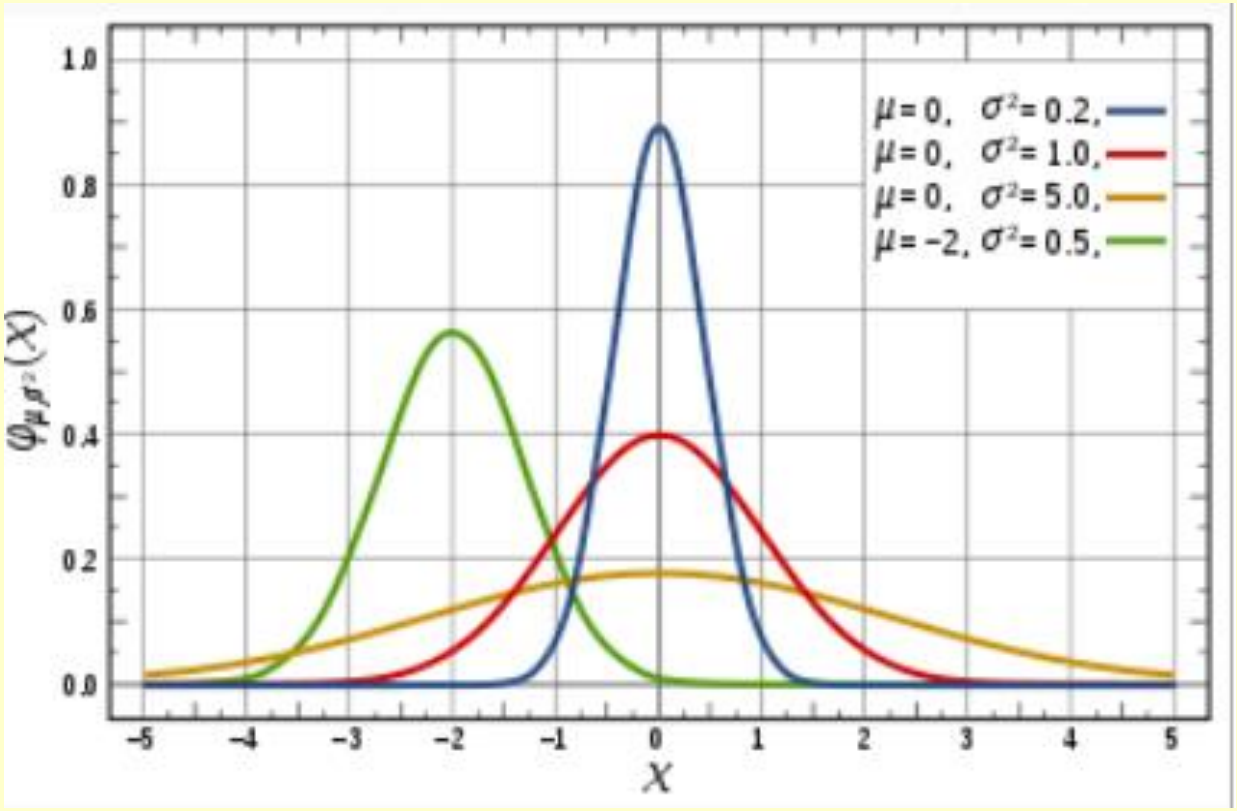
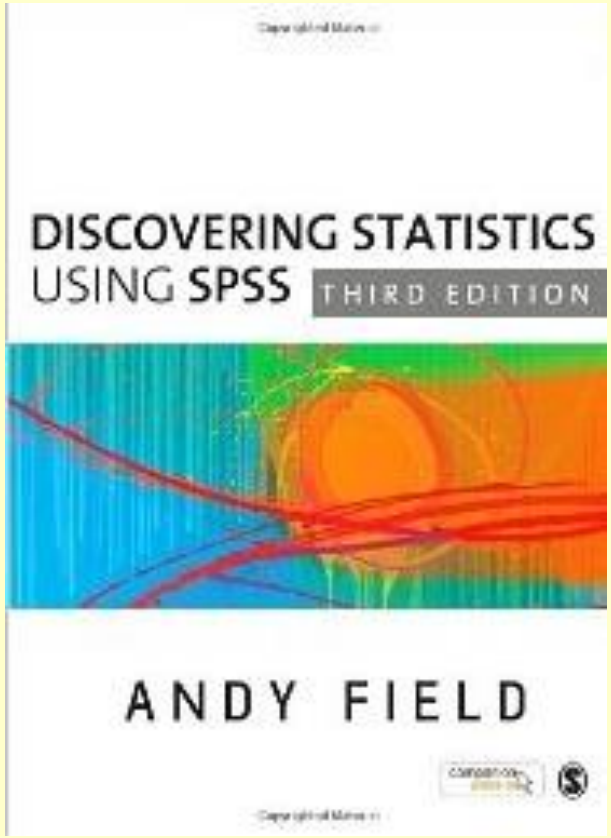
Parametric Statistics: Exploring Assumptions



http://www.pelagicos.net/classes_biometry_fa17.htm

Reading - Field: Chapter 5



R Packages Used in This Chapter

For this chapter, you will use the following packages:

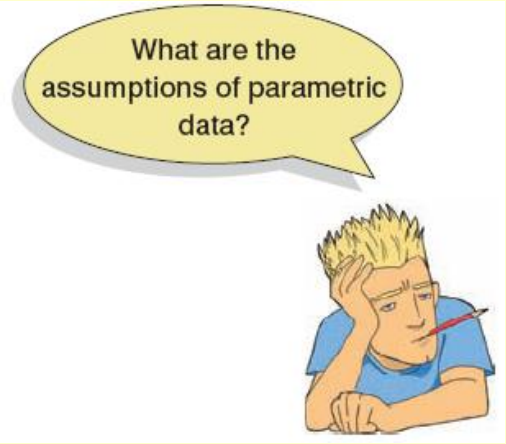
Start Rcmdr

```
install.packages("car");
install.packages("ggplot2");
install.packages("pastecs");
install.packages("psych");

library(car);
library(ggplot2);
library(pastecs);
library(psych);
```

NOTE: red font indicates RCmdr dependencies

Exploring Assumptions



- Assumptions of parametric tests based on the normal distribution
- Aim of this chapter:
- Quantify the assumption of normality
  - Graphical displays
  - Skew
  - Kurtosis
  - Normality tests
- Quantify the homogeneity of variances (when dealing with 2 or more samples): Levene's test

Assessing Normality

- We do not have access to sample the entire biological population, so we test observed data
- 1) Central Limit Theorem
  - If $N < 25$, sampling distribution rarely normal
- 2) Graphical Displays
  - Histogram
  - Q-Q plot
- 3) Skewness / Kurtosis (point estimate +/- SE)
  - Do they overlap with 0 ? (normal distribution)

Assessing Normality

4) Performing Statistical Tests

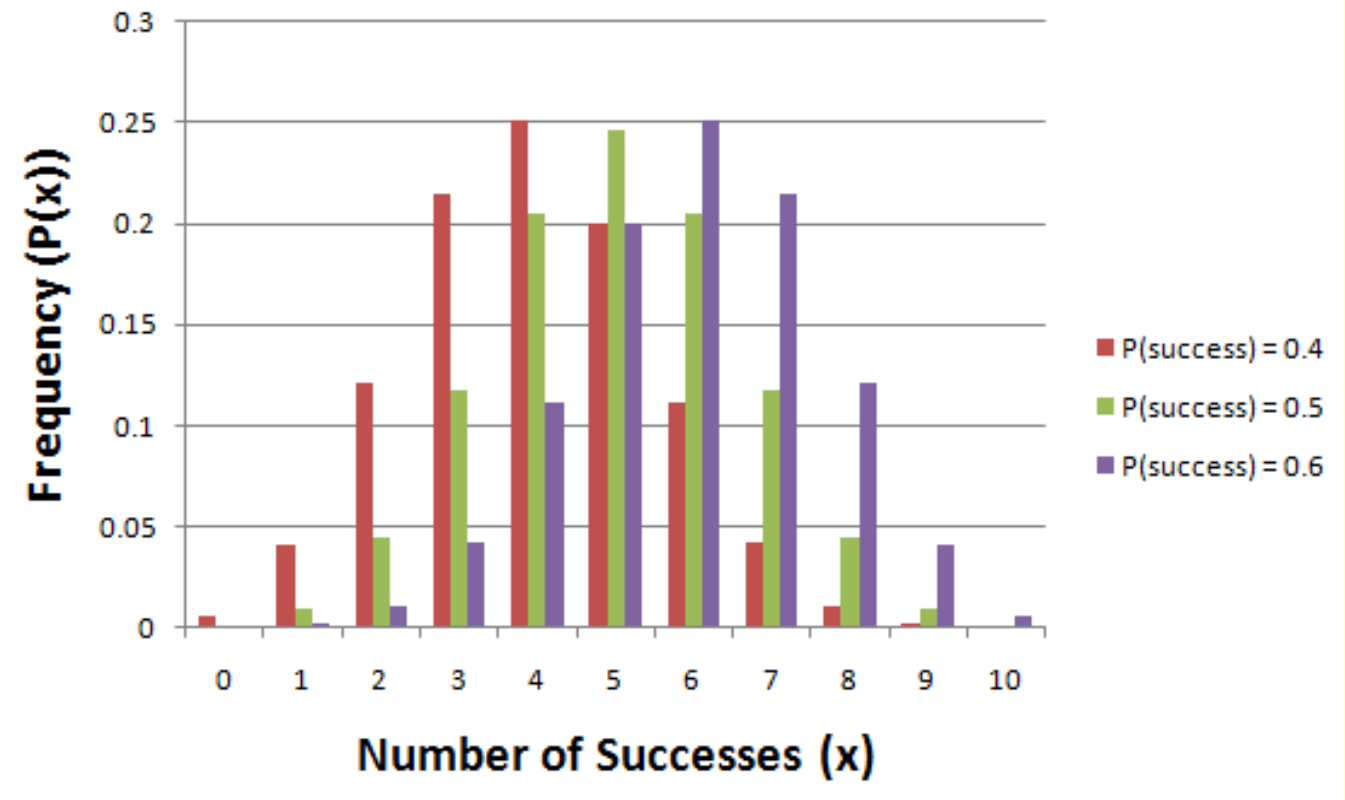
- Shapiro – Wilk Test
  - –Tests if data differ from a normal distribution
  - –Significant = non-Normal data
  - –Non-Significant = Normal data
- Levene's Test (comparing 2 or more samples)

- Tests if the data distributions have equal variances
  - –Significant = different variances
  - –Non-Significant = equal variances

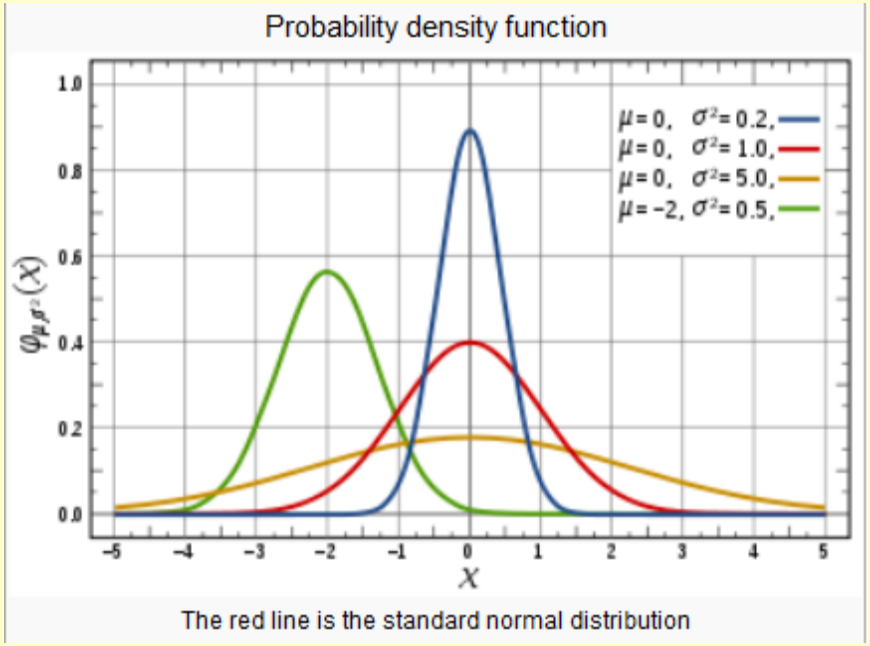
Assessing Normality - Graphically

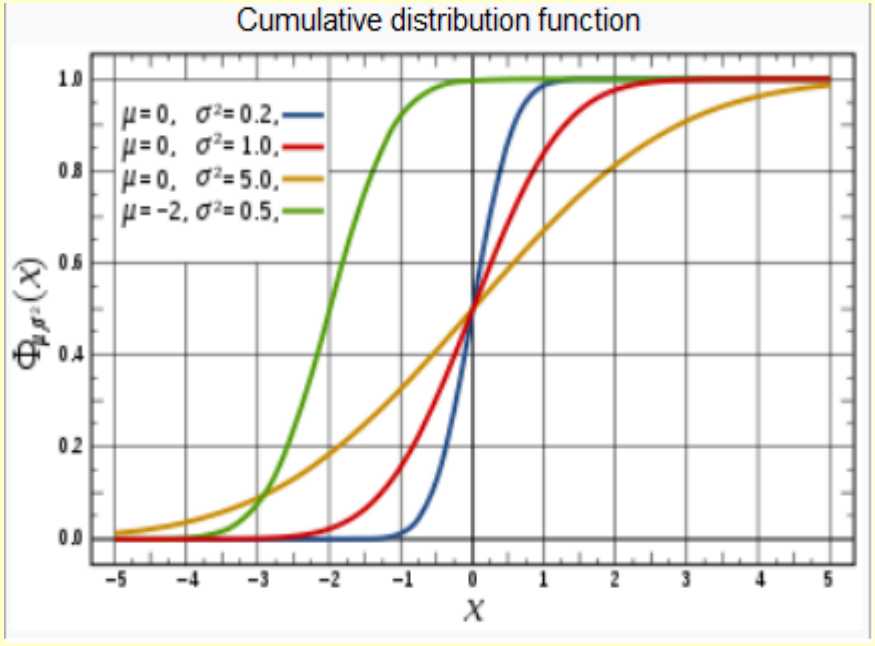
Characteristics of Normal Distributions

Unimodal, Symmetrical, Bell-shaped



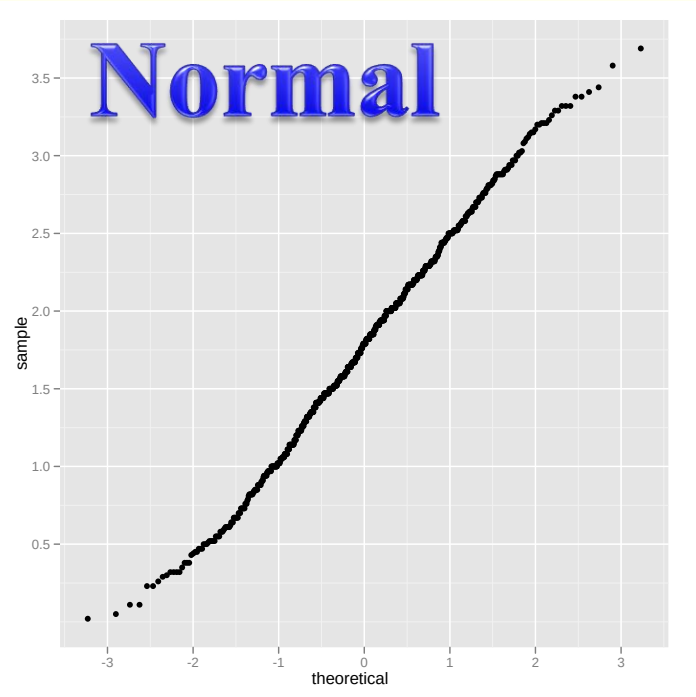
Assessing Normality - Graphically

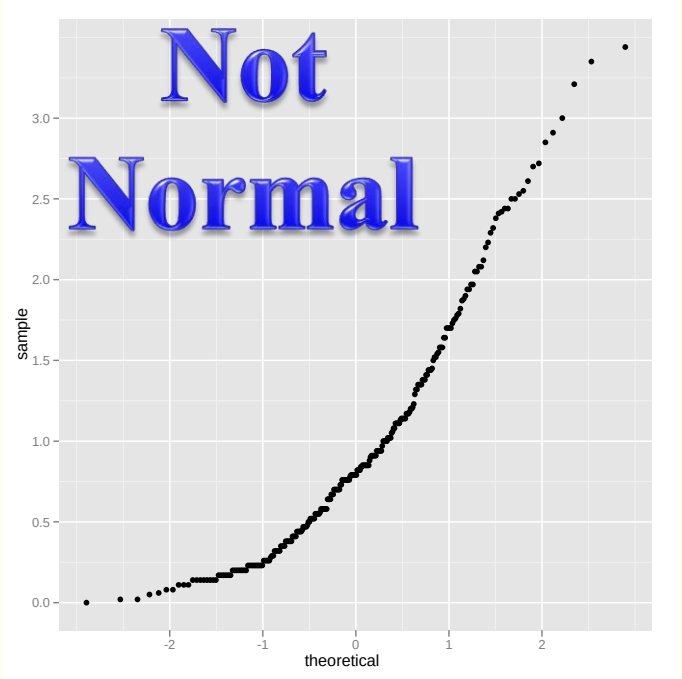




Comparing observations against a cumulative normal distribution (same mean and S.D.)

Assessing Normality - Graphically





A percentile is the proportion of cases (observations) that fall below a certain value.

Each observed percentile compared to the percentile that the value would have in a normal distribution.

Example: Festival Data Set

Biologist worried about potential health effects of music festivals. Measured hygiene of 810 concert-goers over the three days of a music festival.

Hygiene measured using standardized index (from 0 to 4):

0 = you smell terribly 4 = you smell beautifully

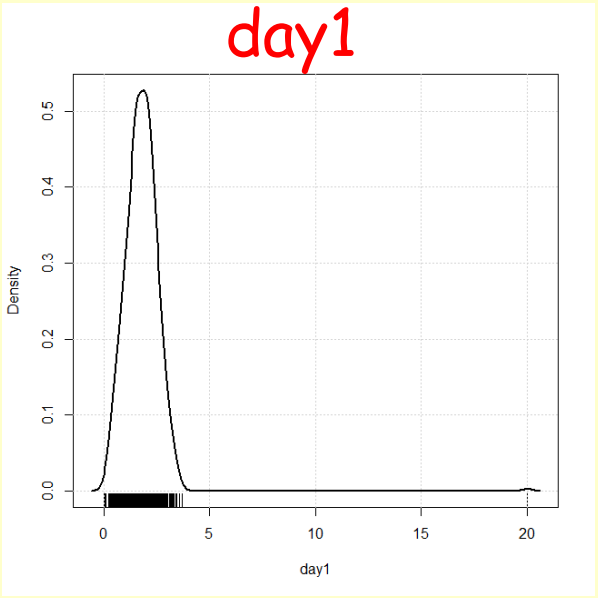
Import Download Festival Data (DownloadFestival.xlsx)

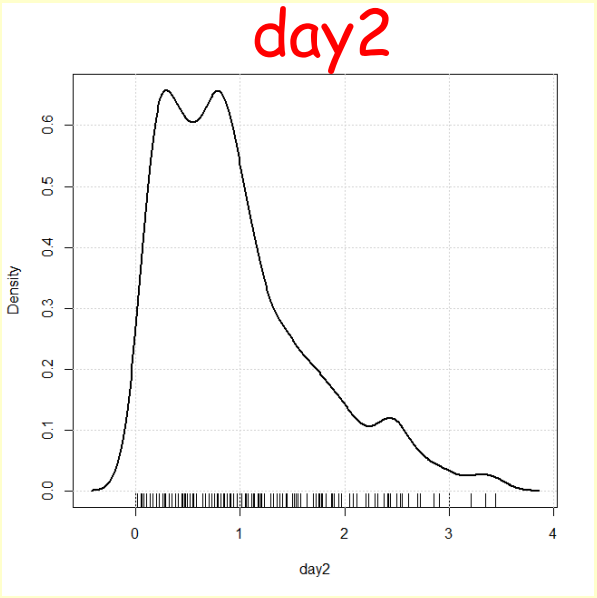
For ease of use, rename the Data Set "Festival"

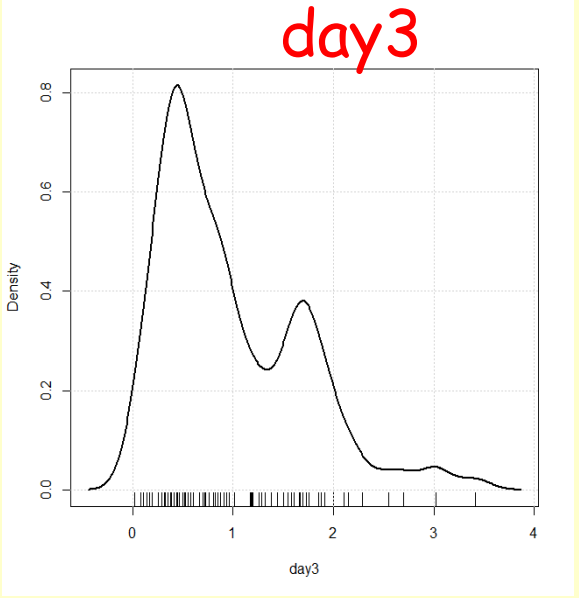
```
> Festival <- DownloadFestival
```

Explore Data Graphically: RCmdr

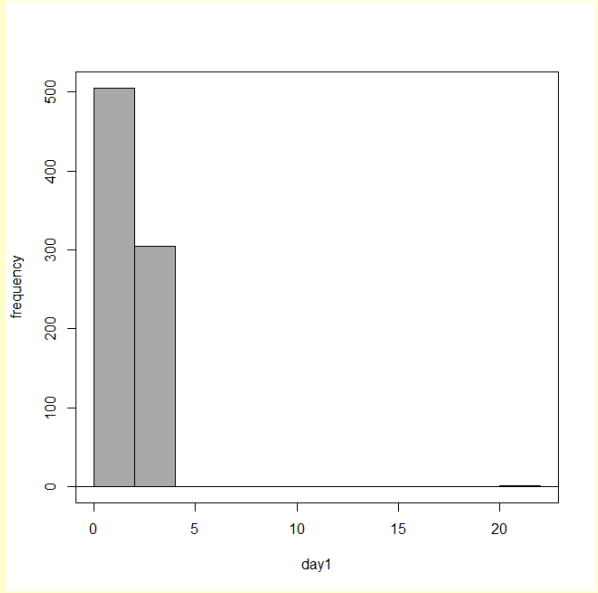
density

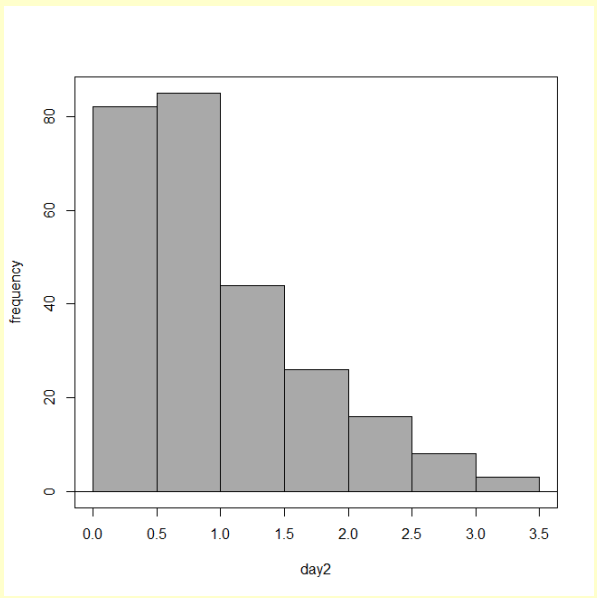


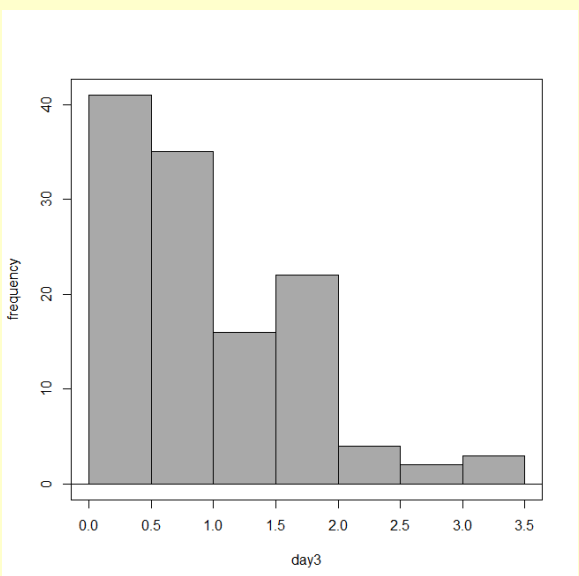




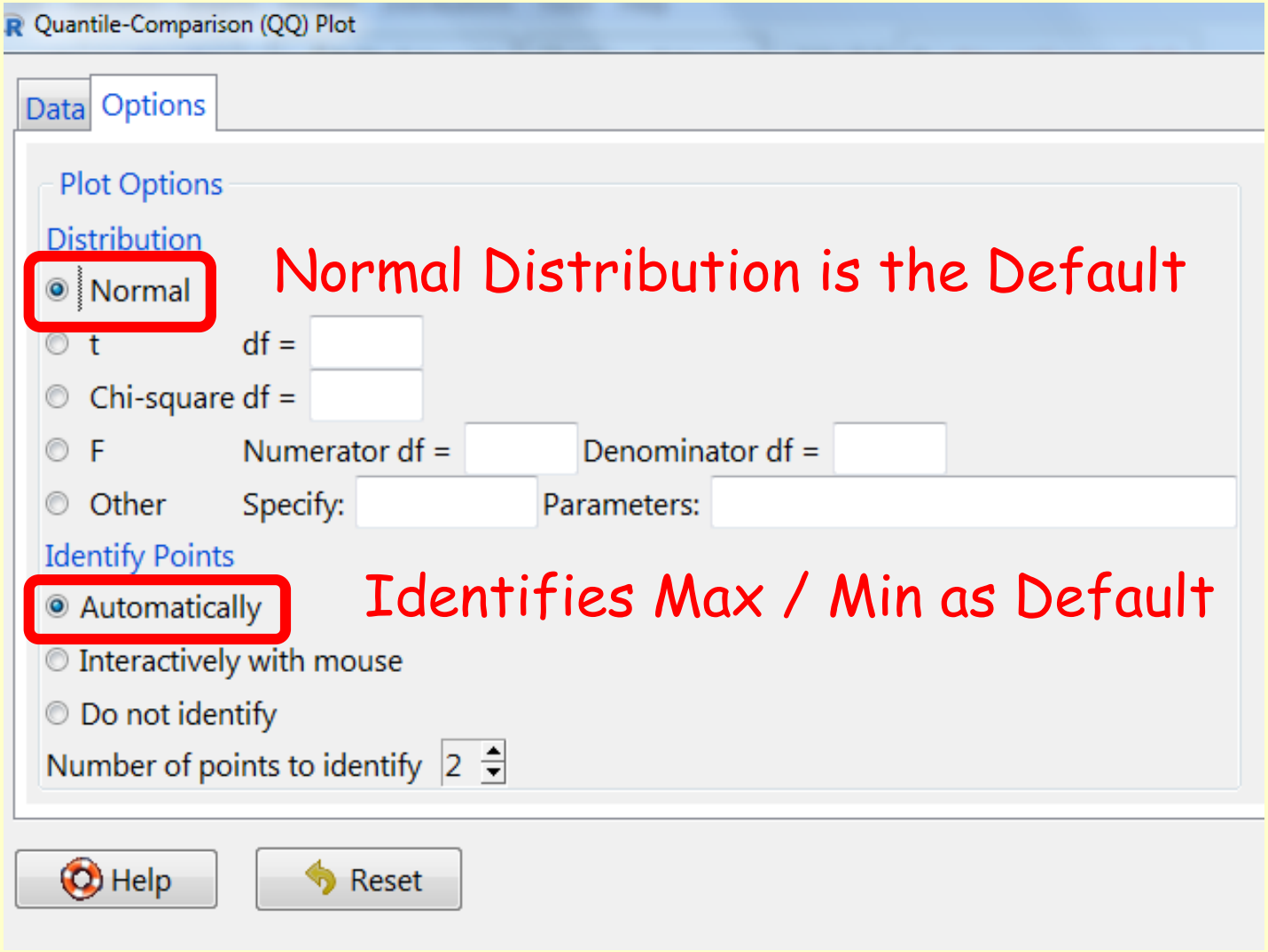
histogram







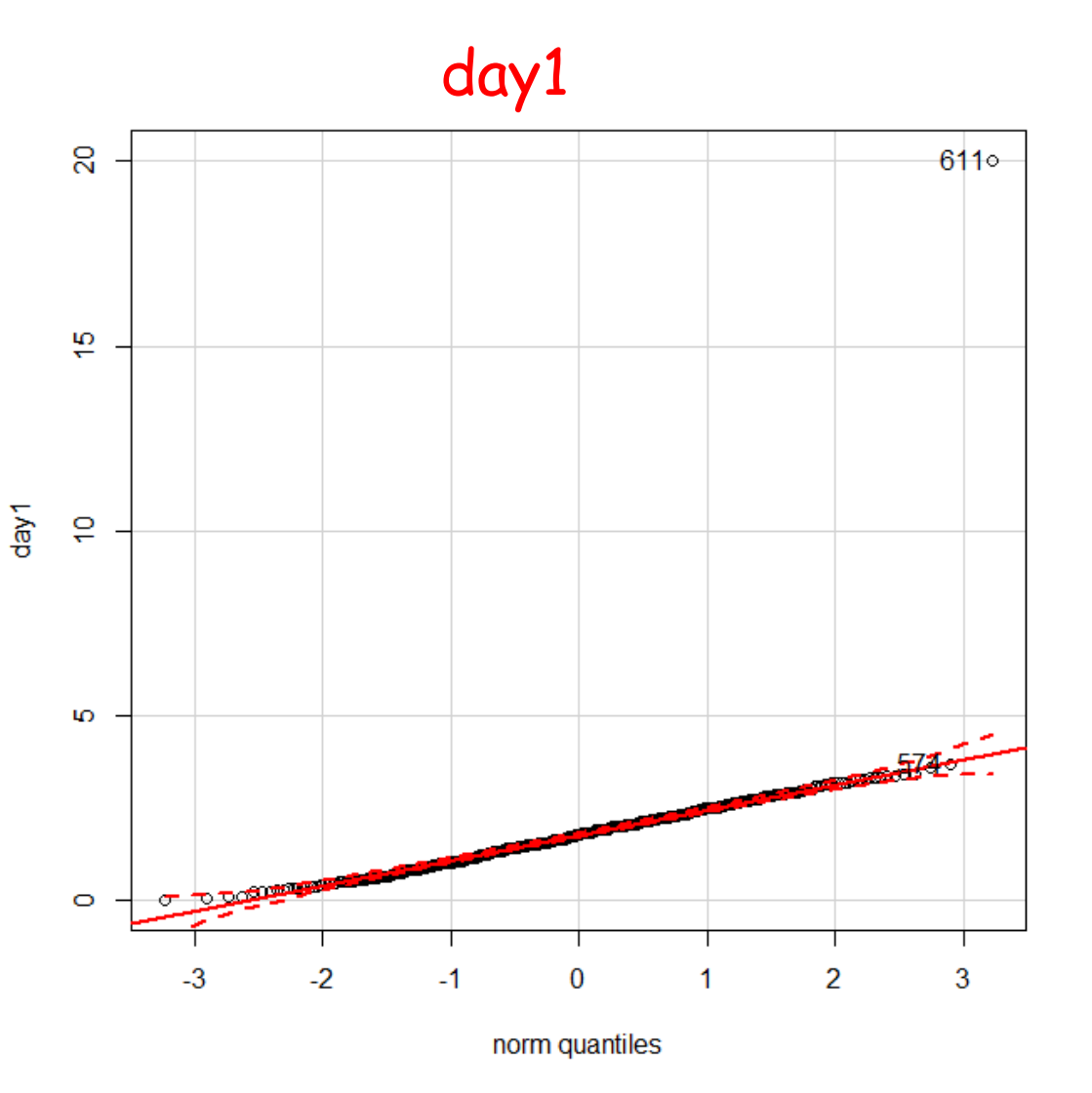
Graphs in Rcmdr – Quantiles



Graphically compares an observed (empirical) distribution (points) with a chosen theoretical expectation (line)

Identify Points: Automatic or Interactively

Graphs in Rcmdr – Quantiles

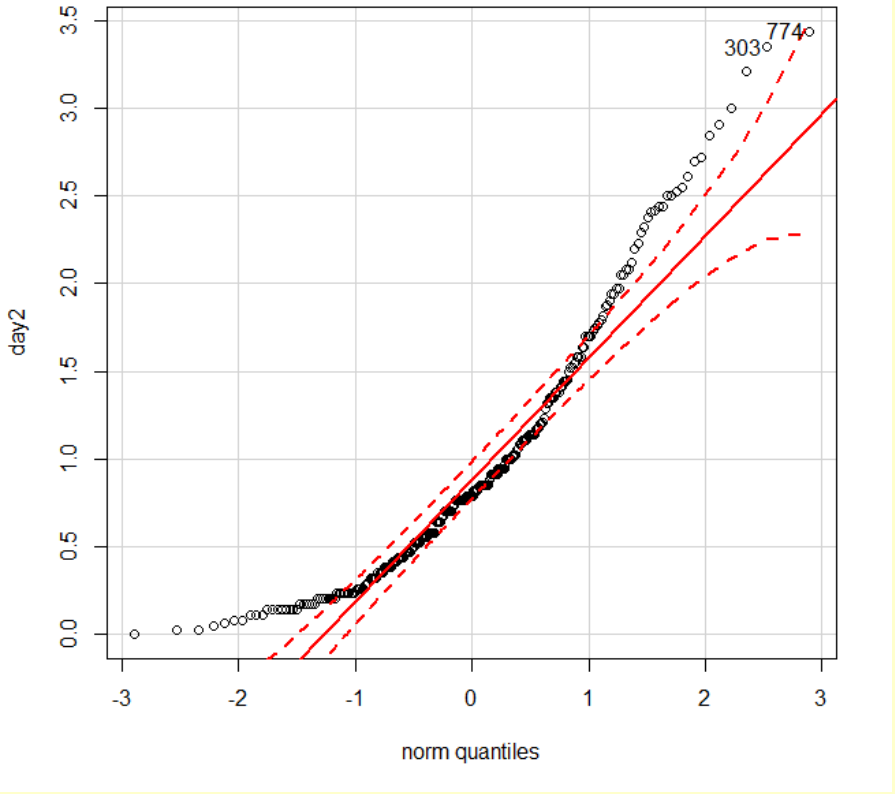


The solid red line is the expected pattern a normal distribution with the same mean and SD and the sampled data.

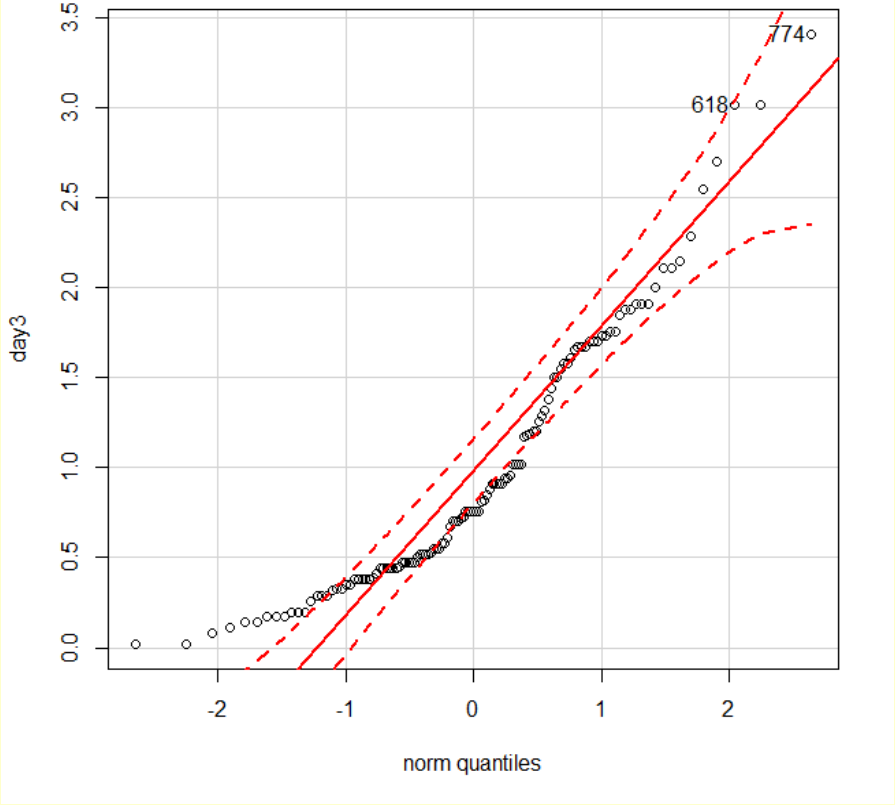
Points outside of the dashed line envelope suggest significant deviations

Graphs in Rcmdr – Quantiles

day 2



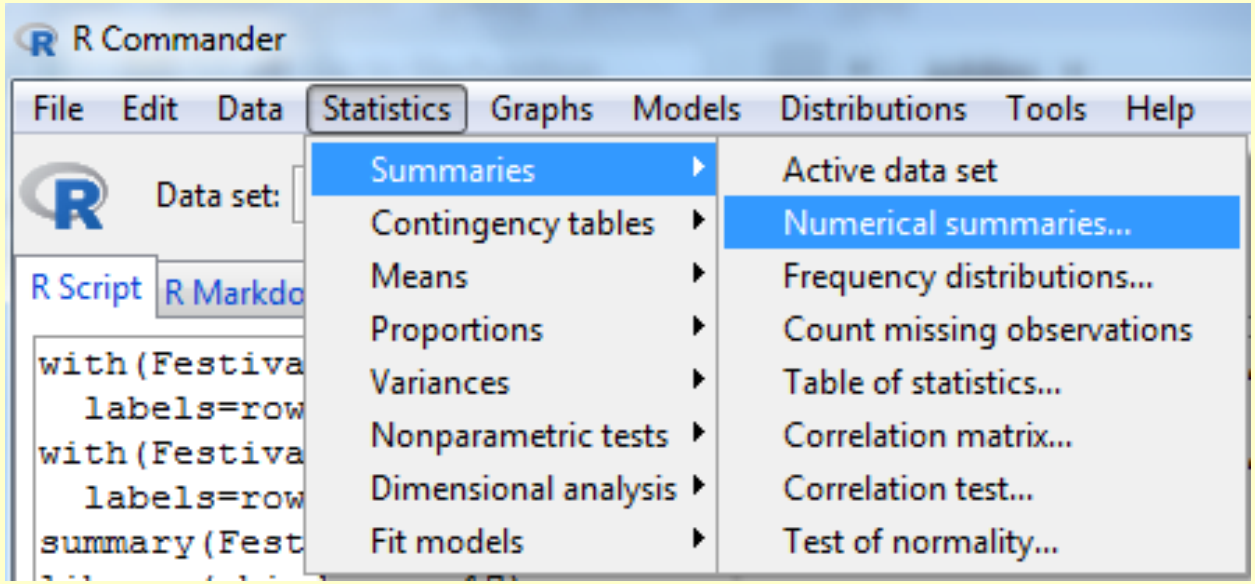
day 3

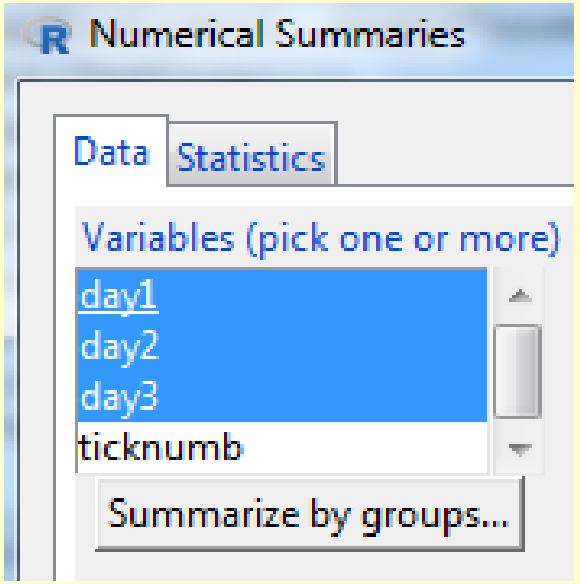


Note: The straight line represents the expected pattern for a normal distribution

Explore Festival Data Set

We can also explore the summary statistics describing the three datasets (day1, day2, day3) using RCmdr:

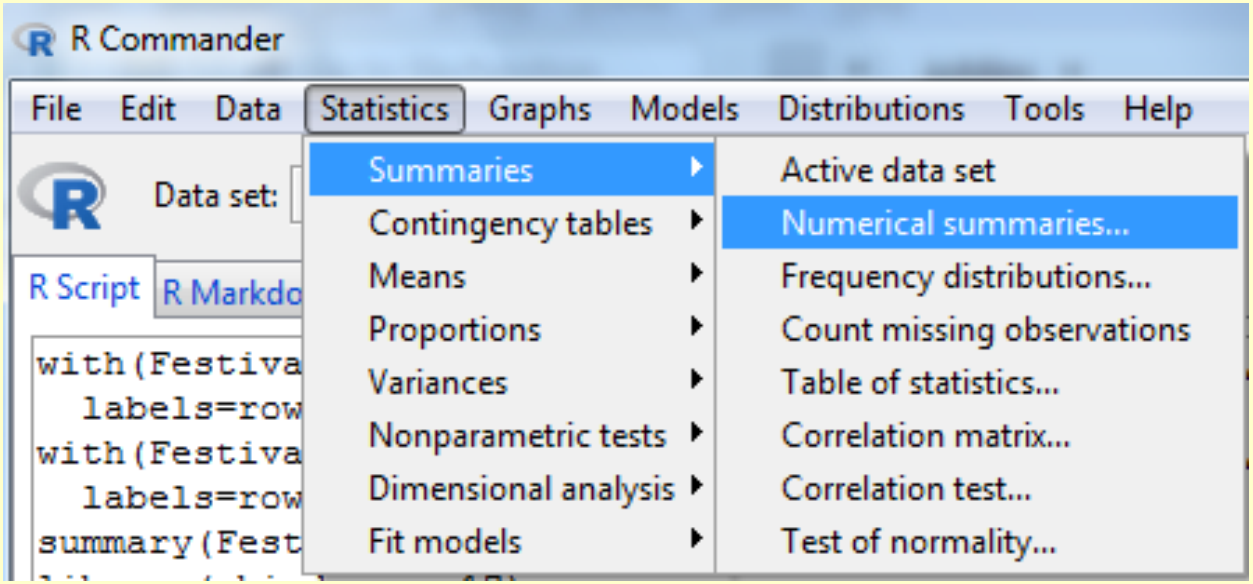


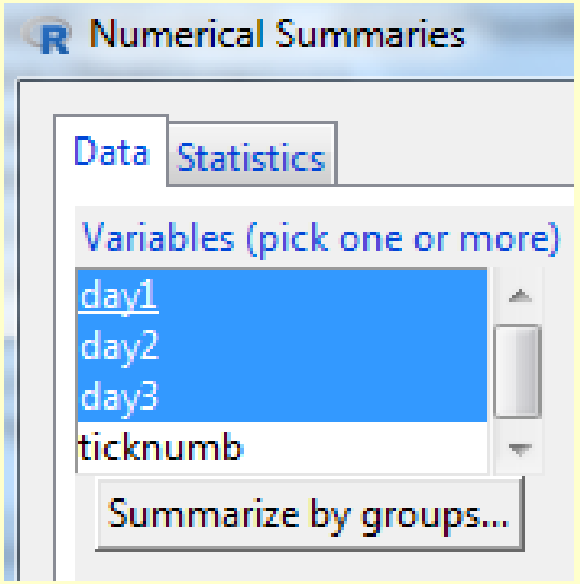



```
> numSummary(Festival[,c("day1", "day2", "day3"), drop=FALSE],
  statistics=c("mean", "sd", "IQR", "quantiles", "skewness", "kurtosis"),
  quantiles=c(0,.25,.5,.75,1), type="2")
```

Explore Festival Data Set

We can also explore the summary statistics describing the three datasets (day1, day2, day3) using RCmdr:





NOTE: multiple datasets can be analyzed at once

What statistics would you use to assess data normality?

Explore Festival Data Set

Exploring the summary statistics describing the three datasets (day1, day2, day3) using RCmdr:

```
> numSummary(Festival[,c("day1", "day2", "day3"),
    drop=FALSE],
    statistics=c("mean", "quantiles", "skewness", "kurtosis"),
    quantiles=c(.5), type="2")
```

	mean	skewness	kurtosis	50%	n	NA
day1	1.7933580	8.865312	170.4502658	1.79	810	0
day2	0.9609091	1.095226	0.8222057	0.79	264	546
day3	0.9765041	1.032868	0.7315003	0.76	123	687

Further Explore Festival Data Set

Exploring additional datasets using other functions:

describe() function in psych package

```
> describe(Festival$day1)
```

vars	n	mean	sd	median	skew	kurtosis
1	810	1.79	0.94	1.79	8.83	168.97

trimmed	mad	min	max	range	se
1.77	0.7	0.02	20.02	20	0.03

Further Explore Festival Data Set

Exploring additional datasets using other functions:

stat.desc() function in psych package

```
> stat.desc(Festival$day1, basic = FALSE, norm = TRUE)
```

basic argument:
Basic statistics included if TRUE
(Note: FALSE is the default)

norm argument:
Statistics relating to normal distribution included if TRUE
(Note: FALSE is the default)

Further Explore Festival Data Set

```
> stat.desc(Festival$day1, basic = FALSE, norm = TRUE)
```

```
    median          mean
1.790000e+00  1.793358e+00

   SE.mean  C.I.mean.0.95
3.318617e-02  6.514115e-02

       var       std.dev
8.920705e-01  9.444949e-01

coef.var
5.266627e-01
```

Further Explore Festival Data Set

```
> stat.desc(Festival$day1, basic = FALSE, norm = TRUE)
```

```
   skewness        skew.2SE
8.832504e+00   5.140707e+01

   kurtosis        kurt.2SE
1.689671e+02   4.923139e+02
```

skew.2SE:
Skew divided by 2 SE

kurtosis.2SE:
Kurtosis divided by 2 SE

- How can we interpret these results?

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Z= (observed value – theoretical value) / (SE of value)

Further Explore Festival Data Set

```
   skewness       skew.2SE
8.832504e+00  5.140707e+01

   kurtosis       kurt.2SE
1.689671e+02  4.923139e+02
```

skew.2SE:
Skew divided by 2 SE

kurtosis.2SE:
Kurtosis divided by 2 SE

What values are needed to have a significant skew / kurtosis significant?

(Different from 0)

Further Explore Festival Data Set

skew.2SE = 5.14
(observed skew) / 2 SE

kurtosis.2SE = 492
(observed skew) / 2 SE

Are skew / kurtosis significant?
(Different from 0)

YES

Rules of thumb to assess significance:

skew.2SE kurtosis.2SE	P value
ABS > 0.98	< 0.05
ABS > 1	< 0.04
ABS > 1.29	< 0.01
ABS > 1.65	< 0.001

Testing Data Normality

```
> stat.desc(Festival$day1, basic = FALSE, norm = TRUE)
```

NOTE:

Because norm argument set to TRUE, stat.desc provided normality test

```
 normtest.W
6.539142e-01

 normtest.p
1.545986e-37
```

Test Statistic

P value

Is this distribution different from a normal distribution ?

YES

How do I know that ?

P < 0.05

NOTE: Null Hypothesis is that data are normal

Testing Data Normality

```
> shapiro.test(Festival$day1)
```

```
Shapiro-Wilk normality test data: Festival$day1

W = 0.65391, p-value < 2.2e-16
```

Is this distribution different from a normal distribution ?

YES

How do I know that ?

P < 0.05

NOTE: Null Hypothesis is that data are normal

Testing Data Normality

```
Shapiro-Wilk normality test data: Festival$day2
W = 0.90832, p-value = 1.282e-11
```

```
Shapiro-Wilk normality test data: Festival$day3
W = 0.90775, p-value = 0.0000003804
```

Is day2 different from a normal distribution ?

How do I know that ?

YES ($P < 0.05$)

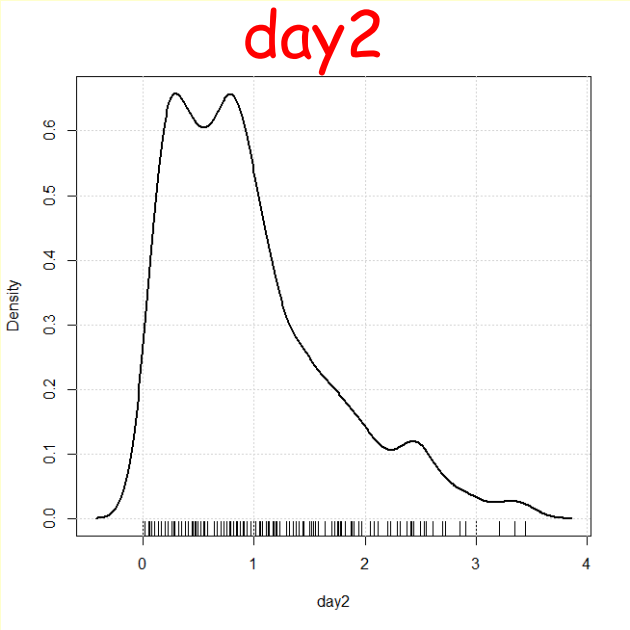
Is day3 different from a normal distribution ?

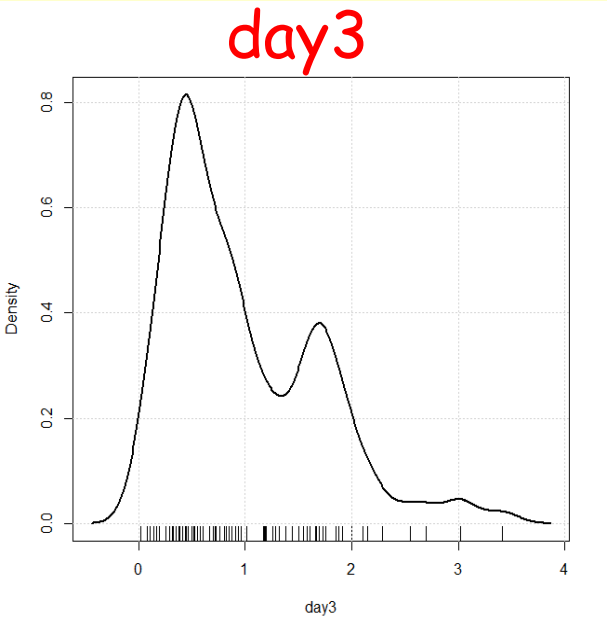
How do I know that ?

YES ($P < 0.05$)

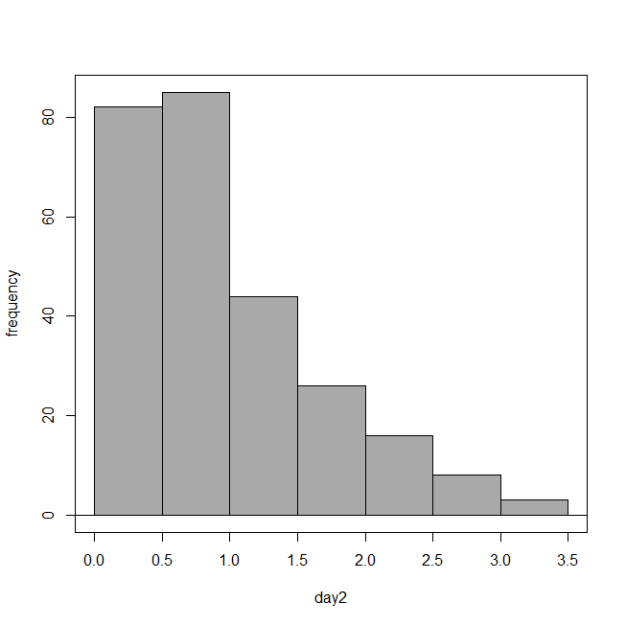
Graphical Data Exploration: RCmdr

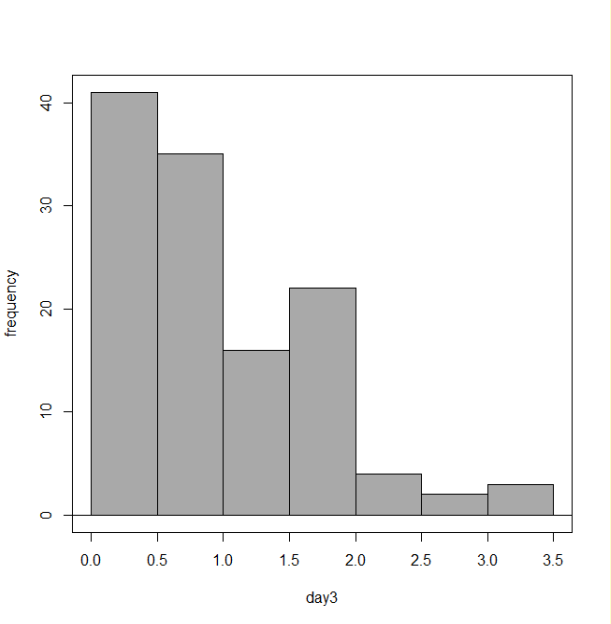
density





histogram





Diagnostics:

Lack of Symmetry

Long tails

Mean > Median

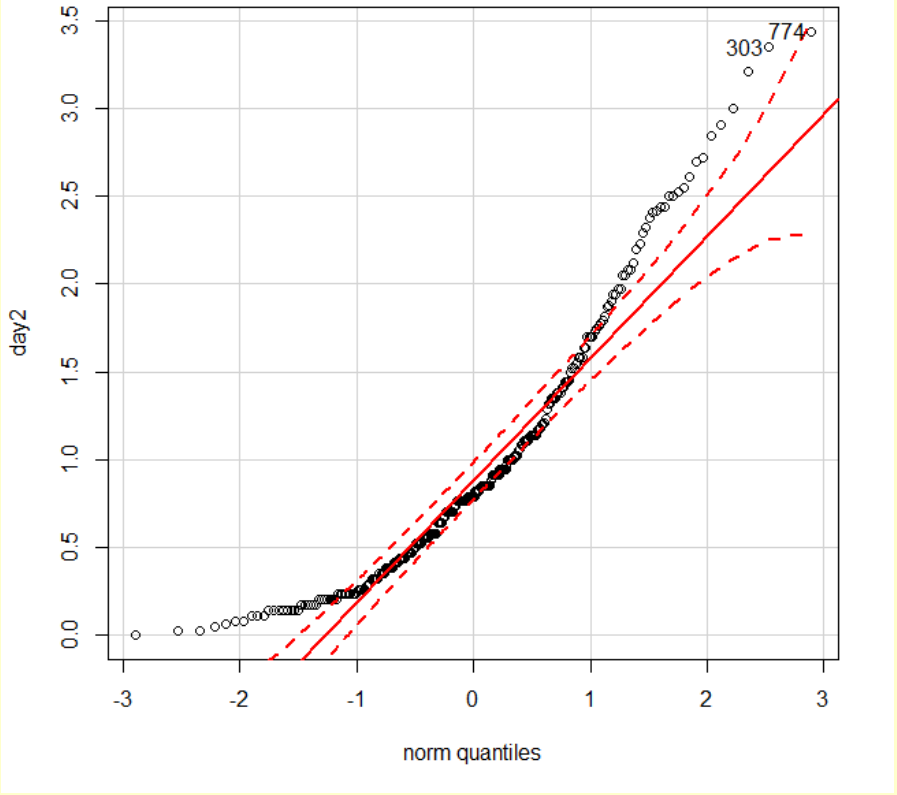
Positive Skew

Positive kurtosis

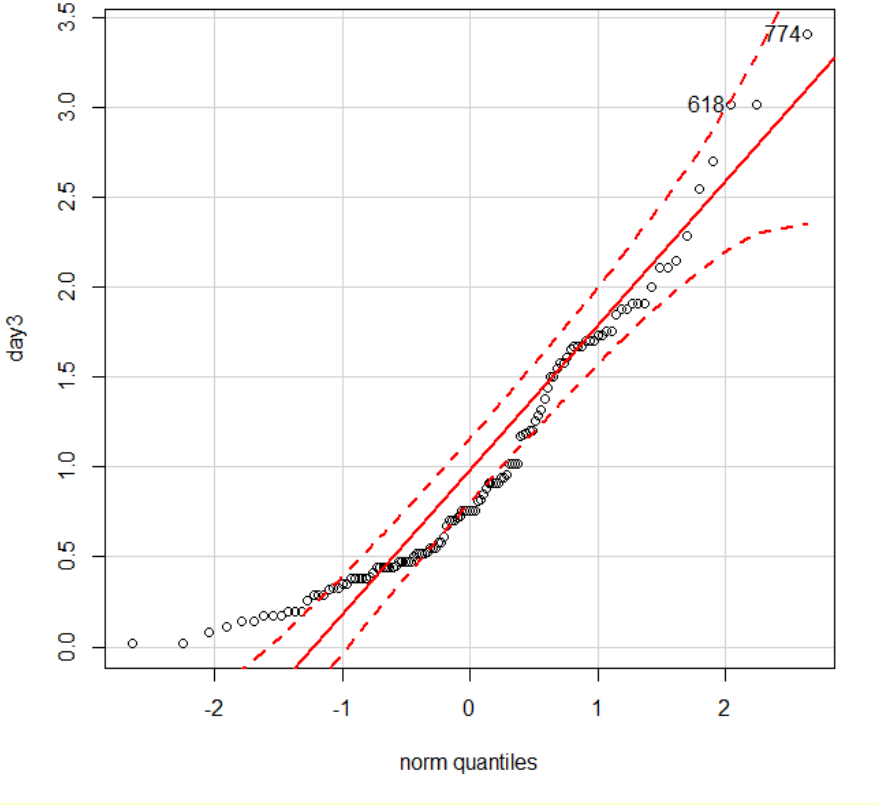
Summary Statistics & Quantiles

	mean	skewness	kurtosis	50%	n
day2	0.9609091	1.095226	0.8222057	0.79	264
day3	0.9765041	1.032868	0.7315003	0.76	123

day 2



day 3

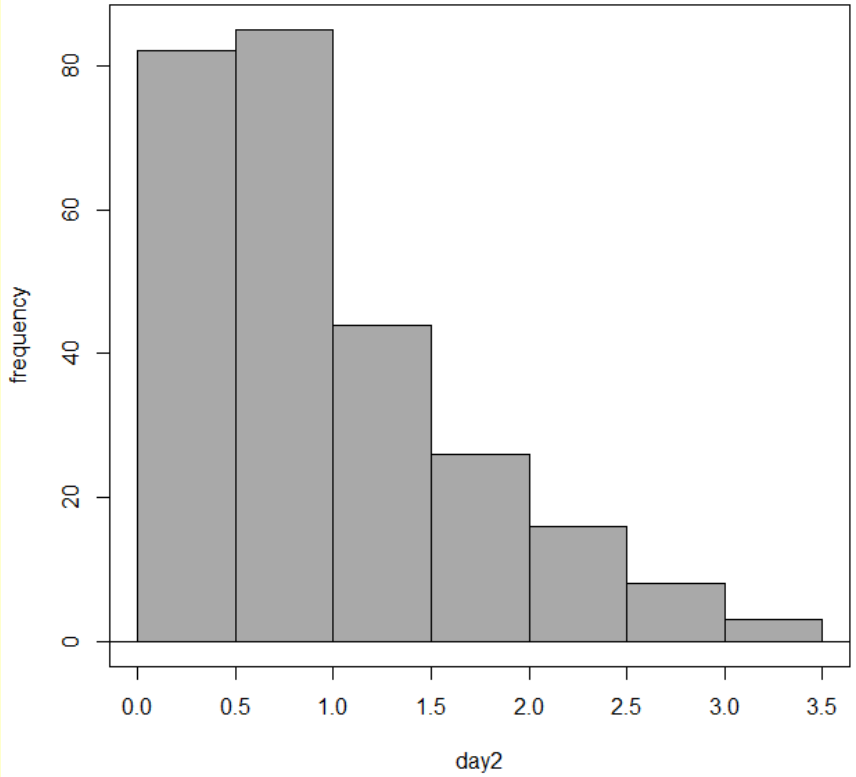


Rule of Thumb (Z scores)

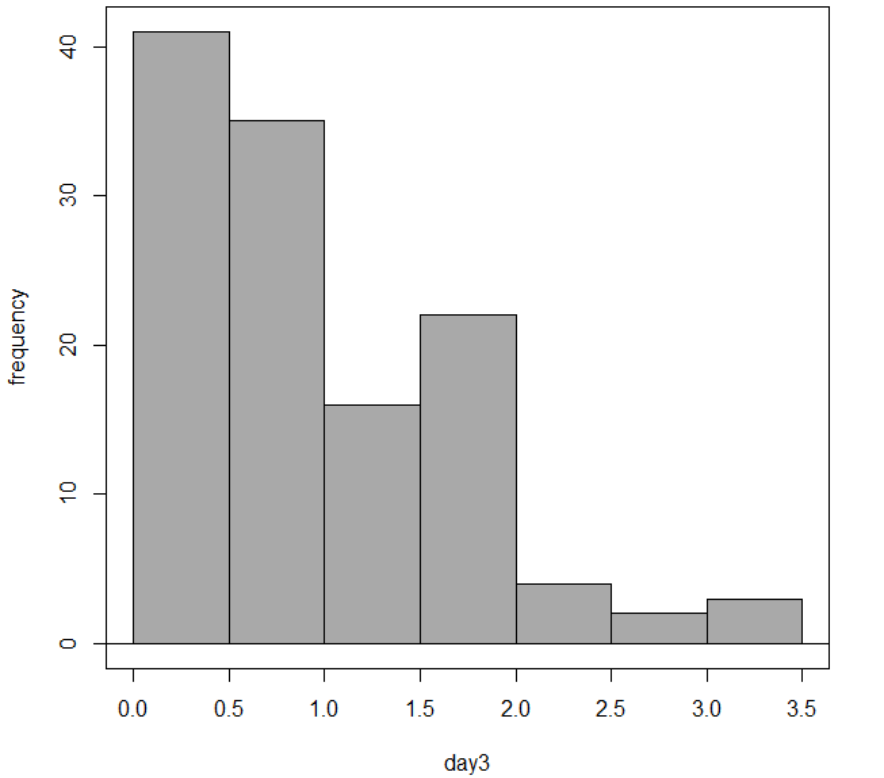
	skewness2.SE	kurtosis.2SE
Day2	3.612	1.265
Day3	2.309	0.686

Significant Results

day 2



day 3



Summary: Normality

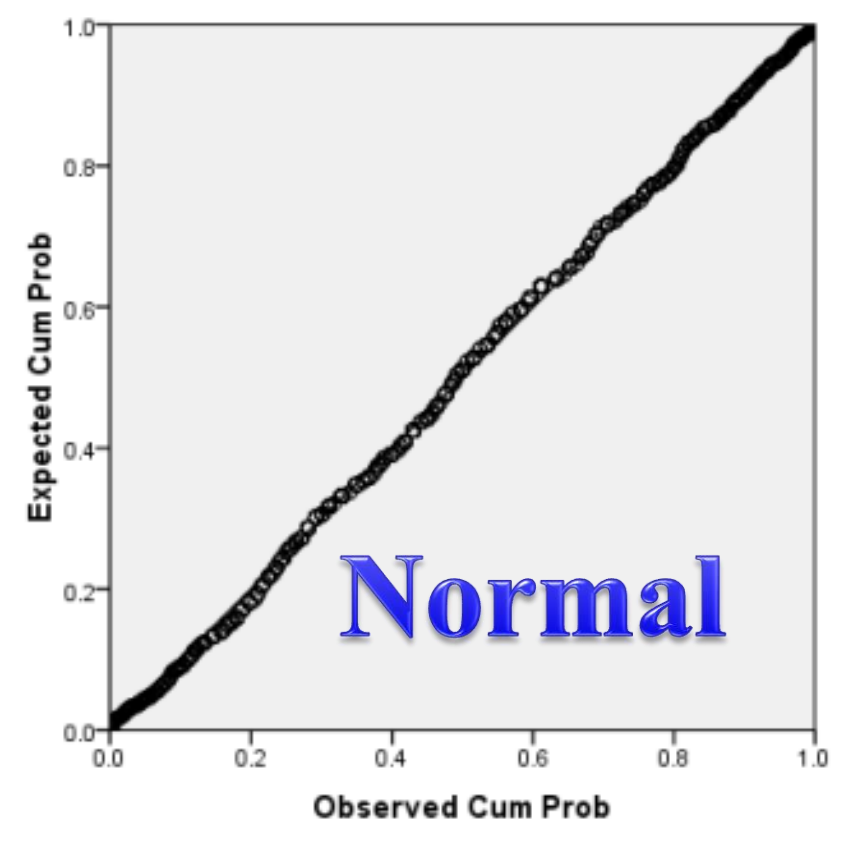
Indicators of a normal (Gaussian) distribution

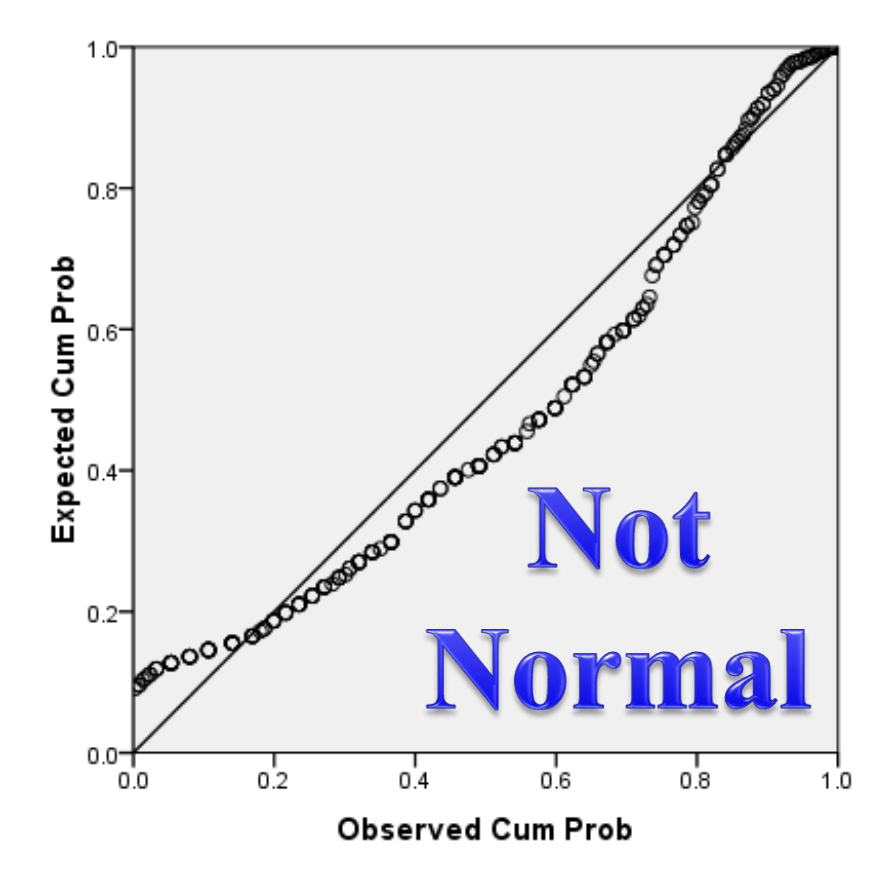
A. Mean = Median = Mode

B. Skewness: measures asymmetry of the distribution. A value of zero indicates symmetry. Symmetry is needed to be a normal distribution. The larger the absolute value the more skewed the distribution.

C. Kurtosis: measures the distribution of mass in the distribution. A value of zero indicates a normal distribution. The larger the absolute value the more distorted the distribution.

1. Assess Normality Graphically





Note: The straight line represents the expected pattern for a normal distribution

2. Assess Skew / Kurtosis

Calculate probability of observed skew / kurtosis, compared to expectation for normal distribution

Use "rule of thumb":

skew.2SE kurtosis.2SE	P value
ABS > 0.98	< 0.05
ABS > 1	< 0.04
ABS > 1.29	< 0.01
ABS > 1.65	< 0.001

3. Use Shapiro-Wilk (S-W) Test

Specific test developed to test null hypothesis that a given sample (x_1, ..., x_n) came from a normally distributed population.

Significant = non-Normal data

Non-Significant = Normal data

Shapiro, SS, Wilk, MB. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52: 591–611.

Summary

- Parametric tests based on normal distributions

- 3 ways of Checking the assumption of normality
  - Graphical displays: Q-Q plots
  - Skew & Kurtosis: Z scores
  - Normality test: S-W

- Next Lecture: When and how to correct problems in the distribution of the data

  - Data Transformations
  - Pitfalls and alternatives