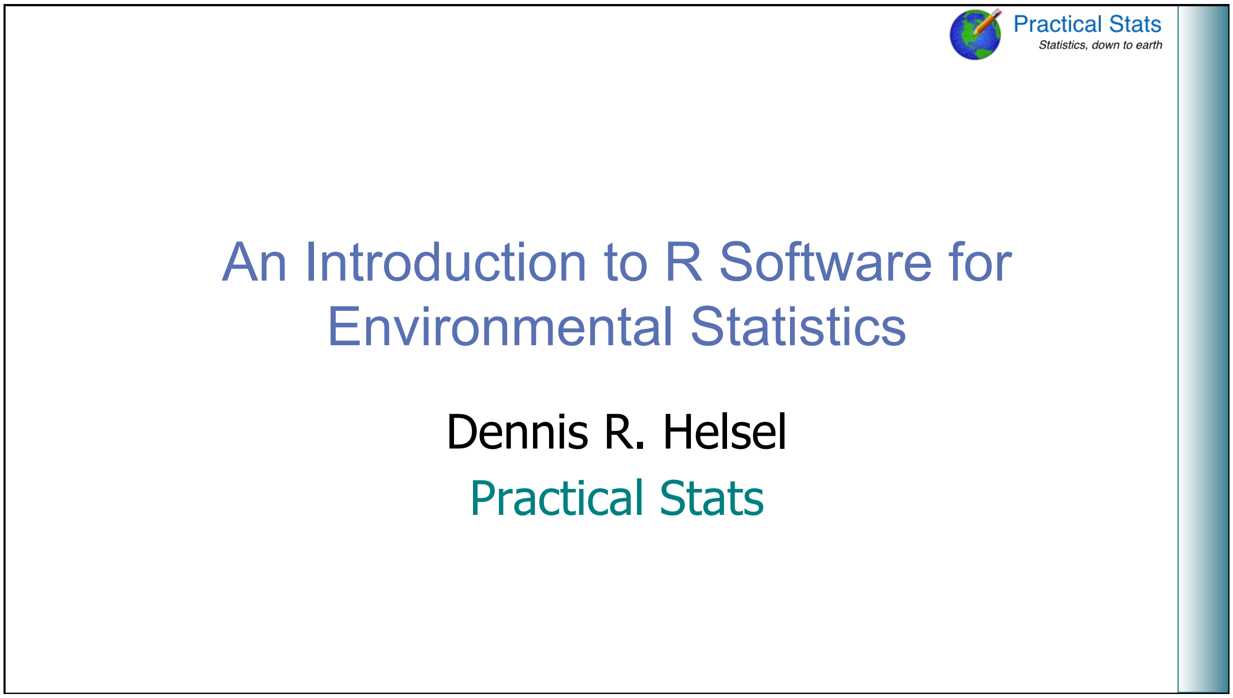
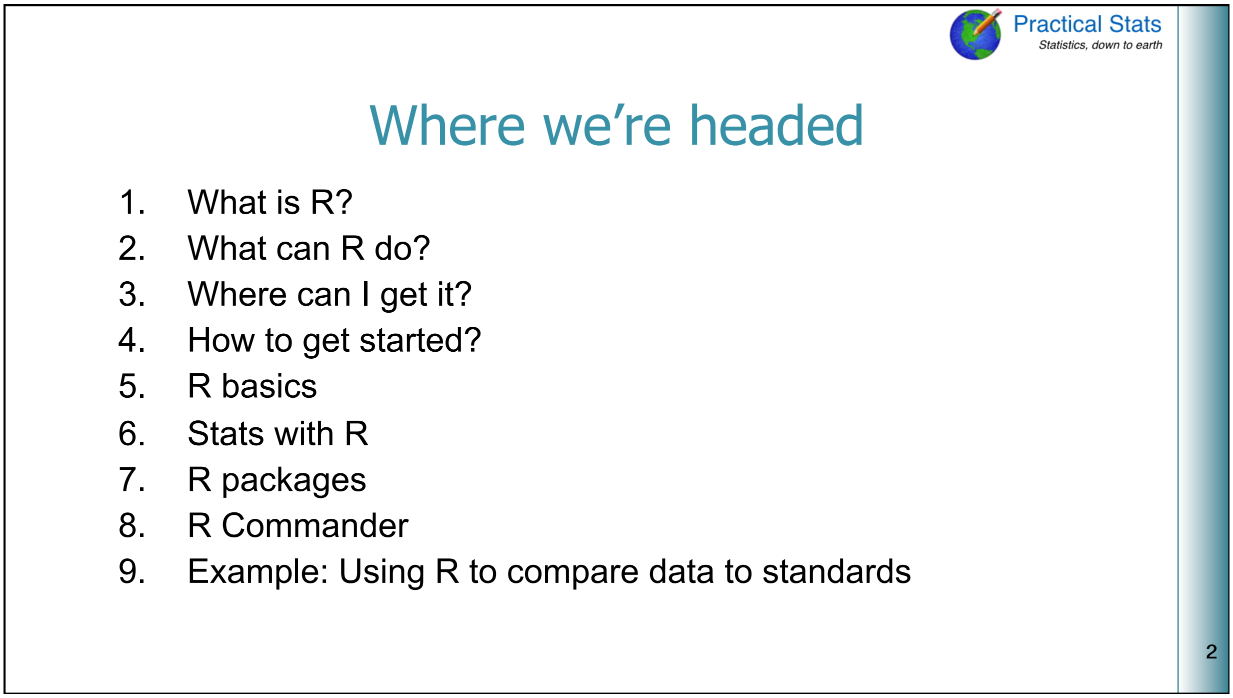


Practical Stats
Statistics, down to earth

An Introduction to R Software for Environmental Statistics

Dennis R. Helsel

Practical Stats

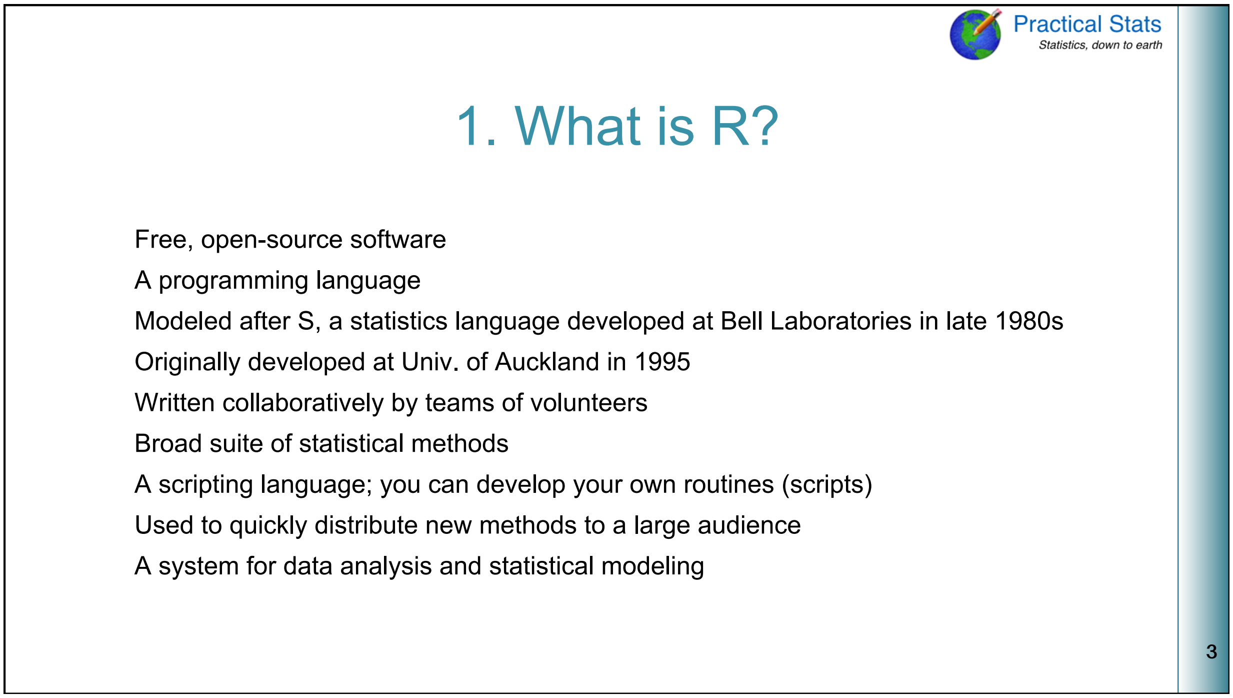


Practical Stats
Statistics, down to earth

Where we’re headed

1. What is R?
2. What can R do?
3. Where can I get it?
4. How to get started?
5. R basics
6. Stats with R
7. R packages
8. R Commander
9. Example: Using R to compare data to standards

2



Practical Stats
Statistics, down to earth

1. What is R?

Free, open-source software

A programming language

Modeled after S, a statistics language developed at Bell Laboratories in late 1980s

Originally developed at Univ. of Auckland in 1995

Written collaboratively by teams of volunteers

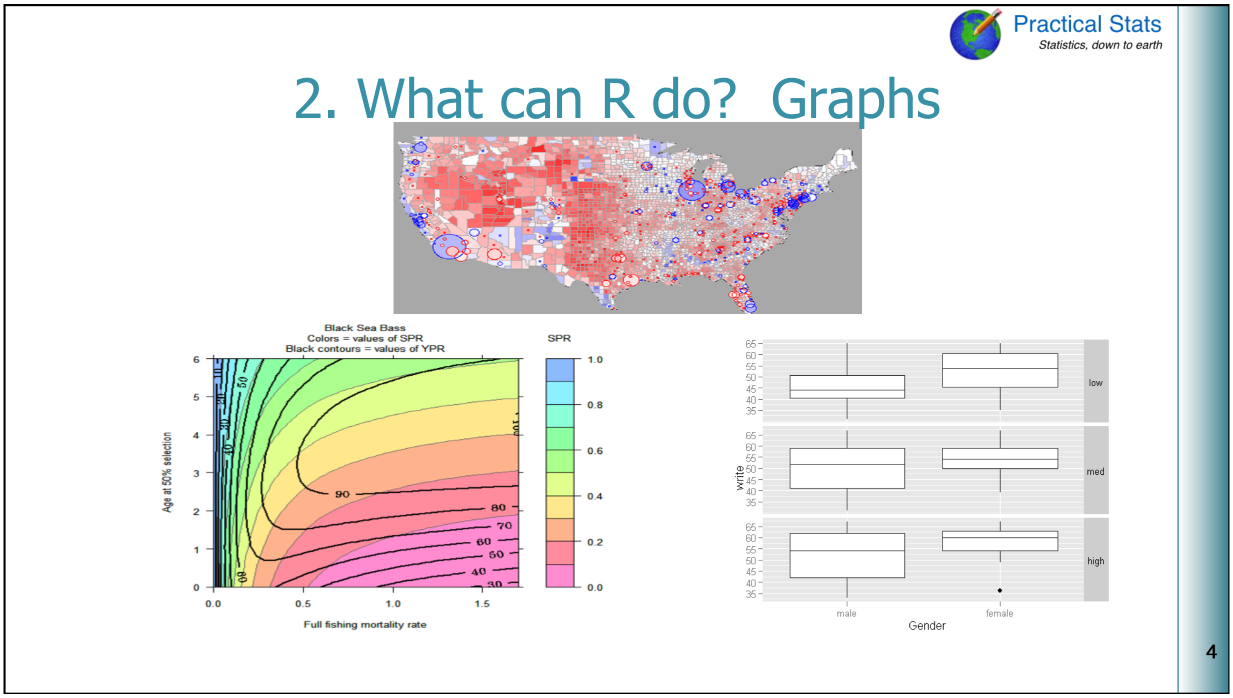
Broad suite of statistical methods

A scripting language; you can develop your own routines (scripts)

Used to quickly distribute new methods to a large audience

A system for data analysis and statistical modeling

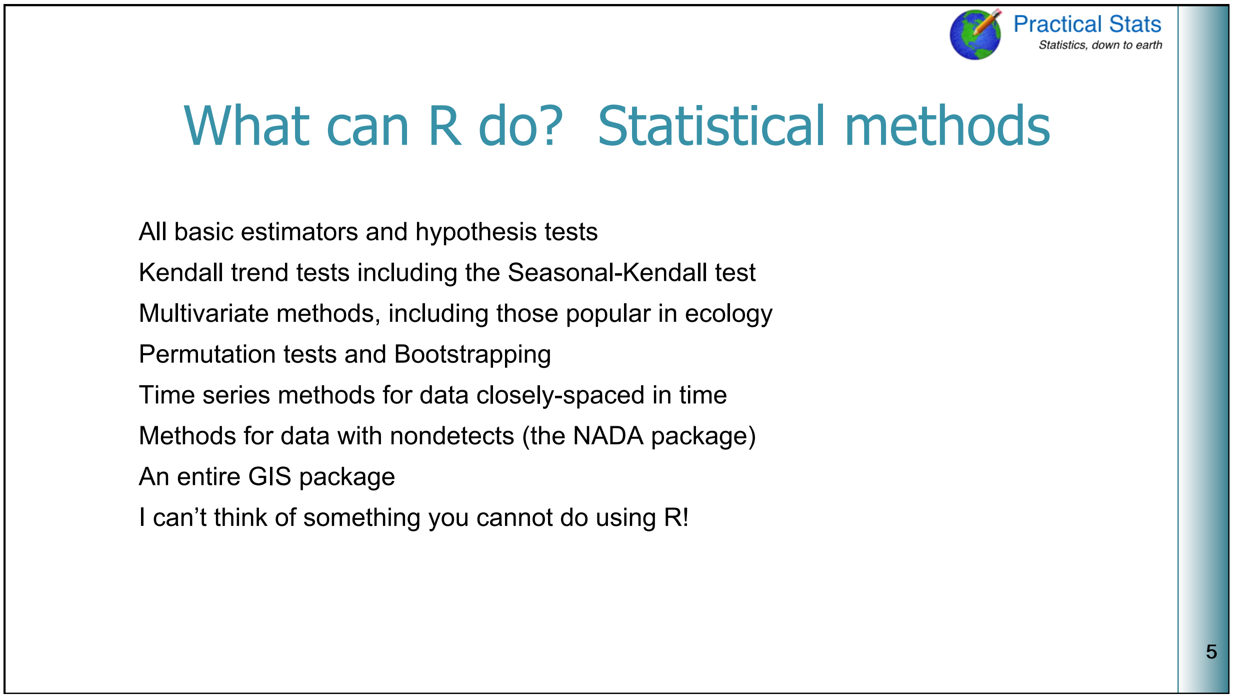
3



Practical Stats
Statistics, down to earth

2. What can R do? Graphs

4



Practical Stats
Statistics, down to earth

What can R do? Statistical methods

All basic estimators and hypothesis tests

Kendall trend tests including the Seasonal-Kendall test

Multivariate methods, including those popular in ecology

Permutation tests and Bootstrapping

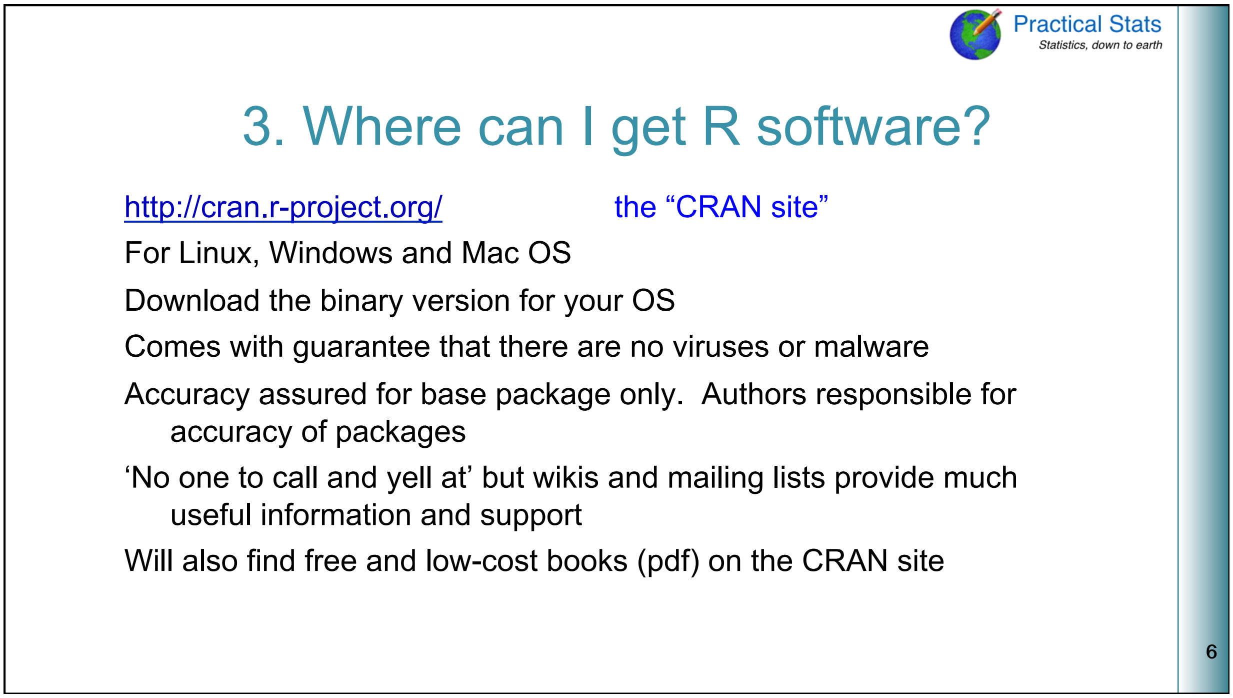
Time series methods for data closely-spaced in time

Methods for data with nondetects (the NADA package)

An entire GIS package

I can’t think of something you cannot do using R!

5



Practical Stats
Statistics, down to earth

3. Where can I get R software?

http://cran.r-project.org/ the “CRAN site”

For Linux, Windows and Mac OS

Download the binary version for your OS

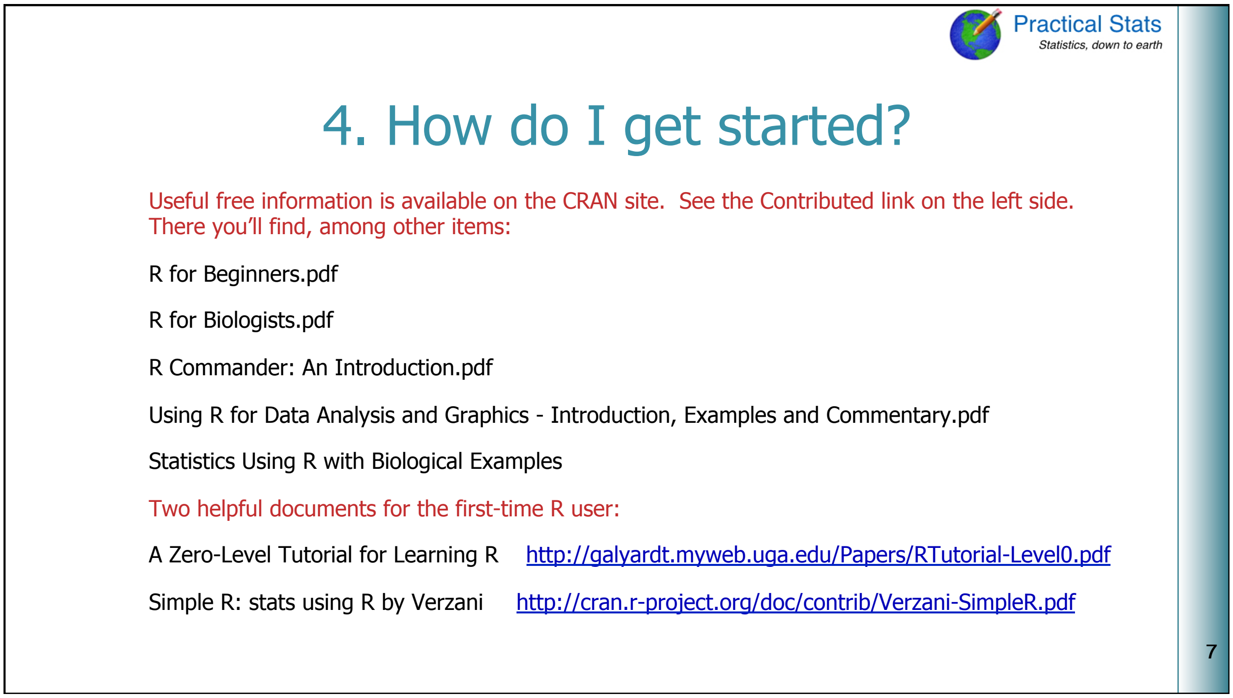
Comes with guarantee that there are no viruses or malware

Accuracy assured for base package only. Authors responsible for accuracy of packages

‘No one to call and yell at’ but wikis and mailing lists provide much useful information and support

Will also find free and low-cost books (pdf) on the CRAN site

6



Practical Stats
Statistics, down to earth

4. How do I get started?

Useful free information is available on the CRAN site. See the Contributed link on the left side. There you'll find, among other items:

R for Beginners.pdf

R for Biologists.pdf

R Commander: An Introduction.pdf

Using R for Data Analysis and Graphics - Introduction, Examples and Commentary.pdf

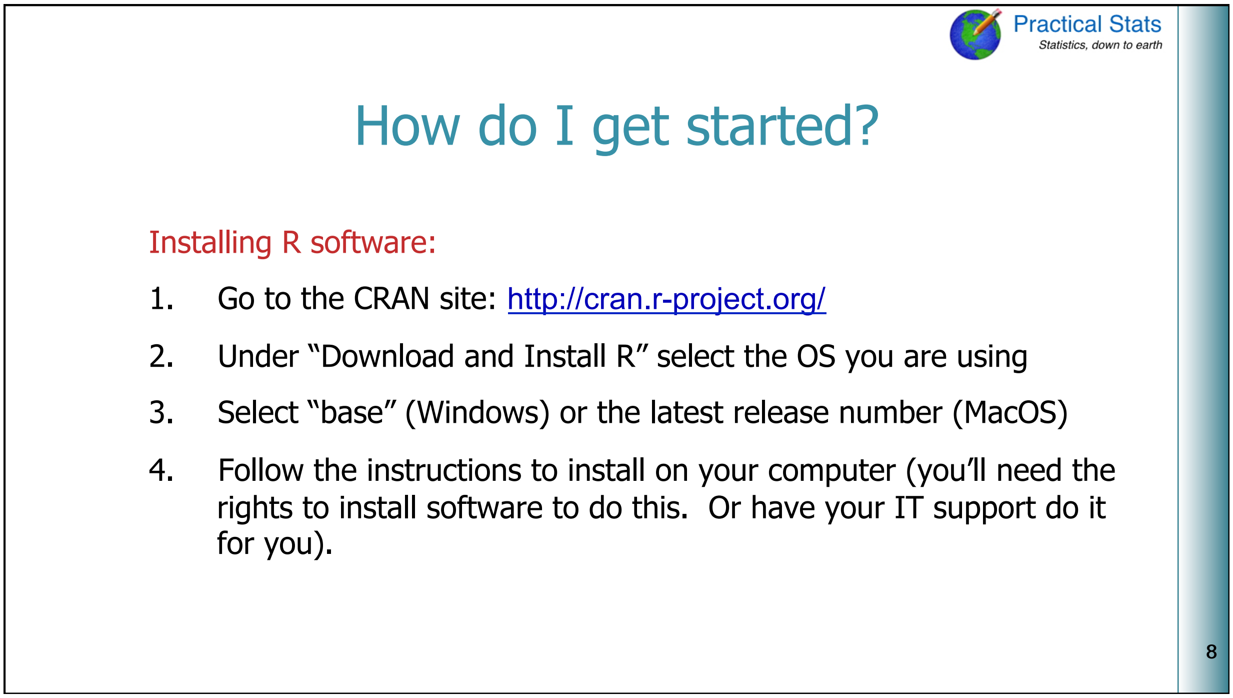
Statistics Using R with Biological Examples

Two helpful documents for the first-time R user:

A Zero-Level Tutorial for Learning R http://galyardt.myweb.uga.edu/Papers/RTutorial-Level0.pdf

Simple R: stats using R by Verzani http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf

7



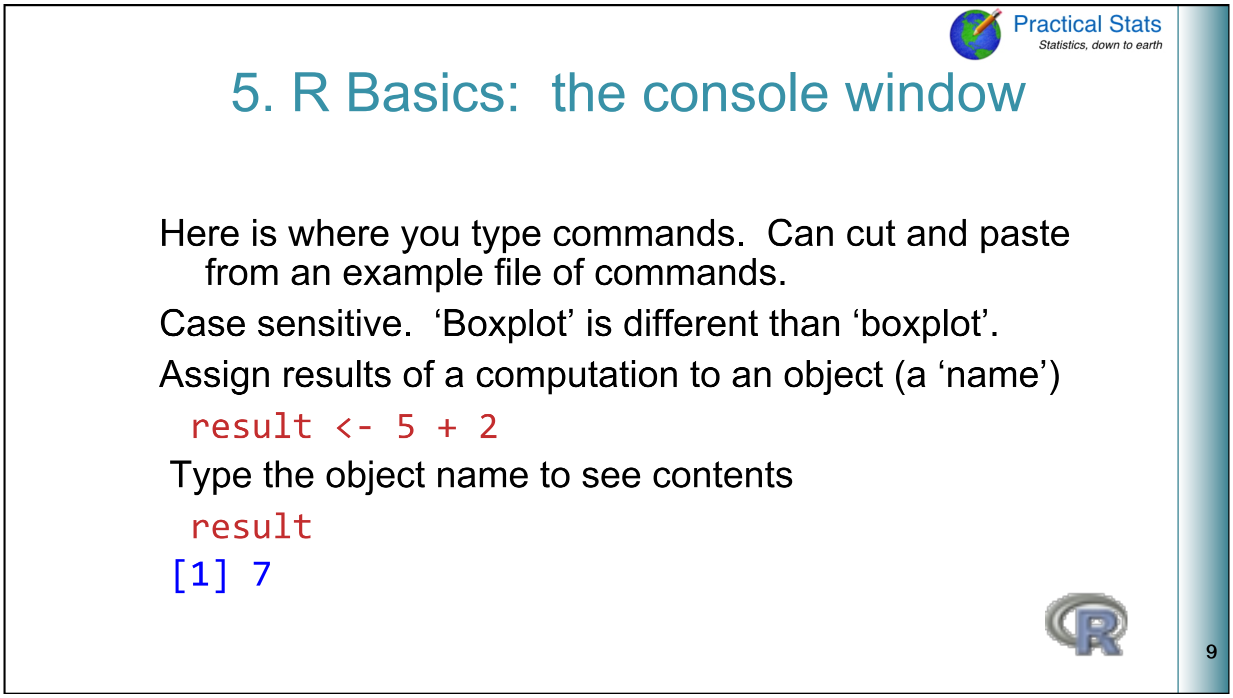
Practical Stats
Statistics, down to earth

How do I get started?

Installing R software:

1. Go to the CRAN site: http://cran.r-project.org/
2. Under "Download and Install R" select the OS you are using
3. Select "base" (Windows) or the latest release number (MacOS)
4. Follow the instructions to install on your computer (you'll need the rights to install software to do this. Or have your IT support do it for you).

8



Practical Stats
Statistics, down to earth

5. R Basics: the console window

Here is where you type commands. Can cut and paste from an example file of commands.

Case sensitive. ‘Boxplot’ is different than ‘boxplot’.

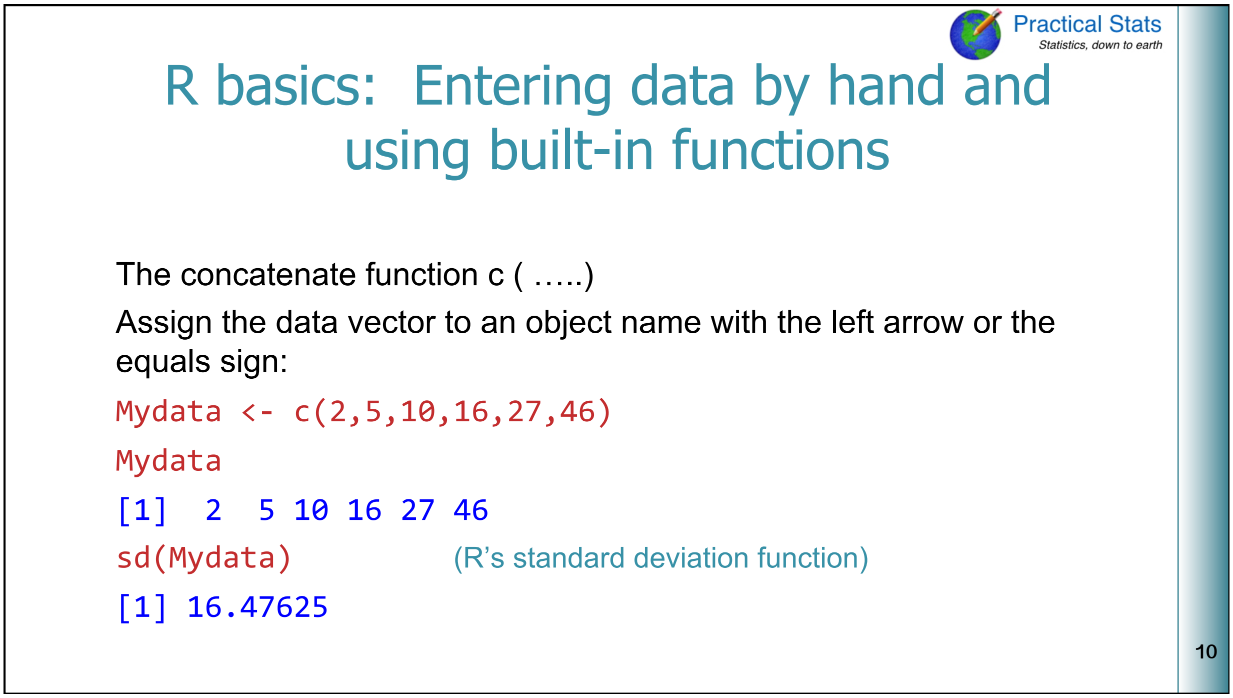
Assign results of a computation to an object (a ‘name’)

```
result <- 5 + 2
```

Type the object name to see contents

```
result
[1] 7
```

9



Practical Stats
Statistics, down to earth

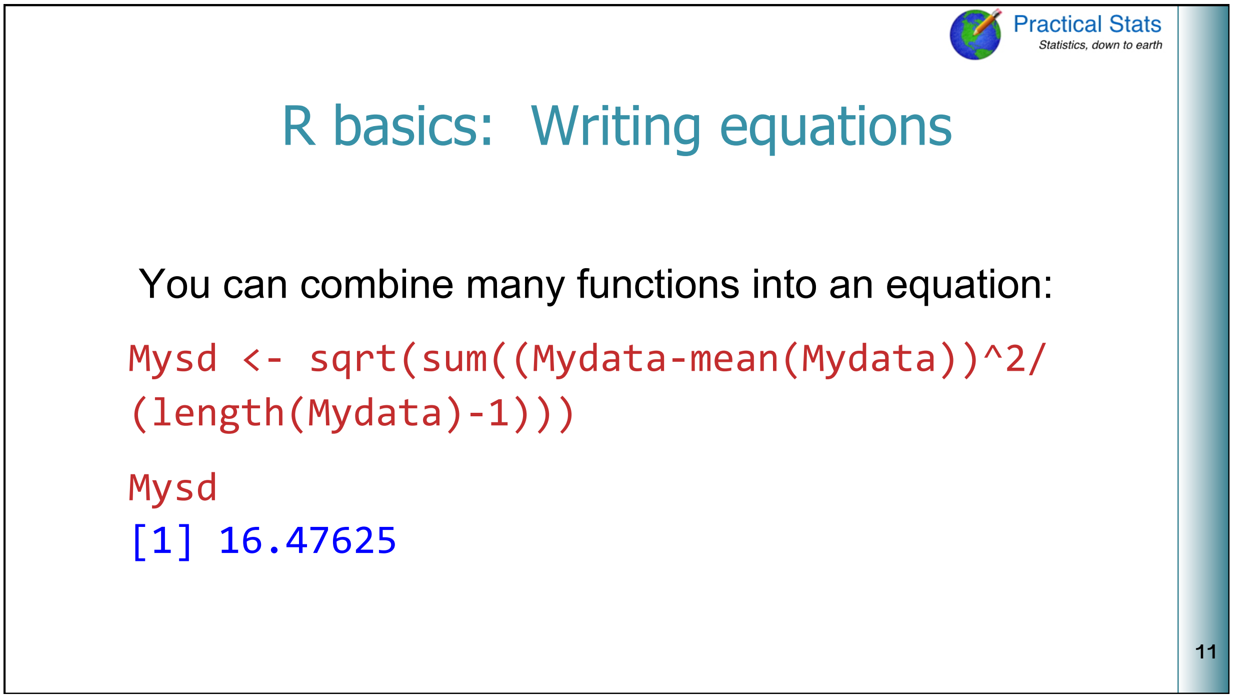
R basics: Entering data by hand and using built-in functions

The concatenate function c (…..)

Assign the data vector to an object name with the left arrow or the equals sign:

```
Mydata <- c(2,5,10,16,27,46)
Mydata
[1]  2  5 10 16 27 46
sd(Mydata)            (R’s standard deviation function)
[1] 16.47625
```

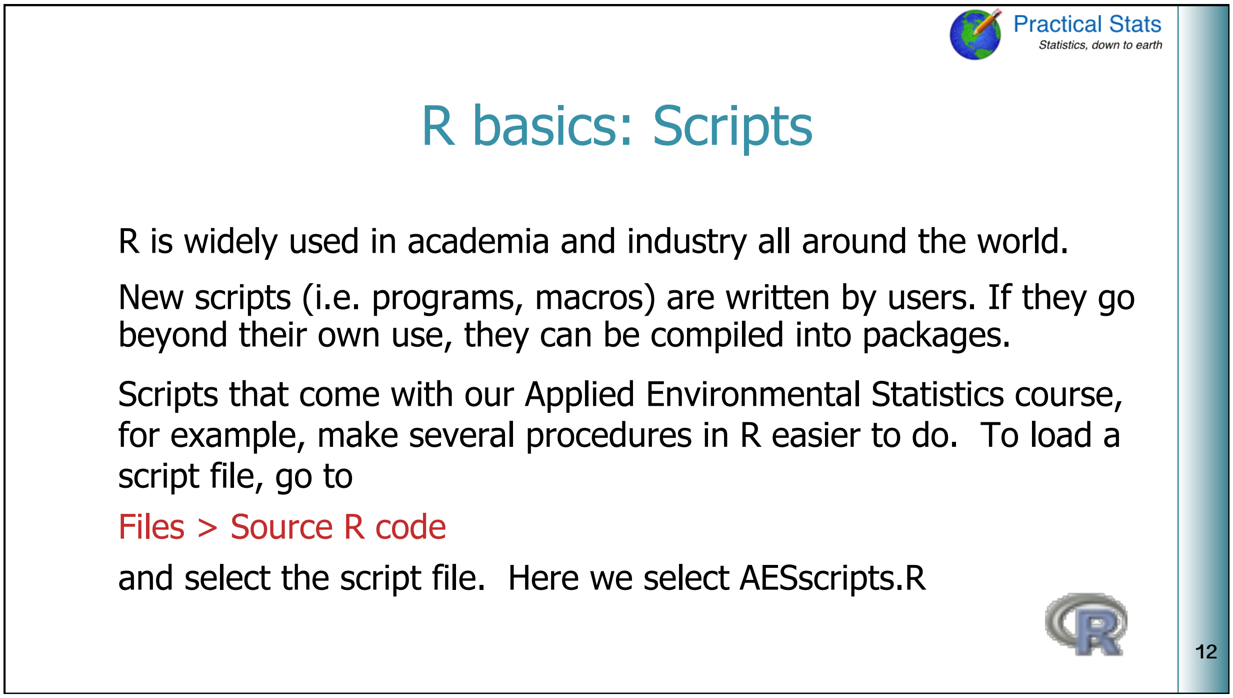
10



R basics: Writing equations

You can combine many functions into an equation:

```
Mysd <- sqrt(sum((Mydata-mean(Mydata))^2/
(length(Mydata)-1)))

Mysd
[1] 16.47625
```



R basics: Scripts

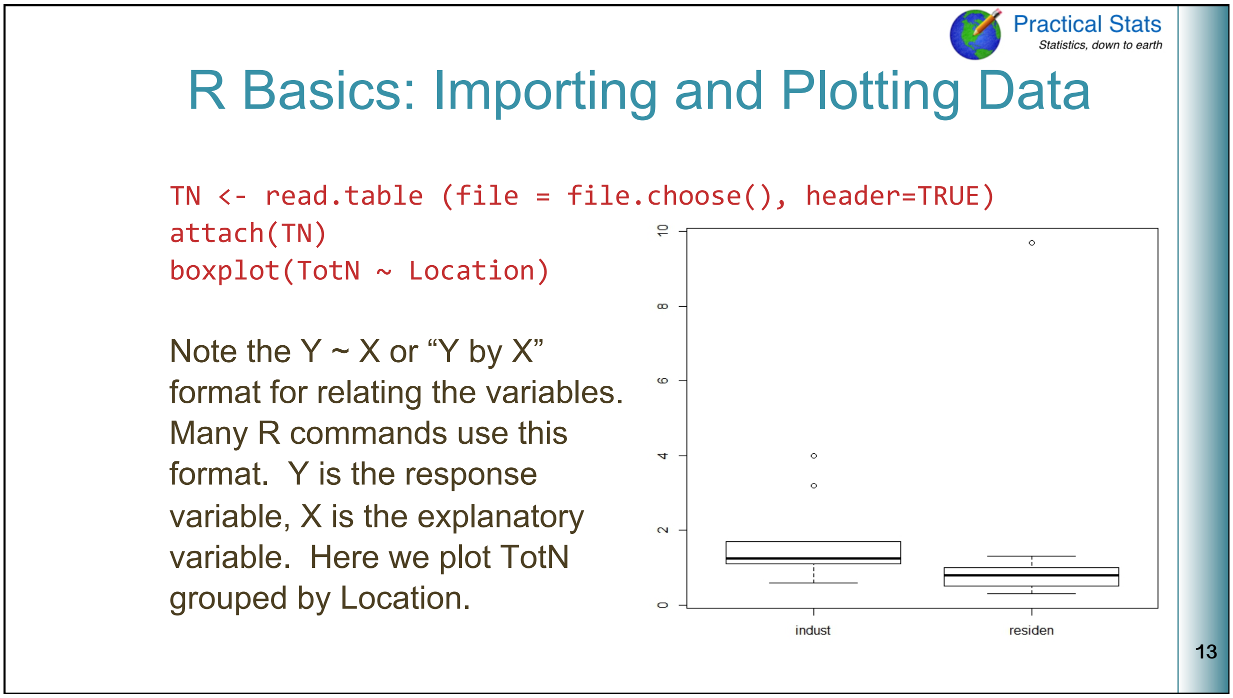
R is widely used in academia and industry all around the world.

New scripts (i.e. programs, macros) are written by users. If they go beyond their own use, they can be compiled into packages.

Scripts that come with our Applied Environmental Statistics course, for example, make several procedures in R easier to do. To load a script file, go to

Files > Source R code

and select the script file. Here we select AESscripts.R



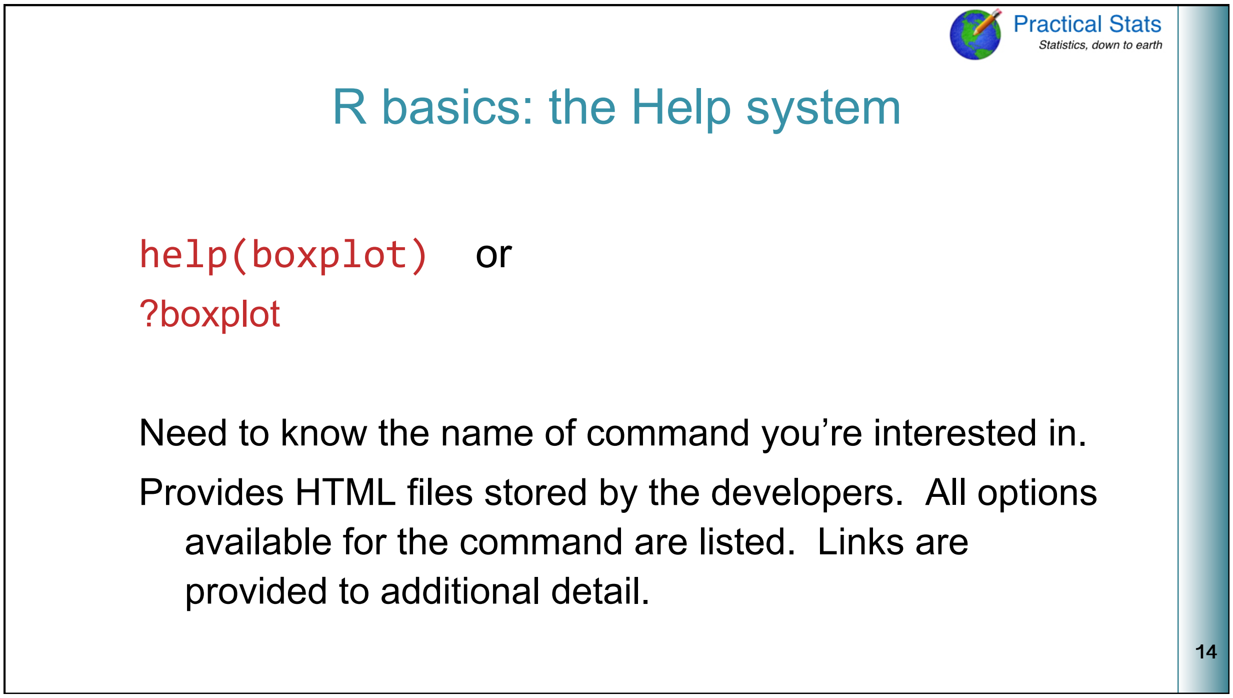
10
8
6
4
2
0
indust
residen

R Basics: Importing and Plotting Data

```
TN <- read.table (file = file.choose(), header=TRUE)
attach(TN)
boxplot(TotN ~ Location)
```

Note the Y ~ X or “Y by X” format for relating the variables. Many R commands use this format. Y is the response variable, X is the explanatory variable. Here we plot TotN grouped by Location.

13



R basics: the Help system

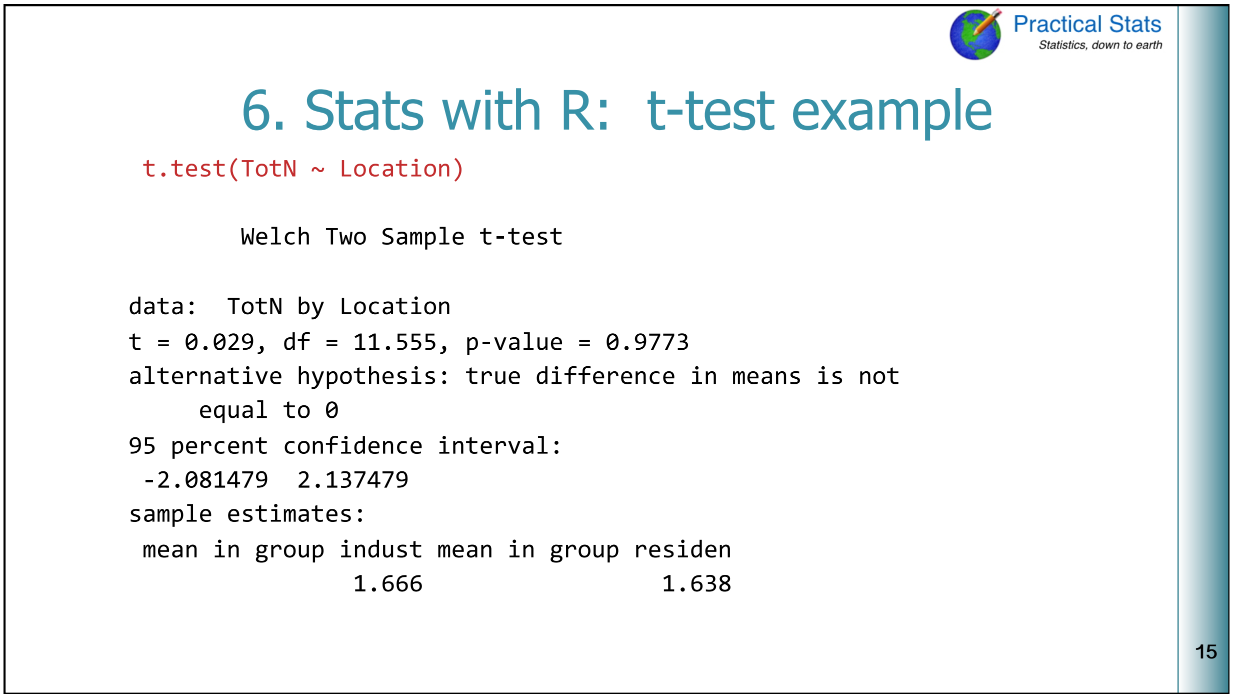
`help(boxplot)` or

`?boxplot`

Need to know the name of command you’re interested in.

Provides HTML files stored by the developers. All options available for the command are listed. Links are provided to additional detail.

14

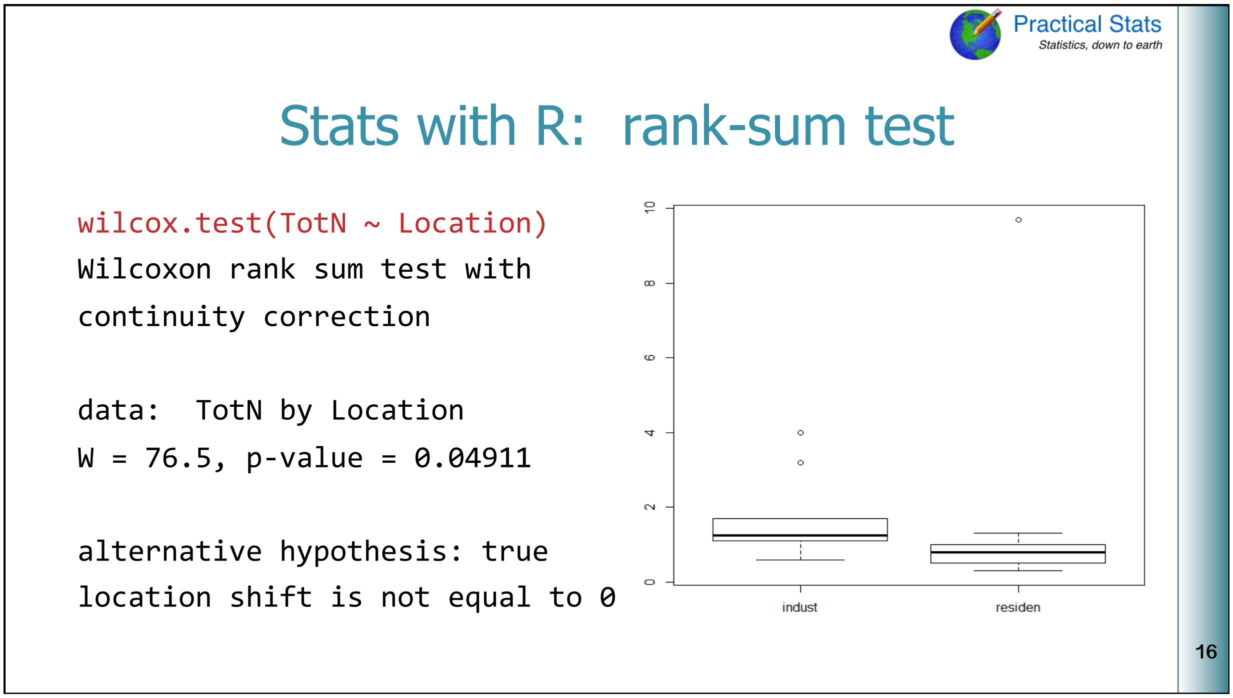


Practical Stats
Statistics, down to earth

6. Stats with R: t-test example

```
t.test(TotN ~ Location)

       Welch Two Sample t-test

data:  TotN by Location
t = 0.029, df = 11.555, p-value = 0.9773
alternative hypothesis: true difference in means is not
     equal to 0
95 percent confidence interval:
 -2.081479  2.137479
sample estimates:
 mean in group indust mean in group residen
                1.666                 1.638
```

15



Practical Stats
Statistics, down to earth

Stats with R: rank-sum test

```
wilcox.test(TotN ~ Location)
Wilcoxon rank sum test with
continuity correction

data:  TotN by Location
W = 76.5, p-value = 0.04911

alternative hypothesis: true
location shift is not equal to 0
```

0
2
4
6
8
10
indust
residen

16

Practical Stats
Statistics, down to earth

7. R Packages: Adding new methods

Packages > Install a package

finds an R package on a mirror site and downloads it into the chosen directory on your computer.

Packages expand the capabilities of R.

Currently (2018) there are over 10,000 add-on packages for R.

17

Practical Stats
Statistics, down to earth

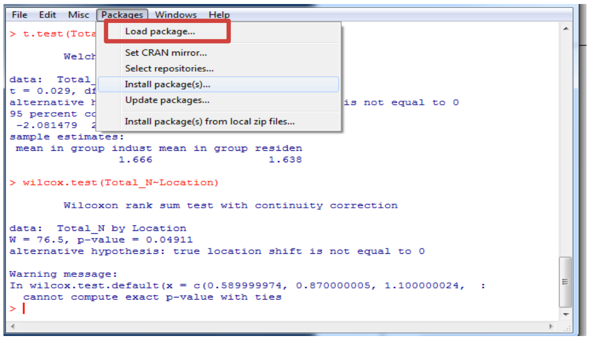
R packages: Loading a package

Packages > Load package

loads a package on your computer (makes the code active).

Choose a package from those you've already installed.

After loading the package, routines inside are available for your use.



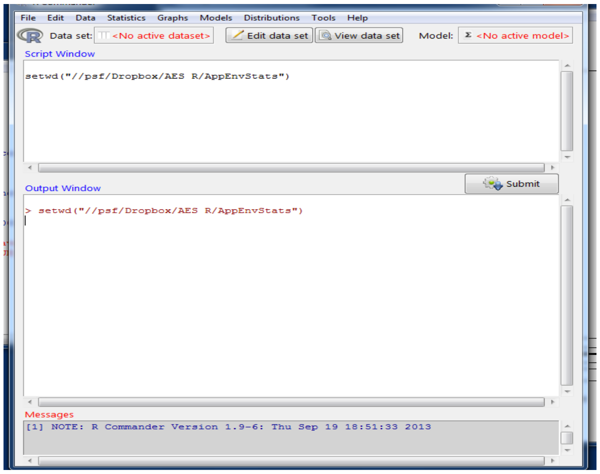
```
> t.test(Tota

        Welch

data:  Total_
t = 0.029, df
alternative h                                        is not equal to 0
95 percent co
 -2.081479  2
sample estimates:
 mean in group indust mean in group residen
               1.666                  1.638

> wilcox.test(Total_N~Location)

        Wilcoxon rank sum test with continuity correction

data:  Total_N by Location
W = 76.5, p-value = 0.04911
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(0.589999974, 0.870000005, 1.100000024,  :
  cannot compute exact p-value with ties
> 
```

18

8. R Commander – GUI interface for R

(Rcmdr package)

A pull-down menu system for common statistical tests and operations


File Edit Data Statistics Graphs Models Distributions Tools Help
Data set: <No active dataset>
Edit data set
View data set
Model: Σ <No active model>
Script Window
Submit
Output Window
Messages

```
setwd("//psf/Dropbox/AES R/AppEnvStats")
```

```
> setwd("//psf/Dropbox/AES R/AppEnvStats")
```

```
[1] NOTE: R Commander Version 1.9-6: Thu Sep 19 18:51:33 2013
```

Rcmdr: Data Menu

To bring in a data set in R's format, select

Data > Load data set

and select the data set name (.RData or .rda)

Arsenic10.RData

This will enable you to use that data set with any of the GUI routines in the R Commander window, and when typing into the console window.

Practical Stats
Statistics, down to earth

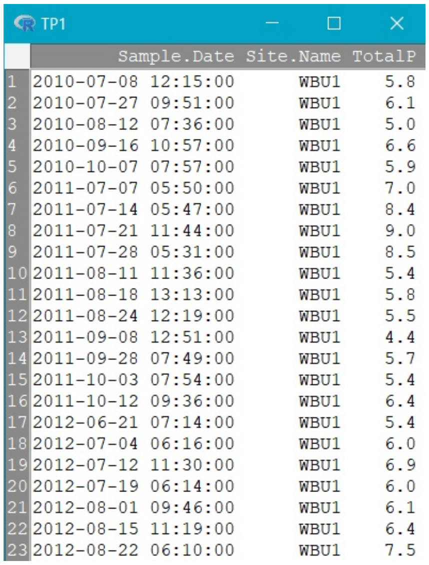
Rcmdr: Importing Excel Datasets

Data > Import Data > Excel

Reads .xls and .xlsx format data

Example: choose the TP1.xlsx dataset

Then look at the data by clicking the View Data button



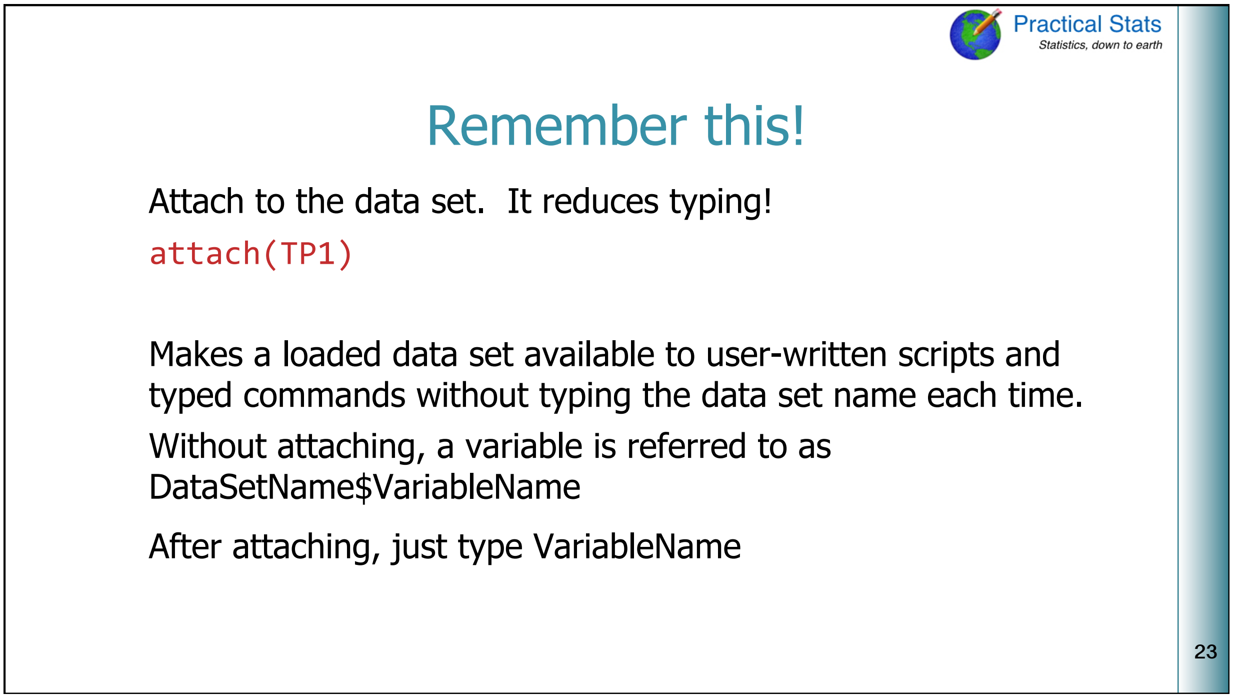
	Sample.Date	Site.Name	TotalP
1	2010-07-08 12:15:00	WBU1	5.8
2	2010-07-27 09:51:00	WBU1	6.1
3	2010-08-12 07:36:00	WBU1	5.0
4	2010-09-16 10:57:00	WBU1	6.6
5	2010-10-07 07:57:00	WBU1	5.9
6	2011-07-07 05:50:00	WBU1	7.0
7	2011-07-14 05:47:00	WBU1	8.4
8	2011-07-21 11:44:00	WBU1	9.0
9	2011-07-28 05:31:00	WBU1	8.5
10	2011-08-11 11:36:00	WBU1	5.4
11	2011-08-18 13:13:00	WBU1	5.8
12	2011-08-24 12:19:00	WBU1	5.5
13	2011-09-08 12:51:00	WBU1	4.4
14	2011-09-28 07:49:00	WBU1	5.7
15	2011-10-03 07:54:00	WBU1	5.4
16	2011-10-12 09:36:00	WBU1	6.4
17	2012-06-21 07:14:00	WBU1	5.4
18	2012-07-04 06:16:00	WBU1	6.0
19	2012-07-12 11:30:00	WBU1	6.9
20	2012-07-19 06:14:00	WBU1	6.0
21	2012-08-01 09:46:00	WBU1	6.1
22	2012-08-15 11:19:00	WBU1	6.4
23	2012-08-22 06:10:00	WBU1	7.5

Practical Stats
Statistics, down to earth

Rcmdr: Data Menu

In the Data Menu you can

- Transform data (take logs, and much more)
- Convert a text variable to a factor
- Merge data sets
- ‘Refresh’ the data set
- Reorder the factor levels to something other than alphabetical order
- ………and more.



Practical Stats
Statistics, down to earth

Remember this!

Attach to the data set. It reduces typing!

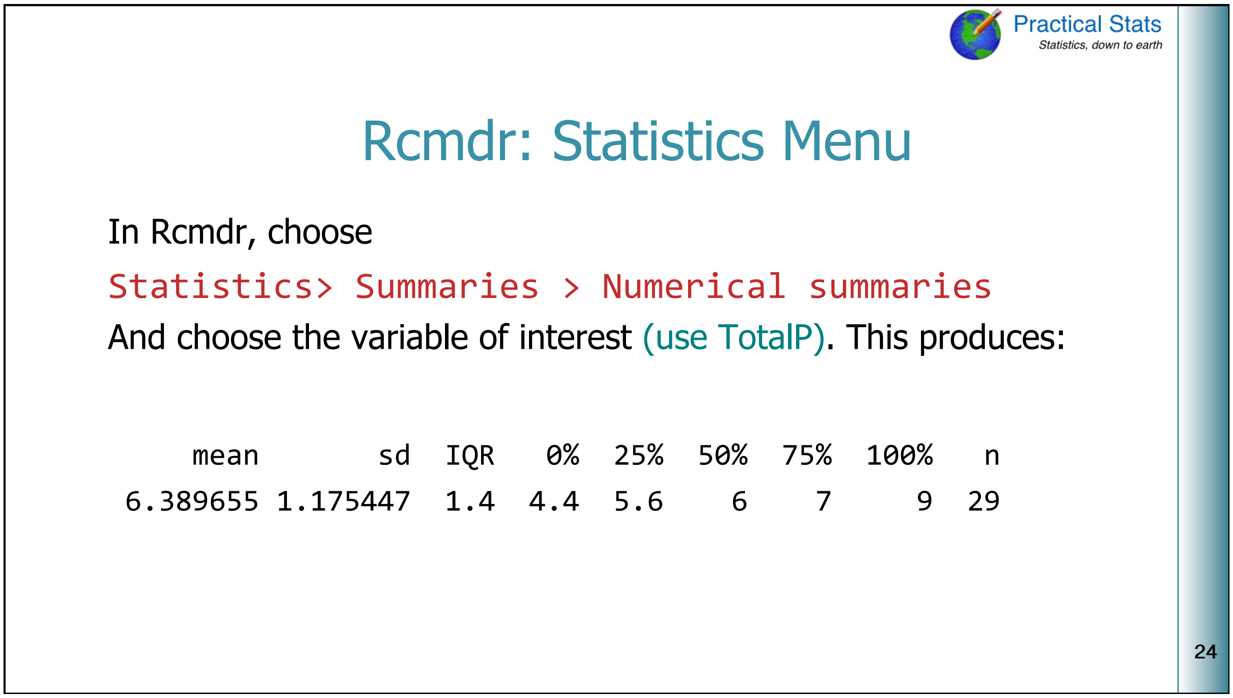
```
attach(TP1)
```

Makes a loaded data set available to user-written scripts and typed commands without typing the data set name each time.

Without attaching, a variable is referred to as DataSetName$VariableName

After attaching, just type VariableName

23



Practical Stats
Statistics, down to earth

Rcmdr: Statistics Menu

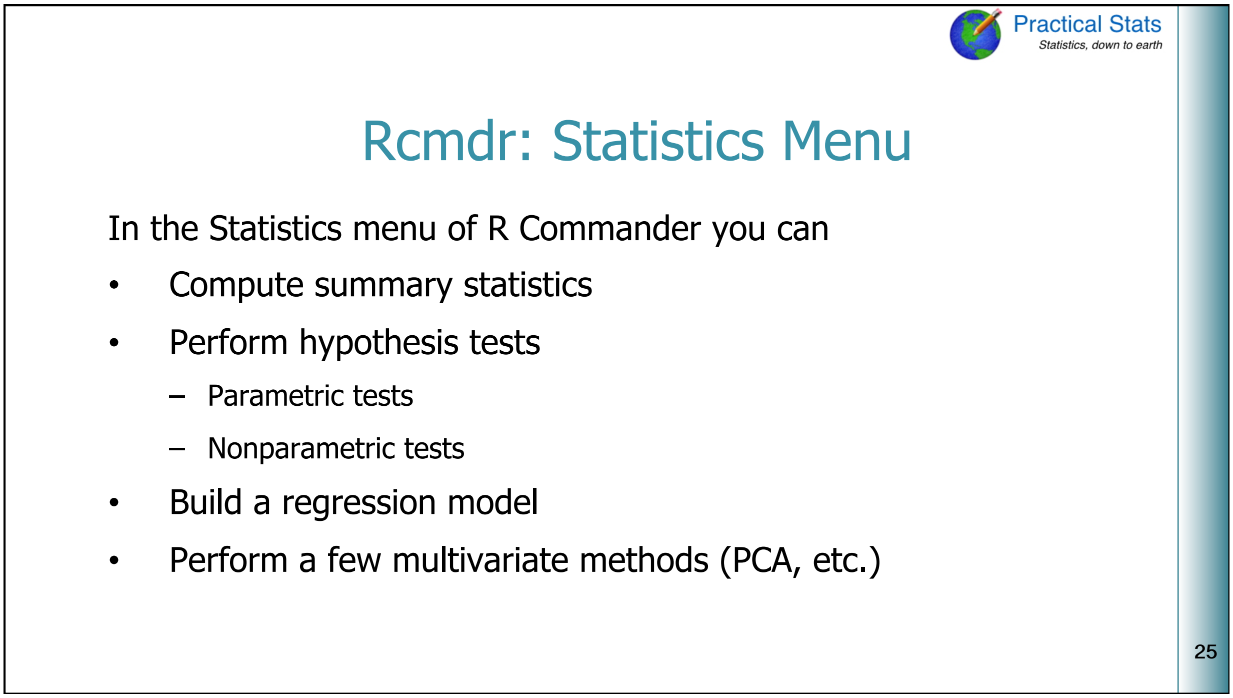
In Rcmdr, choose

`Statistics> Summaries > Numerical summaries`

And choose the variable of interest (use TotalP). This produces:

```
    mean       sd  IQR   0%  25%  50%  75% 100%   n
 6.389655 1.175447  1.4  4.4  5.6    6    7    9  29
```

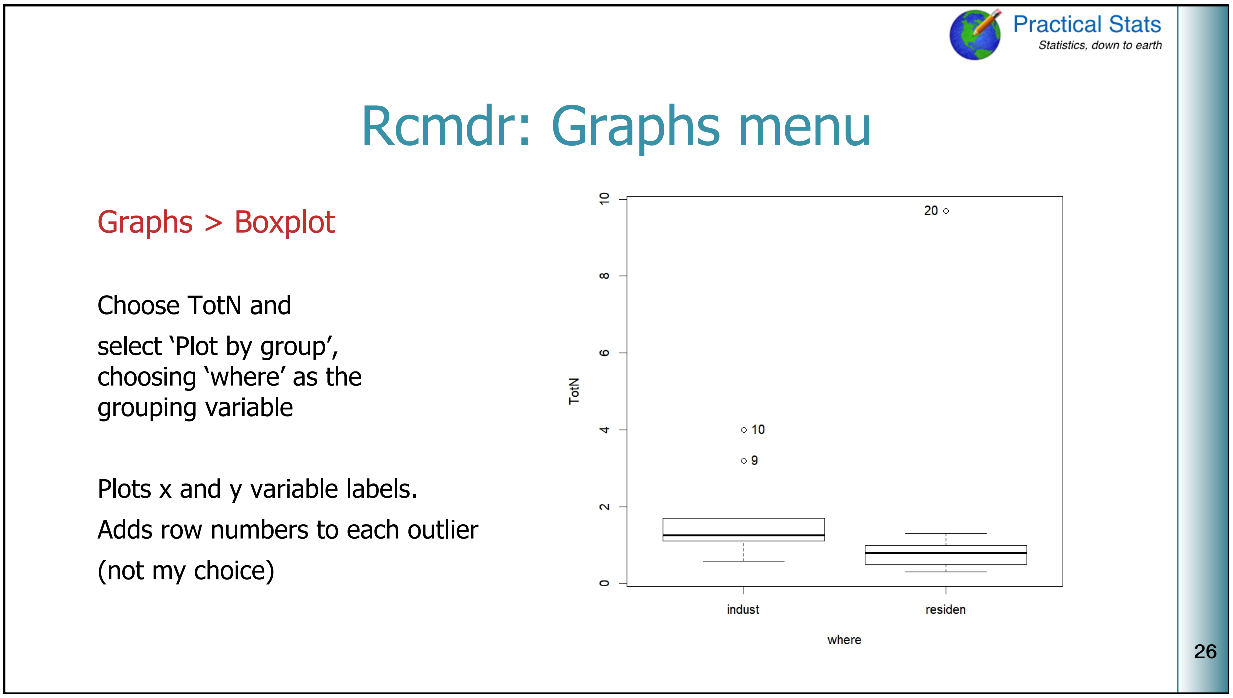
24



Rcmdr: Statistics Menu

In the Statistics menu of R Commander you can

- Compute summary statistics
- Perform hypothesis tests
  - Parametric tests
  - Nonparametric tests
- Build a regression model
- Perform a few multivariate methods (PCA, etc.)



Rcmdr: Graphs menu

Graphs > Boxplot

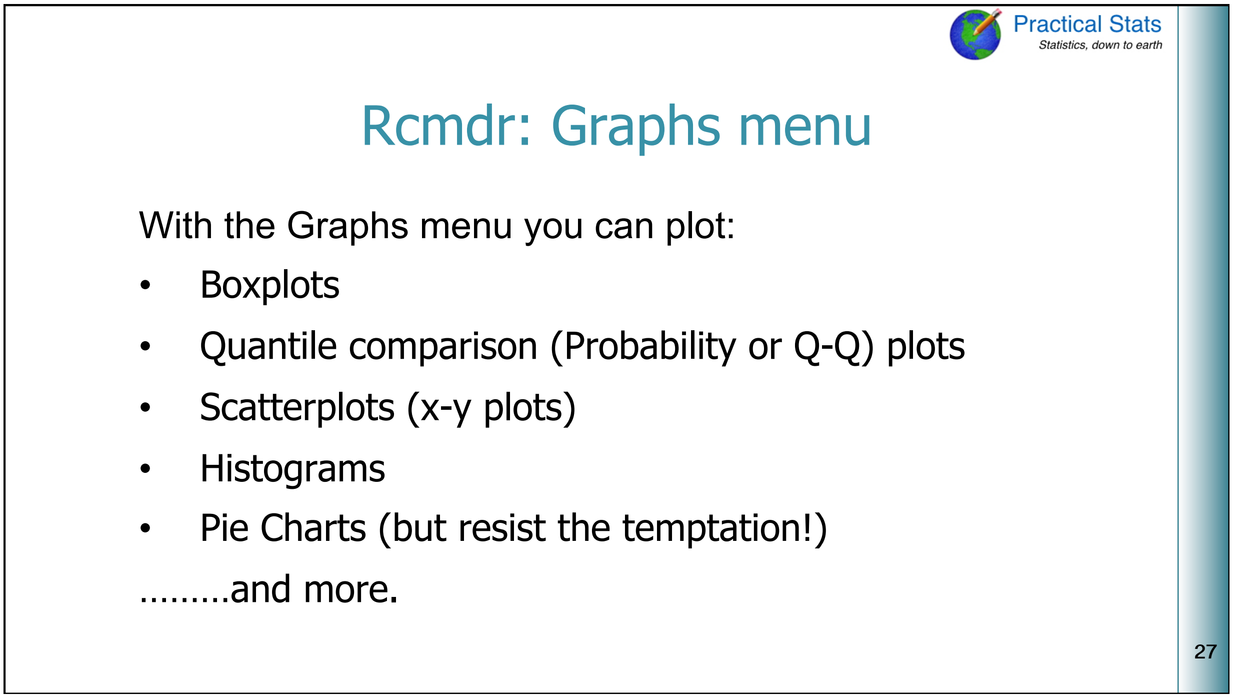
Choose TotN and

select ‘Plot by group’, choosing ‘where’ as the grouping variable

Plots x and y variable labels.

Adds row numbers to each outlier

(not my choice)



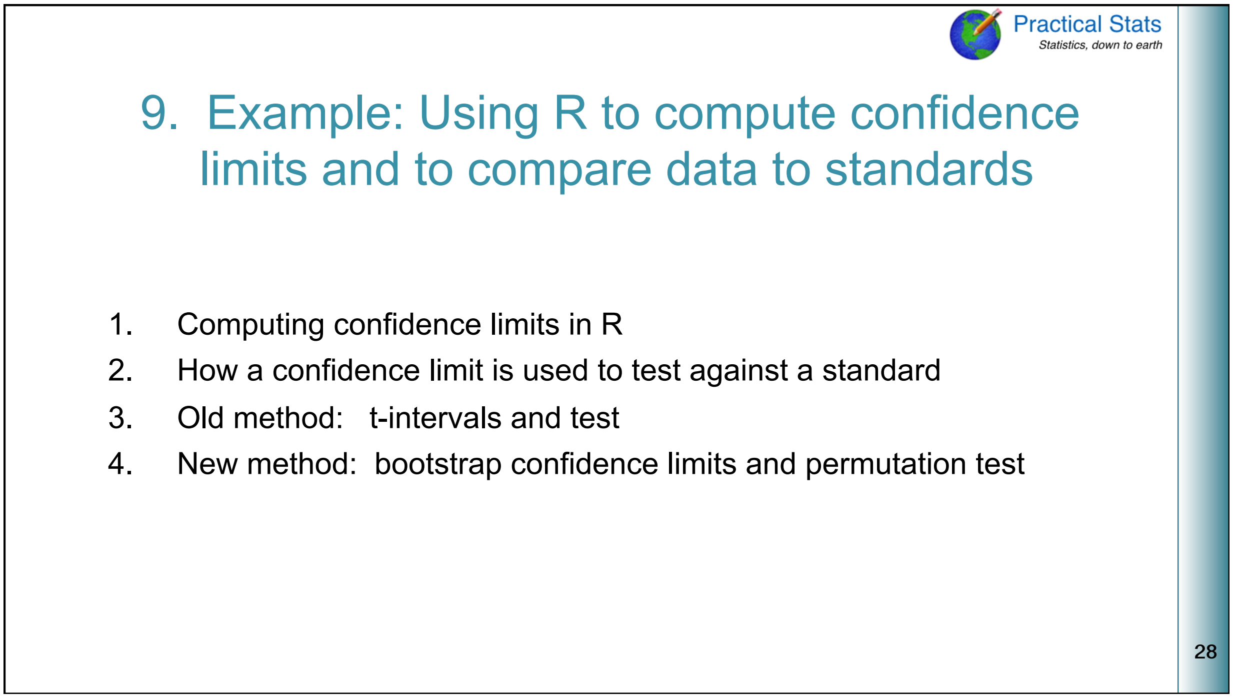
Rcmdr: Graphs menu

With the Graphs menu you can plot:

- Boxplots
- Quantile comparison (Probability or Q-Q) plots
- Scatterplots (x-y plots)
- Histograms
- Pie Charts (but resist the temptation!)

………and more.

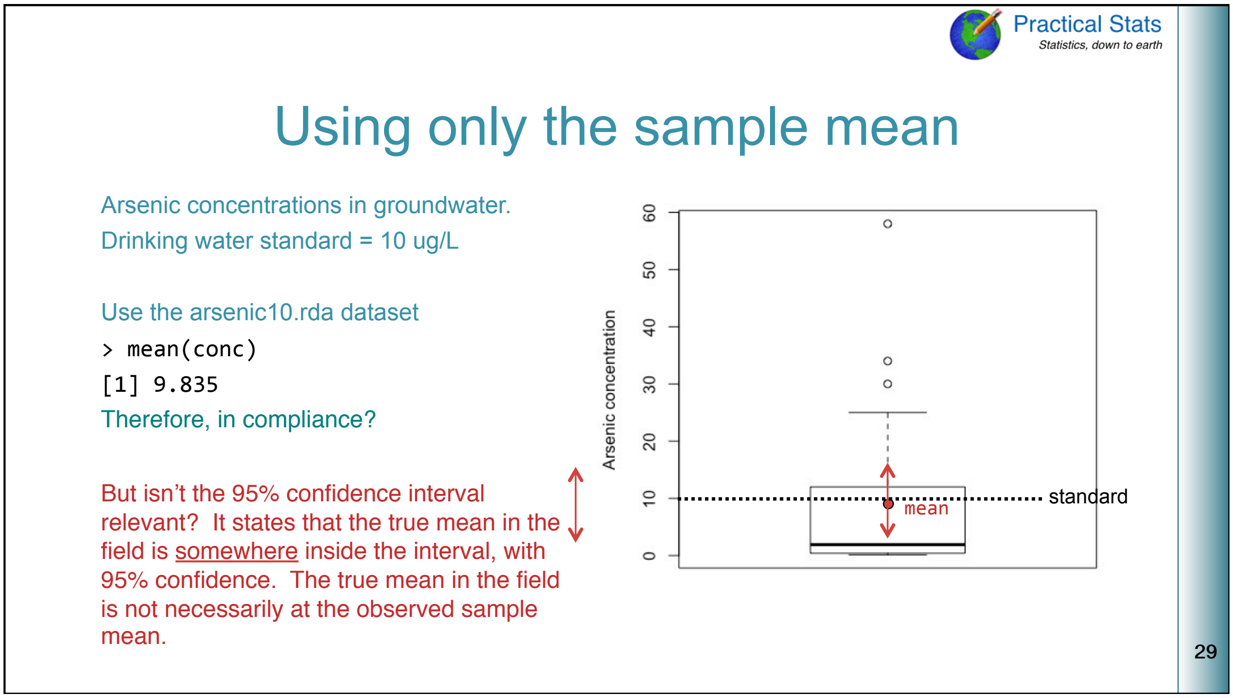
27



9. Example: Using R to compute confidence limits and to compare data to standards

1. Computing confidence limits in R
2. How a confidence limit is used to test against a standard
3. Old method: t-intervals and test
4. New method: bootstrap confidence limits and permutation test

28



Using only the sample mean

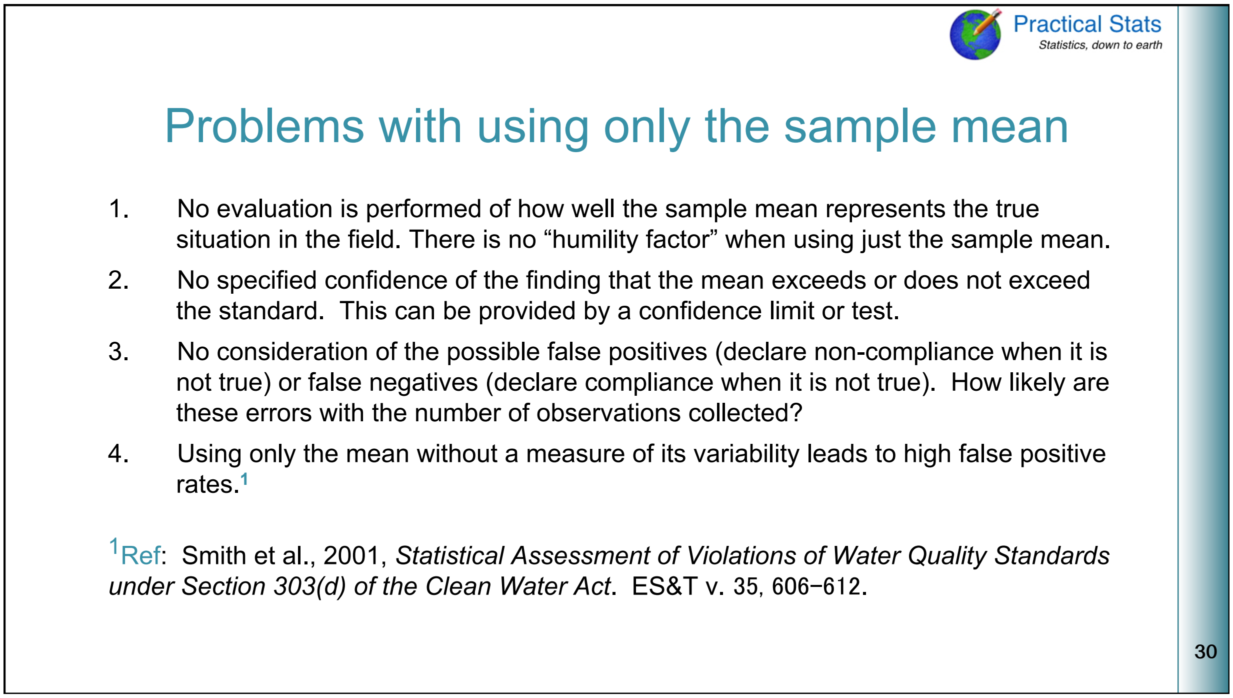
Arsenic concentrations in groundwater.
Drinking water standard = 10 ug/L

Use the arsenic10.rda dataset

```
> mean(conc)
[1] 9.835
```

Therefore, in compliance?

But isn't the 95% confidence interval relevant? It states that the true mean in the field is <u>somewhere</u> inside the interval, with 95% confidence. The true mean in the field is not necessarily at the observed sample mean.

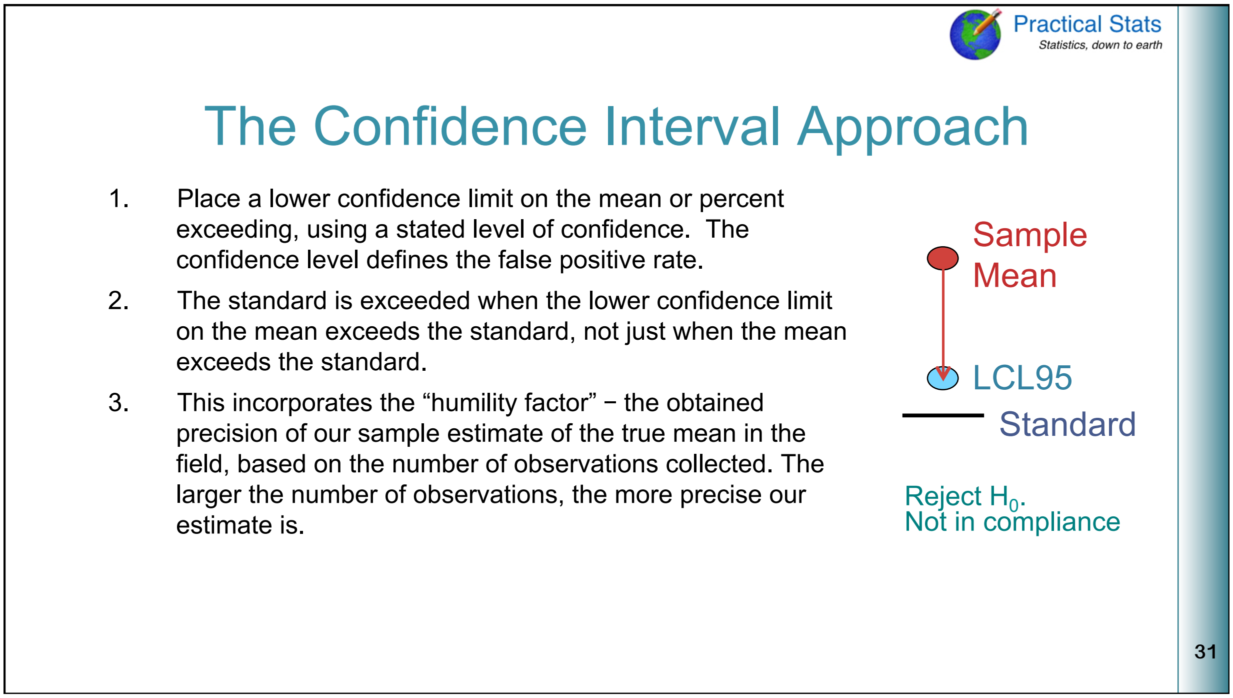


Practical Stats
Statistics, down to earth

Problems with using only the sample mean

1. No evaluation is performed of how well the sample mean represents the true situation in the field. There is no "humility factor" when using just the sample mean.
2. No specified confidence of the finding that the mean exceeds or does not exceed the standard. This can be provided by a confidence limit or test.
3. No consideration of the possible false positives (declare non-compliance when it is not true) or false negatives (declare compliance when it is not true). How likely are these errors with the number of observations collected?
4. Using only the mean without a measure of its variability leads to high false positive rates.[1]

[1] Ref: Smith et al., 2001, *Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act*. ES&T v. 35, 606–612.

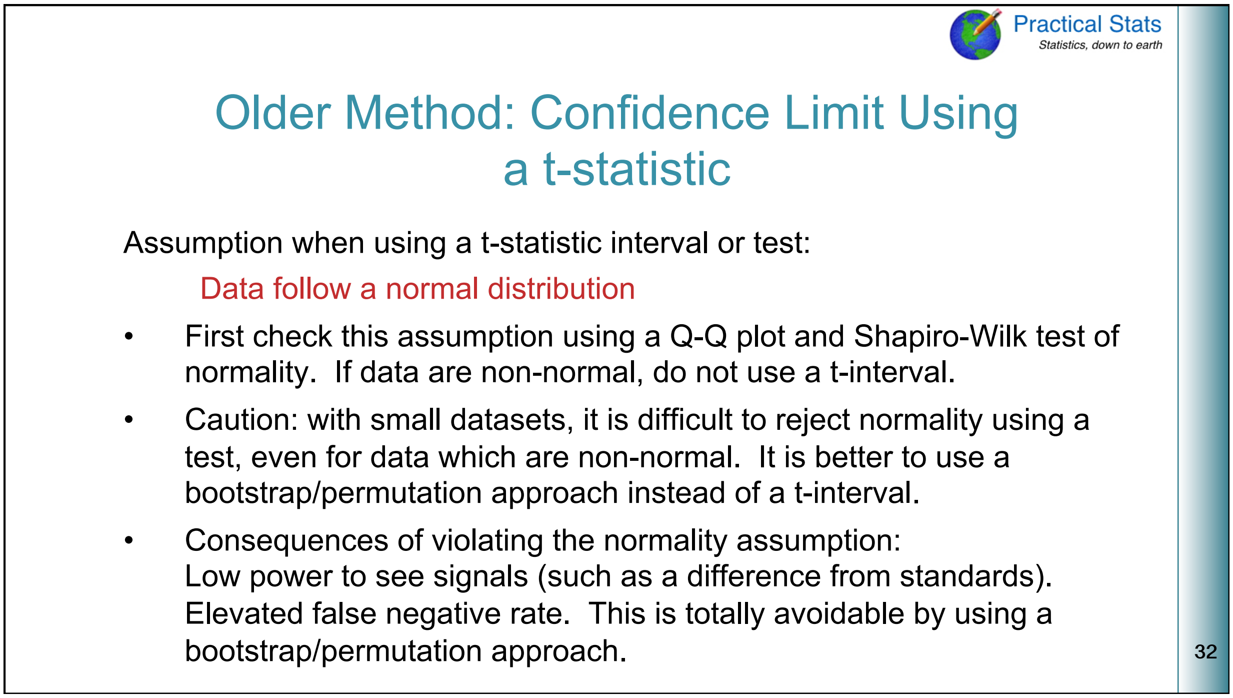


Practical Stats
Statistics, down to earth

The Confidence Interval Approach

1. Place a lower confidence limit on the mean or percent exceeding, using a stated level of confidence. The confidence level defines the false positive rate.
2. The standard is exceeded when the lower confidence limit on the mean exceeds the standard, not just when the mean exceeds the standard.
3. This incorporates the “humility factor” − the obtained precision of our sample estimate of the true mean in the field, based on the number of observations collected. The larger the number of observations, the more precise our estimate is.

Reject H_0.
Not in compliance



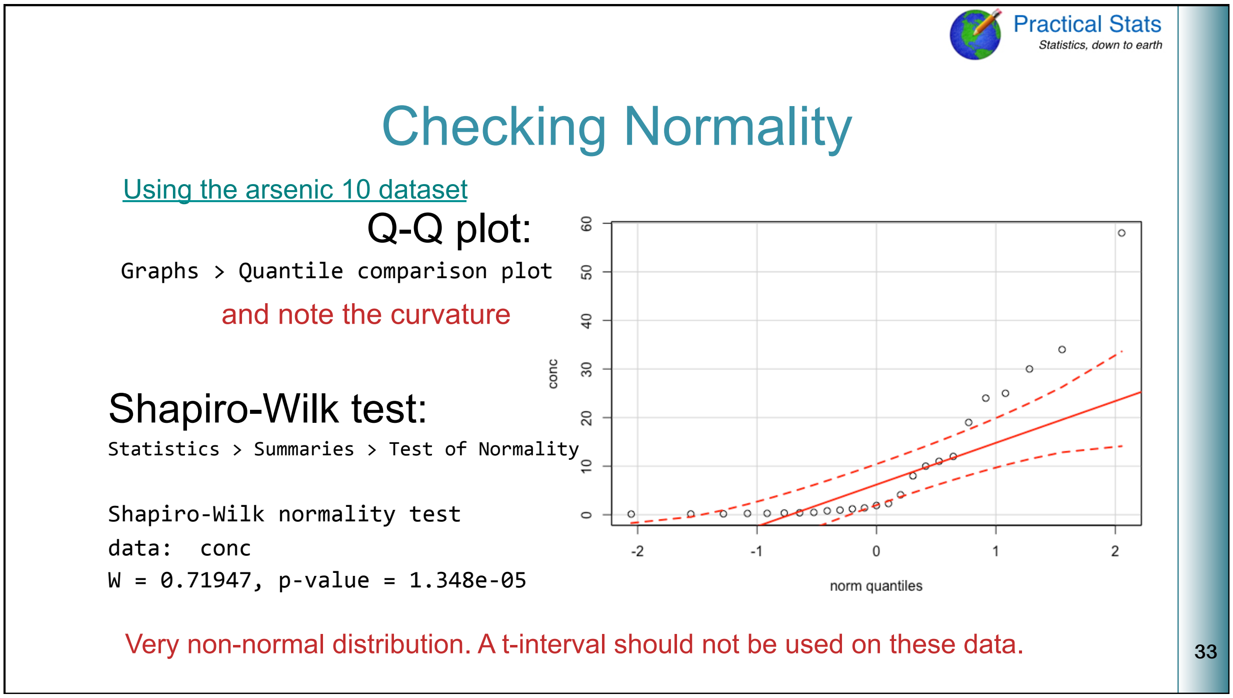
Practical Stats
Statistics, down to earth

Older Method: Confidence Limit Using a t-statistic

Assumption when using a t-statistic interval or test:

Data follow a normal distribution

- First check this assumption using a Q-Q plot and Shapiro-Wilk test of normality. If data are non-normal, do not use a t-interval.
- Caution: with small datasets, it is difficult to reject normality using a test, even for data which are non-normal. It is better to use a bootstrap/permutation approach instead of a t-interval.
- Consequences of violating the normality assumption:
  Low power to see signals (such as a difference from standards).
  Elevated false negative rate. This is totally avoidable by using a bootstrap/permutation approach.



Practical Stats
Statistics, down to earth

Checking Normality

Using the arsenic 10 dataset

Q-Q plot:

`Graphs > Quantile comparison plot`

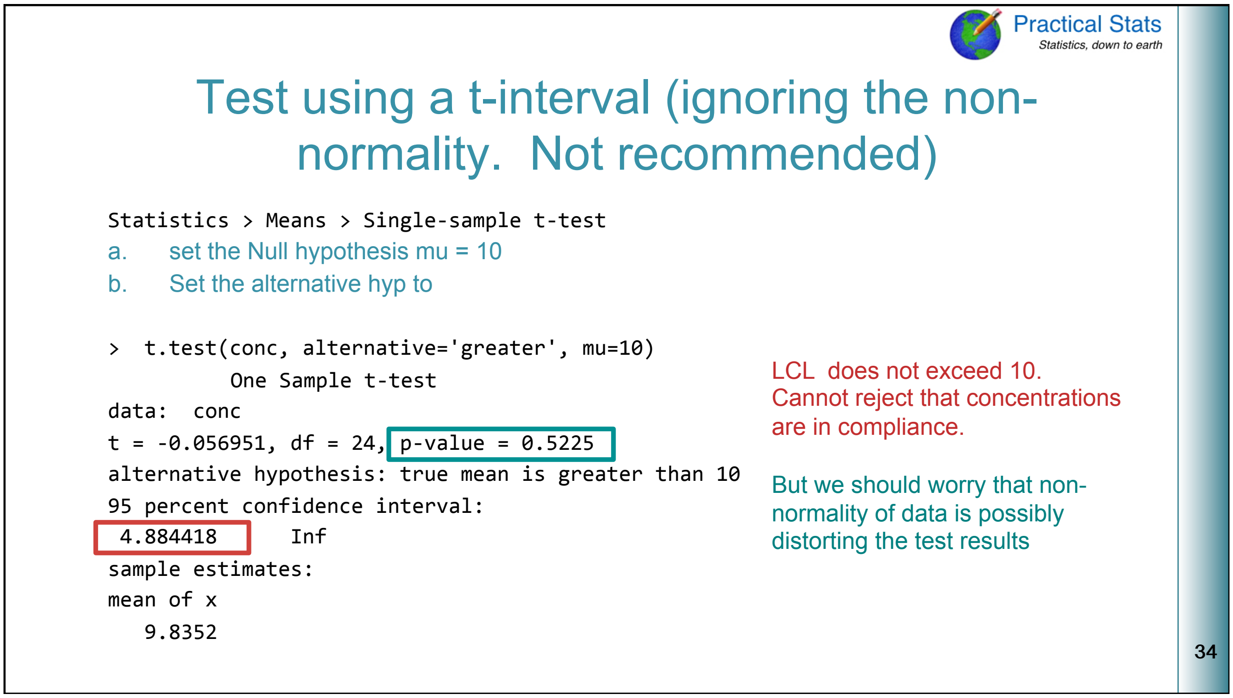
and note the curvature

Shapiro-Wilk test:

`Statistics > Summaries > Test of Normality`

```
Shapiro-Wilk normality test
data:  conc
W = 0.71947, p-value = 1.348e-05
```

Very non-normal distribution. A t-interval should not be used on these data.



Practical Stats
Statistics, down to earth

Test using a t-interval (ignoring the non-normality. Not recommended)

`Statistics > Means > Single-sample t-test`

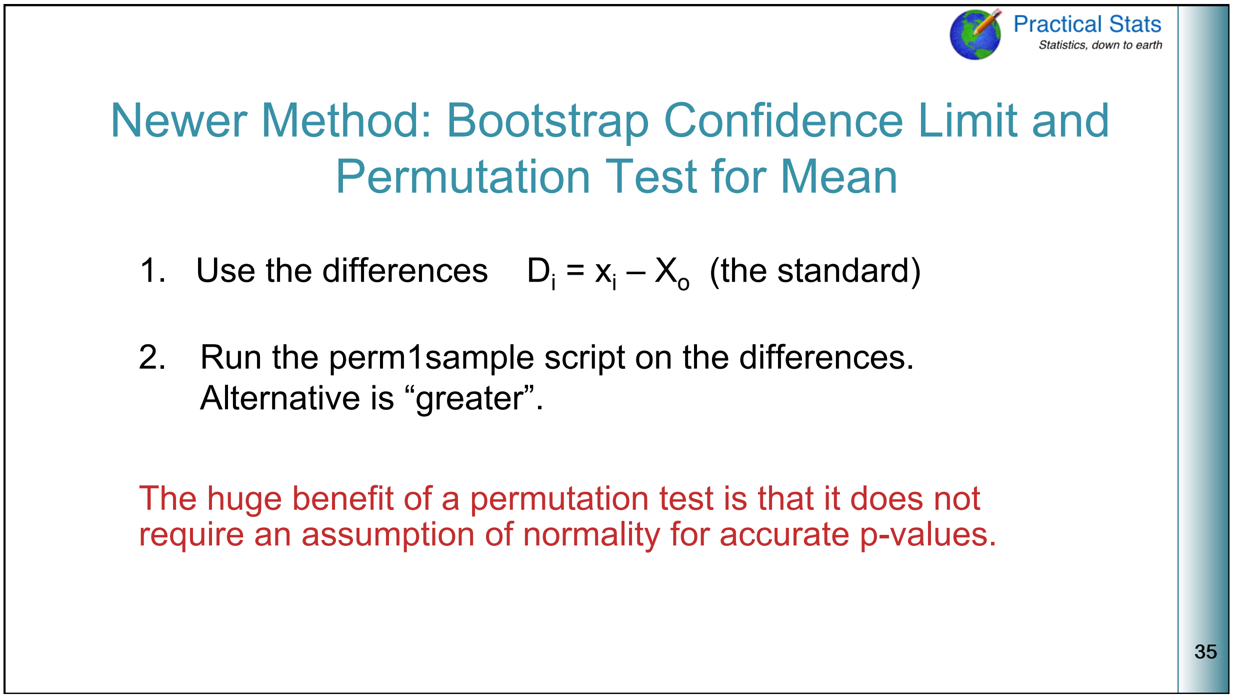
a. set the Null hypothesis mu = 10

b. Set the alternative hyp to

```
>  t.test(conc, alternative='greater', mu=10)
            One Sample t-test
data:  conc
t = -0.056951, df = 24, p-value = 0.5225
alternative hypothesis: true mean is greater than 10
95 percent confidence interval:
 4.884418      Inf
sample estimates:
mean of x
   9.8352
```

LCL does not exceed 10.
Cannot reject that concentrations are in compliance.

But we should worry that non-normality of data is possibly distorting the test results



Practical Stats
Statistics, down to earth

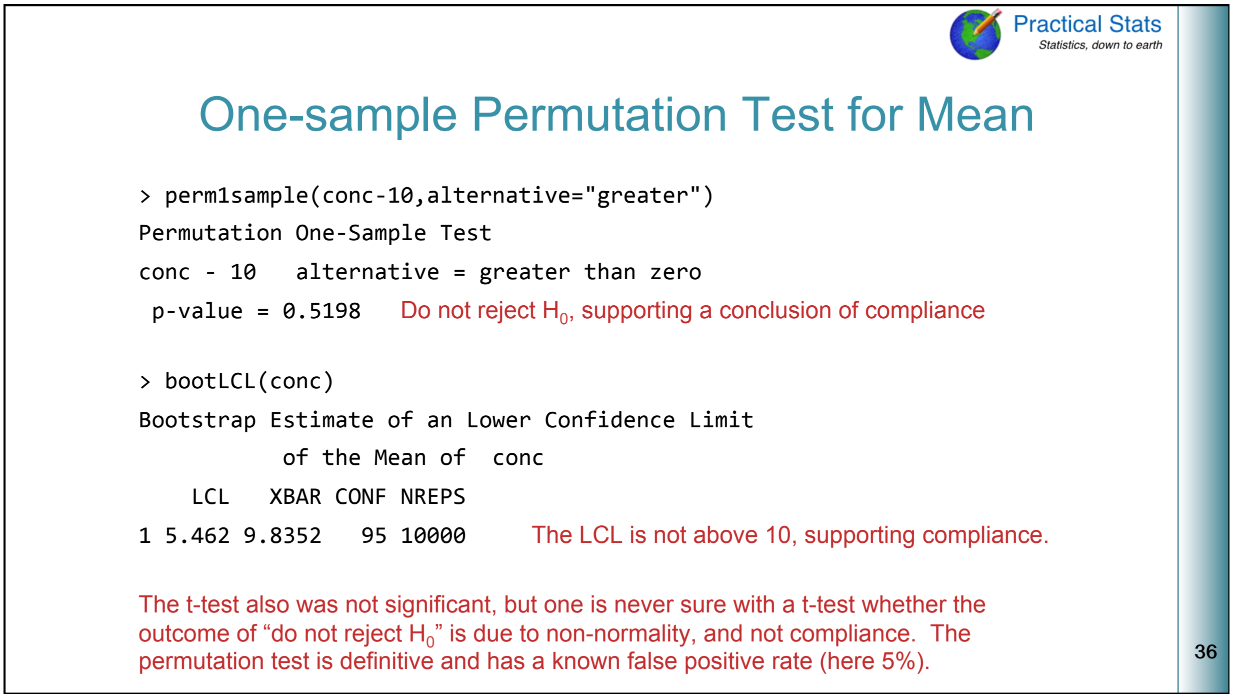
Newer Method: Bootstrap Confidence Limit and Permutation Test for Mean

1. Use the differences $D_i = x_i – X_o$ (the standard)

2. Run the perm1sample script on the differences.
   Alternative is “greater”.

The huge benefit of a permutation test is that it does not require an assumption of normality for accurate p-values.

35



Practical Stats
Statistics, down to earth

One-sample Permutation Test for Mean

```
> perm1sample(conc-10,alternative="greater")
Permutation One-Sample Test
conc - 10   alternative = greater than zero
 p-value = 0.5198
```

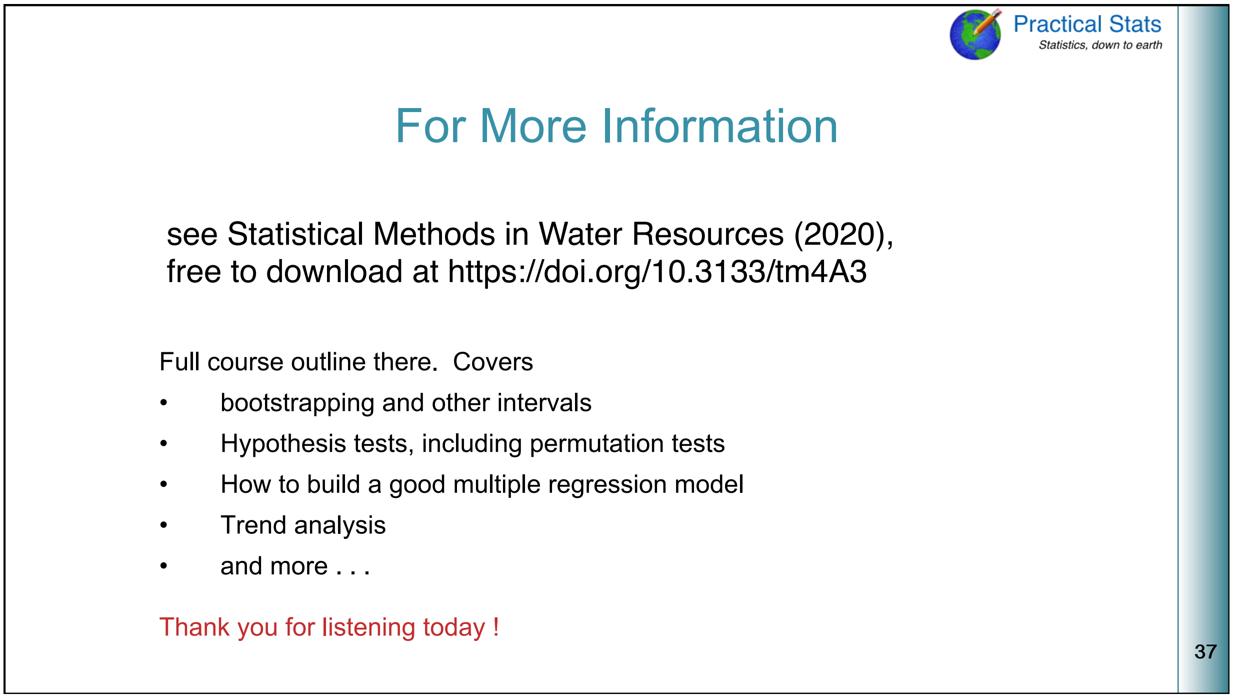
Do not reject H_0, supporting a conclusion of compliance

```
> bootLCL(conc)
Bootstrap Estimate of an Lower Confidence Limit
           of the Mean of  conc
    LCL   XBAR CONF NREPS
1 5.462 9.8352   95 10000
```

The LCL is not above 10, supporting compliance.

The t-test also was not significant, but one is never sure with a t-test whether the outcome of “do not reject H_0” is due to non-normality, and not compliance. The permutation test is definitive and has a known false positive rate (here 5%).

36



Practical Stats
Statistics, down to earth

For More Information

see Statistical Methods in Water Resources (2020),
free to download at https://doi.org/10.3133/tm4A3

Full course outline there. Covers

- bootstrapping and other intervals
- Hypothesis tests, including permutation tests
- How to build a good multiple regression model
- Trend analysis
- and more . . .

Thank you for listening today !

37

Text