Minimal R for Intro Stats

Randall Pruim

July 2, 2014

"Less volume, more creativity."

Mike McCarthy, Head Coach, Green Bay Packers

Mike McCarthy had signs proclaiming his "Less volume, more creativity" mantra hung on the office walls of all of his coordinators during one off-season. When asked about it, he said, "A lot of times you end up putting in a lot more volume, because you are teaching fundamentals and you are teaching concepts that you need to put in, but you may not necessarily use because they are building blocks for other concepts and variations that will come off of that . . . In the offseason you have a chance to take a step back and tailor it more specifically towards your team and towards your players."

Statistics instructors using R face a similar dilemma. R is capable of so much that it is tempting to include this, and then that, and then the other, and then one more thing. Vectors and lists and recycling and coercion and functions and . . . It all seems so fundamental to the way R works. And when mastered, these concepts do become building blocks for other concepts and variations.

But when looking back at the end of a term, we have to admit that some of these things really aren't necessary to get the job done, and may do more harm than good for beginners. We too need to take a step back and tailor things toward our students and their abilities and needs. The colored commands on the next page are sufficient for an Introductory Statistics course that includes ANOVA, regression, and resampling techniques. The others are optional extras. This is followed by a 1-page sampler showing usage examples for some of the functions.

Note: These pages are intended as a guide for instructors, not as a reference card for students. Although they may also be useful for students, they would need supplementing with additional details.

The list of functions we present are not the only sufficient set of functions, but they were carefully chosen to fit as much as possible into a small number of paradigms. In particular,

1. We make use of the "formula interface" whenever possible.

   Students will need the formula interface to do regression an ANOVA. Since we are going to teach it anyway, we use formulas as consistently and as often as we can. In some cases, my colleagues and I have written new functions or expanded the use of existing functions to serve this end. These functions are available in the `mosaic` package and are indicated in the comments in our palette.

2. We use `lattice` graphics.

   R has three separate high level plotting libraries (base, `lattice`, and `ggplot2`). Each has its advantages, but we choose `lattice` because it uses the same formula interface used elsewhere and because it encourages students to things about disaggregating data according to the values of covariates by making this very easy to do.

Whether you use this list or some other list, we encourage you to make a complete list of the commands you want your students to learn over the course of a semester. Organize them by topic. Organize them again by syntactic structure. Ask yourself how they look as a whole. Have you chosen a set of functions that fit well together? And most importantly: What is your creativity to volume quotient?

Help

```
apropos()
?
??
example()
```

Basic Calculations

Basic calculation is very similar to a calculator.

```
# basic ops: + - * / ^ ( )
log()
exp()
sqrt()
```

```
log10()
abs()
choose()
```

Randomization/Simulation

```
rflip()      # mosaic
do()         # mosaic
sample()     # mosaic augmented
resample()   # with replacement
```

```
shuffle()    # mosaic
rbinom()
rnorm()      # etc, if needed
```

Formula Theme

The following syntax (often with some parts omitted) is used for graphical summaries, numerical summaries, and inference procedures.

```
goal( y ~ x | z, data=...,
                 groups=... )
```

For plots:

- `y`: is y-axis variable
- `x`: is x-axis variable
- `z`: conditioning variable (separate panels)
- `groups`: conditioning variable (overlaid graphs)

For other things:

'`y ~ x | z`' can usually be read '`y` is modeled by (or depends on) `x` differently for each `z`'.

See the sampler for examples.

Distributions

```
pbinom(); pnorm();
xpnorm()    # mosaic
pchisq(); pt()
qbinom(); qnorm();
qchisq(); qt()
plotDist()  # mosaic
```

Numerical Summaries

These functions have a formula interface to match plotting.

```
favstats()   # mosaic
tally()      # mosaic
mean()       # mosaic augmented
median()     # mosaic augmented
sd()         # mosaic augmented
var()        # mosaic augmented
```

```
quantile()   # mosaic augmented
prop()       # mosaic
perc()       # mosaic
rank()
IQR()        # mosaic augmented
min(); max() # mosaic augmented
```

Graphics (mostly lattice)

```
bwplot()
xyplot()
histogram() # mosaic augmented
densityplot()
qqmath()
makeFun()   # mosaic
plotFun()   # mosaic
```

```
ladd()      # mosaic
dotPlot()   # mosaic
bargraph()  # mosaic
xqqmath()   # mosaic
```

Interactive Graphics (RStudio)

```
mPlot(data=HELPrct, 'scatter')
mPlot(data=HELPrct, 'boxplot')
mPlot(data=HELPrct, 'histogram')
```

Inference

```
binom.test()  # mosaic augmented
prop.test()   # mosaic augmented
chisq.test()
t.test()      # mosaic augmented
model <- lm() # linear models
anova( model )
summary( model )
makeFun( model )   # mosaic
resid( model )
plot( model )
TukeyHSD( model )  # mosaic aug
plot( TukeyHSD( model ) )
```

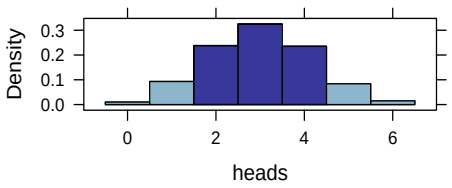
```
confint()      # mosaic augmented
pval()         # mosaic
fisher.test()
xchisq.test()  # mosaic
model <- glm() # logistic regression
```

Data

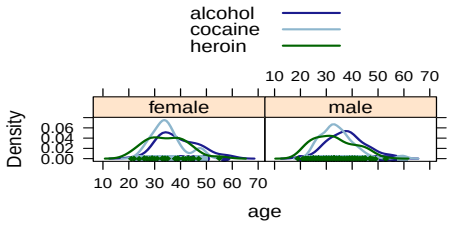
```
read.file()    # mosaic
nrow(); ncol()
summary()
str()
names()
head()
subset()
factor()
c()
cbind(); rbind()
transform()
```

```
merge()
relevel()
ntiles()      # mosaic
cut()
```

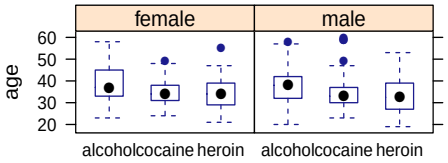
```
rflip(6)

[1] 3
attr(,"n")
[1] 6
attr(,"prob")
[1] 0.5
attr(,"sequence")
[1] "T" "H" "T" "H" "H" "T"
attr(,"verbose")
[1] TRUE
attr(,"class")
[1] "cointoss"

do(2) * rflip(6)

  n heads tails   prop
1 6     3     3 0.5000
2 6     1     5 0.1667

coins <- do(1000) * rflip(6)
tally(~heads, data = coins)


   0    1    2    3    4    5    6
  10   93 238 325 235   84   15

tally(~heads, data = coins, format = "perc")


    0     1     2     3     4     5     6
 1.0   9.3  23.8  32.5  23.5   8.4   1.5

tally(~(heads >= 5 | heads <= 1), data = coins)


 TRUE FALSE
  202   798

histogram(~heads, data = coins, width = 1,
    groups = (heads >= 5 | heads <= 1))
```




```
tally(~sex + substance, data = HELPrct)

        substance
sex      alcohol cocaine heroin
  female      36      41     30
  male       141     111     94

mean(age ~ sex, data = HELPrct)

female   male
 36.25  35.47

diffmean(age ~ sex, data = HELPrct)

diffmean
 -0.7841

favstats(age ~ sex, data = HELPrct)

  .group min Q1 median   Q3 max  mean
1 female  21 31     35 40.5  58 36.25
2   male  19 30     35 40.0  60 35.47
     sd   n missing
1 7.585 107       0
2 7.750 346       0

densityplot(~age | sex, groups = substance,
    data = HELPrct, auto.key = TRUE)
```

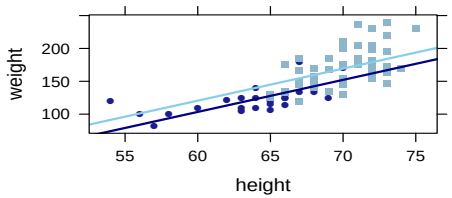



```
bwplot(age ~ substance | sex, data = HELPrct)
```



```
pval(binom.test(~sex, data = HELPrct))

  p.value
1.932e-30

confint(t.test(~age, data = HELPrct))

mean of x     lower     upper     level
    35.65     34.94     36.37      0.95
```

```
model <- lm(weight ~ height + gender,
            data=Heightweight)
wt <- makeFun(model)
wt( height=72, gender="male")

    1
179.1

xyplot(weight ~ height, groups=gender,
       data=Heightweight)
plotFun(wt(h,gender="male") ~ h, add=TRUE,
                col="skyblue")
plotFun(wt(h,gender="female") ~ h, add=TRUE,
                col="navy")
```




```
plotDist("chisq", df = 4)
```