface.

Specific while General?

To keep the discussion concrete, we will focus on two specific systems:

1. R-Commander
2. R with a command-line interface

We will also supplement the above tools with these:

- Spreadsheets, e.g. Google Spreadsheets
- Web applets
- R applets

We chose these because they are all free and readily available and because they illustrate the language-vs-GUI choice.

Resources for Installing the Software

- R locally: `www.r-project.org`
- R-Commander: `www.r-project.org`
- RStudio: `www.rstudio.org`
- Google Docs

Tutorials for getting started

- R Commander:
  - J. Fox, `http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/`
  - Videos: `https://sites.google.com/a/biola.edu/r-commander-demonstrations/`
- R: R. Pruim, N. Horton & D. Kaplan *Teaching Statistics with R and RStudio: An Instructor's Guidebook*
- John Versani, *Getting Started with RStudio* O'Reilly
- Mosaic. All the examples use the `mosaic` package within R. See the vignettes with that package.

```
require(mosaic)
```

Statistical Topics

Seven topics picked to illustrate a range of **computational** issues.

1. Arranging and accessing data
2. Simple descriptive statistics
3. Inference on means
4. Inference on counts
5. Simple regression
6. Multiple regression
7. The "Three R's" of statistical inference

Most of the statistical computing tasks a student encounters, including making graphs of various sorts, are done in a very similar way.

Forms of Data Organization

Some forms of data in statistics courses:

- Tabular: Individual cases and potentially multiple variables. (This paradigm includes relational databases, which involve multiple tables.)
- Summary:

	Group A	Group B
sample size	n_A	n_B
sample mean	m_A	m_B
sample std dev	s_A	s_B

- Word problem format:

  "Data from the National Vital Statistics Report *reveal that the distribution of the duration of human pregnancies is approximately normal with mean* $\mu = 270$ *and* $\sigma = 15$*. Use this normal model to determine ... "*

Assertions

1. Pragmatism calls for simple data to be treated simply, e.g. sets of numbers read from a book or handout.
2. The proper organization of data is a legitimate topic to teach in statistics courses.
3. If we are using computers to analyze data, data ought to be provided in computer readable format or entered by students in a computer readable format.
4. The tabular case/variable format is fundamental and should be emphasized from the start, even when there is only one variable.

Topic 1: Entering Group Data

A psychologist wanted to prove that victims of assault had a lower level of trust in people than those who have not been assaulted. She constructed an inventory to measure level of trust. Fifty victims of assault and fifty people who had not been assaulted took the inventory, with the scores shown below.

Assault Group
20, 21, 21, 22, 24, 24, 29, 29, 31, 32, 32, 32, 32, 33, 34, 34, 34, 35, 35, 36, 36, 36, 36, 36, 36, 36, 37, 39, 39, 39, 40, 40, 41, 41, 42, 43, 44, 44, 45, 45, 45, 46, 47, 47, 47, 50, 53, 53, 56, 58

Control Group
28, 28, 29, 29, 31, 31, 34, 35, 36, 36, 36, 37, 37, 37, 38, 38, 39, 39, 39, 39, 40, 40, 40, 40, 40, 40, 40, 42, 42, 42, 42, 43, 43, 43, 44, 45, 45, 46, 46, 46, 46, 47, 48, 48, 48, 50, 52, 52, 55, 56

Topic 1: Entering Data with Google Spreadsheets

1. Create Google Spreadsheet.

	A	B	C
1	Group	Score	
2	Assault	20	
3	Assault	21	
4	Control	28	

2. Import into the analysis system.

R Commander

Cut and paste ... or as in R generally

R/`mosaic`generally

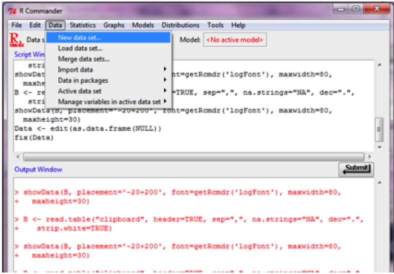
- Export as CSV
- Read with `fetchData()`.

Direct importation from Google will be available through RStudio.

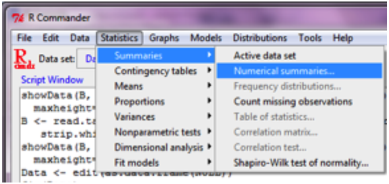
Compute the mean, median, mode, variance, standard deviation, and range of the following dataset (sample). Show your work.

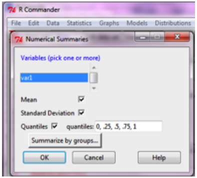
$$2, 12, 9, 3, 9, 7$$

Topic 2: GUI


R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
New data set...
Load data set...
Merge data sets...
Import data
Data in packages
Active data set
Manage variables in active data set
Model: <No active model>
Script Window
Output Window
Submit


Data Editor
File Edit Help
var1 var2 var3 var4 var5 var6 var7


R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set:
Script Window
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
Fit models
Active data set
Numerical summaries...
Frequency distributions...
Count missing observations
Table of statistics...
Correlation matrix...
Correlation test...
Shapiro-Wilk test of normality...


R Commander
File Edit Data Statistics Graphs Models Distributions
Numerical Summaries
Variables (pick one or more)
var1
Mean
Standard Deviation
Quantiles
quantiles: 0, .25, .5, .75, 1
Summarize by groups...
OK
Cancel
Help

Topic 2: Language

1. Quick creation of a small data set.

```
v = c(2, 12, 9, 3, 9, 7)
```

2. Simple descriptive statistics ...

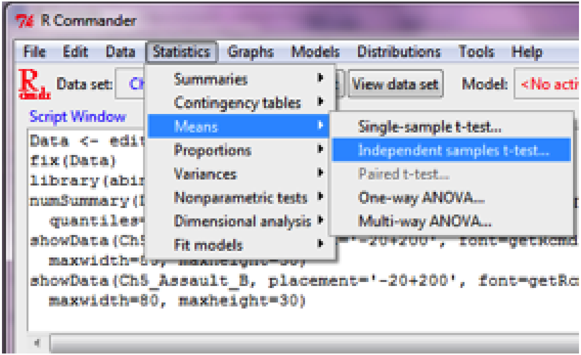
```
mean(v)

[1] 7

median(v)

[1] 8

sd(v)

[1] 3.847

range(v)

[1]  2 12
```

Groupwise statistics:

```
trust = fetchData("Ch5_Assault_B.csv")
mean(Trust_Score ~ Group, data = trust)

Assault Control
  37.74   40.94

sd(Trust_Score ~ Group, data = trust)

Assault Control
  8.962   6.638
```

For the Ch5_Assault data (use Ch5_Assault_B):

1. Compute a 95% confidence interval for the difference between the mean trust scores for the Control and Assault Groups, by hand.
2. Compute a 95% confidence interval for the difference between the mean trust scores for the Control and Assault Groups, using R. Note: It will be slightly larger than the interval obtained by hand, due to the computer's use of the more exact t-scores in place of z-scores.
3. Does your confidence interval contain zero?
4. Interpret the confidence interval. In particular, does it imply that the population of assault victims has a lower mean trust score than controls? Why?


R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set:
View data set
Model:
Script Window
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
Fit models
Single-sample t-test...
Independent samples t-test...
Paired t-test...
One-way ANOVA...
Multi-way ANOVA...
showData(Ch5_Assault_B, placement='-20+200', font=getRc
maxwidth=80, maxheight=30)

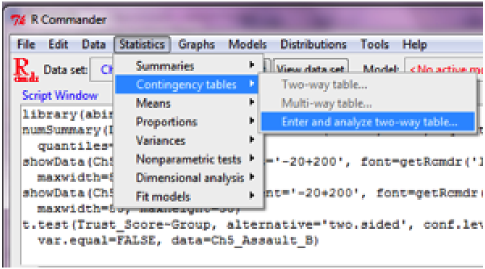
```
trust = fetchData("Ch5_Assault_B.csv")
mean(Trust_Score ~ Group, data = trust)

Assault Control
  37.74   40.94
```

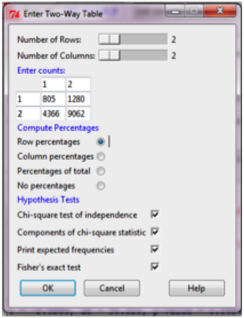
There are better ways to teach this, but ...

```
t.test(Trust_Score ~ Group, data = trust)



Welch Two Sample t-test


data:  Trust_Score by Group
t = -2.029, df = 90.32, p-value = 0.04541
alternative hypothesis: true difference in means is not equ
95 percent confidence interval:
 -6.33321 -0.06679
```

A study of U.S. births found $n = 5171$ babies were born with a finger defect, either syndactyly (fused fingers), polydactyl (extra fingers), or adadctyly (fewer than five fingers). Of these babies with finger defects, it was recorded whether the mother smoked or not. It turned out that 4366 had mothers that did not smoke while pregnant and the remaining 805 did have mothers who smoked while pregnant. In a sample of $10,342$ babies from the same population with normal fingers, 9062 of their mothers did not smoke while pregnant, while the remaining 1280 did smoke while pregnant. Conduct a chi-square test of independence to determine whether finger defect is dependent upon the incidence of mothers smoking at the 1% level of significance. You may R, or do it by hand. (Whitlock and Schluter 2009, p. 226)

Topic 4: GUI


R Commander
File
Edit
Data
Statistics
Graphs
Models
Distributions
Tools
Help
Data set:
View data set
Model:
Script Window
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
Fit models
Two-way table...
Multi-way table...
Enter and analyze two-way table...
t.test(Trust_Score~Group, alternative='two.sided', conf.lev
var.equal=FALSE, data=Ch5_Assault_B)


Enter Two-Way Table
Number of Rows: 2
Number of Columns: 2
Enter counts:
1 2
1 805 1280
2 4366 9062
Compute Percentages
Row percentages
Column percentages
Percentages of total
No percentages
Hypothesis Tests
Chi-square test of independence
Components of chi-square statistic
Print expected frequencies
Fisher's exact test
OK
Cancel
Help

We've got summary data. Since this isn't in case/variable format, we need a different way of reading it in. The instructor could provide it as a table and allow it to be read in — there are several ways to do this.
Here, we'll construct the summary table "by hand" ...

```
fingers = rbind(
  smoke=c(case=805,control=1280),
  nosmoke=c(case=4366,control=9062))
```

Once in tabular form, the χ^2 test is straightforward:

```
chisq.test(fingers)


Pearson's Chi-squared test with Yates' continuity correctio

data:  fingers
X-squared = 29.9, df = 1, p-value = 4.558e-08
```

Perhaps it's more authentic to start with the "raw" data, in case-variable table format.

For the Instructor ...

Many examples are based on summary data. You may want to simulate a "raw" data file that's consistent with the summary.

```
d = expandTable(fingers,vnames=c("Smoker","Group"))
```

```
   Smoker   Group
1 nosmoke control
2 nosmoke    case
```

You can save this as a spreadsheet file

```
write.csv(d,"eCOTS_FingerData.csv",row.names=FALSE)
```

From the student's point of view

The process is as always:

1. Fetch the data.
2. Carry out the analysis.

```
fingers = fetchData("eCOTS_FingerData.csv")
twoway = table(fingers)
```

But is this what you want?

The finger-defect study has a case/control design.

- You had better point this out, or the students will think that the prevalence rate of finger defects is 5171 out of $5171 + 10342$. In fact, the prevalence rate is not indicated by the data.
- The question is not just whether there is evidence, but how big is the effect. With a χ^2 test, you don't get an effect size.
- The standard statistic to use here is an odds ratio.

The Right Test

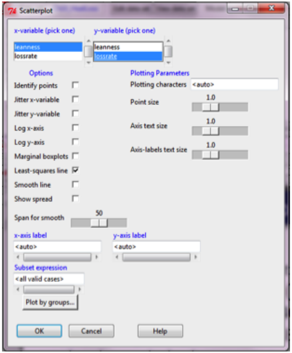
```
fisher.test(twoway)


Fisher's Exact Test for Count Data

data:  twoway
p-value = 5.853e-08
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6958 0.8438
sample estimates:
odds ratio
    0.7661
```

Based on these data, smoking during pregnancy increases the risk of finger defects by between 19 and 44 percent.

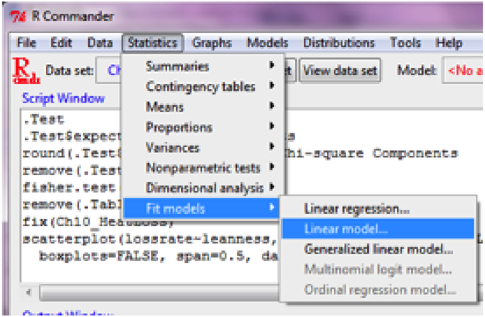
To investigate whether subcutaneous fat provides insulation in humans, Sloan and Keatinge (1973) measured the rate of heat loss by boys swimming for up to 40 min. in water at 20.3 $^\circ C$ and expending energy at about 4.8 kcal/min. Heat loss was measured by change in body temperature, recorded using a thermometer under the tongue, divided by time spent swimming, in minutes. The authors measured an index of body "leanness" on each boy as the reciprocal of the skin-fold thickness adjusted for total skin surface area (in meters squared) and body mass (in kg). (Whitlock and Schluter, 2009, p. 499). The dataset is in the Ch10_HeatLoss.

Leanness (m^2/kg)	7.0	7.0	6.2	5.0	4.4	3.3
Heat loss ($^\circ C$/ min)	0.103	0.097	0.090	0.091	0.071	0.024
Leanness (m^2/kg)	3.6	2.8	2.4	2.1	2.1	1.7
Heat loss ($^\circ C$/ min)	0.014	0.041	0.031	0.010	0.006	0.002

Topic 5: GUI



Scatterplot
x-variable (pick one)
y-variable (pick one)
leanness
lossrate
leanness
lossrate
Options
Identify points
Jitter x-variable
Jitter y-variable
Log x-axis
Log y-axis
Marginal boxplots
Least-squares line
Smooth line
Show spread
Span for smooth
50
Plotting Parameters
Plotting characters
<auto>
Point size
1.0
Axis text size
1.0
Axis-labels text size
1.0
x-axis label
<auto>
y-axis label
<auto>
Subset expression
<all valid cases>
Plot by groups...
OK
Cancel
Help

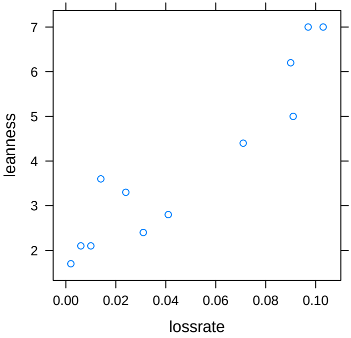


R Commander
File
Edit
Data
Statistics
Graphs
Models
Distributions
Tools
Help
Data set:
View data set
Model:
Script Window
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
Fit models
Linear regression...
Linear model...
Generalized linear model...
Multinomial logit model...
Ordinal regression model...
.Test
scatterplot(lossrate~leanness,
boxplots=FALSE, span=0.5,

Topic 5: Language

Instructors should have a direct pipeline to their students. Using the Internet, this is easily arranged:

```
heat = fetchData("Ch10_HeatLoss.csv")
xyplot(leanness ~ lossrate, data = heat)
```

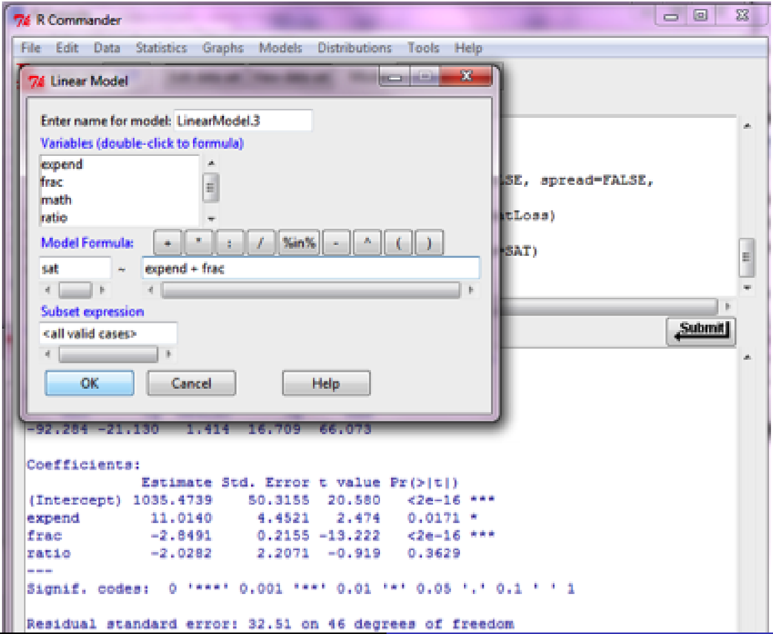



```
lm(leanness ~ lossrate, data = heat)
```

Example: State-by-State SAT Scores and School Spending

Questions:

1. What is the association between statewide mean SAT scores and school spending?
2. What covariates are there?
3. How to deal with the covariates?


R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Linear Model
Enter name for model: LinearModel.3
Variables (double-click to formula)
expend
frac
math
ratio
Model Formula:
+ * : / %in% - ^ ( )
sat
~
expend + frac
Subset expression
<all valid cases>
OK
Cancel
Help
Submit
SE, spread=FALSE,
tLoss)
SAT)
-92.284 -21.130 1.414 16.709 66.073
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1035.4739 50.3155 20.580 <2e-16 ***
expend 11.0140 4.4521 2.474 0.0171 *
frac -2.8491 0.2155 -13.222 <2e-16 ***
ratio -2.0282 2.2071 -0.919 0.3629

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 32.51 on 46 degrees of freedom

Simple regression gives a significant result ...

```
states = fetchData("SAT.csv")
confint(lm(sat ~ expend, data = states))

               2.5 %   97.5 %
(Intercept) 1000.04 1178.546
expend        -35.63   -6.158
```

Expenditure is negatively correlated with SAT scores!
But this result is significantly misleading. There are covariates, especially the fraction of students taking the SAT, that are important.

```
confint(lm(sat ~ expend + frac, data = states))

                2.5 %   97.5 %
(Intercept) 949.909 1037.754
expend        3.788   20.785
frac         -3.284   -2.418
```

After adjusting for `frac`, spending is positively (and significantly) correlated with SAT scores.

Topic 7: The Three R's of Statistical Inference

George Cobb has described the process of statistical inference in terms of the "Three R's":

Randomize, Repeat, Reject

A GUI enables you to simplify referring to an operation, but doesn't necessarily display the logic of the process.
A Language, if properly concise, provides a notation for doing so.

```
heat = fetchData("Ch10_HeatLoss.csv")
coef(lm(lossrate ~ leanness, data = heat))

(Intercept)    leanness
   -0.02691     0.01897
```

Now permute the explanatory variable, implementing the null hypothesis:

```
coef(lm(lossrate ~ shuffle(leanness), data = heat))
```

```
  (Intercept) shuffle(leanness)
1     0.07134           -0.0058
```

```
coef(lm(lossrate ~ shuffle(leanness), data = heat))
```

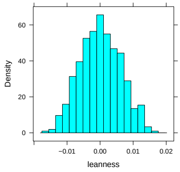
```
  (Intercept) shuffle(leanness)
```

This can be automated:

```
do(5) * lm(lossrate ~ shuffle(leanness), data = heat)

  Intercept   leanness   sigma r-squared
1   0.03707  0.0028394 0.04078  0.019559
2   0.05625 -0.0019969 0.04098  0.009674
3   0.05088 -0.0006432 0.04116  0.001004
4   0.07206 -0.0059827 0.03935  0.086832
5   0.02810  0.0051021 0.03986  0.063152
```

Many iterations ...

```
s = do(1000) * lm(lossrate ~ shuffle(leanness),
     data = heat)
xhistogram(~leanness, data = s)
```




```
prop(~abs(leanness) >= 0.1897, data = s)

TRUE
   0
```

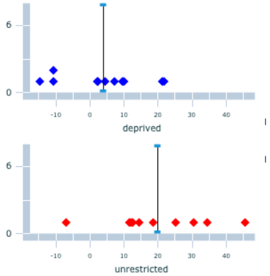
The p-value is very small.

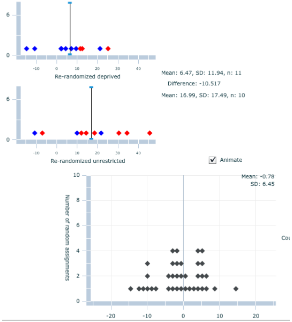
Independent Applets

Animation has its benefits, but you don't need to build your course around animation software.

Example: The Rossman/Chance Applet Collection

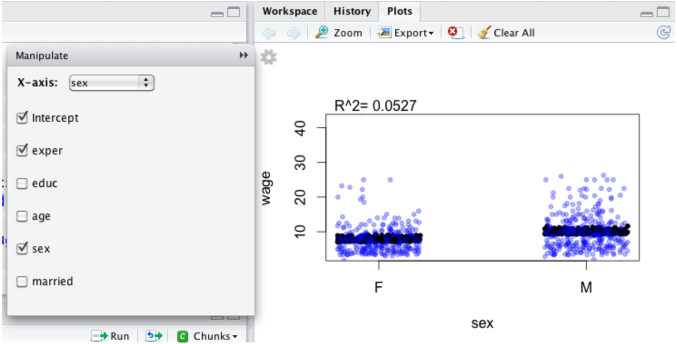
http://www.rossmanchance.com/applets/randomization20/Randomization.html





Combining a GUI with a Language

```
require(mosaicManip)
wages = fetchData("cps.csv")
mLM(wage ~ exper + educ + age + sex + married,
     data = wages)
```



A Proverb ...

> *Give me a fish, and I eat for a day. Teach me to fish, and I eat for a lifetime.*

For statistical computing ...

> *Give me a GUI, and I'll compute for this course. Give me a language and I'll compute for a lifetime.*

For some students, computing for the course may be sufficient; the GUI may be enough.
Choose an appropriate tool for the task that your students face.

But don't be scared of using a language.