R for Statistics and Graphics

Session 2

R for Descriptive Statistics

Mehmet Tevfik DORAK, MD PhD

School of Life Sciences, Pharmacy & Chemistry

Kingston University London

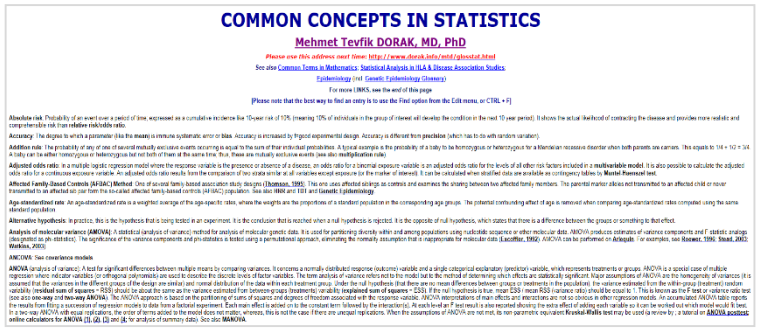
Istanbul University, Capa Faculty of Medicine

19 April 2019

Kingston University London

Introduction to the Crash Course

WEB-BASED SOURCES


COMMON CONCEPTS IN STATISTICS

Mehmet Tevfik DORAK, MD, PhD

Please use this address next time: http://www.dorak.info/mtd/glosstat.html

See also Common Terms in Mathematics - Statistical Analysis in SLA & Phrase Association Studies

Epidemiology (ref. Genetic Epidemiology Glossary)

For more LINKS, see the end of this page

[Please note that the best way to find an entry is to use the Find option from the Edit menu, or CTRL + F]

Absolute risk: Probability of an event over a period of time, expressed as a cumulative incidence like 10-year risk of 10% (meaning 10% of individuals in the group of interest will develop the condition in the next 10 year period). It shows the actual likelihood of contracting the disease and provides more realistic and comprehensive risk than relative risk/odds ratio.

Accuracy: The degree to which a parameter (like the mean) is immune systematic error or bias. Accuracy is increased by tryout experimental design. Accuracy is different from precision (which has to do with random variation).

Addition rule: The probability of any of one of several mutually exclusive events occurring is equal to the sum of their individual probabilities. A typical example is the probability of a baby to be homozygous or heterozygous for a Mendelian recessive disorder when both parents are carriers. This equals to 1/4 + 1/2 = 3/4. A baby can be either homozygous or heterozygous but not both of them at the same time; thus, these are mutually exclusive events (see also multiplication rule).

Adjusted odds ratio: In a multiple logistic regression model where the response variable is the presence or absence of a disease, an odds ratio for a binomial exposure variable is an adjusted odds ratio for the levels of all other risk factors included in a multivariable model. It is also possible to calculate the adjusted odds ratio for a continuous exposure variable. An adjusted odds ratio results from the comparison of two strata similar in all variables except exposure (or the marker of interest). It can be calculated when stratified data are available as contingency tables by Mantel-Haenszel test.

Affected Family-Based Controls (AFBAC) Method: One of several family based association study designs (Thomson, 1995). This one uses affected siblings as controls and examines the sharing between two affected family members. This parental marker alleles not transmitted to an affected child or never transmitted to an affected sib pair from the so-called affected family-based controls (AFBAC) population. See also HBM and TDT and Genetic Epidemiology.

Age-standardized rate: An age-standardized rate is a weighted average of the age-specific rates, where the weights are the proportions of a standard population in the corresponding age groups. The potential confounding effect of age is removed when comparing age-standardized rates computed using the same standard population.

Alternative hypothesis: In practice, this is the hypothesis that is being tested in an experiment. It is the conclusion that is reached when a null hypothesis is rejected. It is the opposite of null hypothesis, which states that there is a difference between the groups or something to that effect.

Analysis of molecular variance (AMOVA): A statistical (analysis of variance) method for analysis of molecular genetic data. It is used for partitioning diversity within and among populations using nucleotide sequence or other molecular data. AMOVA produces estimates of variance components and F statistic analogs (also greatest as phi-statistics). The significance of the variance components and phi-statistics is tested using a permutational approach, eliminating the normality assumption that is inappropriate for molecular data (Excoffier, 1992). AMOVA can be performed on Arlequin. For examples, see Roewer, 1996; Stoad, 2002; Watkins, 2003.

ANCOVA: See covariance models

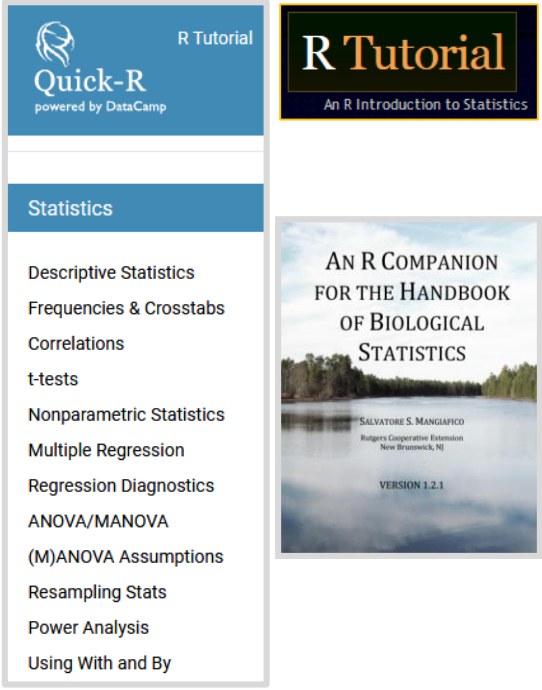
ANOVA (analysis of variance): A test for significant differences between multiple means by comparing variances. It concerns a normally distributed response (outcome) variable and a single categorical explanatory (predictor) variable, which represents treatments or groups. ANOVA is a special case of multiple regression where indicator variables (or orthogonal polynomials) are used to describe the discrete levels of factor variables. The term analysis of variance refers not to the model but to the method of determining which effects are statistically significant. Major assumptions of ANOVA are the homogeneity of variances (it is assumed that the variances in the different groups of the design are similar) and normal distribution of the data within each treatment group. Under the null hypothesis (that there are no mean differences between groups or treatments in the population), the variance estimated from the within-group (treatment) random variability (residual sum of squares = RSS) should be about the same as the variance estimated from between-groups (treatments) variability (explained sum of squares = ESS). If the null hypothesis is true, mean ESS / mean RSS (variance ratio) should be equal to 1. This is known as the F test or variance ratio test (see also one-way and two-way ANOVA). The ANOVA approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable. ANOVA interpretations of main effects and interactions are not so obvious in other regression models. An accumulated ANOVA table reports the results from fitting a succession of regression models to data from a factorial experiment. Each main effect is added on to the constant term followed by the interaction(s). At each level an F test result is also reported showing the extra effect of adding each variable so it can be worked out which model would be best. In a two-way ANOVA with equal replications, the order of terms added to the model does not matter, whereas, this is not the case if there are unequal replications. When the assumptions of ANOVA are not met, its non-parametric equivalent Kruskal-Wallis test may be used (a review by; a tutorial on ANOVA posthoc online calculators for ANOVA (1), (2), (3) and (4); for analysis of summary data). See also MANOVA.

http://www.dorak.info/mtd/glosstat.html


ENDMEMO

§§ Statistical Analysis

» ANOVA
» Binomial Test
» Chi Square Test
» Clustering Trees
» Coefficient
» F-test
» Microarray Analysis
» Normality Test
» Standard Deviation
» T-test
» Wilcoxon Rank Test
» Z-test


R Tutorial

Quick-R

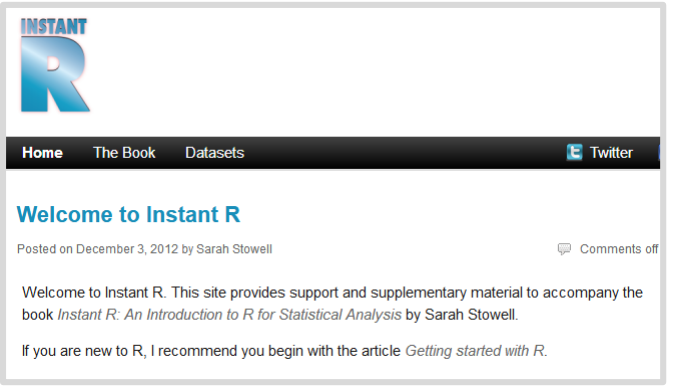
powered by DataCamp

Statistics

Descriptive Statistics
Frequencies & Crosstabs
Correlations
t-tests
Nonparametric Statistics
Multiple Regression
Regression Diagnostics
ANOVA/MANOVA
(M)ANOVA Assumptions
Resampling Stats
Power Analysis
Using With and By

R Tutorial

An R Introduction to Statistics


INSTANT R

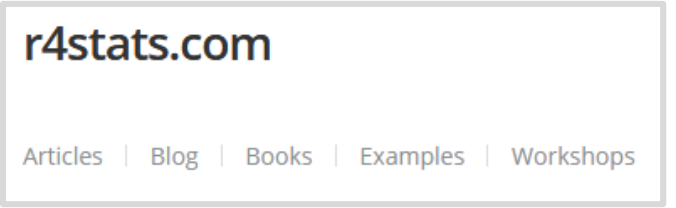
Home | The Book | Datasets | Twitter

Welcome to Instant R

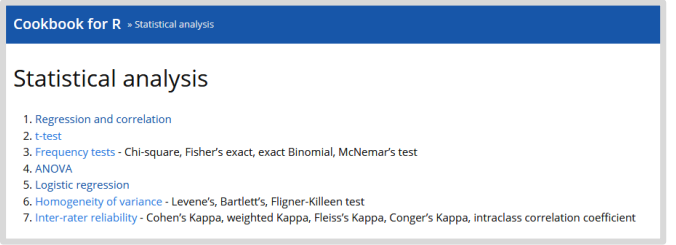
Posted on December 3, 2012 by Sarah Stowell — Comments off

Welcome to Instant R. This site provides support and supplementary material to accompany the book *Instant R: An Introduction to R for Statistical Analysis* by Sarah Stowell.

If you are new to R, I recommend you begin with the article *Getting started with R*.


r4stats.com

Articles | Blog | Books | Examples | Workshops


Cookbook for R » Statistical analysis

Statistical analysis

1. Regression and correlation
2. t-test
3. Frequency tests - Chi-square, Fisher's exact, exact Binomial, McNemar's test
4. ANOVA
5. Logistic regression
6. Homogeneity of variance - Levene's, Bartlett's, Fligner-Killeen test
7. Inter-rater reliability - Cohen's Kappa, weighted Kappa, Fleiss's Kappa, Conger's Kappa, intraclass correlation coefficient

R Tutorials @ ListenData:
https://www.listendata.com/p/r-programming-tutorials.html


how to in R


r-statistics.co by Selva Prabhakaran


BU

Boston University School of Public Health

Kingston University London

Basic Statistical Analysis Using the R Statistical Package

Descriptive Statistics
Visual Exploration

Enter your vector name or a column of a dataset (as dataset$column) within the brackets:

```
hist()     # like: hist(iris$Sepal.Width, col = "darkblue")
plot()     # like: plot(iris$Sepal.Width, col = "darkblue")
boxplot()  # like: boxplot(iris$Sepal.Width, col = "darkblue")
```

Try the following:

```
boxplot(iris$Sepal.Length, iris$Sepal.Width, iris$Petal.Length, iris$Petal.Width)
boxplot(iris[-5])
boxplot(iris[-5], col = "red")
boxplot(iris$Sepal.Width, col="orchid2", pch=24, main="My First Boxplot",
                col.axis="blue", col.lab="red", col.main="darkblue", boxwex =0.2)
# Repeat the last command by changing the numbers for pch (plot symbol) and boxwex (box
  width expansion)
boxplot(iris$Sepal.Width ~ iris$Species, col = "orchid2", border = "red")
        # OR: boxplot(iris$Sepal.Width ~ iris$Species, col = "azure", border = "blue")
abline(h = 3.1, col = "blue")
```

More variations in the script "`s2.R`"

Kingston University London

Script: `s2.R`

Descriptive Statistics
Visual Exploration

Function arguments for base R plots:

Many of the basic plotting commands in base R will accept the same options to control axis limits, labeling, print a title, change the plotting symbol, change the size of the plotting symbols and text, and change the line types. Use them inside the parentheses of a plotting command to have their effect.

```
main = "Eureka"     # add a title above the graph
pch = 16            # set plot symbol to a filled circle
color = "red"       # set the item colour
xlim = c(-10,10)    # set limits of the x-axis (horizontal axis)
ylim = c(0,100)     # set limits of the y-axis (vertical axis)
lty = 2             # set line type to dashed
las = 2             # rotate axis labels to be perpendicular to axis
cex = 1.5           # magnify the plotting symbols 1.5-fold
cex.lab = 1.5       # magnify the axis labels 1.5-fold
cex.axis = 1.3      # magnify the axis annotation 1.3-fold
xlab = "Body size" # label for the x-axis
ylab = "Frequency" # label for the y-axis
```

Example:

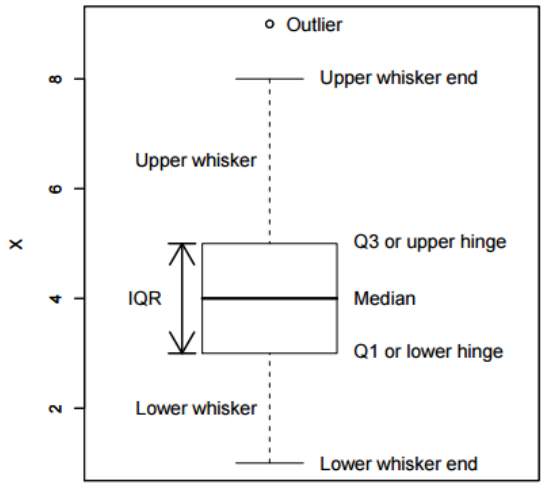
```
boxplot(iris$Sepal.Width, col="red", pch=24, main="My First Boxplot",
col.axis="blue", col.lab="red", col.main="darkblue", boxwex =0.2)
```

Kingston University London

Descriptive Statistics
Boxplot

- ***Boxplots***- boxplot consists of a rectangular box bounded above and below by "hinges" that represent the quartiles Q3 and Q1 respectively, and with a horizontal "median" line through it. You can also see the upper and lower "whiskers", and a point marking a **potential** "outlier".

> *IQR (interquartile range) = Q3—Q1, (the box in the plot)*
>
> *whiskers = ±1.58IQR/√ n ∗ IQR, where n is the number of samples. (datapoints)*




```
# Adding a line to the boxplot:
data(iris)
boxplot(iris[-5], col="orchid2", border="blue")
abline(h = 2.5, col = "red")

# OR, all in one line separated by (;):
data(iris); boxplot(iris[-5], col="orchid2",
border="blue"); abline(h=2.5, col="red")
```

Kingston University London

Descriptive Statistics

Boxplot

Common Box Plot Arguments

Definitions and examples for a few of the more common arguments are provided in the table below. View the subpages under Data Visualization > Box Plot to learn more about the use of these arguments in box plots.

Argument	Description	Example
border	Single value or vector giving the border color of the boxes. If vector length is less # of boxes, the color values will be repeated. Specifying a single value will border all boxes with that color.	border = "Red" border = c("Red","Blue") border = data$vector
boxwex	A scale factor to be applied to all boxes. When there are only a few groups, the appearance of the plot can be improved by making the boxes narrower.	boxwex = 0.1
col	Single value or vector giving the color of the boxes. If vector length is less # of boxes, the color values will be repeated. Specifying a single value will shade all boxes with that color.	col = "Red" col = c("Red","Blue") col = data$vector
horizontal	A logical value. If TRUE, the bars are drawn horizontally with the first at the bottom.	horizontal = TRUE
log	String specifying if axis scales should be logarithmic. "x" for x-axis, "y" for y axis.	log = "x" log = "y"
names	By default the group labels from the grouping variable are printed. This argument can be used to define custom groups labels.	names = c("a","b","c")
notch	If notch is TRUE, a notch is drawn in each side of the boxes. If the notches of two plots do not overlap this is 'strong evidence' that the two medians differ.	notch = TRUE
outline	If outline is FALSE, the outliers are not drawn (as points whereas S+ uses lines).	outline = FALSE
range	This determines how far the plot whiskers extend out from the box. If range is positive, the whiskers extend to the most extreme data point which is no more than the defined range value times the interquartile range from the box. A value of zero causes the whiskers to extend to the data extremes.	range = 0 range = 1
staplewex	A scale factor to be applied to the whisker caps (staple lines).	staplewex = 2
varwidth	If varwidth is TRUE, the boxes are drawn with widths proportional to the square-roots of the number of observations in the groups.	varwidth = TRUE
width	Vector specifying box widths. If vector length is less than # of boxes, the width values will be repeated.	width = 4 width = c(1,3) width = data$vector

Kingston University London

How to in R

Descriptive Statistics

Statistical Metrics

Enter your vector name or a column of a dataset (as dataset$column) within the brackets:

- `mean()`
- `sd()`
- `var()`
- `median()`
- `mad()` # for median absolute deviation from the median (~SD)
- `IQR()` # for interquartile range (~ SD)
- `skew()` # for skewness

For kurtosis:

- `install.packages("moments"); library("moments");`

```
kurtosis(iris$Sepal.Length)
```

Descriptive Statistics: psych

psych

by William Revelle

describe

Basic Descriptive Statistics Useful For Psychometrics

```
install.packages("psych")
library("psych")
a <- c(2,4,3,4,5,6,5,4,5,6,7,4,3,5,4,6,5,7,6,5,4,5,6,7)
describe(a)
```

```
> library("psych"); a <- c(2,4,3,4,5,6,5,4,5,6,7,4,3,5,4,6,5,7,6,5,4,5,6,7); describe(a)
   vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 24 4.92 1.32      5    4.95 1.48   2   7     5 -0.18    -0.68 0.27
```

After importing an Excel file, try the describe function with a full dataframe (not a vector):

```
describe(iris[-5])
```

```
> describe(iris[-5])
             vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
Sepal.Length    1 150 5.84 0.83   5.80    5.81 1.04 4.3 7.9   3.6  0.31    -0.61 0.07
Sepal.Width     2 150 3.06 0.44   3.00    3.04 0.44 2.0 4.4   2.4  0.31     0.14 0.04
Petal.Length    3 150 3.76 1.77   4.35    3.76 1.85 1.0 6.9   5.9 -0.27    -1.42 0.14
Petal.Width     4 150 1.20 0.76   1.30    1.18 1.04 0.1 2.5   2.4 -0.10    -1.36 0.06
```

mad = median absolute deviation (from the median) ~ SD in normally distributed data

https://www.statisticshowto.datasciencecentral.com/median-absolute-deviation

Kingston University London

Descriptive Statistics: psych

psych

by William Revelle

describe

Basic Descriptive Statistics Useful For Psychometrics

Usage

```
describe(x, na.rm = TRUE, interp=FALSE,skew = TRUE, ranges = TRUE,trim=.1,
             type=3,check=TRUE,fast=NULL,quant=NULL,IQR=FALSE,omit=FALSE)
describeData(x,head=4,tail=4)
describeFast(x)
```

Arguments

x	A data frame or matrix
na.rm	The default is to delete missing data. na.rm=FALSE will delete the case.
interp	Should the median be standard or interpolated
skew	Should the skew and kurtosis be calculated?
ranges	Should the range be calculated?
trim	trim=.1 -- trim means by dropping the top and bottom trim fraction
type	Which estimate of skew and kurtosis should be used? (See details.)
check	Should we check for non-numeric variables? Slower but helpful.
fast	if TRUE, will do n, means, sds, min, max, ranges for an improvement in speed. If NULL, will switch to fast mode for large (ncol * nrow > 10^7) problems, otherwise defaults to fast = FALSE
quant	if not NULL, will find the specified quantiles (e.g. quant=c(.25,.75) will find the 25th and 75th percentiles)
IQR	If TRUE, show the interquartile range
omit	Do not convert non-numerical variables to numeric, omit them instead
head	show the first 1:head cases for each variable in describeData
tail	Show the last nobs-tail cases for each variable in describeData

Descriptive Statistics: psych

psych

by William Revelle

describe

Basic Descriptive Statistics Useful For Psychometrics

```
describe(iris[-5], IQR = TRUE, quant = c(0.10, 0.25, 0.50, 0.75, 0.90))
```

```
> describe(iris[-5], IQR = TRUE, quant = c(.1,.25,.5,.75,.90))
             vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se IQR Q0.1 Q0.25 Q0.5 Q0.75 Q0.9
Sepal.Length    1 150 5.84 0.83   5.80    5.81 1.04 4.3 7.9   3.6  0.31    -0.61 0.07 1.3  4.8   5.1 5.80   6.4 6.90
Sepal.Width     2 150 3.06 0.44   3.00    3.04 0.44 2.0 4.4   2.4  0.31     0.14 0.04 0.5  2.5   2.8 3.00   3.3 3.61
Petal.Length    3 150 3.76 1.77   4.35    3.76 1.85 1.0 6.9   5.9 -0.27    -1.42 0.14 3.5  1.4   1.6 4.35   5.1 5.80
Petal.Width     4 150 1.20 0.76   1.30    1.18 1.04 0.1 2.5   2.4 -0.10    -1.36 0.06 1.5  0.2   0.3 1.30   1.8 2.20
```

Kingston University London

Descriptive Statistics: psych

psych

by William Revelle

describeBy

Basic Summary Statistics By Group

```
describeBy(iris, group = iris$Species)
```

```
 Descriptive statistics by group
group: setosa
             vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
Sepal.Length    1 50 5.01 0.35    5.0    5.00 0.30 4.3 5.8   1.5 0.11    -0.45 0.05
Sepal.Width     2 50 3.43 0.38    3.4    3.42 0.37 2.3 4.4   2.1 0.04     0.60 0.05
Petal.Length    3 50 1.46 0.17    1.5    1.46 0.15 1.0 1.9   0.9 0.10     0.65 0.02
Petal.Width     4 50 0.25 0.11    0.2    0.24 0.00 0.1 0.6   0.5 1.18     1.26 0.01
Species*        5 50 1.00 0.00    1.0    1.00 0.00 1.0 1.0   0.0  NaN      NaN 0.00
------------------------------------------------------------------------------------
group: versicolor
             vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
Sepal.Length    1 50 5.94 0.52   5.90    5.94 0.52 4.9 7.0   2.1  0.10    -0.69 0.07
Sepal.Width     2 50 2.77 0.31   2.80    2.78 0.30 2.0 3.4   1.4 -0.34    -0.55 0.04
Petal.Length    3 50 4.26 0.47   4.35    4.29 0.52 3.0 5.1   2.1 -0.57    -0.19 0.07
Petal.Width     4 50 1.33 0.20   1.30    1.32 0.22 1.0 1.8   0.8 -0.03    -0.59 0.03
Species*        5 50 2.00 0.00   2.00    2.00 0.00 2.0 2.0   0.0   NaN      NaN 0.00
------------------------------------------------------------------------------------
group: virginica
             vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
Sepal.Length    1 50 6.59 0.64   6.50    6.57 0.59 4.9 7.9   3.0  0.11    -0.20 0.09
Sepal.Width     2 50 2.97 0.32   3.00    2.96 0.30 2.2 3.8   1.6  0.34     0.38 0.05
Petal.Length    3 50 5.55 0.55   5.55    5.51 0.67 4.5 6.9   2.4  0.52    -0.37 0.08
Petal.Width     4 50 2.03 0.27   2.00    2.03 0.30 1.4 2.5   1.1 -0.12    -0.75 0.04
Species*        5 50 3.00 0.00   3.00    3.00 0.00 3.0 3.0   0.0   NaN      NaN 0.00
```

Kingston University London

Descriptive Statistics: fBasics

fBasics

by Diethelm Wuertz

basicStatistics

Basic Statistics Summary

```
install.packages("fBasics")
library("fBasics")
a <- c(2,4,3,4,5,6,5,4,5,6,7,4,3,5,4,6,5,7,6,5,4,5,6,7)
basicStats(a)
```

	a
nobs	24.000000
NAs	0.000000
Minimum	2.000000
Maximum	7.000000
1. Quartile	4.000000
3. Quartile	6.000000
Mean	4.916667
Median	5.000000
Sum	118.000000
SE Mean	0.268630
LCL Mean	4.360964
UCL Mean	5.472369
Variance	1.731884
Stdev	1.316011
Skewness	-0.183322
Kurtosis	-0.680157

After importing an Excel file, try the `basicStats()` function with a full dataframe (not a vector):

```
basicStats(iris[-5])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
nobs	150.000000	150.000000	150.000000	150.000000
NAs	0.000000	0.000000	0.000000	0.000000
Minimum	4.300000	2.000000	1.000000	0.100000
Maximum	7.900000	4.400000	6.900000	2.500000
1. Quartile	5.100000	2.800000	1.600000	0.300000
3. Quartile	6.400000	3.300000	5.100000	1.800000
Mean	5.843333	3.057333	3.758000	1.199333
Median	5.800000	3.000000	4.350000	1.300000
Sum	876.500000	458.600000	563.700000	179.900000
SE Mean	0.067611	0.035588	0.144136	0.062236
LCL Mean	5.709732	2.987010	3.473185	1.076353
UCL Mean	5.976934	3.127656	4.042815	1.322313
Variance	0.685694	0.189979	3.116278	0.581006
Stdev	0.828066	0.435866	1.765298	0.762238
Skewness	0.308641	0.312615	-0.269411	-0.100917
Kurtosis	-0.605813	0.138705	-1.416857	-1.358179

Descriptive Statistics: pastecs

pastecs

by Philippe Grosjean

stat.desc

Descriptive Statistics On A Data Frame Or Time Series

Compute a table giving various descriptive statistics about the series in a data frame or in a single/multiple time series

DEFAULTS: `stat.desc(x, basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)`

```
install.packages ("pastecs")
library("pastecs")
x <- c(2,4,3,4,5,6,5,4,5,6,7,4,3,5,4,6,5,7,6,5,4,5,6,7)
stat.desc(x)
```

Try the `stat.desc()` function with a full dataframe (not a vector):

```
stat.desc(iris, norm = TRUE)   # measures of normal distribution included
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
nbr.val	150.00000000	150.00000000	1.500000e+02	1.500000e+02
nbr.null	0.00000000	0.00000000	0.000000e+00	0.000000e+00
nbr.na	0.00000000	0.00000000	0.000000e+00	0.000000e+00
min	4.30000000	2.00000000	1.000000e+00	1.000000e-01
max	7.90000000	4.40000000	6.900000e+00	2.500000e+00
range	3.60000000	2.40000000	5.900000e+00	2.400000e+00
sum	876.50000000	458.60000000	5.637000e+02	1.799000e+02
median	5.80000000	3.00000000	4.350000e+00	1.300000e+00
mean	5.84333333	3.05733333	3.758000e+00	1.199333e+00
SE.mean	0.06761132	0.03558833	1.441360e-01	6.223645e-02
CI.mean.0.95	0.13360085	0.07032302	2.848146e-01	1.229800e-01
var	0.68569351	0.18997942	3.116278e+00	5.810063e-01
std.dev	0.82806613	0.43586628	1.765298e+00	7.622377e-01
coef.var	0.14171126	0.14256420	4.697441e-01	6.355511e-01
skewness	0.30864073	0.31261470	-2.694109e-01	-1.009166e-01
skew.2SE	0.77924478	0.78927812	-6.801988e-01	-2.547904e-01
kurtosis	-0.60581253	0.13870468	-1.416857e+00	-1.358179e+00
kurt.2SE	-0.76961200	0.17620762	-1.799947e+00	-1.725403e+00
normtest.W	0.97609027	0.98491787	8.762681e-01	9.018349e-01
normtest.p	0.01018116	0.10115427	7.412263e-10	1.680465e-08

Kingston University London

Descriptive Statistics: `Hmisc`

Hmisc

by Frank E Harrell Jr

describe

Concise Statistical Description Of A Vector, Matrix, Data Frame, Or Formula

describe() is a generic method that invokes describe.data.frame(), describe.matrix(), describe.vector(), or describe.formula(). describe.vector() is the basic function for handling a single variable.

```
install.packages("Hmisc")
library("Hmisc")
x <- c(2,4,3,4,5,6,5,4,5,6,7,4,3,5,4,6,5,7,6,5,4,5,6,7)
describe(x)
```

Try the `describe()` function with a full dataframe (not a vector):

```
describe(iris)
```

Kingston University London

Descriptive Statistics: epiR

epiR: Tools for the Analysis of Epidemiological Data

epi.descriptives

Descriptive Statistics

Computes descriptive statistics from a vector of numbers.

```
library(epiR)
data(iris)
epi.descriptives(iris$Sepal.Length, conf.level = 0.95)

$`arithmetic`
    n     mean        sd q25 q50 q75   lower    upper min max na
1 150 5.843333 0.8280661 5.1 5.8 6.4 5.70974 5.976927 4.3 7.9  0

$geometric
    n    mean        sd q25 q50 q75    lower    upper min max na
1 150 5.78572 0.1411891 5.1 5.8 6.4 5.655421 5.919022 4.3 7.9  0

$symmetry
   skewness  kurtosis
1 0.3117531 -0.552064
```

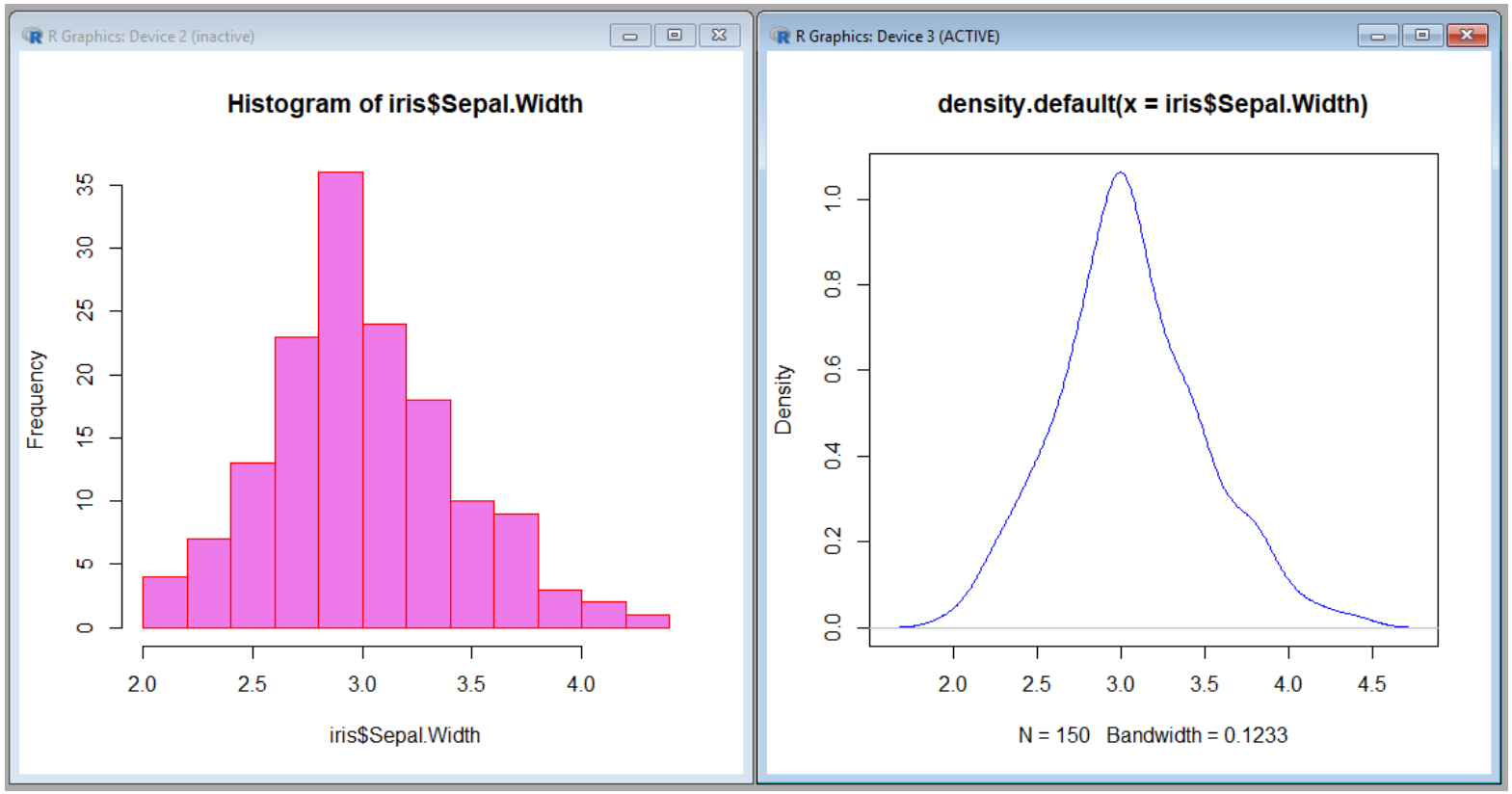
Kingston University London

epiR by Mark Stevenson

Descriptive Statistics: Histogram and Density Plot

Normality checking with histogram and density plot using base R (for vectors)

```
data(iris)
hist(iris$Sepal.Width, col = "orchid2", border = "red")
windows()
d <- density(iris$Sepal.Width); plot(d, col = "blue")
```

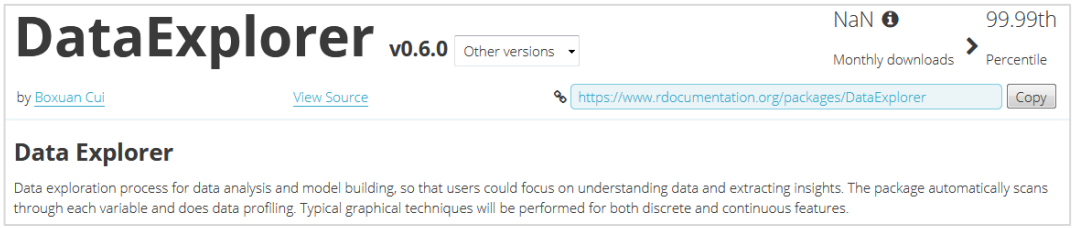


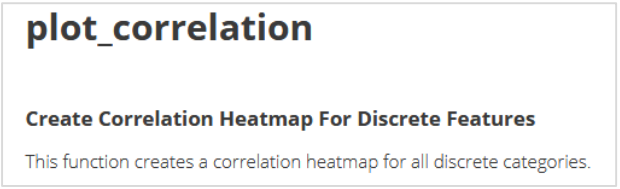
Descriptive Statistics: DataExplorer

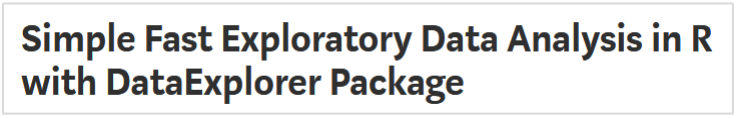
Normality checking with histograms using the R package "*DataExplorer*"

```
install.packages("DataExplorer")
library("DataExplorer")
data(iris)

plot_histogram(iris)          # Ideal for eye test for normal distribution of all variables in one multipanel graph
plot_density(iris)            # Ideal for eye test for normal distribution of all variables in one multipanel graph
```







Kingston University London

Descriptive Statistics: Normality Checking

DataExplorer

by Boxuan Cui View Source

Simple Fast Exploratory Data Analysis in R with DataExplorer Package

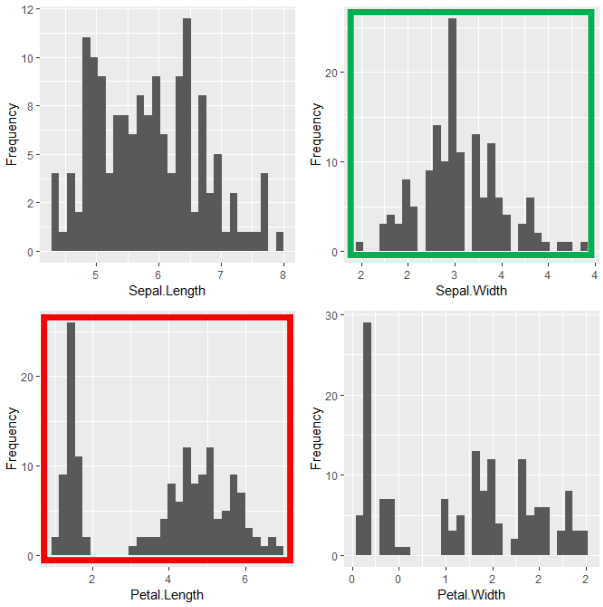
Normality checking with histograms and density plots using the R package "*DataExplorer*"

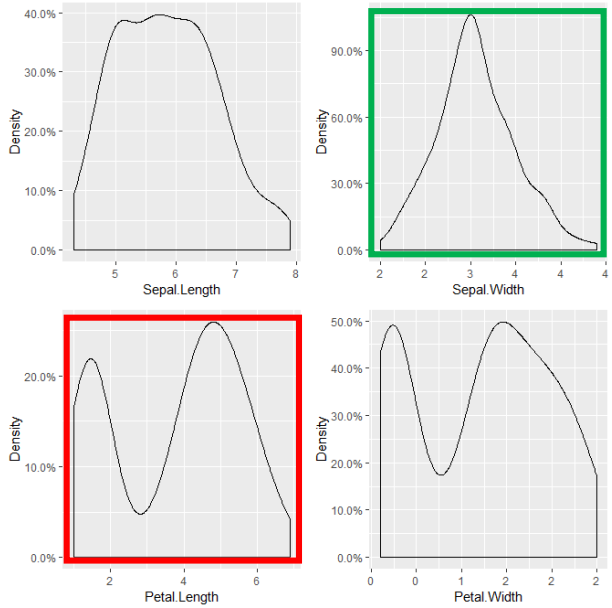
```
install.packages("DataExplorer")
library("DataExplorer")
data(iris)

plot_histogram(iris)          # Ideal for eye test for normal distribution of all variables in one multipanel graph
plot_density(iris)            # Ideal for eye test for normal distribution of all variables in one multipanel graph
```

Can be run like: "`plot_histogram(iris); plot_density(iris)`" in one line

You can have the two plots simultaneously on two windows: "`plot_histogram(iris); windows(); plot_density(iris)`"





Kingston University London

Descriptive Statistics: Normality Checking

Simple Fast Exploratory Data Analysis in R with DataExplorer Package

Simple Fast Exploratory Data Analysis in R with DataExplorer Package

```
install.packages("DataExplorer")
library("DataExplorer")
data(iris)
plot_str(iris)
plot_missing(iris)
plot_histogram(iris)                # Ideal for eye test for normal distribution
of all variables in one multipanel graph
plot_density(iris)                  # Ideal for eye test for normal distribution
of all variables in one multipanel graph
plot_correlation(iris)              # All pairwise correlations with heatmap
# For options of plot_correlation(), see:
https://www.rdocumentation.org/packages/DataExplorer/versions/0.6.0/topics/plot_correlation
plot_bar(iris)                      # Bar graph for categorical variables

# Next command to create a full report requires "installr" and then installation of "pandoc" with
installr:
install.packages("installr")
library("installr")
install.pandoc()
# If successful (may have 32byte - 64byte issues):
create_report(iris)
# Runs all the functions above and creates a full report in HTML format
```

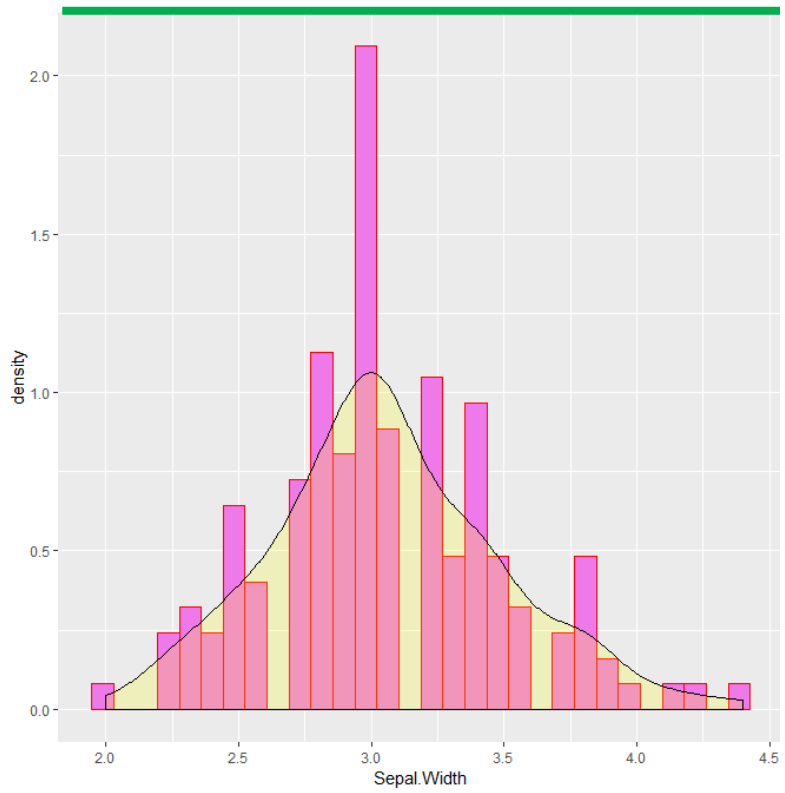
Kingston University London

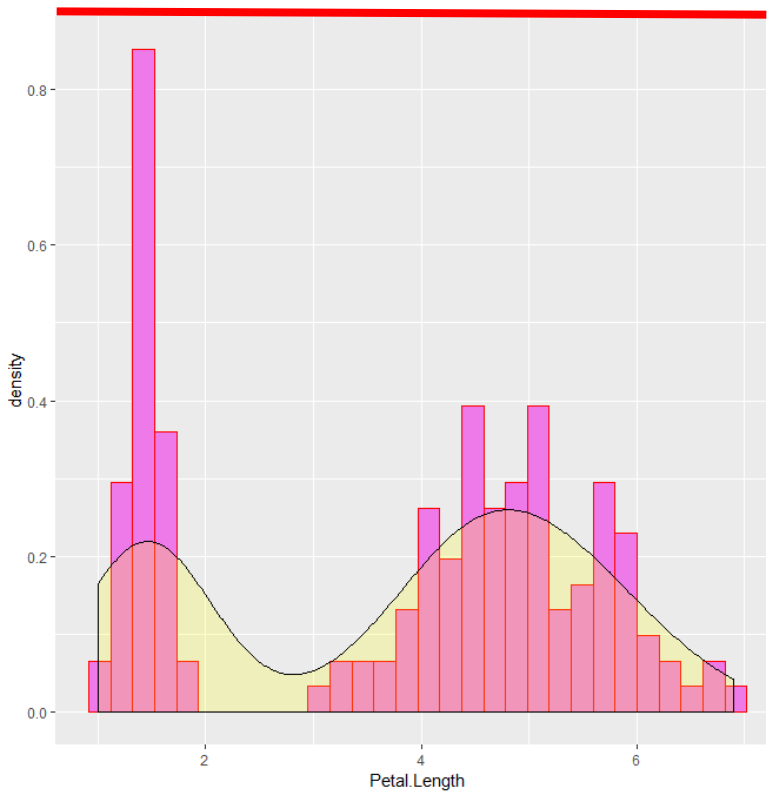
Descriptive Statistics: Normality Checking

Histogram overlaid with density plot with ggplot2:

```
install.packages("ggplot2"); library("ggplot2")
ggplot(iris, aes(x=Sepal.Length)) +  geom_histogram(aes(y=..density..), colour="red",
       fill="orchid2") +  geom_density(alpha=0.2, fill="yellow")
```

https://www.r-bloggers.com/how-to-make-a-histogram-with-ggplot2
http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization



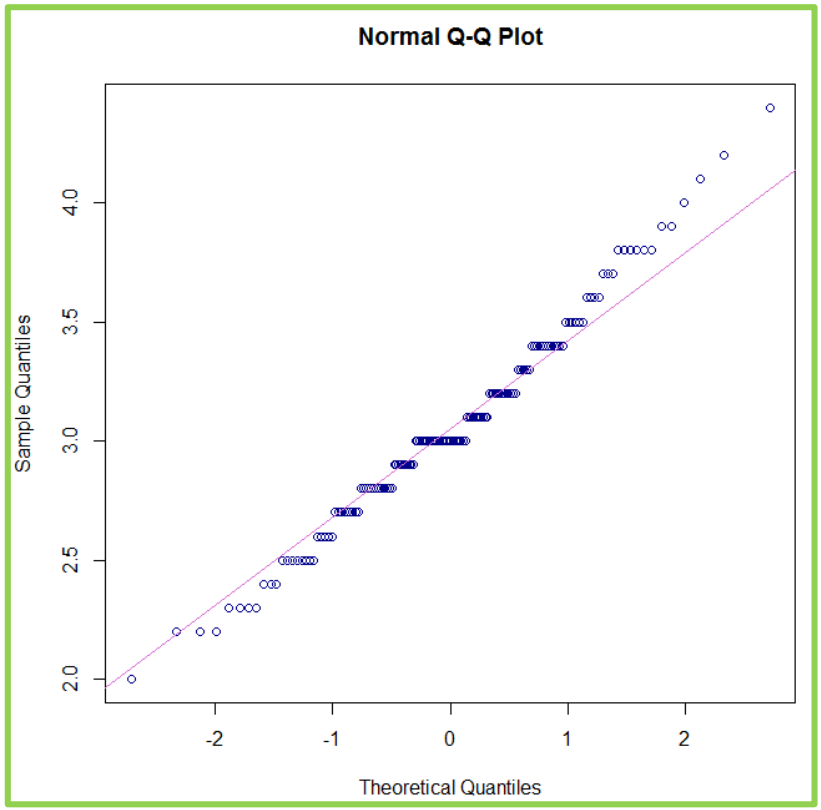


Descriptive Statistics: Normality Checking

qqnorm

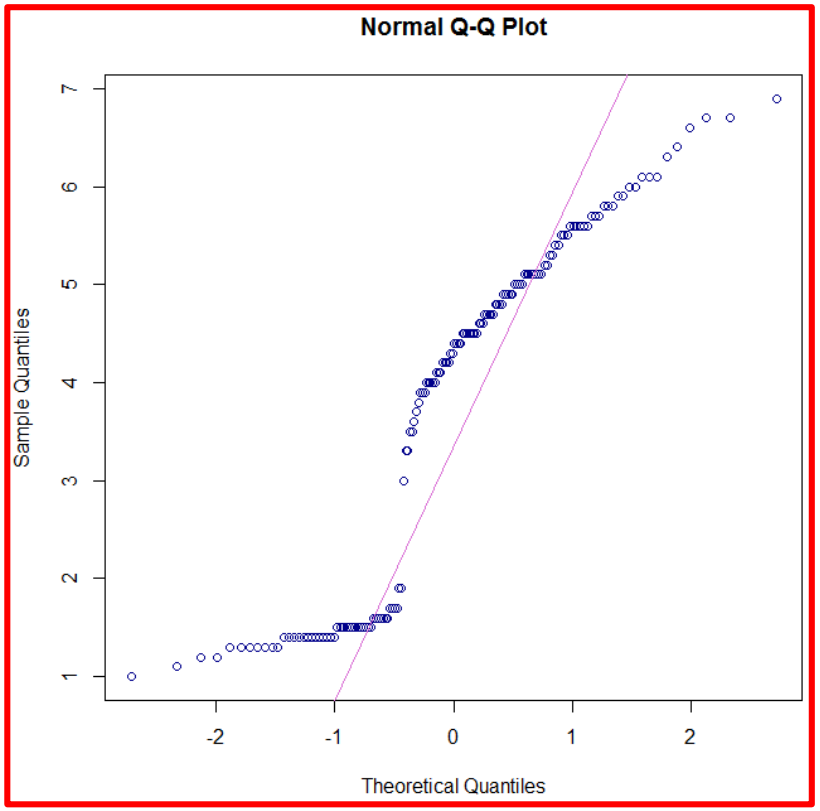
Quantile-Quantile Plots

```
qqnorm(iris$Sepal.Width, col = "darkblue");
qqline(iris$Sepal.Width, col = "orchid")
```



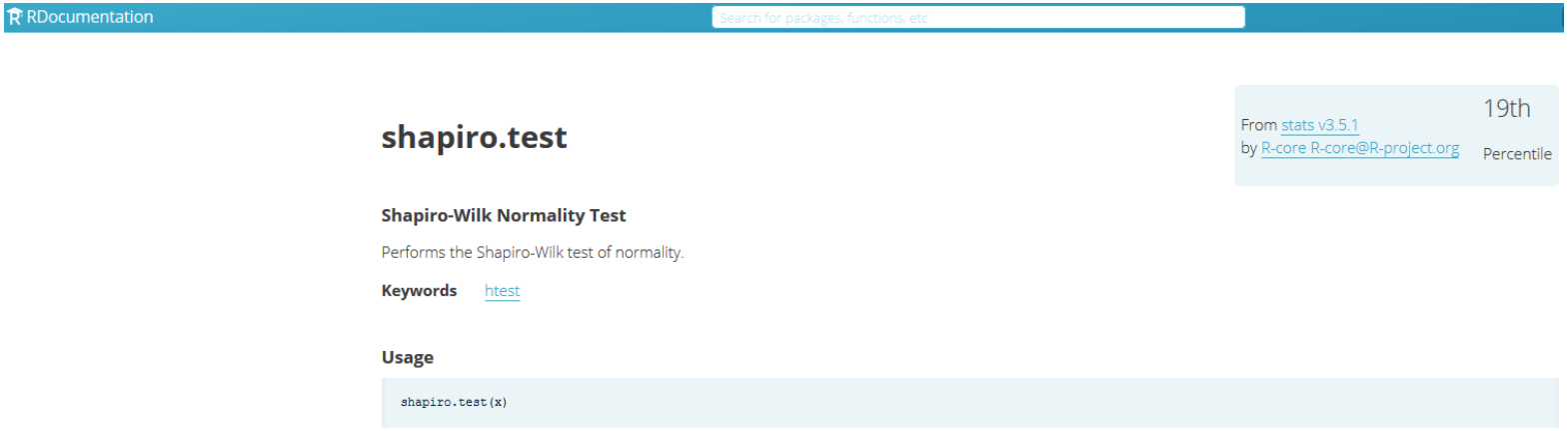
Sepal.Width

```
qqnorm(iris$Petal.Length, col = "darkblue");
qqline(iris$Petal.Length, col = "orchid")
```



Petal.Length

Descriptive Statistics: Normality Checking

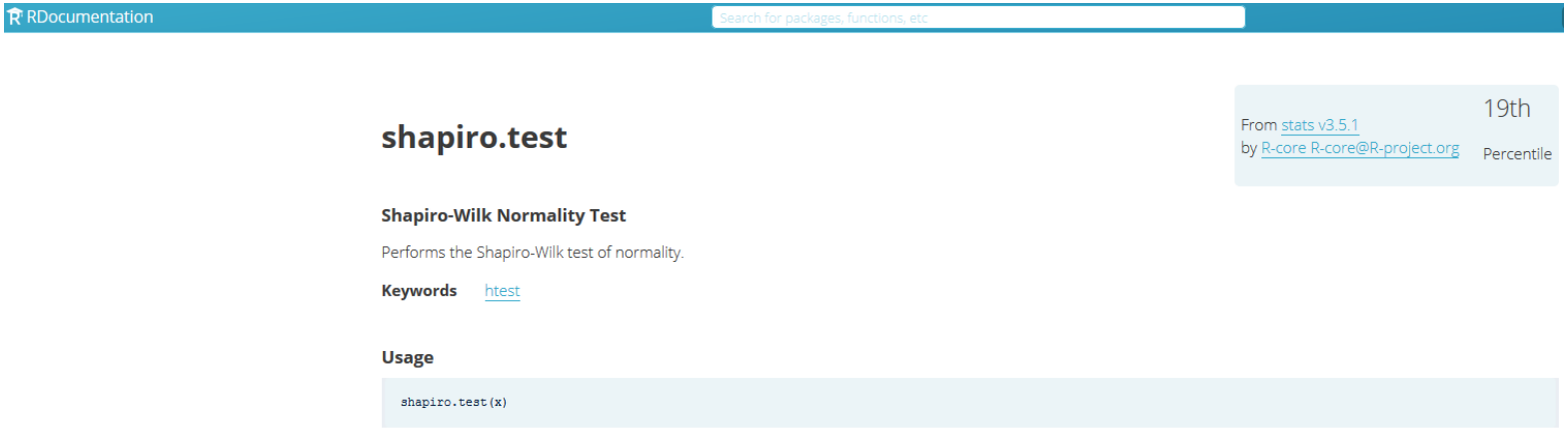



```
data(iris)
shapiro.test(iris$Sepal.Width)    # To run the Shapiro-Wilk test on Sepal.Width
> W = 0.98492, p-value = 0.1012


shapiro.test(iris$Petal.Length)    # To run the Shapiro-Wilk test on Petal.Length
> W = 0.87627, p-value = 7.412e-10
```

Kingston University London

Descriptive Statistics: Normality Checking



Simulated data:

```
set.seed(334)                # To get the same simulated numbers every time
a <- rnorm(100, 30, 5)       # To generate a sample of size 25 from a normal distribution with
                               mean 30 and standard deviation 5

shapiro.test(a)              # To run the Shapiro test on vector "a"
> W = 0.99158, p-value = 0.7899

set.seed(334)
b <- runif(100, min = 20, max = 30)     # To generate a sample size of 25 random
                                          numbers between 20 and 30
shapiro.test(b)              # To run the Shapiro-Wilk test on vector "b"
> W = 0.94666, p-value = 0.000503
```

Kingston University London

Descriptive Statistics: Normality Checking

Kolmogorov-Smirnov Test

ks.test

Kolmogorov-Smirnov Tests

Performs one or two sample Kolmogorov-Smirnov tests.

```
data(iris)
ks.test(iris$Sepal.Width, pnorm, mean(iris$Sepal.Width),
        sd(iris$Sepal.Width))          # To run the K-S test on Sepal.Width
> D = 0.10566, p-value = 0.07023

ks.test(iris$Petal.Length, pnorm, mean(iris$ Petal.Length),
         sd(iris$ Petal.Length))       # To run the K-S test on Petal.Length
> D = 0.19815, p-value = 1.532e-05
```

Alternative tests:
The package "nortest" is dedicated entirely to tests for normality.
This package includes:

- **Anderson–Darling test (ad.test)**
- **Cramer–von Mises test (cvm.test)**
- **Lilliefors test (lillie.test)**
- **Pearson chi-squared test for the composite hypothesis of normality (pearson.test)**
- **Shapiro–Francia test (sf.test)**

Kingston University London

Notes on Graphics

You can save any graphic you generated by code or by the graphics window menu in multiple formats, including TIFF (you can also copy it to the clipboard) - see the script: "tiff.R"

Script: `tiff.R`

You can generate two graphics simultaneously on separate windows

You can fit multiple plots in one window to create a multipanel graphic

You can use a wide variety of options or colors for any element of the graphics

For a full list of graphical parameters, see:

https://www.statmethods.net/advgraphs/parameters.html
https://www.statmethods.net/advgraphs/axes.html

R Color Chart: http://www.endmemo.com/program/R/color.php (see next slide)

PCH Symbols Chart: http://www.endmemo.com/program/R/pchsymbols.php (see the one after next slide)

For advanced graphics, you can use `lattice` (included in base R), or `ggplot2` (see script `s2.R` for examples)

Kingston University London

Notes on Graphics

Graphical Parameters

You can customize many features of your graphs (fonts, colors, axes, titles) through graphic options.

One way is to specify these options in through the **par( )** function. If you set parameter values here, the changes will be in effect for the rest of the session or until you change them again. The format is **par(*optionname*=*value*, *optionname*=*value*, ...)**

```
# Set a graphical parameter using par()

par()              # view current settings
opar <- par()      # make a copy of current settings
par(col.lab="red") # red x and y labels
hist(mtcars$mpg)   # create a plot with these new settings
par(opar)          # restore original settings
```

Base R: Colors

R color cheatsheet

Finding a good color scheme for presenting data can be challenging. This color cheatsheet will help!

R uses hexadecimal to represent colors

Hexadecimal is a base-16 number system used to describe color. Red, green, and blue are each represented by two characters (#rrggbb). Each character has 16 possible symbols: 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F:

"00" can be interpreted as 0.0 and "FF" as 1.0
i.e., red= #FF0000 , black=#000000, white = #FFFFFF

Two additional characters (with the same scale) can be added to the end to describe transparency (#rrggbbaa)

R has 657 built in color names

To see a list of names:
colors()
These colors are displayed on P. 3.

Example:
peachpuff4

R translates various color models to hex, e.g.:

- RGB (red, green, blue): The default intensity scale in R ranges from 0-1; but another commonly used scale is 0-255. This is obtained in R using maxColorValue=255. *alpha* is an optional argument for transparency, with the same intensity scale.
  rgb(r, g, b, maxColorValue=255, alpha=255)
- HSV (hue, saturation, value): values range from 0-1, with optional alpha argument
  hsv(h, s, v, alpha)
- HCL (hue, chroma, luminance): hue describes the color and ranges from 0-360; 0 = red, 120 = green, blue = 240, etc. Range of chroma and luminance depend on hue and each other
  hcl(h, c, l, alpha)

A few notes on HSV/HLC
HSV is a better model for how humans perceive color.

R Color Palettes
This is for all of you who don't know anything about color theory, and don't care but want some nice colors on your map or figure….NOW!

TIP: When it comes to selecting a color palette, **DO NOT** try to handpick individual colors! You will waste a lot of time and the result will probably not be all that great. R has some good packages for color palettes. Here are some of the options

Packages: grDevices and colorRamps

grDevices comes with the base installation and colorRamps must be installed. Each palette's function has an argument for the number of colors and transparency (*alpha*):

grDevices palettes
cm.colors
topo.colors
terrain.colors
heat.colors
rainbow
see P. 4 for options

heat.colors(4, alpha=1)
> #FF0000FF" "#FF8000FF" "#FFFF00FF" "#FFFF80FF"

For the rainbow palette you can also select start/end color (red = 0, yellow = *1/6*, green = *2/6*, cyan = *3/6*, blue = *4/6* and magenta = *5/6*) and saturation (s) and value (v):
rainbow(n, s = 1, v = 1, start = 0, end = max(1, n - 1)/n, alpha = 1)

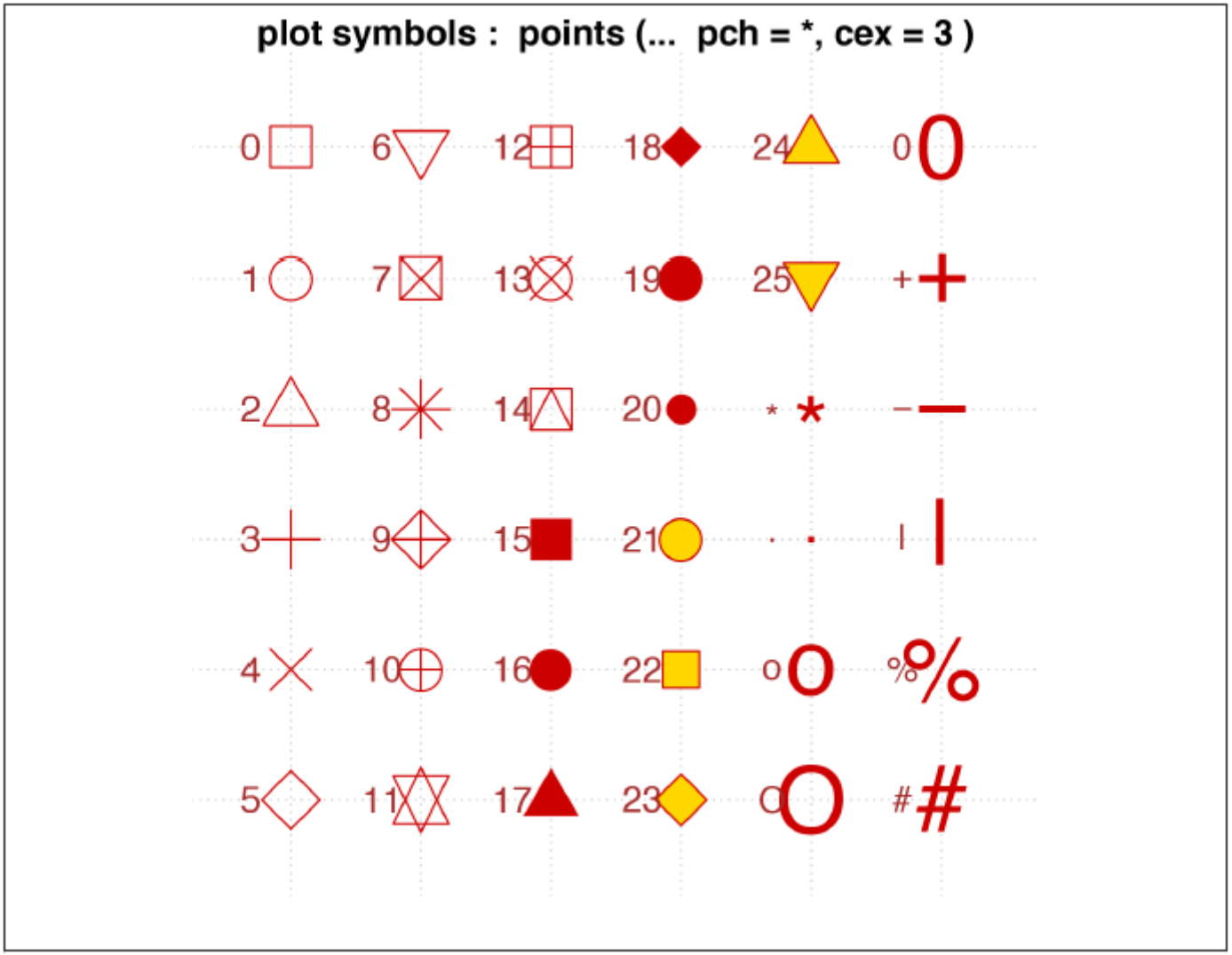
Package: RcolorBrewer

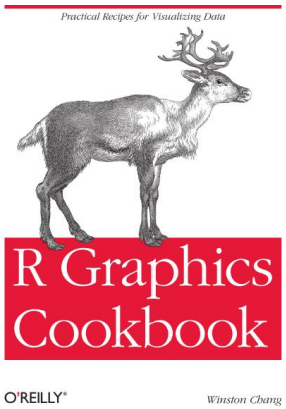
This function has an argument for the number of colors and the color palette (see P. 4 for options).
brewer.pal(4, "Set3")
> "#8DD3C7" "#FFFFB3" "#BEBADA" "#FB8072"

To view colorbrewer palettes in R: display.brewer.all(5)
There is also a very nice interactive viewer:
http://colorbrewer2.org/

Kingston University London

Base R: Shapes





Kingston University London

Advanced Graphics Example

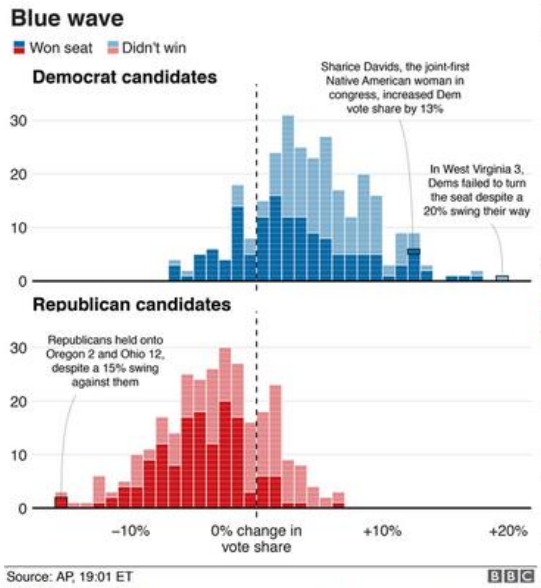
BBC Visual and Data Journalism cookbook for R graphics

Last updated: 2019-01-24

How to create BBC style graphics

At the BBC data team, we have developed an R package and an R cookbook to make the process of creating publication-ready graphics in our in-house style using R's ggplot2 library a more reproducible process, as well as making it easier for people new to R to create graphics.

The cookbook below should hopefully help anyone who wants to make graphics like these:



We'll get to how you can put together the various elements of these graphics, but **let's get the admin out of the way first...**

Next

R for Inferential Statistics

I. Categorical Data Analysis

Kingston University London