R for statistics: first steps

CAR meeting, Raleigh, Feb 26 2011

Peter Aldhous, San Francisco Bureau Chief



NewScientist

peter@peteraldhous.com



A note of caution before using any stats package: Beware running with scissors!

You can bamboozle yourself, and your readers, by misusing statistics.

Make sure you understand the methods you are using, in particular the assumptions that must be met for them to be valid (e.g. normal distribution for many common tests).

So before rushing in, consult:

- IRE tipsheets (e.g. Donald & LaFleur, #2752; Donald & Hacker, #2731)
- Statistical textbooks (e.g. <u>http://www.statsoft.com/textbook/</u> is free online)
- Experts who can provide a reality check on your analysis!

Getting started:

Download R, instructions at: <u>http://www.r-project.org/</u>

Start the program:

R Console	
File Edit Misc Packages Windows Help	
R version 2.10.1 (2009-12-14) Copyright (C) 2009 The R Foundation for Statistical Computing ISBN 3-900051-07-0	*
R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details.	
Natural language support but running in an English locale	
R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.	
Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.	
Loading required package: rcom Loading required package: rscproxy >	
4	b.

Prepare the data:

1) Save as a csv file

	А	В	С	D	E
1	Age	Salary	Sector		
2	74	8	Financial s	services	
3	70	2.13	Manufact	uring	
4	69	7.5	Retail		
5	69	4.82	Manufact	uring	
6	63	1.49	Manufact	uring	
7	62	3.96	Manufact	uring	
8	62	2.34	Retail		
9	61	5.43	Manufact	uring	
10	61	2.54	Financial s	services	
11	61	1.98	Retail		
12	60	7.26	Manufact	uring	
13	59	2.96	Manufact	uring	
14	58	2.17	Retail		
15	58	0.21	Retail		
16	57	11.03	Financial s	services	
17	57	3.17	Financial s	services	
18	56	8.02	Financial s	services	
19	56	3.5	Manufact	uring	
20	56	2.82	Retail		
21	56	2.04	Manufact	uring	
22	55	7.36	Manufact	uring	
23	55	4.98	Retail		
24	55	4.24	Financial s	services	
25	53	4.06	Retail		

2) Point R at the data

R Console	
File Edit Misc Packages W	/indows Help
Source R code New script Open script Display file(s)	·12-14) R Foundation for Statistical Computing
Load Workspace Save Workspace	comes with ABSOLUTELY NO WARRANTY. stribute it under certain conditions. cence()' for distribution details.
Load History Save History	ort but running in an English locale
Change dir	or more information and cite R or R packages in publications.
Print	
Save to File	demos, 'help()' for on-line help, or TML browser interface to help.
Exit	
Loading required packag Loading required packag >	-
•	

First steps in the R command line:

Load and examine the data:

```
> CEOs <- read.csv("CEOs.csv", header=T)
> ls()
[1] "CEOs"
> str(CEOs)
'data.frame': 59 obs. of 3 variables:
$ Age : int 74 70 69 69 63 62 62 61 61 61 ...
$ Salary: num 8 2.13 7.5 4.82 1.49 3.96 2.34 5.43 2.54 1.98 ...
$ Sector: Factor w/ 3 levels "Financial services",..: 1 2 3 2 2 2 3 2 1 3 ...
```

View a basic summary:

> summar	y(CEOs)				
A	ge	Sal	ary	Sector	
Min.	:32.00	Min.	: 0.210	Financial service	es:15
1st Qu.	:45.50	1st Qu.	: 2.500	Manufacturing	:20
Median	:50.00	Median	: 3.500	Retail	:24
Mean	:51.54	Mean	: 4.042		
3rd Qu.	:57.00	3rd Qu.	: 5.395		
Max.	:74.00	Max.	:11.030		

What is the mean age and salary for CEOs in each sector?

> tapply(CEOs\$Age,	CEOs\$Sector, m	ean)
Financial services	Manufacturing	Retail
51.06667	54.30000	49.54167
> tapply(CEOs\$Sala	ry, CEOs\$Sector,	mean)
Financial services	Manufacturing	Retail
5.192667	4.236000	3.160417

An alternative, computing mean and standard deviation in one go:

<pre>> numSummary(CEOs\$Salary,</pre>	groups=CEOs\$Sector,	<pre>statistics=c("mean",</pre>	"sd"))

	mean	sd	n	
Financial services	5.192667	2.361245	15	
Manufacturing	4.236000	2.157409	20	
Retail	3.160417	1.821692	24	



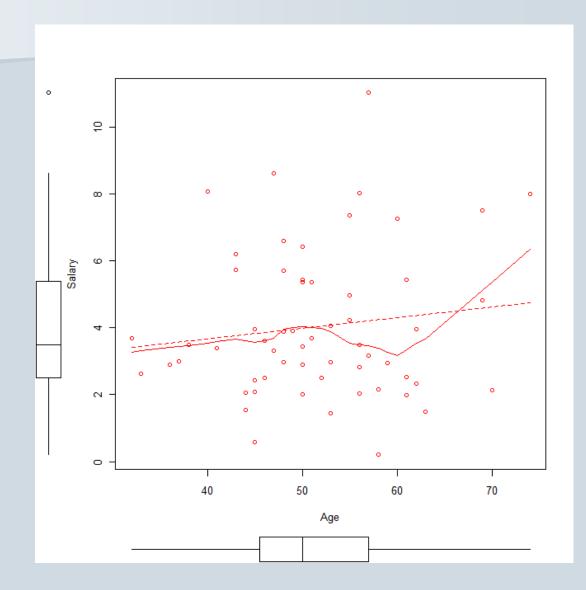
Is there a correlation between age and salary?

```
> cor.test(CEOs$Age, CEOs$Salary, method="pearson")
```

```
Pearson's product-moment correlation
```


Draw a graph to explore the correlation analysis:

> scatterplot(Salary ~ Age, data=CEOs)



Does mean CEO age differ significantly across sectors?

> AnovaAg	ge <-	aov(Age ~ s	Sector, data=Cl	EOs)		
> summar	y(An	ovaAge)				
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Sector	2	251.6	125.776	1.5917	0.2126	
Residuals	56	4425.1	79.019			

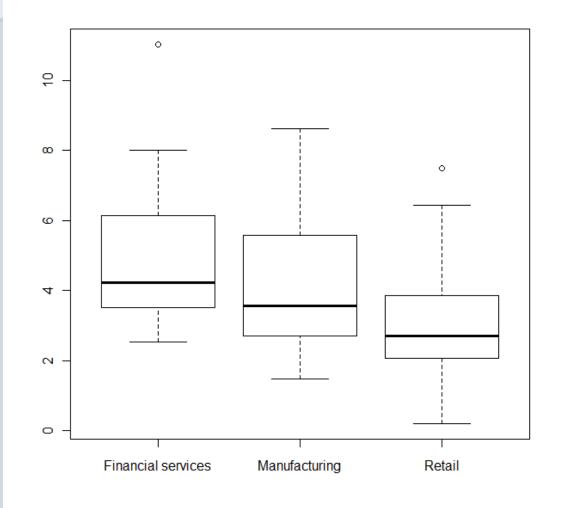
Does mean CEO salary differ significantly across sectors?

> AnovaSalary	<- ao	v(Salary ~ S	Sector, data	=CEOs)	
> summary(An	ovaSa	lary)			
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sector	2	39.266	19.633	4.5279	0.01504 *
Residuals	56	242.817	4.336		
Signif. codes:	0 '***	' 0.001 '**'	0.01 '*' 0.0	5'.'0.1''	1



Draw a graph to explore the distribution of salary by sector:

> boxplot(Salary ~ Sector, data=CEOs)

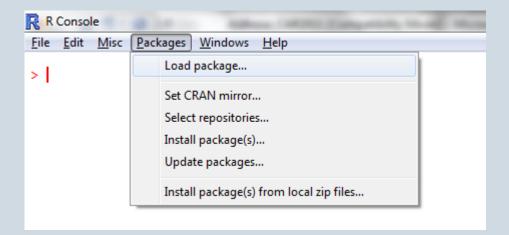


Moving beyond the basic functions: R packages

The R community has written 2800+ extensions to R's basic statistical and graphical functions, available from the CRAN repository: <u>http://cran.r-project.org/web/packages/</u>

Follow the links to find a PDF Reference Manual for each package.

To see what packages you have available:



If you can't find the package you need, select "Install package(s)" and download from the nearest mirror of the CRAN repository

Some examples of R packages:

• **<u>ggplot2</u>**: for sophisticated statistical graphics

• <u>survival</u>: for analysis of data where the dependent variable is the time that elapses to a particular event. Used in medical research to see if drugs prolong life; you might use these methods to analyze the time taken to process permit applications for different groups of people

• <u>rgdal</u>: for geospatial data analysis; allows you import shapefiles and other forms of geodata and perform GIS-type analyses

Back to our analysis of CEO salary by sector:

Previous summary:

> AnovaSalary <	<- ao	v(Salary ~ S	ector, data:	=CEOs)	
> summary(Ano	vaSa	lary)			
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sector	2	39.266	19.633	4.5279	0.01504 *
Residuals	56	242.817	4.336		
Signif. codes: 0	! ***	' 0.001 '**'	0.01 '*' 0.0	5'.'0.1''1	

Load multcomp package using "Load package" or:

>require(multcomp)
Loading required package: multcomp
Loading required package: mvtnorm
Loading required package: survival
Loading required package: splines

Perform multiple comparisons:

```
>Comparisons <- glht(AnovaSalary, linfct = mcp(Sector = "Tukey"))
> summary(Comparisons)
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = Salary ~ Sector, data = CEOs)

Linear Hypotheses:

```
Estimate Std. Error t value Pr(>|t|)

Manufacturing - Financial services == 0 -0.9567 0.7112 -1.345 0.3759

Retail - Financial services == 0 -2.0322 0.6854 -2.965 0.0121 *

Retail - Manufacturing == 0 -1.0756 0.6305 -1.706 0.2114

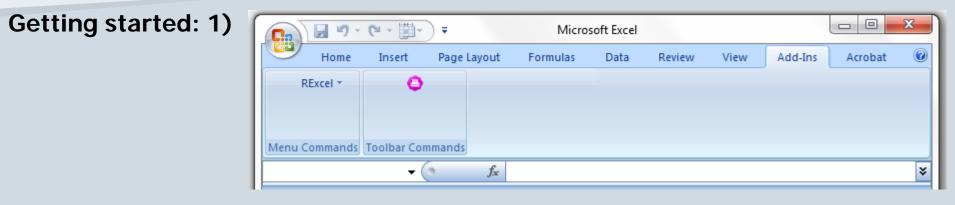
----

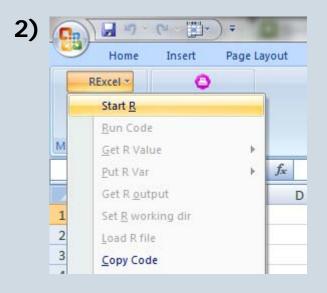
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Adjusted p values reported -- single-step method)
```

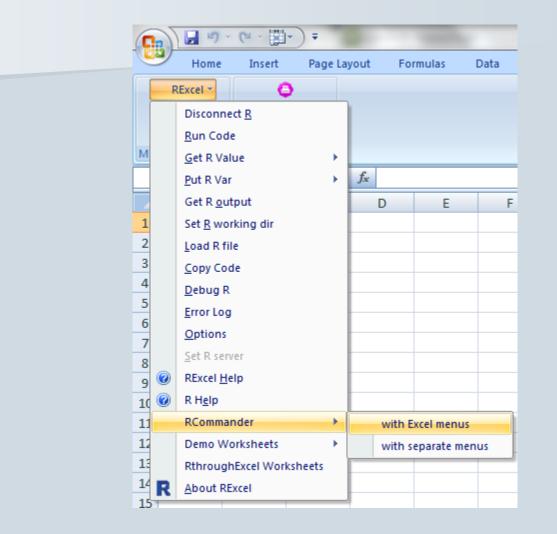
If the command line isn't for you: R through Excel

Download RExcel, instructions at: http://rcom.univie.ac.at/download.html





Getting started: 3)



Now we're ready to go:

0	100	- (2 - E] -) ₹	-		-	and the second			Boo	KI 7% R Commander
	Home	Insert	Page L	ayout Fo	ormulas	Data	Review	View	Add-Ins	Acrobat	Script Window
Men	RExcel *			R File - D Dataset: No		* Model		node *	ributions *	Tools + Help +	
	A1		. (*	f_{x}							
	A	В	С	D	E	F	G		н	I J	
1											
2											4
3											Output Window
4											
5											
6								_			
7											
8											
9 10								_			
10											
12											
13											
14											
15											
16											
17											
18											
19											
20											

Load the data into R:

Highlight the data range, right click and select "Put R DataFrame"

6) 🖬 🤊	- (2 - 📑	 →) =		and Installed		-		
6	Home		Page l	avout	t Formulas	Data	Review	View	Add-Ins
	RExcel *	10)		File - Data - Sta	tistics x (Tranhs x Mode	ale y Diet	ributions x
	REACCI		-		iset: No active dat				inducions
				Data	iser. No active dat		uei. No active	noue ·	
Me	nu Command	s Toolbar (ommands	Calil	bri - 11 - A		- % • 🚿	irs	
	A1	~	• (•						_
				В				<u></u>	
4	A	B	C		D E Run <u>c</u> ode in Rcn		F G	1	H
1	Age 74	Salary	Sector	R	-	Iur		H	
2 3	74		Financia Manufac		<u>R</u> un code			H	
3 4	69		Retail		<u>G</u> et R Value			H	
5	69		Manufac		<u>P</u> ut R Var			H	
6	63		Manufac		Get R <u>D</u> ataFrame	:			
7	62		Manufac		Put R D <u>a</u> taFrame				
8	62		Retail	1	Rcmdr Get		•		
9	61		Manufac		Get R <u>O</u> utput				
10	61		Financia		Name Ran <u>g</u> e				
11	61		Retail		Prettyformat Nu	mbers			
12	60		Manufad	X	Cu <u>t</u>				
13	59		Manufad		<u> </u>				
14	58		Retail	1	Paste				
15	58		Retail		Paste <u>S</u> pecial				
16	E7	11.00	Financia		ruste <u>special</u>			-	

Compute summary statistics:

(5-	(2 - 10		CEOs - Microsoft Excel									
<u> </u>	9	Home	Insert	Page La	yout	Formula	s	Data	Review	Vi	ew	Add-Ins	Acrobat	
		RExcel *	0		R File	∗ Data ∗	Stati	stics 💌	Graphs * Mod	dels *	Dist	ributions * T	ools * Help *	
					Dataset:	CEOs		Summ	aries	•		Active data	set	
								Conti	ngency tables	•		Numerical s	ummaries	
M	lenu	Commands	Toolbar Cor	mmands				Mean	s	•		Frequency	distributions	
		CEOs	- (∫x A	\ge		Propo	rtions	•		Count missi	ing observations	s
	4	А	В	С	D			Varian	nces	•		Table of sta	tistics	ĸ
3	8	48	6.59 F	inancial	service	s		Nonpa	arametric tests	•		Correlation	matrix	
3	9	48	5.72 F	inancial	service	s		Dimen	sional analysi	s 🕨		Correlation	test	
4	0	48	3.88 F	inancial	service	s		Fit mo	dels	►		Shapiro-Wi	lk test of norma	lity
4	1	48	2.98 N	/lanufact	uring									

Variables (pick one or more)
Age <u>Salary</u>
Mean 🔽
Standard Deviation 🔽
Quantiles 🔽 quantiles: 0, .25, .5, .75, 1
Summarize by groups
OK Cancel Help



Output in R Commander:

```
- -
                                                                            X
74 R Commander
 Script Window
 numSummary(CEOs[,c("Age", "Salary")], groups=CEOs$Sector,
   statistics=c("mean", "sd", "guantiles"), guantiles=c(0,.25,.5,.75,1))
                                                                      Submit
 Output Window
 > numSummary(CEOs[,c("Age", "Salary")], groups=CEOs$Sector,
   statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
 Variable: Age
                      mean sd 0% 25% 50% 75% 100% n
 Financial services 51.06667 9.924477 32 47.50 50.0 56.50 74 15
 Manufacturing 54.30000 8.227681 40 48.75 53.5 60.25 70 20
            49.54167 8.747567 33 44.00 50.0 55.25 69 24
 Retail
 Variable: Salary
                      mean sd 0% 25% 50% 75% 100% n
 Financial services 5.192667 2.361245 2.54 3.5100 4.24 6.155 11.03 15
 Manufacturing 4.236000 2.157409 1.49 2.8075 3.56 5.505 8.62 20
 Retail
                3.160417 1.821692 0.21 2.0750 2.72 3.775 7.50 24
```

ANOVA of CEO salary by sector:

(5	9 -	(* - 📳					-			-	-	(EOs - N	Microsoft Ex	cel
4	9	Home	Insert	Page La	ayout	Formula	as	Data	Review	Vi	ew	Add-Ins	Acrobat			
	1	RExcel *	Ç		R File	• Data •	Stati	istics 🔻 G	raphs * Mod	dels -	Dist	ributions *	Tools 🕆 Help	-		
					Dataset:	CEOs		Summa	ries	►	e 🕆					
								Contin	gency tables	►						
М	enu	Commands	Toolbar Co	ommands				Means		•		Single-san	nple t-test			
		CEOs	•	(•	∫x A	ge		Propor	tions	•		Independ	ent samples t-	test		
	4	А	В	С	D			Varian	ces	•		Paired t-te	st		К	
38	3	48	6.59	Financial	service	s		Nonpa	rametric tests	•		One-way /	NOVA			
39	Э	48	5.72	Financial	service	s		Dimens	sional analysi	s 🕨		Multi-way	ANOVA			
4(48		Financial		s		Fit mod	fels	►					_	
4	1	48	2.98	Manufact	uring		_				0					

7⁄2 One-Way Analysis of	Variance
Enter name for model:	AnovaSalary
Groups (pick one)	Response Variable (pick one)
Sector	Age Salary
Pairwise comparisons o	f means 🔽
ОК С	Cancel Help

Output in R Commander:

summary(AnovaSalary)		
numSummary(CEOs\$Salary , groups=CEOs\$Se	ctor, statistics=c("mean", "	sd"))
.Pairs <- glht (AnovaSalary, linfct = mo	p(Sector = "Tukey"))	
confint(.Pairs) # confidence intervals		
cld(.Pairs) # compact letter display		
old.oma <- par(oma=c(0,5,0,0))		
plot(confint(.Pairs))		
par(old.oma)		
remove(.Pairs)		
4		k.
		Cubmit
Output Window		Submit
> AnovaSalary <- aov(Salary ~ Sector, d	lata=CEOs)	
> summary(AnovaSalary) Df Sum Sq Mean Sq F value		
Sector 2 39.266 19.633 4.5279		
Residuals 56 242.817 4.336	0.01301	
Signif. codes: 0 '***' 0.001 '**' 0.01	. '*' 0.05 '.' 0.1 ' ' 1	
Linear Hypotheses:		
hinear hypotheses.	Estimate lwr upr	
Manufacturing - Financial services == (-	
Retail - Financial services == 0		
Retail - Manufacturing == 0	-1.0756 -2.5922 0.4411	
> cld(.Pairs) # compact letter display		



Statistical graphics with R made easy:

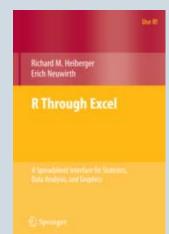
Web application for ggplot2 package: http://www.yeroon.net/ggplot2/



R tutorials:

We've only scratched the surface with these examples. To learn how to perform other common statistical analyses in R (e.g. linear and logistic regression; t-tests; tests for normality; non-parametric methods), here are some resources:

- <u>Quick-R</u>: intended for people switching from SAS/SPSS/Stata, but easy to follow even if you haven't used these packages
- **<u>Rtips</u>**: comprehensive, but not quite so easy to navigate
- **<u>Burns Statistics</u>**: put together by a specialist in quantitative finance; tutorials on the menu at left
- <u>R Through Excel</u>: well-illustrated guide to using the add-in



R for statistics: first steps

CAR meeting, Raleigh, Feb 26 2011

Slides: <u>http://www.peteraldhous.com/CAR/Aldhous_CAR2011_RforStats.pdf</u>

Simulated data used here: <u>http://www.peteraldhous.com/Data/CEOs.csv</u>

Peter Aldhous, San Francisco Bureau Chief



NewScientist

