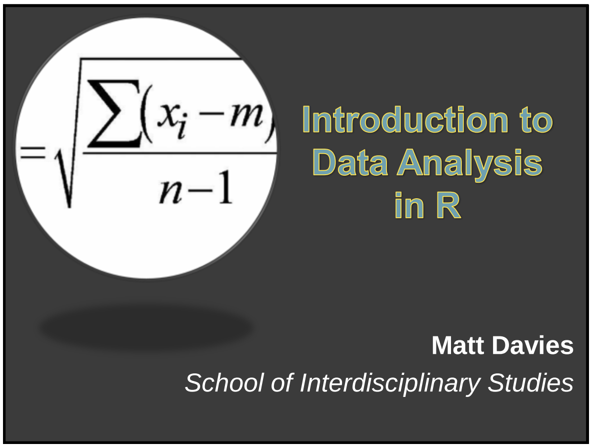
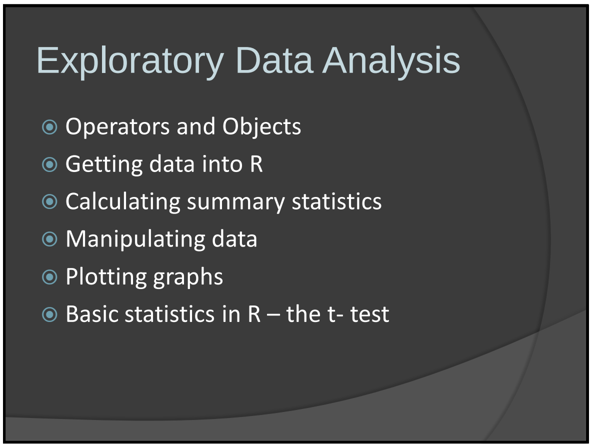


$$= \sqrt{\frac{\sum(x_i - m)}{n-1}}$$

Introduction to Data Analysis in R

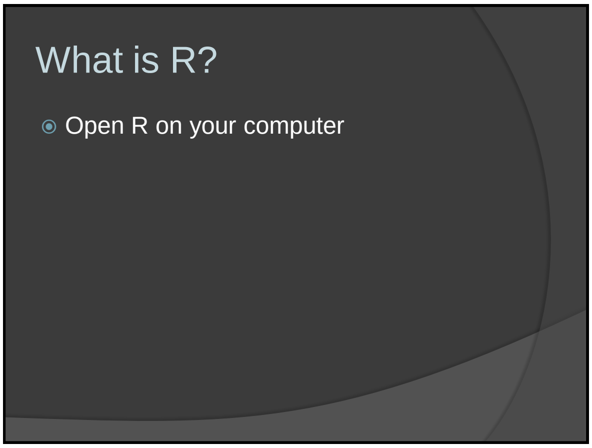
Matt Davies

School of Interdisciplinary Studies



Exploratory Data Analysis

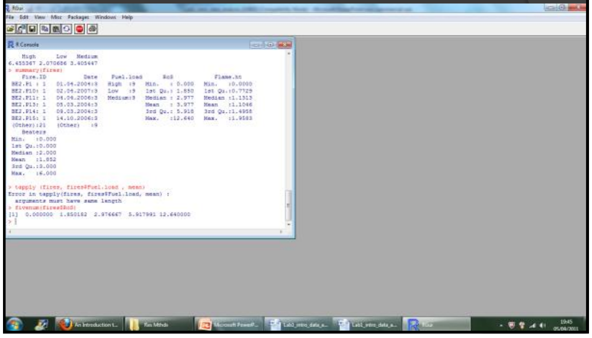
- Operators and Objects
- Getting data into R
- Calculating summary statistics
- Manipulating data
- Plotting graphs
- Basic statistics in R – the t- test



What is R?

- Open R on your computer

R



```
   High      Low   Medium
6.455367 2.070686 3.405447
> summary(fires)
   Fire.ID            Date     Fuel.load      RoS            Flame.ht
 BE2.P1 : 1   01.04.2004:3   High  :9   Min.   : 0.000   Min.   :0.0000
 BE2.P10: 1   02.04.2007:3   Low   :9   1st Qu.: 1.850   1st Qu.:0.7729
 BE2.P11: 1   04.04.2006:3   Medium:9   Median : 2.977   Median :1.1313
 BE2.P13: 1   05.03.2004:3              Mean   : 3.977   Mean   :1.1046
 BE2.P14: 1   09.03.2004:3              3rd Qu.: 5.918   3rd Qu.:1.4958
 BE2.P15: 1   14.10.2006:3              Max.   :12.640   Max.   :1.9583
 (Other):21   (Other)   :9
   Beaters
 Min.   :0.000
 1st Qu.:0.000
 Median :2.000
 Mean   :1.852
 3rd Qu.:3.000
 Max.   :6.000

> tapply (fires, fires$Fuel.load , mean)
Error in tapply(fires, fires$Fuel.load, mean) :
  arguments must have same length
> fivenum(fires$RoS)
[1]  0.000000  1.850182  2.976667  5.917991 12.640000
> |
```

R –important notes

- Data format
  - Every individual observation (known as a "case") must be a unique line in the data table
- R is case sensitive
  - In object names or lists of factors "Fire.1" is different to "fire.1"
- There's always more than one way of doing things!

R – what can it do?

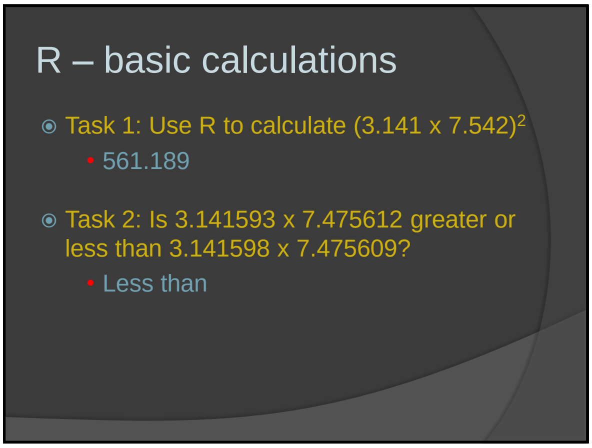
- Acts as a simple calculator using "operators"

  + - / * ^

- Includes "logical operators"

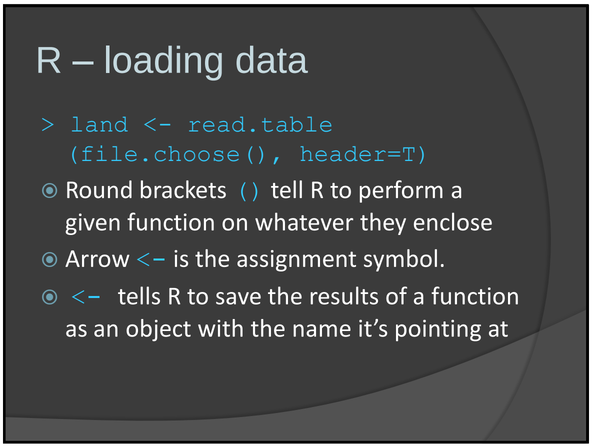
  < <= > >= == !=

- Contains pre-programed functions for running a HUGE variety of statistical tests
- Can create very flexible graphs



R – basic calculations

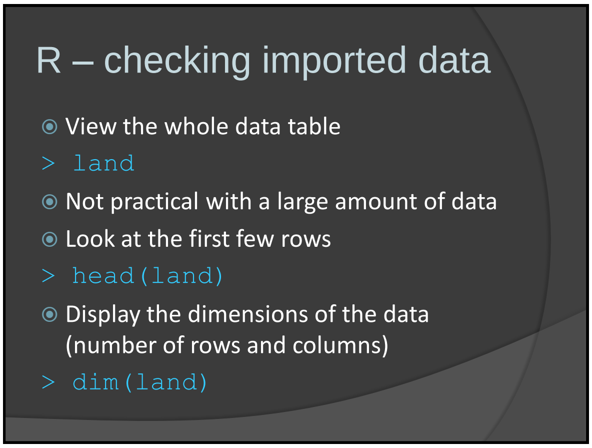
- Task 1: Use R to calculate (3.141 x 7.542)^2
  - 561.189
- Task 2: Is 3.141593 x 7.475612 greater or less than 3.141598 x 7.475609?
  - Less than



R – loading data

```
> land <- read.table
  (file.choose(), header=T)
```

- Round brackets `()` tell R to perform a given function on whatever they enclose
- Arrow `<-` is the assignment symbol.
- `<-` tells R to save the results of a function as an object with the name it’s pointing at



R – checking imported data

- View the whole data table

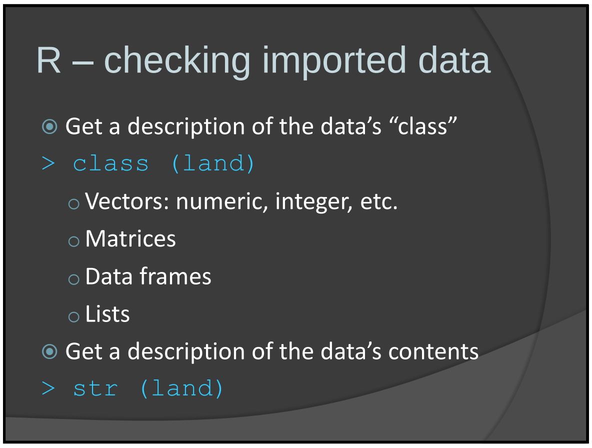
```
> land
```

- Not practical with a large amount of data
- Look at the first few rows

```
> head(land)
```

- Display the dimensions of the data (number of rows and columns)

```
> dim(land)
```



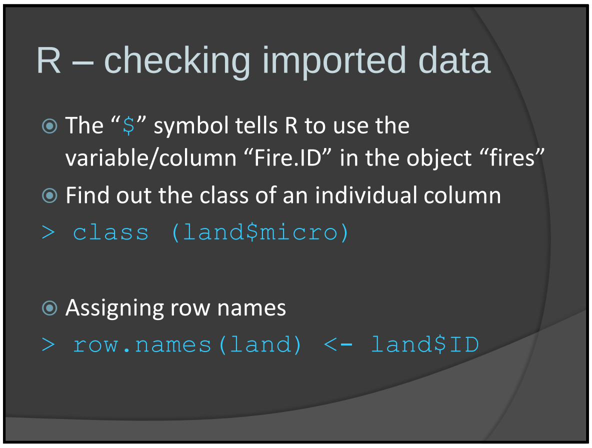
R – checking imported data

- Get a description of the data’s “class”

```
> class (land)
```

  - Vectors: numeric, integer, etc.
  - Matrices
  - Data frames
  - Lists
- Get a description of the data’s contents

```
> str (land)
```



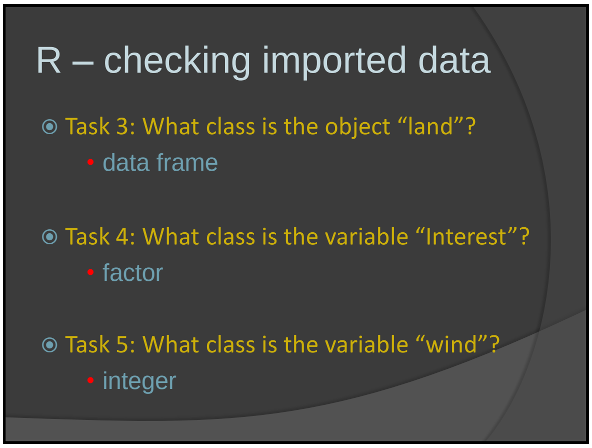
R – checking imported data

- The “$” symbol tells R to use the variable/column “Fire.ID” in the object “fires”
- Find out the class of an individual column

```
> class (land$micro)
```

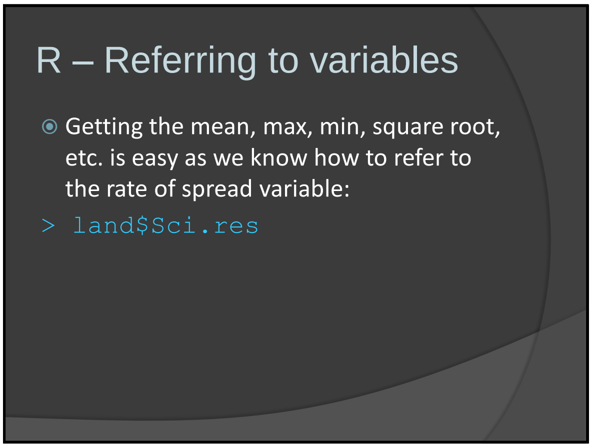
- Assigning row names

```
> row.names(land) <- land$ID
```



R – checking imported data

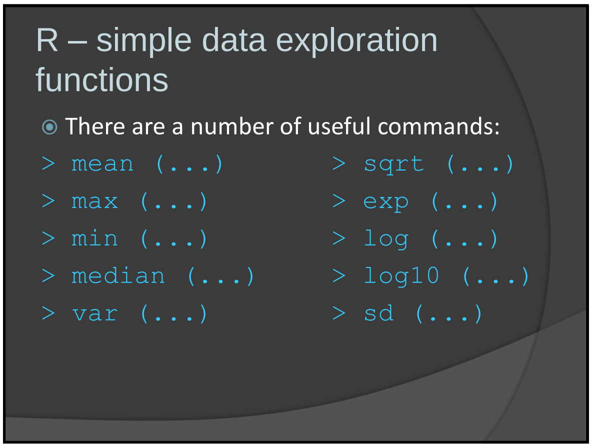
- Task 3: What class is the object “land”?
  - data frame
- Task 4: What class is the variable “Interest”?
  - factor
- Task 5: What class is the variable “wind”?
  - integer



R – Referring to variables

- Getting the mean, max, min, square root, etc. is easy as we know how to refer to the rate of spread variable:

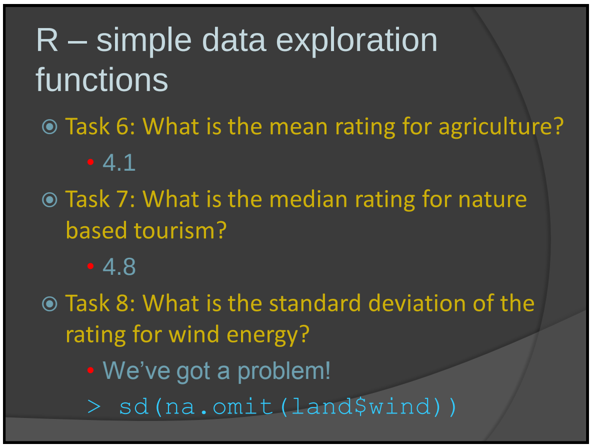
```
> land$Sci.res
```



R – simple data exploration functions

- There are a number of useful commands:

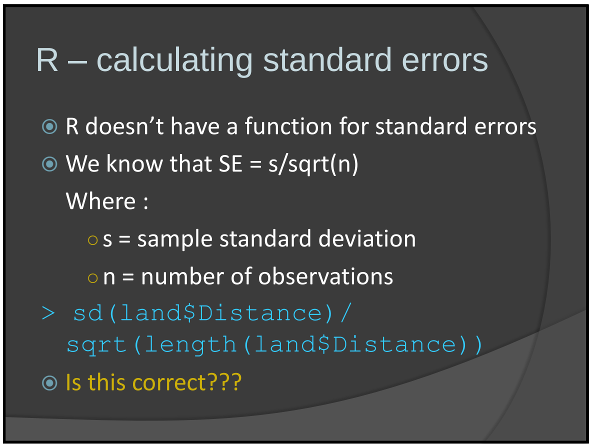
```
> mean (...)        > sqrt (...)
> max (...)         > exp (...)
> min (...)         > log (...)
> median (...)      > log10 (...)
> var (...)         > sd (...)
```



R – simple data exploration functions

- Task 6: What is the mean rating for agriculture?
  - 4.1
- Task 7: What is the median rating for nature based tourism?
  - 4.8
- Task 8: What is the standard deviation of the rating for wind energy?
  - We've got a problem!

```
> sd(na.omit(land$wind))
```



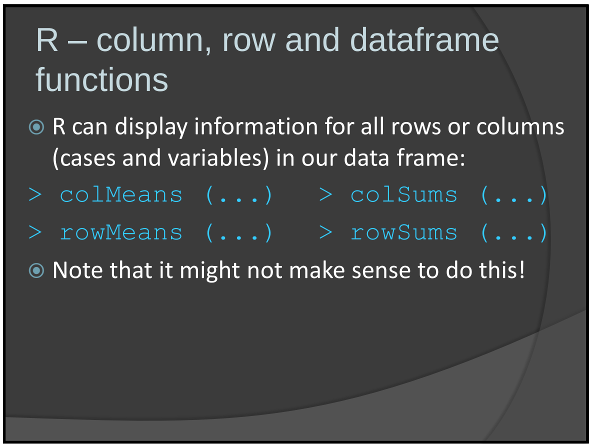
R – calculating standard errors

- R doesn't have a function for standard errors
- We know that SE = s/sqrt(n)

  Where :
  - s = sample standard deviation
  - n = number of observations

```
> sd(land$Distance)/
  sqrt(length(land$Distance))
```

- Is this correct???

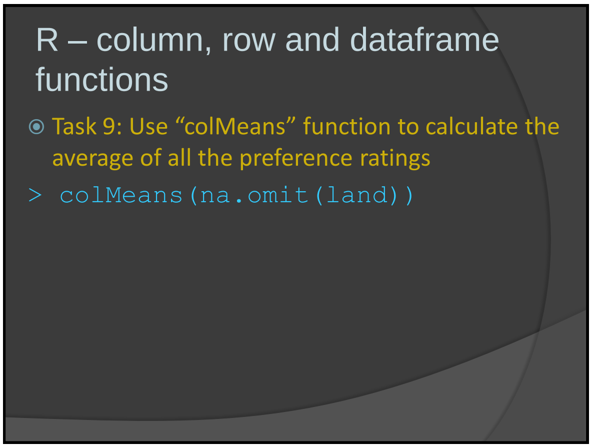


R – column, row and dataframe functions

- R can display information for all rows or columns (cases and variables) in our data frame:

```
> colMeans (...)    > colSums (...)
> rowMeans (...)    > rowSums (...)
```

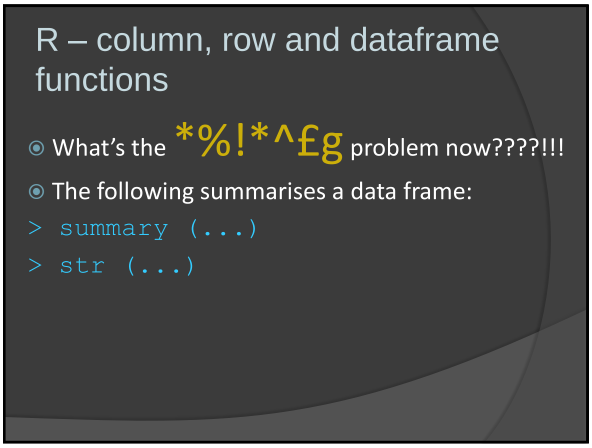
- Note that it might not make sense to do this!



R – column, row and dataframe functions

- Task 9: Use “colMeans” function to calculate the average of all the preference ratings

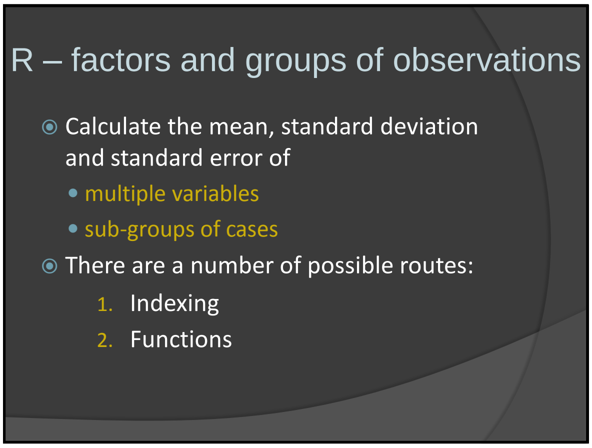
```
> colMeans(na.omit(land))
```



R – column, row and dataframe functions

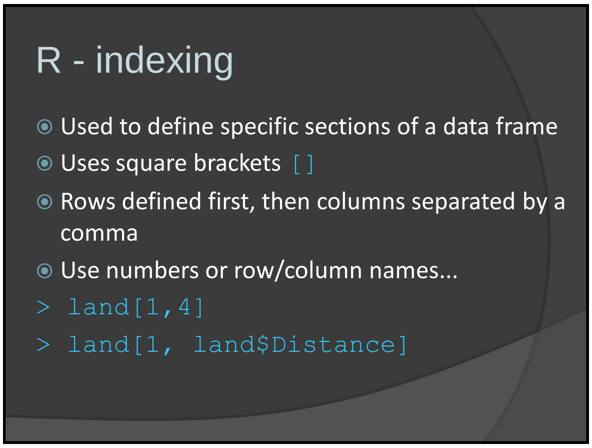
- What's the *%!*^£g problem now????!!!
- The following summarises a data frame:

```
> summary (...)
> str (...)
```



R – factors and groups of observations

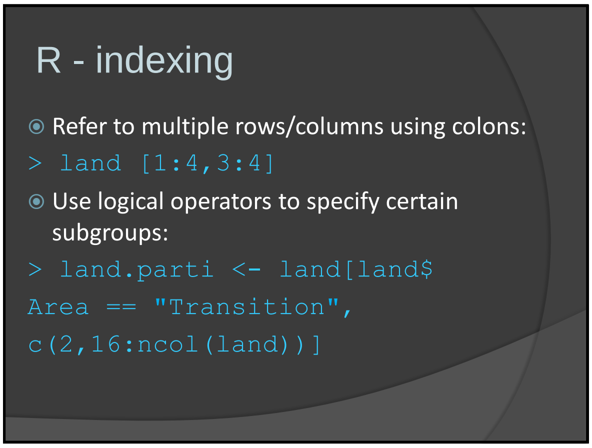
- Calculate the mean, standard deviation and standard error of
  - multiple variables
  - sub-groups of cases
- There are a number of possible routes:
  1. Indexing
  2. Functions



R - indexing

- Used to define specific sections of a data frame
- Uses square brackets [ ]
- Rows defined first, then columns separated by a comma
- Use numbers or row/column names...

```
> land[1,4]
> land[1, land$Distance]
```



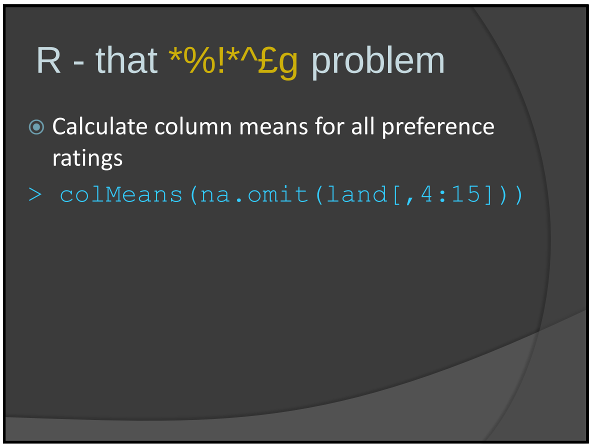
R - indexing

- Refer to multiple rows/columns using colons:

```
> land [1:4,3:4]
```

- Use logical operators to specify certain subgroups:

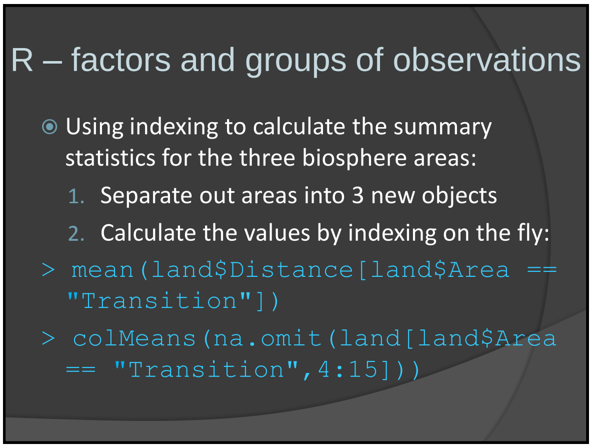
```
> land.parti <- land[land$
Area == "Transition",
c(2,16:ncol(land))]
```



R - that *%!*^£g problem

- Calculate column means for all preference ratings

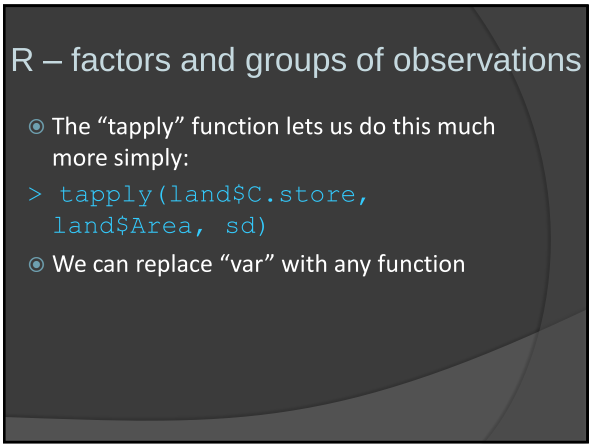
```
> colMeans(na.omit(land[,4:15]))
```



R – factors and groups of observations

- Using indexing to calculate the summary statistics for the three biosphere areas:
  1. Separate out areas into 3 new objects
  2. Calculate the values by indexing on the fly:

```
> mean(land$Distance[land$Area ==
  "Transition"])
> colMeans(na.omit(land[land$Area
  == "Transition",4:15]))
```

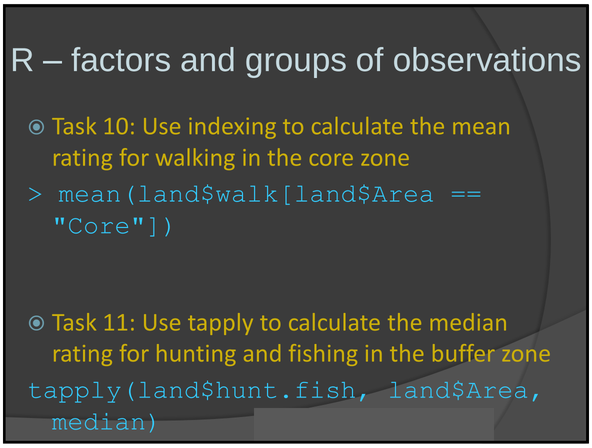


R – factors and groups of observations

- The “tapply” function lets us do this much more simply:

```
> tapply(land$C.store,
  land$Area, sd)
```

- We can replace “var” with any function



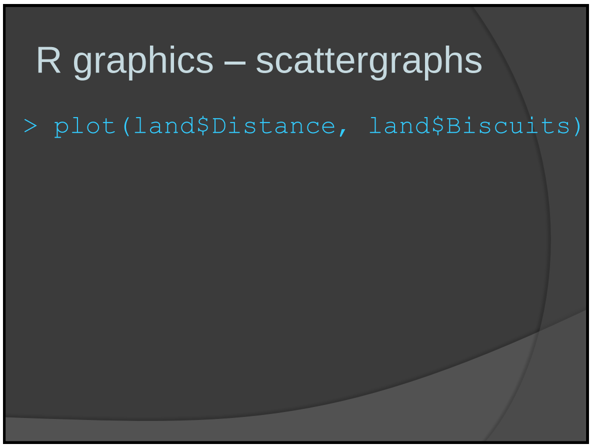
R – factors and groups of observations

- Task 10: Use indexing to calculate the mean rating for walking in the core zone

```
> mean(land$walk[land$Area ==
  "Core"])
```

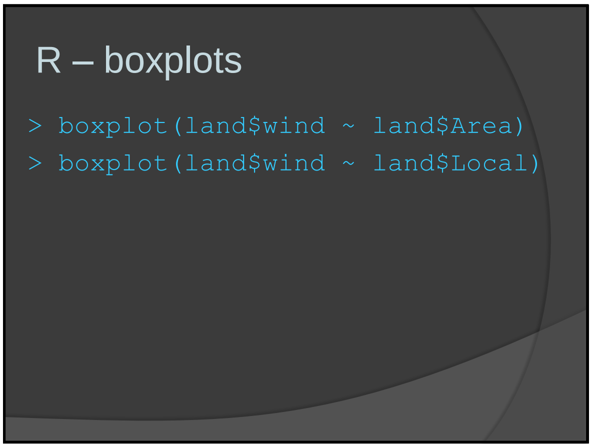
- Task 11: Use tapply to calculate the median rating for hunting and fishing in the buffer zone

```
tapply(land$hunt.fish, land$Area,
  median)
```



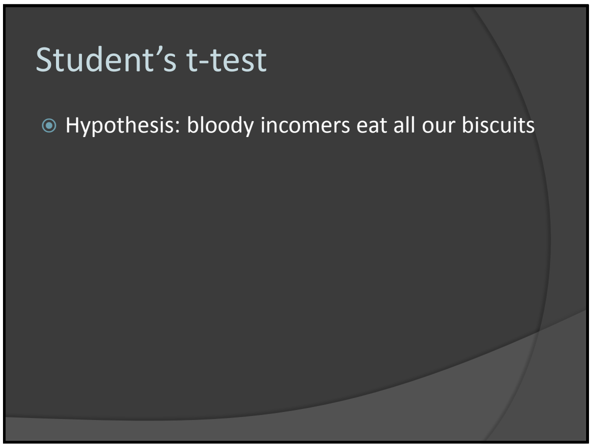
R graphics – scattergraphs

```
> plot(land$Distance, land$Biscuits)
```



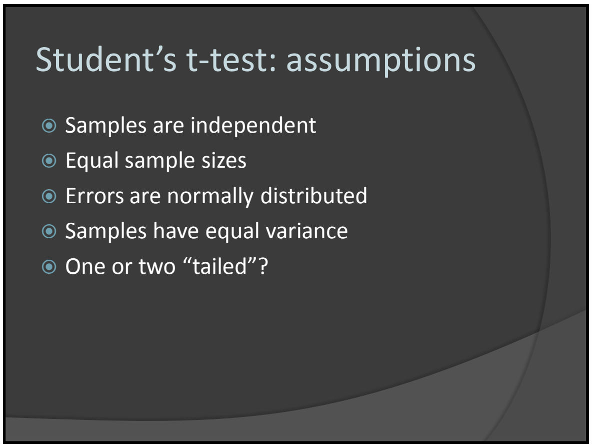
R – boxplots

```
> boxplot(land$wind ~ land$Area)
> boxplot(land$wind ~ land$Local)
```



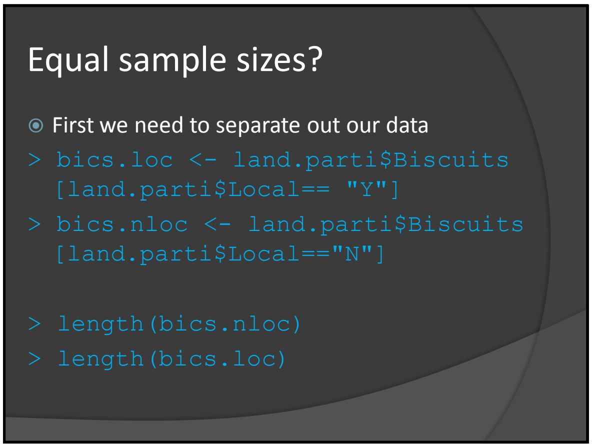
Student’s t-test

- Hypothesis: bloody incomers eat all our biscuits



Student’s t-test: assumptions

- Samples are independent
- Equal sample sizes
- Errors are normally distributed
- Samples have equal variance
- One or two “tailed”?

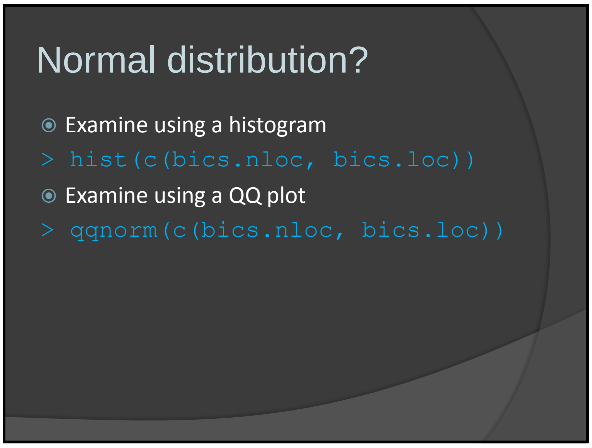


Equal sample sizes?

- First we need to separate out our data

```
> bics.loc <- land.parti$Biscuits
  [land.parti$Local== "Y"]
> bics.nloc <- land.parti$Biscuits
  [land.parti$Local=="N"]

> length(bics.nloc)
> length(bics.loc)
```



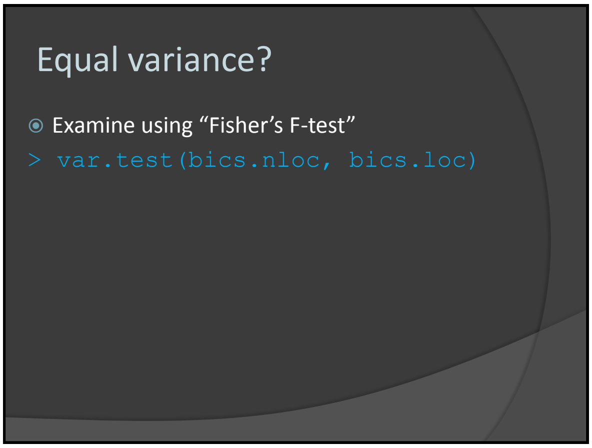
Normal distribution?

- Examine using a histogram

```
> hist(c(bics.nloc, bics.loc))
```

- Examine using a QQ plot

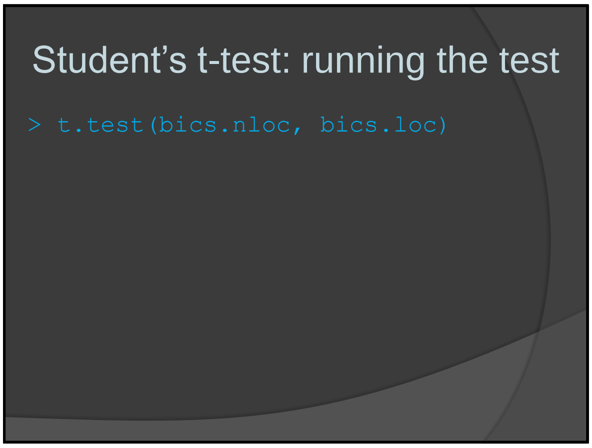
```
> qqnorm(c(bics.nloc, bics.loc))
```



Equal variance?

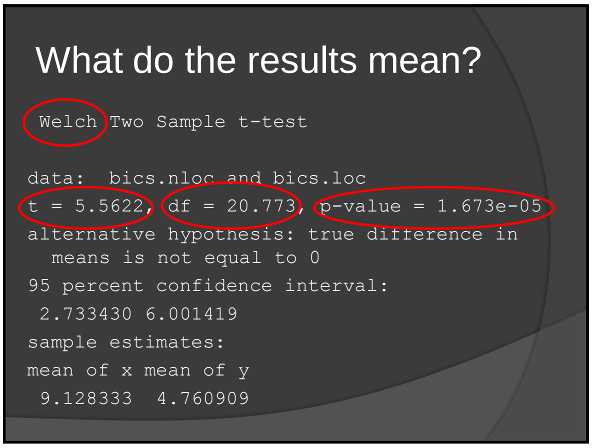
- Examine using “Fisher’s F-test”

```
> var.test(bics.nloc, bics.loc)
```



Student’s t-test: running the test

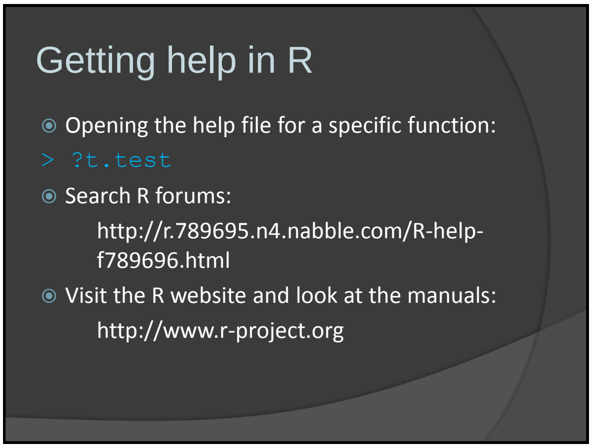
```
> t.test(bics.nloc, bics.loc)
```



What do the results mean?

```
  Welch Two Sample t-test

data:  bics.nloc and bics.loc
t = 5.5622, df = 20.773, p-value = 1.673e-05
alternative hypothesis: true difference in
    means is not equal to 0
95 percent confidence interval:
  2.733430 6.001419
sample estimates:
mean of x mean of y
  9.128333  4.760909
```



Getting help in R

- Opening the help file for a specific function:

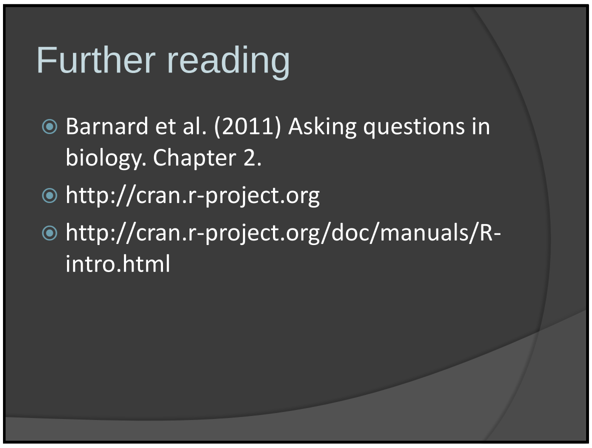
```
> ?t.test
```

- Search R forums:

  http://r.789695.n4.nabble.com/R-help-f789696.html

- Visit the R website and look at the manuals:

  http://www.r-project.org



Further reading

- Barnard et al. (2011) Asking questions in biology. Chapter 2.
- http://cran.r-project.org
- http://cran.r-project.org/doc/manuals/R-intro.html