The R Project
Ross Ihaka
Department of Statistics
The University of Auckland

What is R?
- A statistical software system created at the University of Auckland in the early to mid 1990s.
  - Originally, created as a testbed for ideas.
  - Later, developed as a teaching tool.
  - Finally, adopted by a wider community of developers and users.
- A free software system created and maintained by an international collaboration of statisticians (from academia and industry).
  - A core “language” and set of libraries.
  - A very large collection of user-contributed libraries (currently more than 1800).

Getting Started
- R was created by Robert Gentleman and Ross Ihaka to explore some statistical computing ideas.
- It combines ideas from the “Scheme” and “S” programming languages.
- The name “R” is a play on the name “S” and the first initials of the authors.
- Success of the initial ideas lead to a goal of using R to teach elementary statistics courses.
- The system was introduced to a wider audience in a 1996 paper in The Journal of Computational Statistics and Graphics.

Wider Distribution
- Word about R began to spread to colleagues overseas.
- Requests for copies soon followed.
- Versions were distributed under the Free Software Foundation’s GNU Public License (Version 2).
- A “Core Development Team” of volunteers formed to support the further development of R.
- A network of “CRAN” websites was formed to distribute the software.
- It is now estimated that there are over a million R users worldwide (estimate by Intel Capital).

Simple Dot Charts
The R philosophy is “make standard things easy.”
The data set “animalsSpeed” contains observations of land animal speeds in mph.
The following statement converts mph to kph and sorts the values into ascending order.
> speed = sort(1.6 * animalsSpeed)
With the units converted we can display the data in a Cleveland “dot chart”
> dotchart(speed,
main = "Land Animal Speeds",
xlab = "Speed (km/hr)")
Customised Dot Charts

A second part of the R philosophy is that the software should not unduly restrict what you can do.

Here is a highly customised dotchart:

```r
dotchart(speed, pch = 19,
          limits = lim(c(0, 2), c(12.5, 112.5)),
          width = c(1, 3),
          shortticks = c(1, seq(12.5, 11.25, by = 20)),
          labelticks = c(0, 2, seq(5, 100, by = 25)),
          main = "Land Animal Speeds",
          xlab = "Speed (km/hr)")
```

Statistics with R

```r
> res <- lm(Volume ~ Height + Girth, data = trees)
> summary(res)
```

Call:
```
lm(formula = Volume ~ Height + Girth, data = trees)
```

Residuals:
```
Min 1Q Median 3Q Max
-6.090 -2.919 0.259 2.648 8.487
```

Coefficients:
```
Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.80877  4.6892  1.243 0.210523
Height  0.33930  0.1832  1.857 0.072974 .
Girth  4.70582  0.3943 11.986  < 2e-16 ***
```

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.902 on 28 degrees of freedom
Multiple R-squared: 0.946,  Adjusted R-squared: 0.942
```

```r
> plot(res)
```

Capabilities

- R is the most fully-featured statistical software system available.
- In addition to the base-level functionality there are over 1800 extensions packages available, including many “hot off the press” statistical techniques.
- It comes with the ability to communicate with other systems like `Minitab`, `S`, `SAS`, `SPSS`, `Stata`, `Systat`, etc., as well as machine learning systems such as `KEA`.
- The graphics system is considered one of the best available.
- Support is available from some of the world’s best statisticians on the `R-help` mailing list.
- Best of all — it’s all free!

Critique

- R was designed for working with quite small data sets.
- It works well for data sets with tens of variables and thousands of observations.
- It does not handle large datasets well.
- Offloading data processing to a specialist database system can help a little.
- Database support can be found in the packages `DBI`, `RJDBC`, `RODBC`, `RMySQL`, `ROracle`, `RPostgreSQL`, etc.

The Future

- R is a stable, widely-used software system which is still undergoing incremental improvement.
- Because it is widely used and depended on by a large user base, it is no longer a suitable platform for experimentation.
- The need to deal with much larger datasets indicates that a “new” system is needed.
- Initial work on this is being conducted by Duncan Temple Lang, Brendan McArdle and Ross Ihaka.
- Initial indications are that it will be possible to build a system which is hundreds of times faster and which can handle much larger datasets.

Resources

- [www.r-project.org](http://www.r-project.org) — The main R Project website.
- [cran.r-project.org](http://cran.r-project.org) — The Comprehensive R Archive Network.
- [cran.stat.auckland.ac.nz](http://cran.stat.auckland.ac.nz) — A local CRAN mirror.
- [journal.r-project.org](http://journal.r-project.org) — The R Journal.