Statistics and Programming in R

David Stephens and Niall Adams

[d.stephens,n.adams]@imperial.ac.uk

Department of Mathematics, Imperial College London

29/30 September 2005

Timetable

	Day 1	Day 2
10.00 - 12.00	Introduction, Demo, Essentials	Programming Constructs
13.00 - 14.30	Basic analysis techniques	Writing Functions
15.00 - 16.30	Regression ANOVA	Efficiency Issues Automation

All sessions will include lectures, demos and tutorials.

We will be using R for Windows, v2.0.1.

Some familiarity with probability and statistics is assumed.

R

From the R documentation:
R is a GNU project that provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.

One of R 's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R

From the R documentation

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

Much statistical functionality is provided by the user community.

New methods are often implemented and distributed in R .

History

R is readily traced back to the S language – “a language for programming with data” – developed primarily by John Chambers at Bell labs.

The S language was developed into a commercial product called Splus by a company now called Insightful (`http://www.insightful.com/`). This development has been mainly in the user interface and graphics and interoperability functionality. Splus uses version 4 of the S language.

R started out as an Open Source system “not unlike” version 3 of the S language, developed by Ross Ihaka and Robert Gentleman, Based on a different implementation. R is now developed and maintained by a core group.

Getting R

Can always download the most recent release (currently 2.1.1) either as source or pre-compiled of R from a Comprehensive R Archive Network (CRAN) site (`http://cran.uk.r-project.org/` for UK).

R consists of a *base system*, providing the language and interface and contributed *packages* providing the enhanced statistical and graphical functionality.

In general, it is better to upgrade the whole system on an occasional basis than to update individual packages on the release of new versions. Different releases are deliberately installed in separate locations, so it is possible to have concurrent versions.

Install R

To obtain and install a pre-compiled windows binary:

1. `rwXXXX.exe` from CRAN (currently `rw2011.exe`).
2. Execute, and follow instructions. Defaults are sensible.
3. Start `R` .
4. Select `packages|install`
5. Select a CRAN mirror
6. Select the required packages

Note that (i) installing packages can be slow, (ii) pre-requisites are automatically satisfied, and (iii) can experience problems with version numbers.

Demo: Starting and Packages

What is R ?

R should be regarded as an implementation of the S *programming language* - not a statistical package with limited programming tagged on (like SAS, SPSS, Minitab). As such, it provides

- Programming language constructs
- Data structures
- Functions for mathematics, statistics and graphics.

In particular, note that *everything* in R is an *object*...it will become clear what this means later!

Demo: Simple Data Analysis

To illustrate ideas, let us conduct some simple data analysis, involving a regression model. Data are distance, climb and record times for 35 Scottish hill races. Note the following features:

- Interaction via a command line interface (scripts later)
- The value of a function may be assigned a name.
- Graphics occurs as side effects to function calls.
- Functions can operate on arguments of different types.
- Prompt to save workspace with `q()`

R Essentials: Data

Consider

```
hills <- read.table("hills.txt")
```

We have called the function `read.table`, with a single, unnamed argument, and assigned, via the assignment operator `<-`, the value of the call to the object `hills`.

All entities in R are *objects* and all objects have a *class*. The class dictates what the object can contain, and what many functions can do with it. Thus, `read.table` is an object of class `function`, and `hills` is an object of a type determined by the function, in this case, a `data.frame`. Everything we do will involve creating and manipulating objects.

R Essentials: Data

Objects are accessed by name. Objects that we create should have names that are meaningful to us, and should follow the `R` syntax rules: letters and numbers (except first position) and periods (avoid underscore).

A common problem is naming objects that are already in use by `R` , reserved words like `if` or `break` or system objects like `mean`, `TRUE`, and `pi`.

Another common irritation for new users is that `R` is cases sensitive, so `T` (a system representation of logical true) is different to `t` (function for matrix transpose).

R Essentials: Functions

Functions are called by name, followed by a bracketed list of arguments

```
plot(x,y)
plot(lm(time~dist))
```

Functions return a value. Graphics functions in addition have side effects, that create and modify plots. The argument list can take various formats, and not all arguments always need be specified.

As with the example above, functions behave differently according to the class of their arguments.

Demo: Vectors

Consider the quadratic equation

$$x^2 + x + 1 = 0$$

In usual notation, we have coefficients $a = b = c = 1$. This equation has real roots if the *discriminant*

$$b^2 - 4ac > 0$$

- Store coefficients as an object
- Compute discriminant
- Construct a plot

R Essentials: Data Vectors

Consider

```
coeffs <- c(1,1,1)
class(coeffs)
length(coeffs)
names(coeffs)
```

Here `coeffs` is a *vector* of type `numeric`. Vectors are a fundamental class in `R` – there are no scalars. All objects also have a `length`, which is only informative for certain classes.

The `c` (for combine) function is a basic tool for creating vectors.

Vectors can have names, queried and modified by the `names` function. Note the function call occurring on the right hand side of the assignment operator.

R Essentials: Data: Vectors

The classes available for vectors are

- `numeric` - real numbers.
- `character` - character strings, arbitrary length.
- `logical` - vector of logical `TRUE` and `FALSE` values.
- (signed) `integer` - integer values.
- `complex` - for complex numbers $a + bi$, $i = \sqrt{-1}$.
- `list` - a 'clothesline' object, each numbered element can contain any `R` object.

R Essentials: Data: Vectors

Vectors can be created other ways

```
seq(-3,3,length=200) # regular spaced points
-2:2 # sequence of integers
x <- vector("numeric",length=20) # create
c(x,x) # combine two vectors
```

Care must be taken in combining vectors of different types. R will deploy a set of internal rules to resolve the class of the combined vector.

Note that # is the comment operator.

R Essentials: Data: Vectors

The classes available for vectors are

- `numeric` - real numbers.
- `character` - character strings, arbitrary length.
- `logical` - vector of logical `TRUE` and `FALSE` values.
- (signed) `integer` - integer values.
- `complex` - for complex numbers $a + bi$, $i = \sqrt{-1}$.
- `list` - a 'clothesline' object, each numbered element can contain any `R` object.

R Essentials: Data: Vectors

Access the elements of a vector by number, or name

```
coeffs[2]
coeffs["2"]
```

We may wish to remove the names from a vector

```
names(coeffs) <- NULL
```

The object `NULL` represents nothing, the empty set. `NULL` has length zero, and `R` will deploy special rules when `NULL` occurs.

R Essentials: Arithmetic

Simple arithmetic on constants follows the usual precedence rules

```
ax <- 1 ; bx <- 2 ; cx <- 3
x <- ax+bx*cx # x =7
x <- bx/bx*cx/ax+bx # x=5
x <- (bx/bx)*(cx/ax)+bx # better
axˆbx # 2 raised to 3
10 %% 9 # 10 mod 9
```

Use parentheses to simply expressions. Note that these must balance. If they do not R responds with either a syntax error, or a continuation request (a + prompt).
Recall that vectors are a fundamental class of object in R . The examples above therefore involve arithmetic on vectors, like ax.

R Essentials: Arithmetic

Things get more complicated. Consider

```
x <- seq(-3,3,length=200)
y <- coeffs[1]*x^2+coeffs[2]*x+coeffs[3]
```

Here we are multiplying a vector of length 1 by a vector of length 200. This problem is dealt with by *recycling* the shorter vector until it matches the length of the longer vector. Fractional recycling will result in a warning message.

A vector can have zero length. This is represented as

`numeric(0)` As a function call, this creates a zero length

vector. This can be useful if we need to construct a vector of unknown length.

R Essentials: Special Values

With computer arithmetic we require extra symbols to represent missing values and mathematical pathologies.

- Missing values are represented as `NA`. The IEEE special values for floating point arithmetic are also used: `Inf`, `-Inf` and `NaN`. `NaN` is used for indefinite results like `Inf/Inf`.
- There are functions for elementwise testing of vectors for the presence of special values, of the form `is.XX`, where `XX` can be `na`, `nan`, `finite`, `infinite`.
- Special values can cause problems in programming, and if we are being careful, we should check for their presence.

R Essentials: Simple Functions

There are a large collection of functions that operate on numeric vectors.

```
round(x,2) ; trunc(x); ceiling(x);
abs(x); log(x); log10(x) ; sqrt(x) ; exp(x);
sin(x) ; acos(x); tanh(x) ; # radians
```

In each case, the result of the function is a vector of the same length as the argument. Note that `log` can take a second argument, the base of the logarithm. The default base is e. Note that the following are equivalent

```
log(x); log(x,exp(1))
log(x=x,exp(1)) ; log(x=x,base=exp(1))
```

R Essentials: Simple Functions

Standard functions for reducing numeric vectors

```
min(x); max(x)
sum(x); prod(x)
```

and the *cumulative* equivalents

```
cumsum(x); cumprod(x)
```

Can already see how to usefully combine functions

```
sum(x)/length(x) # mean
prod(1:5) # factorial
sum(x-mean(x))^2/(length(x)-1) # sample
variance.
```

R Demo: Logic

A full set of logical operators and functions are available.

R Essentials: Logic

Logical conditions can be applied to numeric vectors

```
x <- seq(-3,3,length=200) > 0
```

Now x is a logical vector of length 200. The condition $>$ has been applied elementwise.

The other comparison operators are >=, < , <=, ==,and !=. The final two are *exact* equality and inequality, respectively. As such, they should only be applied to entities that are represented exactly, like integers.

Logical vectors are subject to the usual recycling rules.

R Essentials: Logic

Logical values can be combined and modified with `!`, the negation operator, `&` the intersection operator (logical AND) and `|`, the union operator (logical OR).

Truth tables

A	B	A AND B	A OR B
T	T	T	T
T	F	F	T
F	F	F	T
F	T	F	F

```
c(T,T,F,F) & c(T,F,F,T)
```

Be careful with negation, the symbol `!` is overloaded,and may be interpreted as a shell escape. Again, brackets are used to resolve such problems.

R Essentials: Logic

A handy feature of logical vectors is that they can be used in ordinary arithmetic.

```
A <- c(T,F,T)
A + 1
```

The resulting vector is `c(2,1,1)`. R has noted the combination of logical and numeric vectors, and *coerced* the logical vector to numeric, by mapping TRUE to 1 and FALSE to zero. There are a range of *coercion* functions, pre-fixed with "as", like `as.logical`.

The use of logical vectors in ordinary arithmetic means we can easily count numbers of TRUE or FALSE in a comparison `sum(x > 0.5)`

R Essentials: Functions for Logical Vectors

The function `any` returns value 1 if at least one element of its logical argument is `TRUE`. The function `all` returns value 1 if all elements are `TRUE`. Note we can implement similar functionality directly

```
sum(x)> 1 #x logical
sum(x) == length(x)
```

Sometimes useful are the functions for set operations : `union` - $A \cup B$,`intersect` - $A \cap B$ and `setdiff` - $A \cup \bar{B}$.
I have found `setdiff` very useful in classification experiments.

R Essentials: Factors

`factors` are vector like objects (class?) used to store sets of categorical data. For example

```
drinks <-
factor(c("beer","beer","wine","water")
```

Here, we have constructed a vector of class character, and converted it to a factor. The factor is displayed without quotes. It is informative to examine how factor is stored, using

```
unclass(drinks)
```

Refer to individual elements in the same way as a vector.

R Demo: Data Frames and Matrices

The most commonly used class for routine data analysis is the `data.frame`, suitable for displaying tabular data. The `matrix` class is suitable for linear algebra.

R Essentials: Data Frames

The typical class used for data analysis in R is the *data frame*. These are suitable for storing the usual observations in rows, variables in columns data format. The advantage of this class is that the columns can be of different classes: numeric, logical, character and so on.

We have already seen an example of a data frame, the object `hills`.

Data frames can have row names, common to all columns.

```
row.names(hills)
```

For a data frame, the column names are accessed with the function `names`. Strictly, a data frame is a *list*, where all elements are required to have the same length.

R Essentials: Data Frames

Usually, a data frame will be created by a suitable call to a data import function. It is also possible to combine vectors into a data frame. For example

```
data.frame(X1=1:10,X2=I(letters[1:10]),X3=factor(letters[1:10]))
```

```
data.frame(1:10,I(letters[1:10]),factor(letters[1:10]))
```

By default, R will attempt to pick row names from the constituent vectors, and otherwise will use numeric row names, and guess at column names if they are not given. Row names can be provided as an extra argument.

Note the use of `I()` to override the default behaviour of converting character vectors to factors.

R Essentials: Data Frames

Earlier, we used the attach command to make the columns of the hills data frame available by name. This is sometimes useful, although there is a risk of masking other objects. To undo, use `detach(hills)`.
If a data frame has names, we can refer to the columns using the `$` operator as follows

```
hills$time
```

This is now simply a vector,and can be manipulated as such. The real power of factors arises when they are constituents of data frames. In a statistical model, the factor will be treated in an appropriate manner.

R Essentials: Matrices

While a data frame can collect together vectors of different class, and be displayed in a matrix-like manner, to explicitly operate on mathematical matrices we use the `matrix` class, which requires that all elements have the same type.
Create a matrix with something like

```
matrix(1:12,nrow=3,ncol=4)
```

Note the use of named arguments in the function call. A warning occurs if the vector being made into a matrix cannot recycle to `nrow` $\times$ `ncol`. A matrix has dimensions, accessed with the `dim` function, that returns the number of rows and columns.
Names can be associated with rows and columns using the function `dimnames`. Here, the names are a list, with a component for each dimension.

R Essentials: Combining Data

We have seen that vectors can be combined with the c function.
Matrices and data frames be combined using he function `rbind` and `cbind`, for row and columnwise combination respectively. For example

```
xx <- cbind(x,x)
xxx <- rbind(x,x)
rbind(xx,xxx) # Error
```

Note that the dimensions of the objects being combined must be compatible.

R Essentials: Indexing

We have seen that we can refer to individual components of vectors. More general facilities are available for selecting components from vectors, matrices and data frames.

To refer to the ith row, jth column element of a matrix or data frame, use `x[i,j]`.

There are more general ways of indexing objects, such that `i` and `j` can be: a vector of positive or negative integers, a logical vector, a vector of character strings, or empty.

R Essentials: Indexing

Examples

```
x <- 1:10
names(x) <- letters[x]
x[1:3], # elements 1,2,3
x[c(1,10)] # elements 1 and 10
x[c(-1,-2)] # remove 1 and 2
x[ x > 5] # elements > 5
x[c("a","d")] # elements 1 and 4
x[] # all elements
jj1 <- matrix(1:100,ncol=10)
jj1[1:5,] # first 5 rows, all cols
jj1[1:4,x[x <3]]
```

R Essentials: Manipulating DFs

Applying transformations to a single column in a data frame is straightforward

```
hills$time <- round(hills$time*60)
```

Groups of columns can be handled similarly, by an appropriate indexing operation.

```
x.df[,1:3] <- x.df[,1:3]/2
```

Note that the division operator here is applied element wise to each element.

R Essentials: Operations on DFs

As we saw above, data frames can participate in arithmetic like operations. The usual rules apply, vectors will be recycled - sometimes giving strange results.

Some functions will operate elementwise on data frames, like `log`.

Other times we need to be operate on columns only, and the functions `lapply` and `sapply` provide functionality for such a procedure. The difference between the two function is that the former returns a list, while the latter attempts to simply the result into a vector or matrix.

```
x <- matrix(1:10,ncol=2)
lapply(x,max)
sapply(x,max)
```

R Essentials: Lists

We have mentioned lists a few times. Lists are the most general class in R . A list is simply a numbered collection of objects, of *any* class.

```
x.lis <-
list(a=1:10,b=letters[1:3],b=matrix(1:10,ncol=2))
x.lis$1
x.lis[[2]]
```

We have already seen the use of the $ operator. Elements of a list can also be accessed by their index number, using the *double* square brackets operator. The usual indexing operations can also be applied to a list.

The c function can also be used with lists.

R Essentials: Operating on Matrices

We distinguish computation with data frames from computation with matrices. We have elementwise computations

```
x.mat <- matrix(1:10,ncol=2)
x.mat+1
x.mat + x.mat
```

As usual we need to be careful about how recycling rules (which are complex for such situations) will apply. We also have matrix multiplication from linear algebra

```
x.mat %*% t(x.mat)
```

where `t` is the matrix transpose function. If the matrix and vector dimensions do not conform, an error message results.

R Essentials: Operating on Matrices

To compute

$$X^\top y$$

where X is a vector and y is matrix, could consider

```
t(X) %*% y
crossprod(X,y)
```

Note that the latter is more efficient. The function `crossprod` with a single matrix argument X computes $X^\top X$.

Typical linear algebra functions are available: `eigen`, `svd`, `qr`, `solve`, and so on.

R Essentials: Operating on Matrices

We use the `apply` function to do an identical computations on rows or columns of a matrix. The function prototype is

```
apply(X, MARGIN, FUN, ...)
```

where `X` is a matrix, `MARGIN` refers to rows ($=1$) or columns($=2$), `FUN` is the name of the function to be applied to each row or column, and `...` is a special symbol meaning extra optional arguments, in this case, for the function `FUN`.

```
apply(x,1,sum) # rows
apply(x,2,sum) # columns
```

Where possible we prefer using `apply` (and the related function `sweep`) to explicit looping, for efficiency reasons.

R Essentials: Object Attributes

An attribute is an R object attached to another R object by name. Objects can have any number of attributes, which are represented as a list.

For example, the dimensions (and dimnames) of a matrix are attributes.

```
attributes(x.mat)
```

Attributes are often used for storing ancillary information, derivative information in optimisation problems, for example.

R Essentials: Function Arguments

To examine the arguments for a function use `args`.

```
args(c)
args(pmax)
args(lm)
```

In the first case `NULL` is returned, meaning unspecified arguments, where an arbitrary number of arguments can be given.

In the second and thirds cases, the arguments include `...`, referring to unspecified arguments.

Argument lists include named values with specified defaults, in the format `name=value`.

R Essentials: Calling Functions

We have seen function calls with both specified and unspecified arguments. In calling functions arguments can either be specified by

- Order. Provide arguments in the order given by the function prototype.
- Name. Provide arguments explicity by name, as `name=value`. Only sufficient letters of the name to uniquely identify it are required.

These two approaches can be mixed. For example

```
plot(x,y,type="l")
```

R Demo: Usage Issues

rm ls and so on. Tactics
data entry - also input output
Naming conventions and tactics
storage
scripting
getting stuff in

R Object Storage

R objects that we create occupy a *workspace* that we can examine with

```
ls(all=T)
objects()
```

Note that names that start with a period are hidden from the `ls` command, and are useful system objects, like `.Random.seed`.

There are a collection of databases that R uses to store objects. This collection is maintained as a list called the *search path*, accessed with the `search` function.

R Object Storage

The default behavior of `attach` puts a list in the second position of the path, such that its elements can be accessed by name. The database in position 1 is the working database.

All objects in the workspace are stored in memory. When we exit `R` , we are prompted to save a workspace image. Doing this means that the workspace is stored in its current state, in a file (called `.RData` by default), and can be recovered for further work at a future date.

It is possible to have multiple `.Rdata` files, and switch between them. This provides a convenient mechanism for collecting together different projects.

R Object Storage

Storing everything in memory has some implications for how we work, especially if we are doing memory intensive computations. In such cases keep the number of duplicate and intermediate results should be kept to a minimum. A handy trick is to use names like `jj1`, `jj2` for such results, then routine delete them with

```
rm(list=objects(pattern="jj*"))
```

The function `object.size` provides an estimate, in bytes of the memory allocation for an object.

R Essentials: Scripts

The R console provides a convenient interface for simple commands. For more complicated work, such as programming, R provides scripts, where a sequence of R command can be entered, and processed later. To start a new script choose

```
File -> New Script
```

Commands are entered here, and selected for execution by highlighting followed by <CTRL> R.

Scripting is a very useful feature, and for most tasks will be the default work mode.

R Simple Data IO

If data is represented in a file in a simple delimited tabular format, it is easiest to use `read.table`. Use `write.table` to write a data frame to a file.

The `scan` function is sometimes useful for numeric vectors. This can be used both interactively, for entering numbers, or for reading a stream of numbers from a file.

To manually enter data, create a dummy data frame, and invoke the data editor, as follows

```
a.df <- data.frame()
fix(a.df)
```

Essentials: Import from other systems

The R Data Import/Export document says ". . . reading a binary data file written by another statistical system. This is often best avoided. . . ". This is an area where R is not as evolved as Splus. R has limited functionality for reading binary objects from EpiInfo, Minitab, S-PLUS, SAS, SPSS.

Import is possible from spreadsheet style regular grids in text formats. Direct access to a .xls file can be managed, but is not recommended. Better to output the desired parts of the sheet in simple delimited tabular format.

Some limited access to RDBMS systems is possible, using appropriate packages.

R Getting Help

R has various interactive help facilities. The most useful, to access the manual pages for a specific command, is simply to use the `?` function. For example

```
?mean
```

Like UNIX manual pages, the R manual pages include a "See Also" and "Examples" section, which can be very useful. To conduct a more general search, akin to unix `apropos`, use `help.search`. For example

```
help.search("regression")
```

The function `help.start` will fire up the HTML help system. Lots of good stuff here!

R : Demo: Simple Statistics

R : Statistics: Numeric Summaries

Finally, we can start to look at some statistical functions. A full range of summary statistics are available, including

```
mean(x) ; mean(x,trim=0.95)
median(x)
sqrt(var(x)); var(x,y); var(x.mat)
range(x)
cor(x); cor(x,y); cor(x.mat); cor(x.mat,y.mat)
```

Note different behaviour for `cor` and `cov` depending on the argument list. Some functions have functionality for dealing with missing values (`Na`). The `mean` function, for example, has argument prototype `na.rm=FALSE`.

R Statistics: Summary Function

The summary function is particularly useful with many of the statistical functions. When applied to a numeric vector

```
summary(1:20)
```

the function produces a 6 point summary, comprising the minimum, maximum, lower and upper quartiles, and the median and mean. Applied to a matrix, the summary function generates this data for each column.

If the arguments include missing values, `summary` additionally counts the number.

R : Statistics: Probability Distributions

R includes functions for computing quantities associated with a variety of distributions. The generic prototype for these function is

`*dist(args)` where `*` is one of

- `p` (probability),
- `d` (density),
- `q` (quantile),
- `r` (random number)

and `dist` is the nickname of a distribution.

For example, to generate 20 (pseudo) random variates for a specific Normal distribution

```
rnorm(20,mean=2,var=3)
```

whereas for a Chi-squared distribution

```
rchisq(20,df=5)
```

Extra distributions are available in the package `SuppDists`.

R Statistics: Random number generation

R provides a collection of sophisticated random number generators. The default is the "Mersenne Twister". This requires a start value, called the *seed*. The random number sequence is completely specified by this. It is sometimes useful to fix the seed so as to achieve a repeatable sequence. Do this with

```
set.seed(1)
```

To sample from a finite population, use the function `sample`. For example, to sample from a biased coin experiment

```
sample(c("Head","Tail"), 10,probs=c(0.3,0.7),replace=T)
```

By default, all elements of the population are sampled with equal probability.

R : Statistics: Graphical Summaries

R has impressive graphics functionality - except for dynamic graphics. A potential downside for new users is that there is no GUI for graph construction.
A frequent data analytic task is comparing a sample with a distribution. This is often done with a QQ plot. Typically we plot theoretical quantiles on the horizontal axis, and empirical quantiles on the vertical axis. For example to compare a sample with an Exp(1) distribution

```
plot(qexp(ppoints(x),1),sort(x))
```

The function `ppoints` generates a sequence of probability points at which to evaluate the theoretical distribution.

R : Statistics: Graphical Summaries

If our data is consistent with the theoretical distribution, the points should fall on a straight line through the origin

```
abline(0,1)
title("QQ plot") # add a title
```

Such plots are most frequently used in residual analysis.
Note that R provides function `qqnorm` for comparing against a standard normal distribution, and `qqplot` for comparing two samples.
Another simple way of comparing two samples is as follows

```
plot(c(x,y),rep(0:1,c(length(x),length(y))),xlab="",ylab="")
```

Note the use of the extra arguments to override the default axis labelling.

R : Statistics: Graphical Summaries

Other useful graphical tools include the histogram, and the box and whisker plot.

```
hist(x,prob=T)
boxplot(x)
```

Histogram are generated using a default binning strategy. To specify the number of bins, modify the argument `nbreaks`. Specify the breakpoints by providing a vector for the argument `breaks`.

A box and whisker plot displays the median, upper and lower quartiles, and whiskers extending to the normal distribution based 5% and 95% points. All observations beyond these points are flagged as stars.

R Statistics: Graphical Summaries

Box and whisker plots, and to a lesser extent histograms, are useful for comparing multiple samples. Box plots can take a model specification (of which more later), or a list. For example, suppose vector `x` contains observations of three groups, and vector `ind` is a code, with a value 1,2, or 3 representing the group, then

```
boxplot(split(x,ind))
```

will produce a display with one box plot for each class. The function `split` divides an object according to values in an indicator vector.
Displaying overlaid histograms is more problematic, since we must match the breakpoints for each display. An alternative is generate three histograms in the same figure.

R Statistics: Graphical Summaries

The `R` function for general control of graphical parameters is `par`. This has many (!) arguments. The argument `mfrow` determines how many plots will be placed on the figure. For example, to display two separate histograms, one above the other

```
par(mfrow=c(2,1))
hist(x)
hist(y)
```

The vector valued argument `mfrow` is the number of rows and columns of plots to be placed in the figure. By default, the plots are displayed in row major order.

R Statistics: Graphical summaries

The `plot` function can is most useful for displaying how one variable changes as a function of another. We have seen various example already. Let us look at adding things to a plot.
Consider the Pima Indians data: a collection of variables observed on a particular group of native American Indians who are healthy or have diabetes. This data includes measurements of tricep skinfold and blood glucose level. After attaching the data frame

```
plot(triceps,glucose,type="n",xlab="Tricep",ylab="glucose")
plot(triceps[diabetes=="neg"],glucose[diabetes=="pos"])
points(triceps[diabetes=="pos"],glucose[diabetes=="pos"],col=2,pch=2)
```

We use the `type="n"` argument to set up a plot to accommodate all the data, without displaying anything, then put down the pieces separately.

R Graphical Summaries

The `col` argument specifies a colour, and the `pch` argument specifies the plotting symbol. It may be useful to add a legend, with the `legend` function. This function has (simplified) prototype

```
function (x, y = NULL, legend, ...
```

So we need to specify coordinates for the legend. For this example, a reasonable choice is

```
legend(40,50,c("Diabetes","Healthy"),pch=1:2)
```

If we had been plotting data connecting with lines we would use the argument `lty` which refers to line style.
Add text to a plot with the `text` function.

R Graphics: Portability

There are a variety of formats R graphics can be exported to. One option is to create the figure, then use `File -> Copy to the Clipboard`, and select appropriately for the target application.
Another option is to save the file in a specific format. Again, this can be done via the `File` menu. This can also be achieved with commands, by embedding the graphics commands between a call to a driver and a driver termination call. For example

```
jpeg("file.jpg")
...graphics commands
dev.off()
```

creates a JPG file containing the result of the graphics commands.

Statistical Objectives

Suppose that we have observed experimental outcomes

- $x_1, ..., x_n$ on the n trials
- we have observed $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$, termed a **random sample**.

This sample can be used to answer qualitative and quantitative questions about the nature of the experiment being carried out.

The objectives of a statistical analysis can be summarized as follows. We want to, for example,

- **SUMMARY : Describe** and **summarize** the sample $\{x_1, ..., x_n\}$ in such a way that allows a specific probability model to be proposed.
- **INFERENCE : Deduce** and **make inference about** the parameter(s) of the probability model θ.
- **TESTING : Test** whether θ is "**significantly**" larger/smaller/different from some specified value.
- **GOODNESS OF FIT : Test** whether the probability model encapsulated in the mass/density function f, and the other model assumptions are **adequate** to explain the experimental results.

Types Of Study

The data in an experimental study can be obtained in a number of different situations that can be classified as follows:

- one sample
- two independent samples
- two related samples (“within individuals”)
- two related samples (predictor and response)
- k independent samples
- k related samples (multivariable, within individuals)

One Sample

- repeated, independent observations of some phenomenon
- aim to summarize “location/scale” of sample
- test hypothesized *target* values
- test distributional summaries

ONE SAMPLE ANALYSIS

Two Independent Samples

- repeated, independent observations under different conditions *(fixed effects)*
- control/treatment
- healthy/affected
- aim to compare two samples
- same mean level ?
- same variability ?
- same distribution ?

TWO SAMPLE ANALYSIS

Two Related Samples I

Paired Analysis

- two repeated observations on same experimental units
- two observations on different but related *(matched)* experimental units
- start/end of trial
- matched/paired analysis
- any change in mean level ?

TWO SAMPLE PAIRED ANALYSIS

Two Related Samples II

Predictor / Response

- two related observations on different features of same experimental units
- predictor/response
- objective is to predict response
- normal data/non-normal data
- correlation ?
- any predictive ability ?
- classification ?

REGRESSION ANALYSIS

K Independent Samples

- $k \geq 2$ sets of independent observations (fixed effects)
- different experimental conditions (control, level 1,...,level $k-1$)
- ordered levels ?
- normal/non-normal data ?
- any change in mean measure across treatment levels ?

ANOVA ANALYSIS

K Related Samples

- $k \geq 2$ sets of observations (on same experimental units)
- time dependent
- same feature, different experimental conditions (fixed effects)
- different (related) features
- normal/non-normal data ?
- regression/correlation ?
- comparison of fixed effects ?

REPEATED MEASURES/MULTIVARIATE ANALYSIS

Exploratory Data Analysis

The four principal features that we need to assess in the data sample are

1. The **location**, or “average value” in the sample.
2. The **mode**, or “most common” value in the sample.
3. The **scale** or **spread** in the sample.
4. The **skewness** or **asymmetry** in the sample.

These features of the sample are important because we can relate them directly to features of probability distributions.

Numerical Summaries

- Sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample variance: either (S^2 or s^2 may be used)

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Sample skewness

$$\kappa = \frac{1}{nS^2}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

- Sample quantiles: suppose that the sample has been sorted into ascending order $x_{(1)} < ... < x_{(n)}$. The pth quantile, $0 < p < 100$ is $x^{(p)} = x_{(k)}$ where k is the nearest integer to $pn/100$. Special cases include

Median	m	$= x^{(50)}$, the 50th quantile
Lower quartile	q_{25}	$= x^{(25)}$, the 25th quantile
Upper quartile	q_{75}	$= x^{(75)}$, the 75th quantile
Inter-quartile range	IQR	$= q_{75} - q_{25}$
Sample minimum	$x_{\min}$	$= x_{(1)}$
Sample maximum	$x_{\max}$	$= x_{(n)}$
Sample range	R	$= x_{(n)} - x_{(1)}$

Key aspects of the sample can be summarized using the first four sample moments and their transformations

- 1st Moment⇀**LOCATION** : $\frac{1}{n}\sum_{i=1}^{n} x_i$
- 2nd Moment⇀**SCALE** : $\frac{1}{n}\sum_{i=1}^{n} x_i^2$
- 3rd Moment⇀**SKEWNESS** : $\frac{1}{n}\sum_{i=1}^{n} x_i^3$
- 4th Moment⇀**KURTOSIS** (“heavy-tailedness”) : $\frac{1}{n}\sum_{i=1}^{n} x_i^4$

Graphical Summaries

- histogram
- boxplot
- scatterplot

Transformations

It may be necessary or advantageous to consider data **transformations;**

- $y_i = \log_{10} x_i$
- $y_i = \log x_i = \ln x_i$
- $y_i = \sqrt{x_i} = x_i^{1/2}$
- $y_i = x_i^{\alpha}$ some α
- $y_i = \log\left(\frac{x_i}{1 - x_i}\right)$

NOTE: This is not any form of statistical trickery, but may be necessary to allow formal statistical assessment

Transformations

In R , transformations are straightforward using the algebraic operators outlined above:

```
# Square
y<-x^2

# Exponential
y<-exp(-x)
```

Transformations

```
#log (to various bases)
#Base e (natural log)
y<-log(x)

# Base 10
y<-log10(x)

# Base 2
y<-log2(x)

# Base 7.2
y<-log(x,base=7.2)


# Square root
y<-sqrt(x)
```

Resources

- WEB
  - `http://www.r-project.org/`
  - `http://cran.uk.r-project.org/`
- BOOKS
  - John M. Chambers. *Programming with Data*. Springer, New York, 1998.
  - William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Fourth Edition. Springer, 2002.
  - William N. Venables and Brian D. Ripley. *S Programming*. Springer, 2000.
  - Peter Dalgaard. *Introductory Statistics with R*. Springer, 2002.

Resources

- Julian J. Faraway. *Linear Models with R*. Chapman & Hall/CRC, Boca Raton, FL, 2004
- Paul Murrell. *R Graphics*. Chapman & Hall/CRC, Boca Raton, FL, 2005.

available on web