Online Academic Data Analysis Bootcamp Using Open-Access Program R

Parametric Statistics: Exploring Assumptions

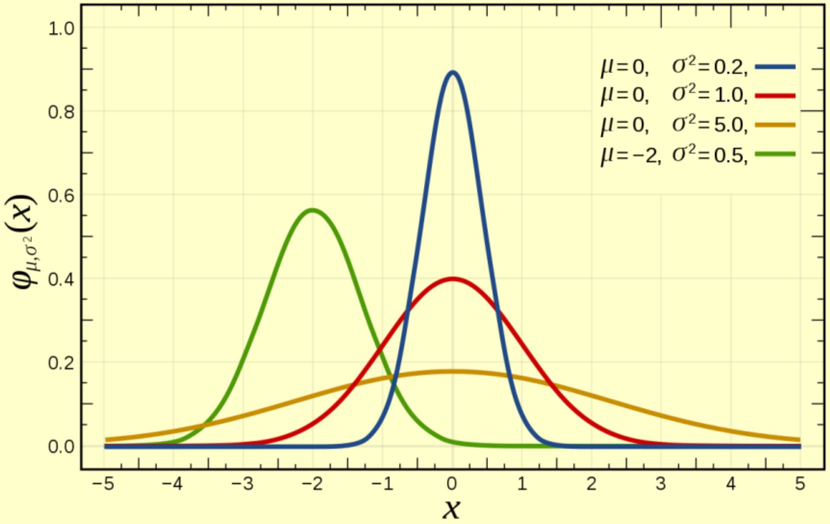
Patrick Njage (Ph.D)
Technical University of Denmark and
Academic Data Analysts (https://www.academicdataanalysts.org/)

Outline

- Quantify the assumption of normality
  - Graphical displays
  - Skew
  - Kurtosis
  - Normality tests
- Quantify the homogeneity of variances
  - Two-variance F-test: compares two samples
  - Bartlett's Test: compares two or more samples
  - Levene's Test: compares two or more samples
- When and how to correct problems in the distribution of the data
  - Data Transformations
  - Pitfalls and alternatives

The normal distribution

Data are ideally distributed symmetrically around the centre of all scores.
A vertical line through the centre of the distribution should look the same on both sides.
This is commonly known as a normal distribution and is characterized by the bell-shaped curve.



Assessing Normality

- We do not have access to sample the entire biological population, so we test the 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

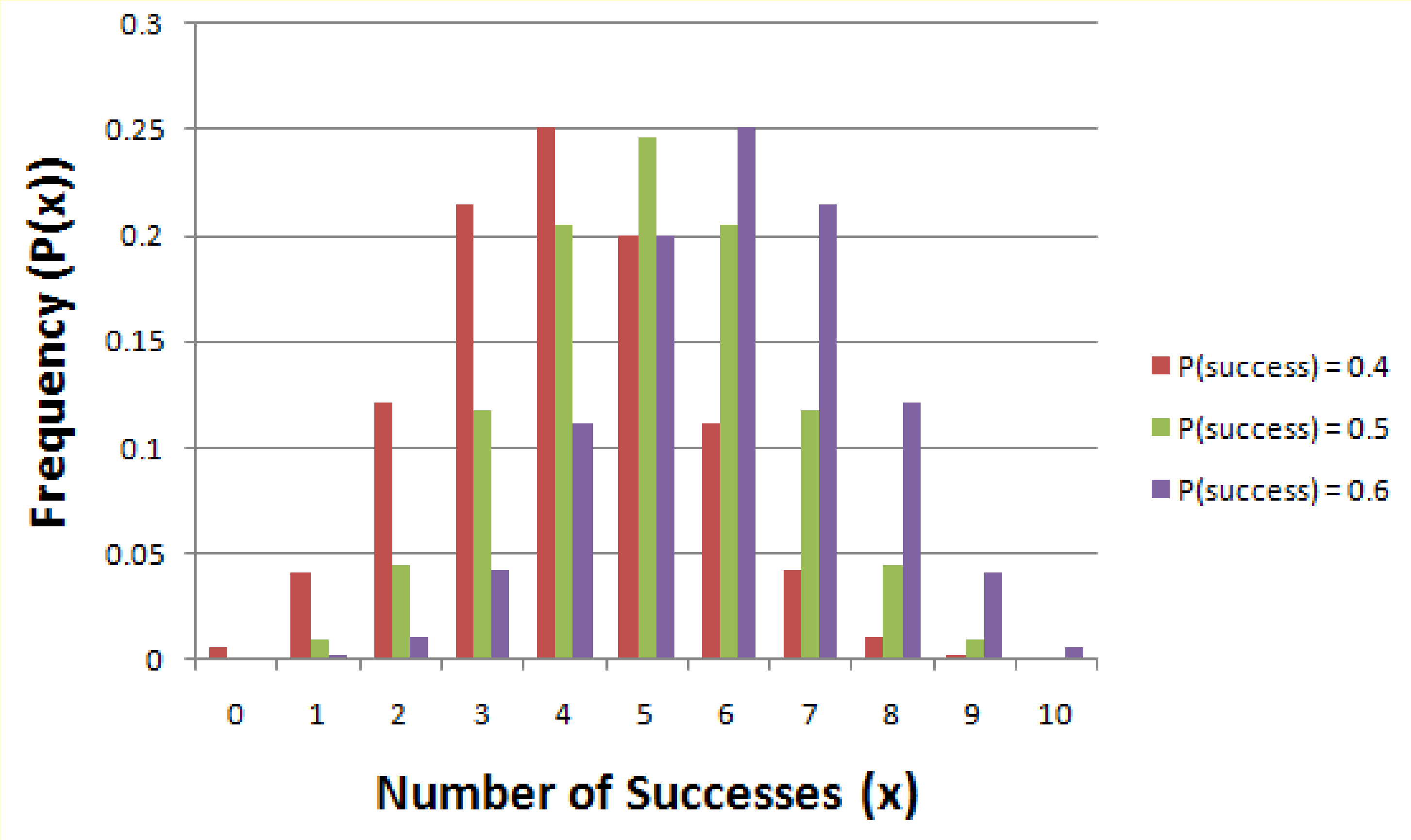
    Significant = non-Normal data

    Non-Significant = Normal data

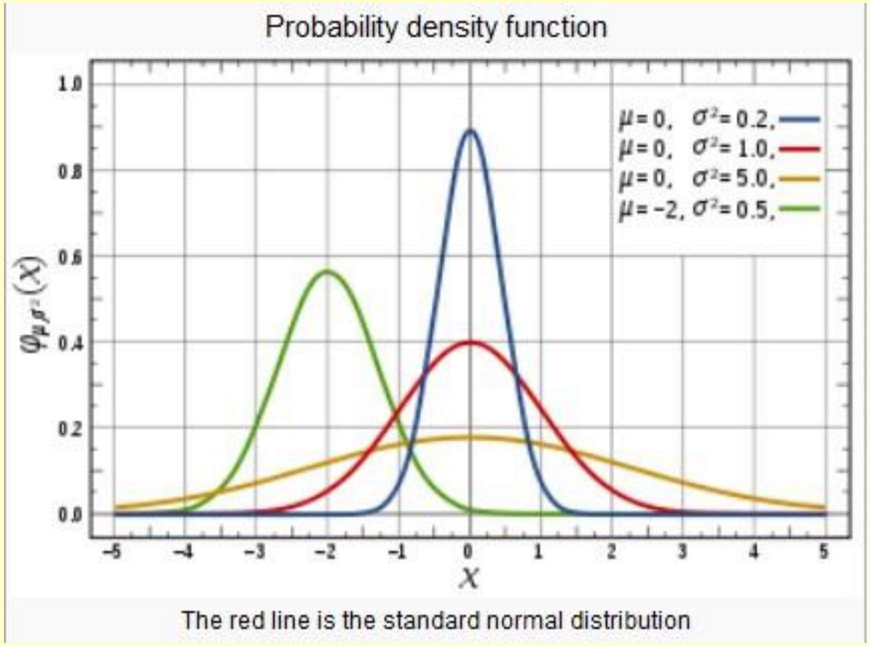
Assessing Normality - Graphically

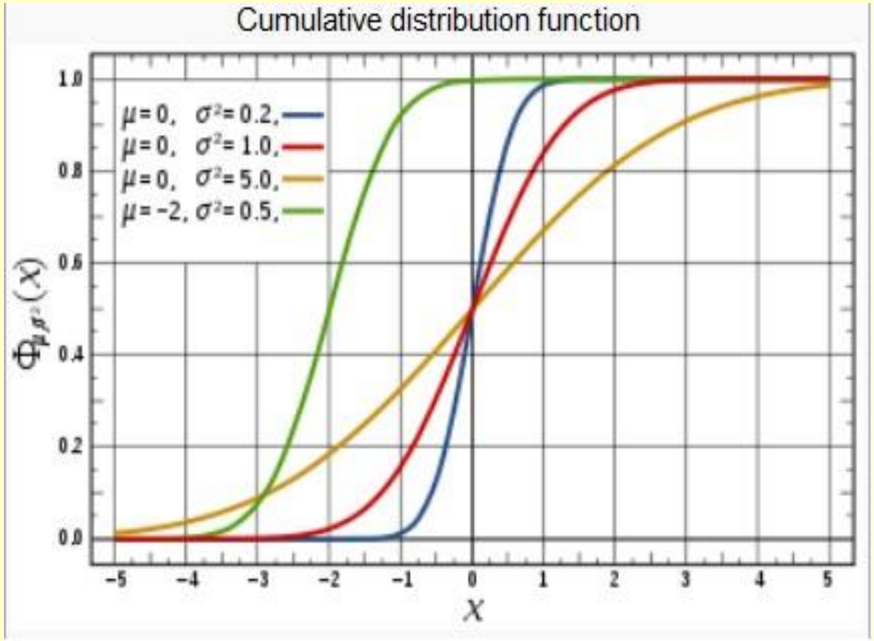
Characteristics of Normal Distributions

Unimodal, Symmetrical, Bell-shaped



Assessing Normality - Graphically

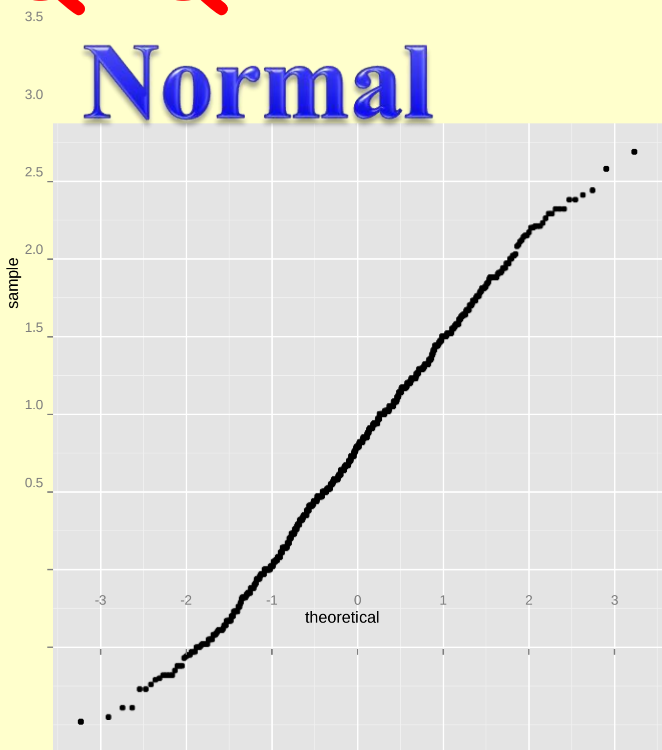


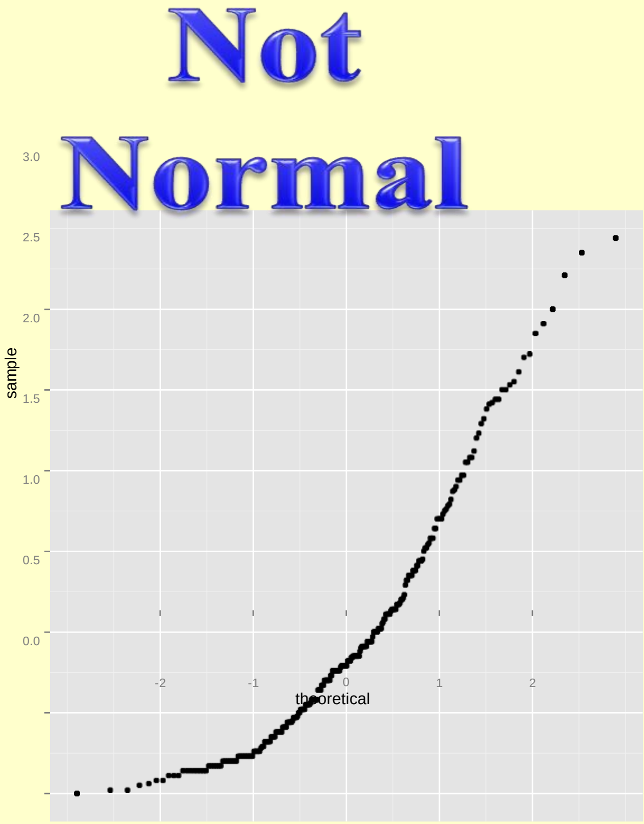


Comparing observations against a cumulative normal distribution (same mean and S.D.)

Assessing Normality – Graphically

Q-Q Plots





The percentiles denote the proportion of cases (observations) that fall below a certain value.

Compared observed percentiles to percentiles we would expect from 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):

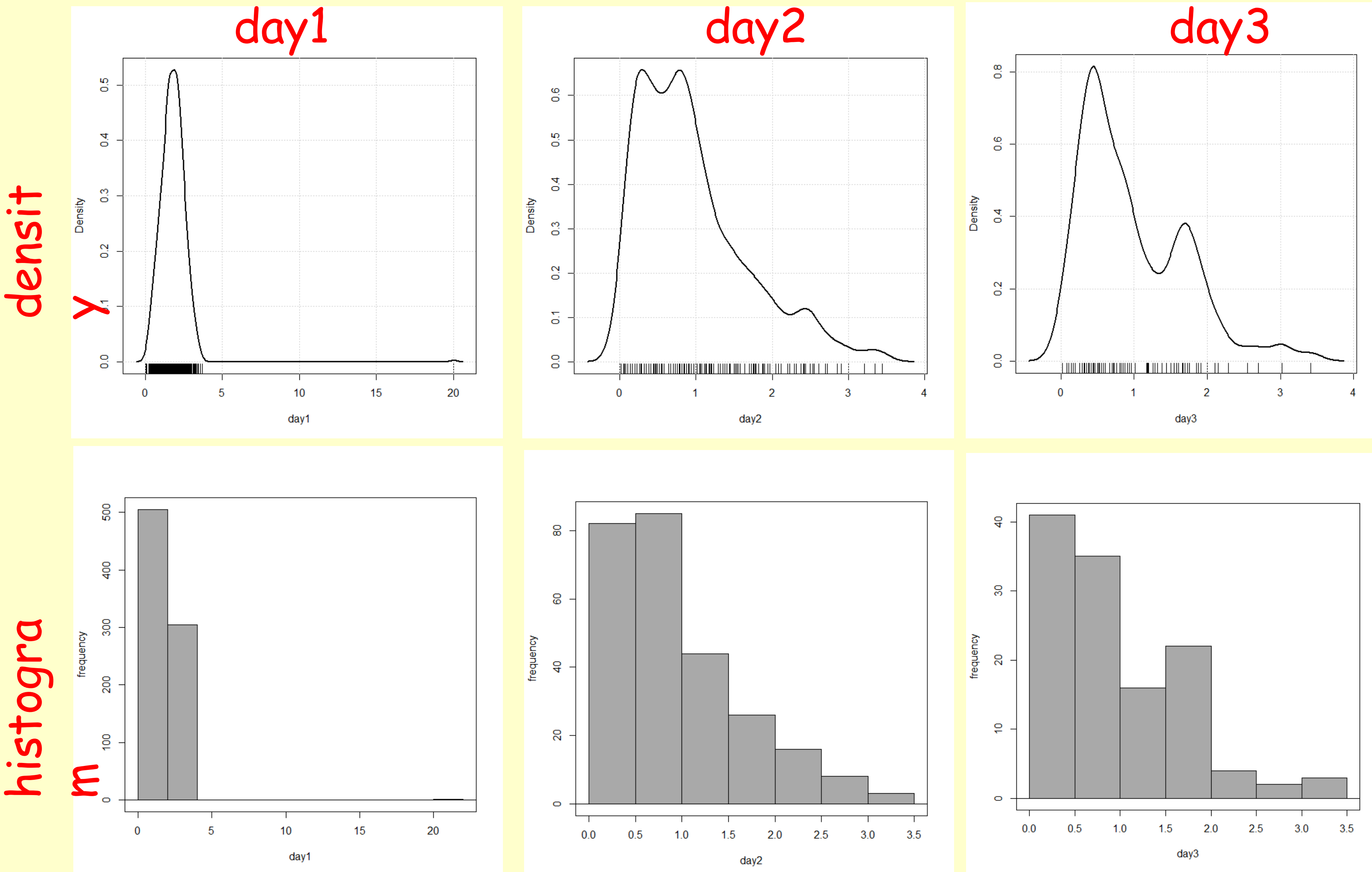
0 = you smell terribly   4 = you smell beautifully

Import Download Festival Data (MusicFestival.xlsx)

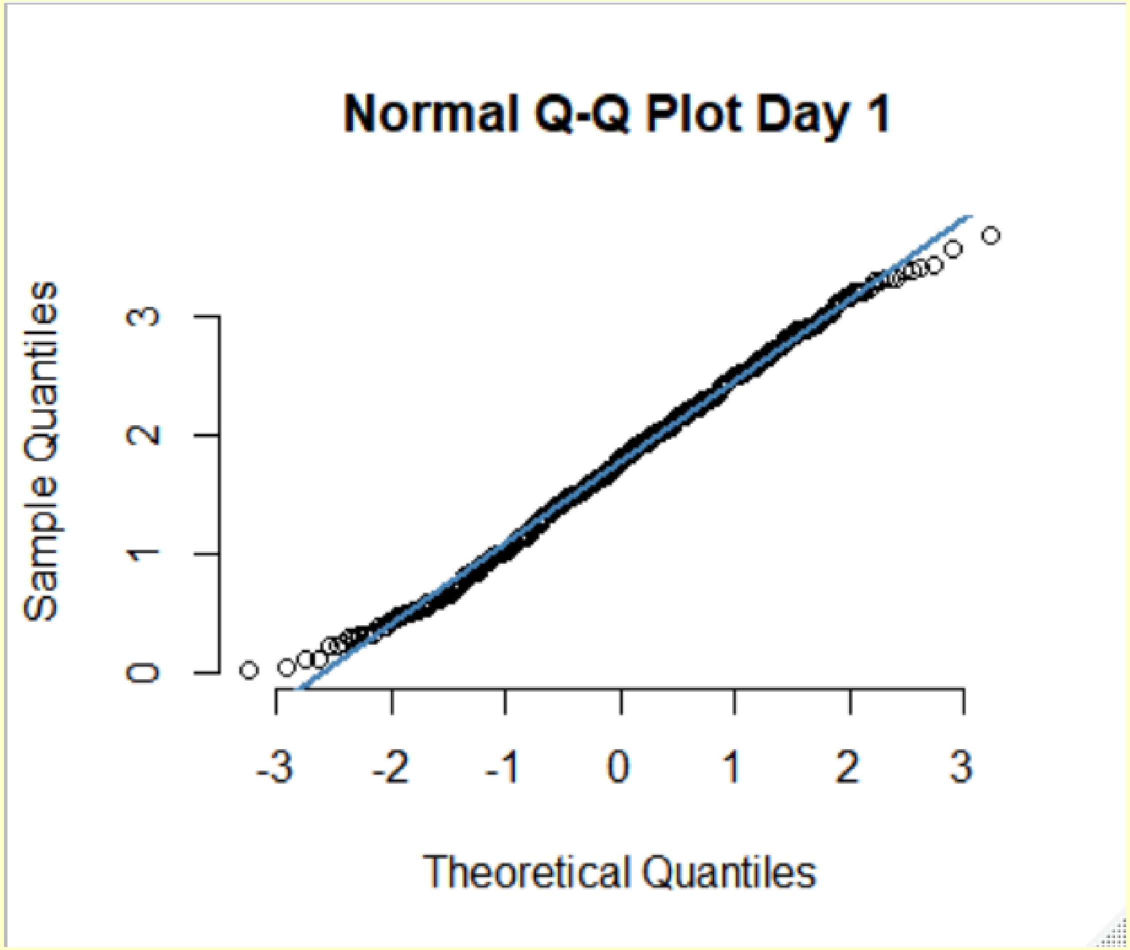
For ease of use, rename the Data Set "Festival"

```
> Festival <- DownloadFestival
```

Explore Data Graphically: RCmdr



Graphs: Q-Q Plots – Quantiles

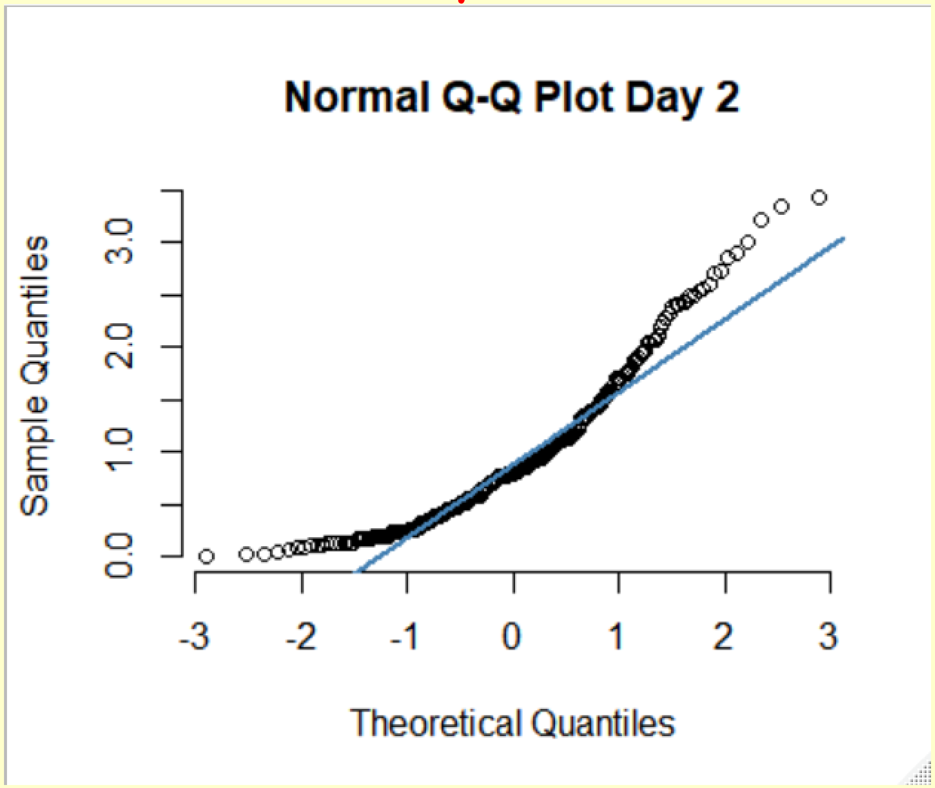


The solid blue line is the expected pattern a normal distribution with the same mean and SD and the sampled data.

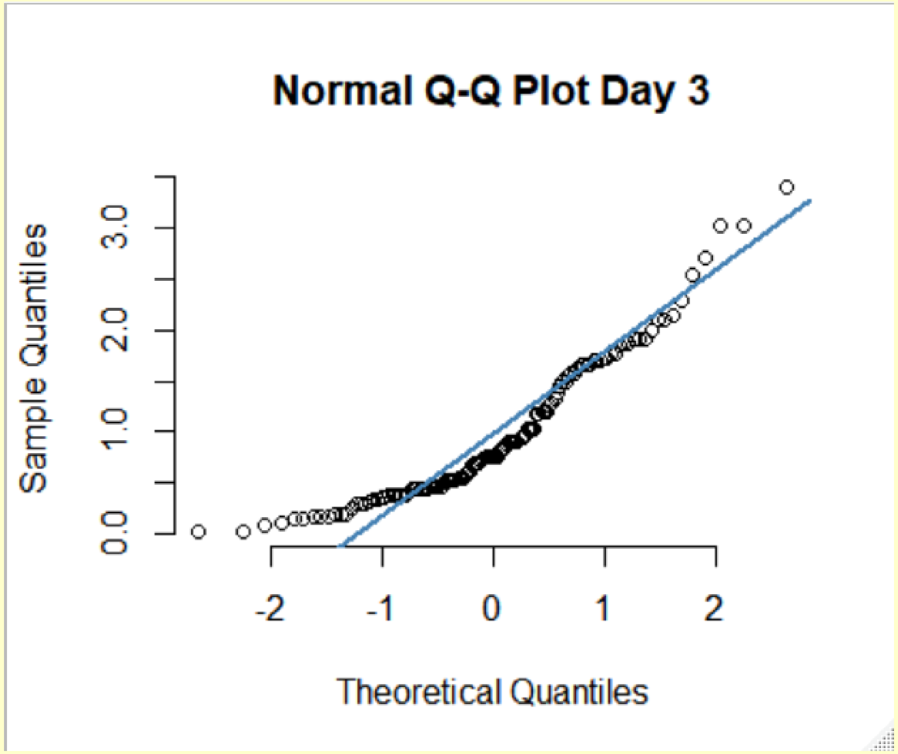
Points outside of the dashed line envelope suggest significant deviations

Graphs in Rcmdr – Quantiles

day 2



day 3



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

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	Test Statistic
6.539142e-01	

normtest.p	P value
1.545986e-37	

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 ?

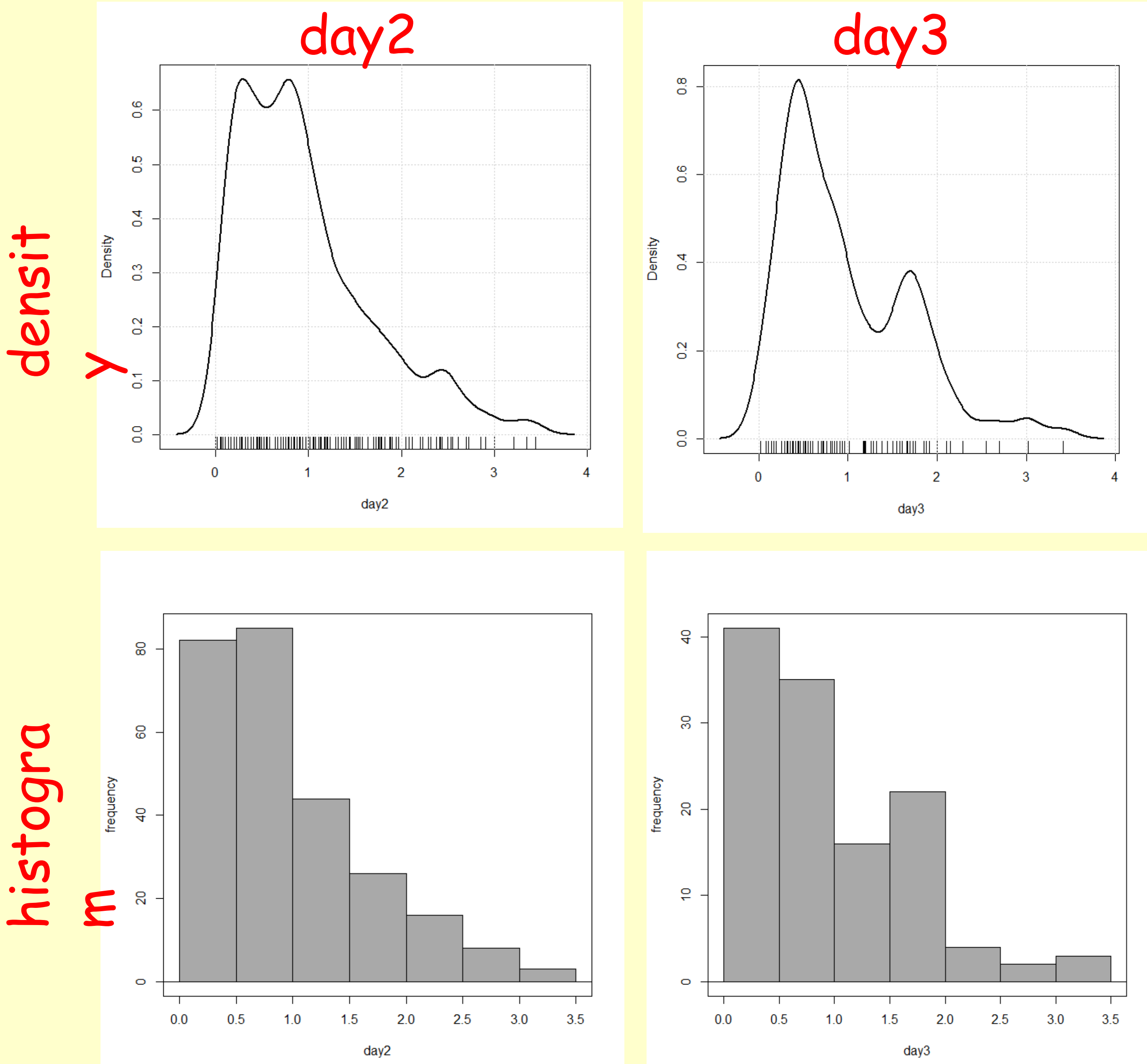
YES ($P < 0.05$)

Is day3 different from a normal distribution ?

How do I know that ?

YES ($P < 0.05$)

Graphical Data Exploration: RCmdr



Diagnostics:

Lack of Symmetry

Long tails

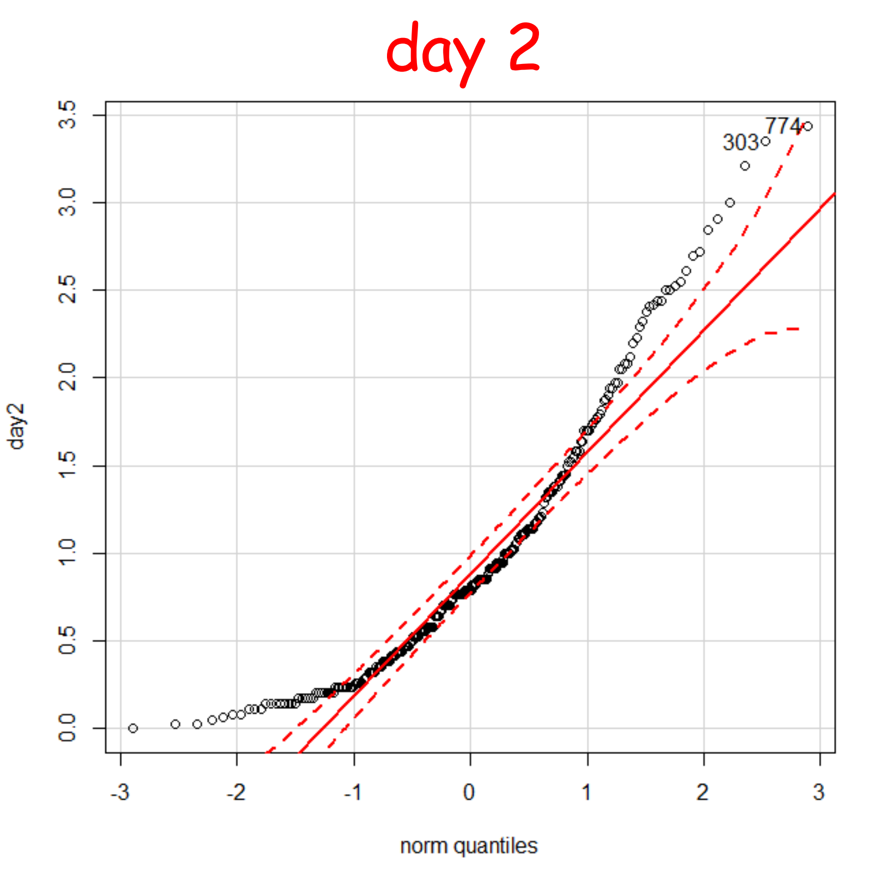
Mean > Median

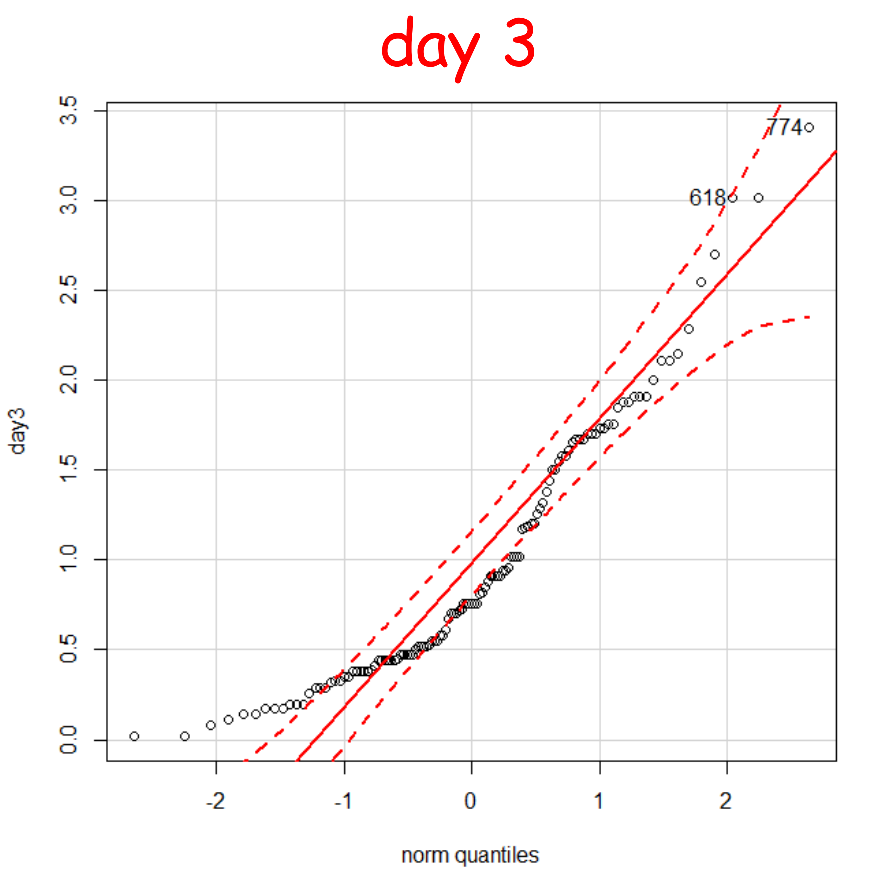
Positive Skew

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

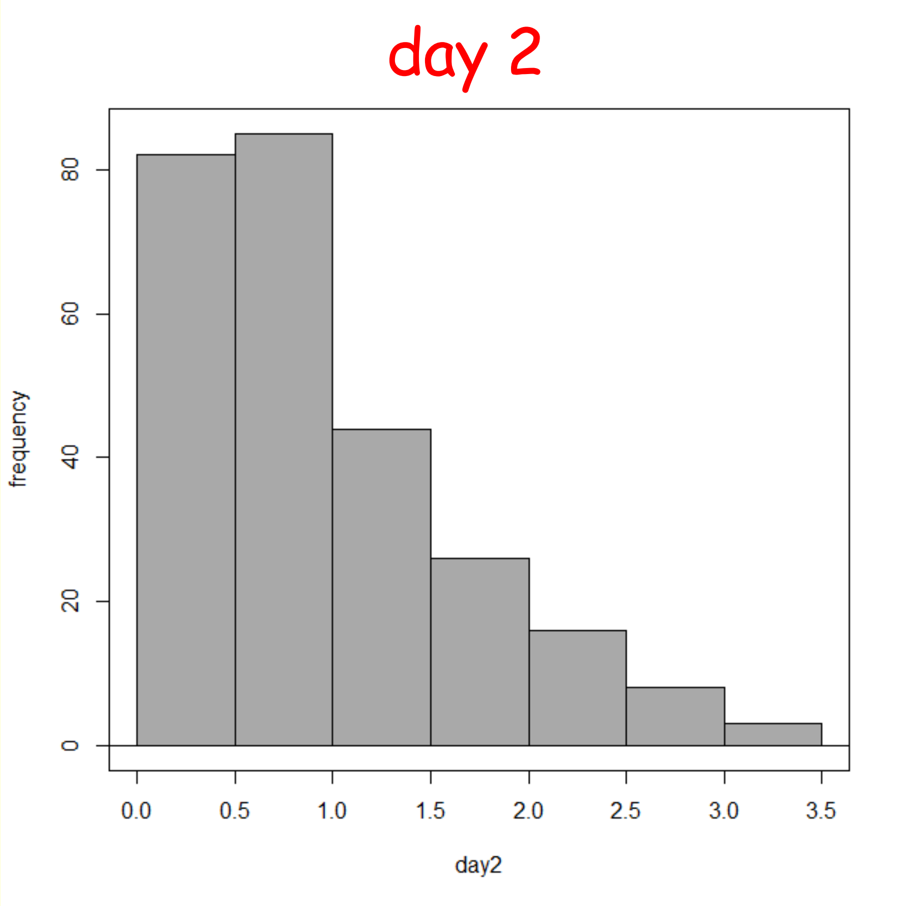


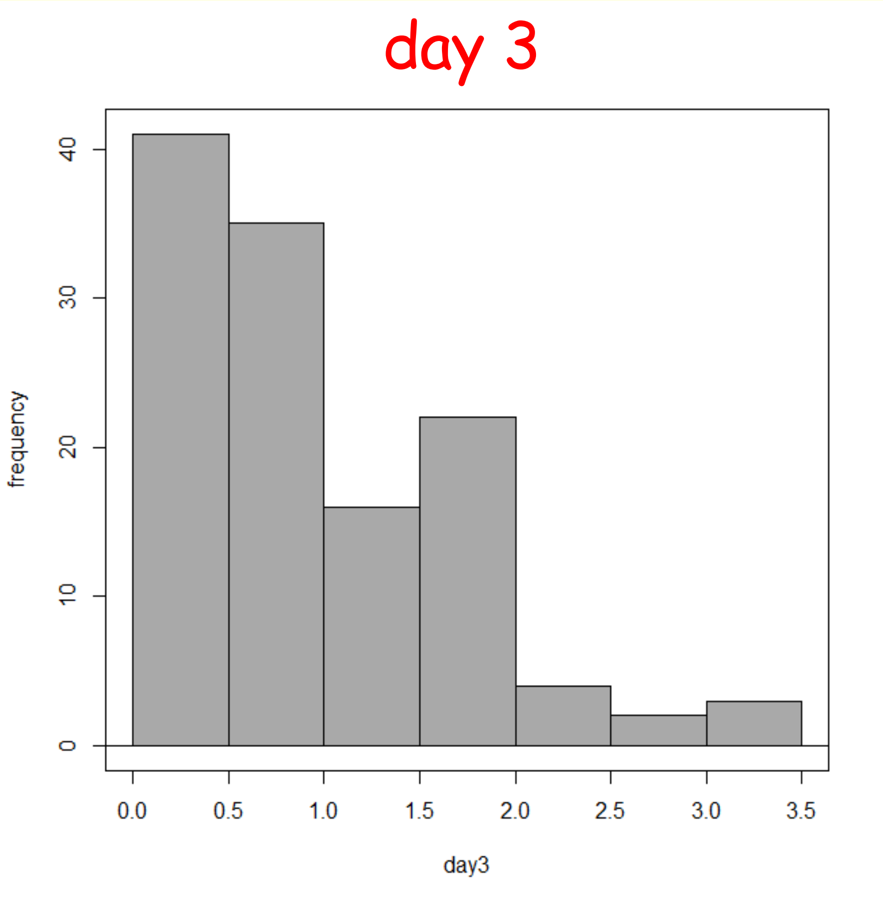


Rule of Thumb (Z scores)

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

Significant Results





Summary: Normality

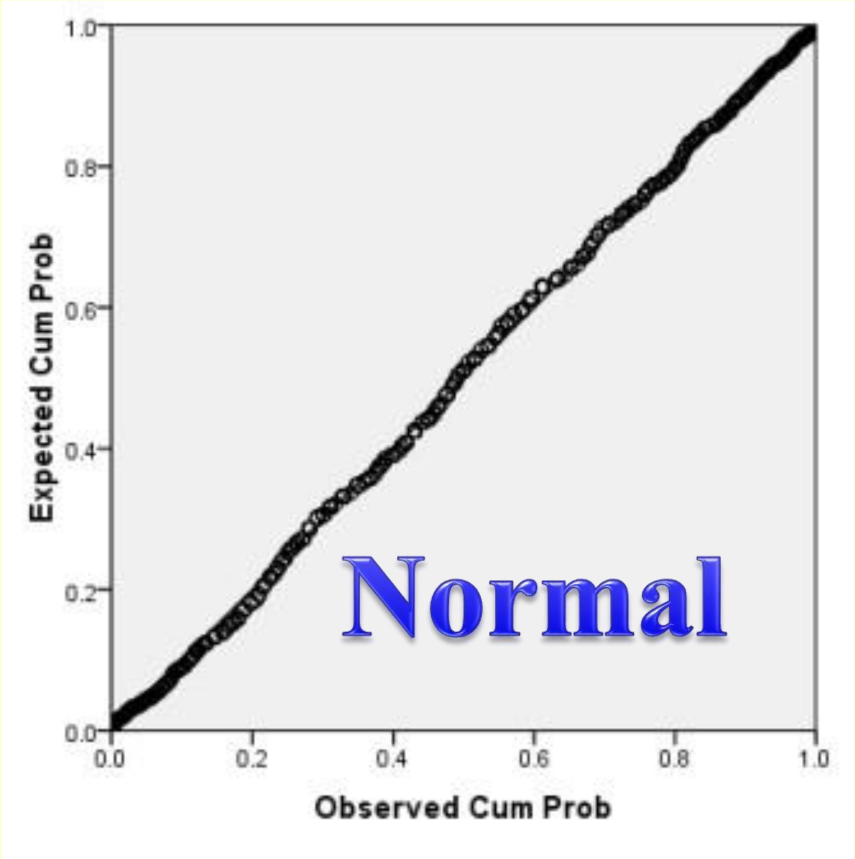
Indicators of a normal (Gaussian) distribution

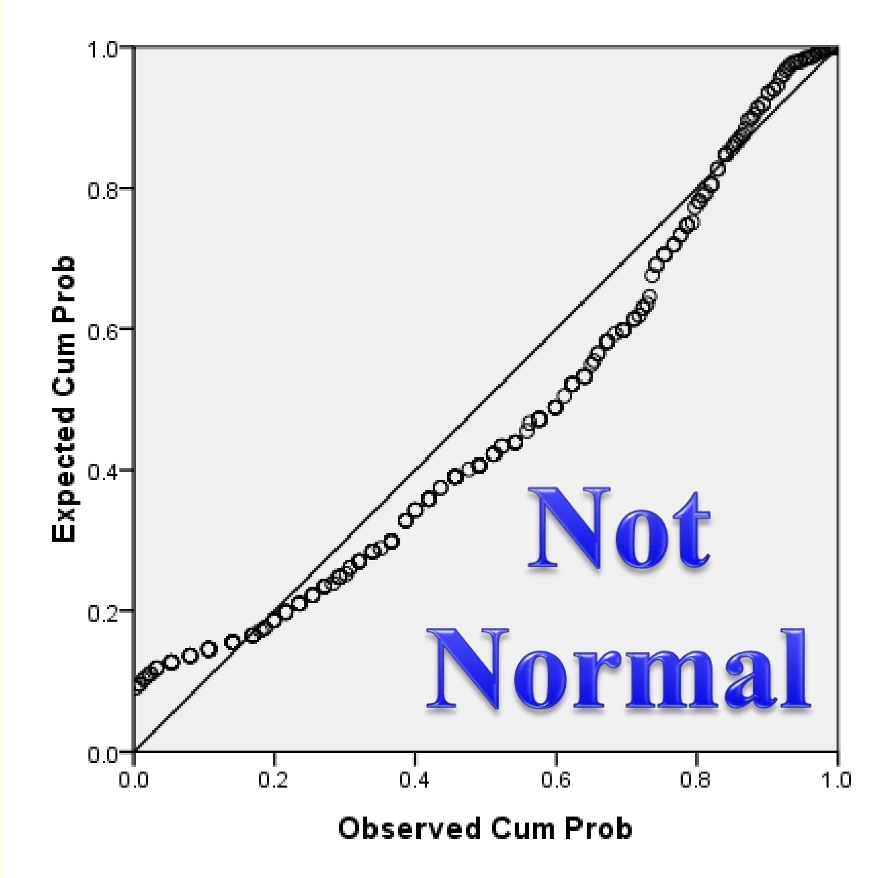
A. Mean = Median = Mode

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”:

ABS: absolute

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

Homogeneity of Variance

Assessing Variance Homogeneity

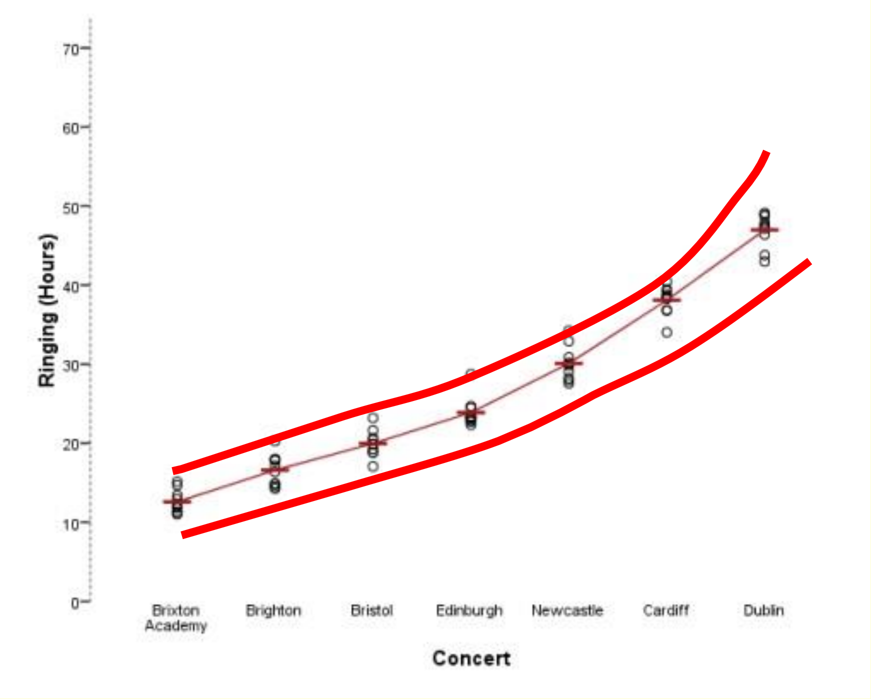
Recommendations:

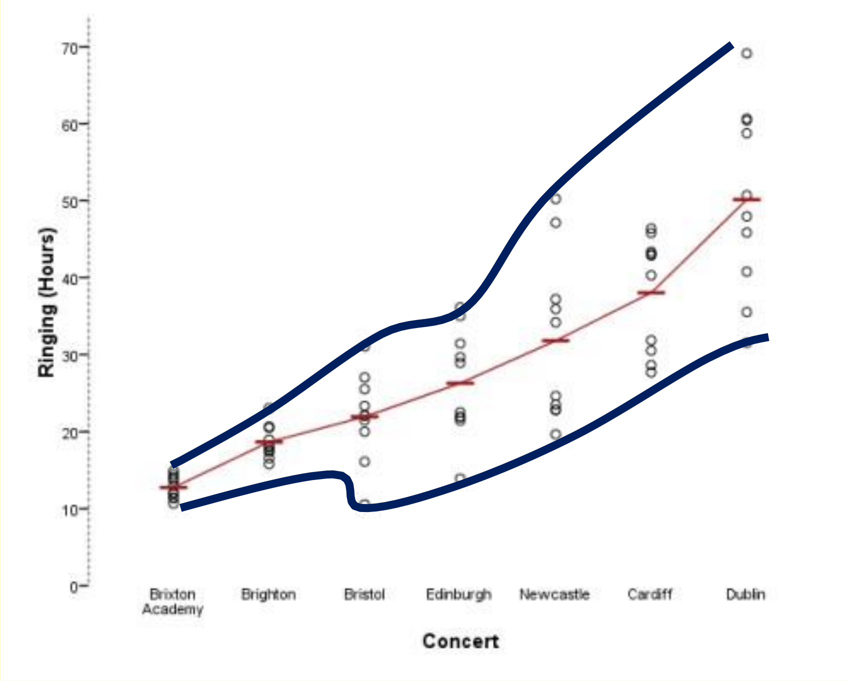
- Visualize data using scatterplots
- Variance Ratio: (with 2 or more groups) VR = Largest variance / Smallest variance If VR < 2, can assume homogeneity
- Levene's Test OR Bartlett's Test:

Significant = Variances are not equal

Non-Significant = Variances are equal

Variance Homogeneity - Graphic





Homogeneous **Heterogeneous**

Graphs illustrating data with homogeneous (left) and heterogeneous (right) variances

Assessing Variance Homogeneity

Graphs: Scatterplots (e.g., Regressions)

Variance Ratio: (with 2 or more groups)

VR = Largest variance / Smallest variance

If VR < 2, can assume homogeneity

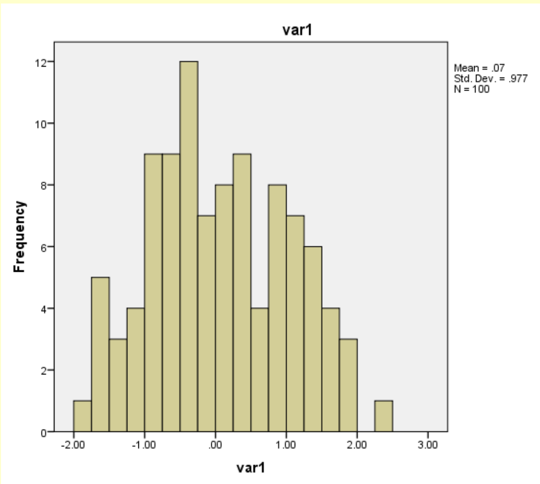
Levene's Test OR Bartlett's Test

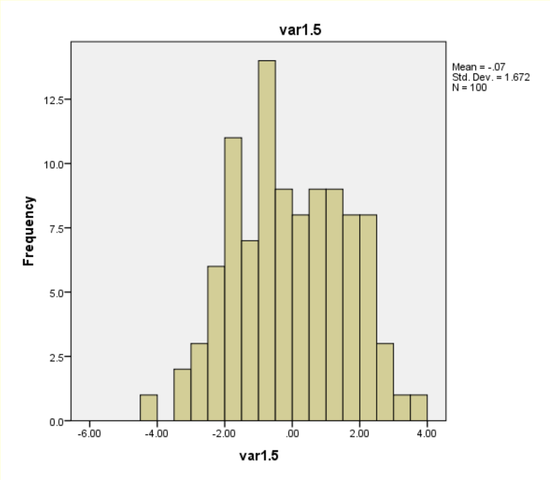
Significant = Variances are not equal

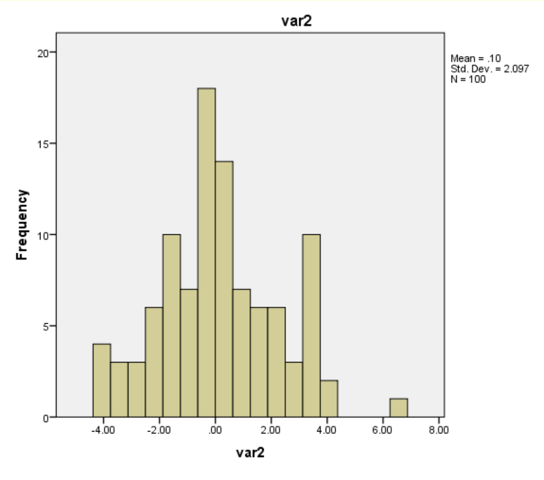
Non-Significant = Variances are equal

Variance Homogeneity – Ratio

Comparing three normal distributions with
the same means and different variances: 1, 2.25, 4







Pairwise Variance Comparisons

Larger Ratio / Smaller Ratio

4 /1 = 4

Rule of Thumb:

Ratio> 2

Variance Homogeneity Tests

Options available

- Two-variance F-test: compares two samples
- Bartlett's Test: compares two or more samples
- Levene's Test: compares two or more samples

Variance Homogeneity - Test

In R, use leveneTest() function in package car

leveneTest (outcome variable, group definition, center = median OR mean);

Default:

Center is the Median (better than mean)

leveneTest (exam$rexam, rexam$uni); OR
leveneTest (exam$rexam, rexam$uni, center = mean);

Variance Homogeneity - Test

Exam

Levene's Test for Homogeneity of Variance (center = median)

	Df	Fvalue	Pr(>F)
group	1 98	2.0886	0.1516

Total Degrees of Freedom = 100 -1 (N – 1)

Numeracy

Levene's Test for Homogeneity of Variance (center = median)

	Df	Fvalue	Pr(>F)
group	1 98	5.366	0.02262 *

--- Signif. codes:
0 '***'
0.001 '**'
0.01 '*'
0.05 '.'

Reporting Levene's Test Results

The Levene's test statistic is denoted with the letter F. Because there are two different degrees of freedom: the numerator (groups - 1) and the denominator (N – groups), it takes the form: F (df1, df2).

Exam	Numeracy
Levene's Test, $F(1, 98) = 2.09$, *ns**	Levene's Test, $F(1, 98) = 5.366$, $p = 0.023$

**ns : not significant*

Variance Homogeneity –Bartlett's Test

Bartlett test of homogeneity of variances
data: exam by uni Bartlett's
K-squared = 2.122, df = 1, p-value = 0.1452

```
bartlett.test(exam ~ uni, data=rexam)
```

Bartlett test of homogeneity of variances data: numeracy by uni
Bartlett's K-squared =
7.4206, df = 1, p-value = 0.006448

```
bartlett.test(numeracy ~ uni, data=rexam)
```

Assessing Variance Homogeneity

Bartlett's Test:

The Bartlett's test is used to test if *k* samples have equal variances (homogeneity of variances).

The Bartlett's test is sensitive to departures from normality. That is, if your samples come from non-normal distributions, then a significant Bartlett's test may simply reflect the lack of normality (typeI error)

Dixon, W. J. and Massey, F.J. (1969). *Introduction to Statistical Analysis*, McGraw-Hill, New York.

Assessing Variance Homogeneity

Levene's Test:

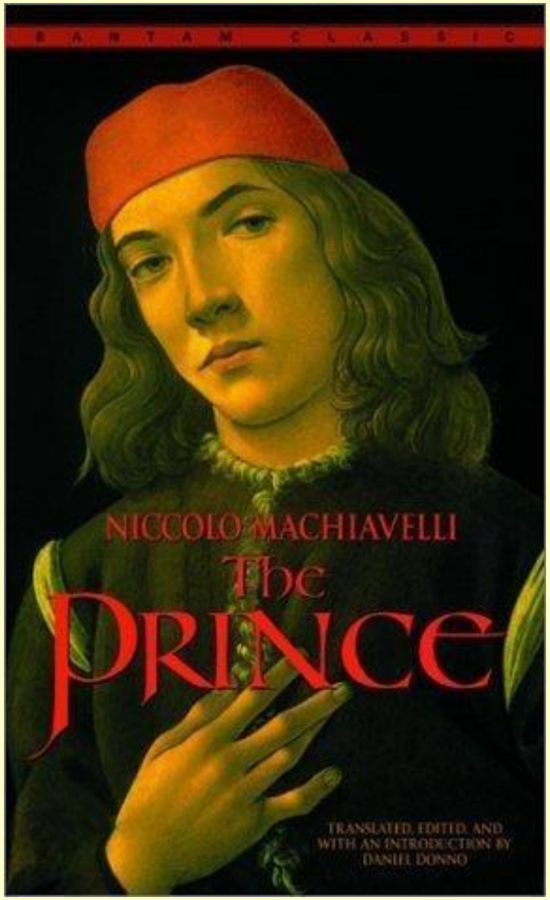
The Levene's test is used to test if *k* samples have equal variances (homogeneity of variances).

The Levene test is less sensitive than the Bartlett test to departures from normality. If you have strong
evidence that your data do in fact come from a normal, distribution, then Bartlett's test has more power.

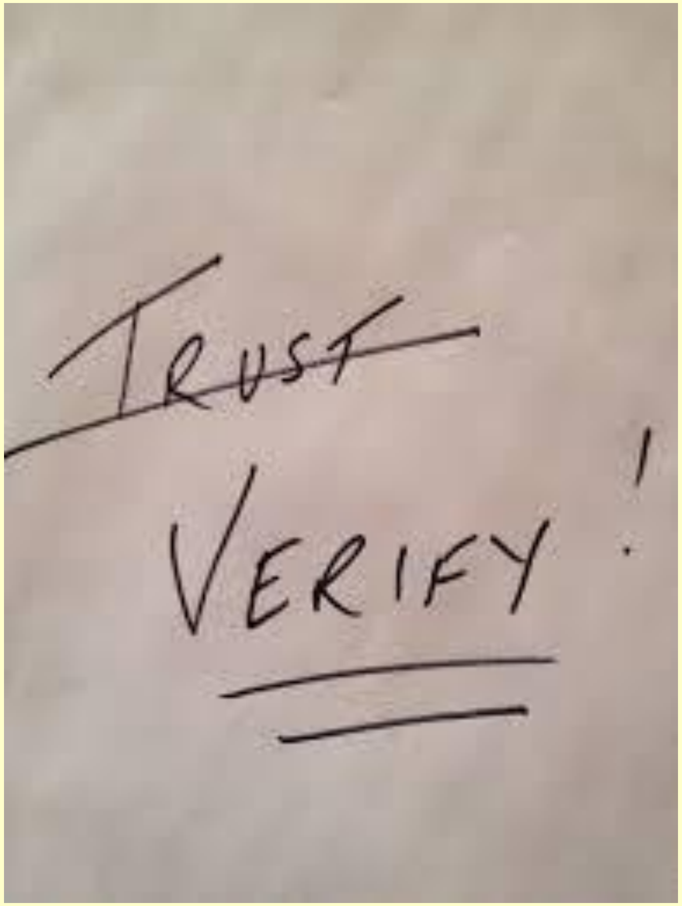
Levene, H. (1960). In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., Stanford University Press, pp. 278-292.

Why Have Non-normal Data…

What Next ?



Transform …



and Verify

Monotonic Data Transformations

Monotonic: Data values are changed but value ranks are not changed

Non-monotonic: Data values are changed and value ranks are also changed

NOTE: non-monotonic transformations change the signal in the data. Only use monotonic transformations, if possible.

Log Transformation

Logarithmic transformation fx = ln(x) OR log(x)

TRANSFORMATION	Reasonable and acceptable domain of x	Range of $f(x)$
$\log(x)$	positive	all

(x) → **f(x)**

- This transformation is useful when:
- High degree of variation within samples (e.g., Chl Conc.)
- Large outliers (tails) and lots of zeros

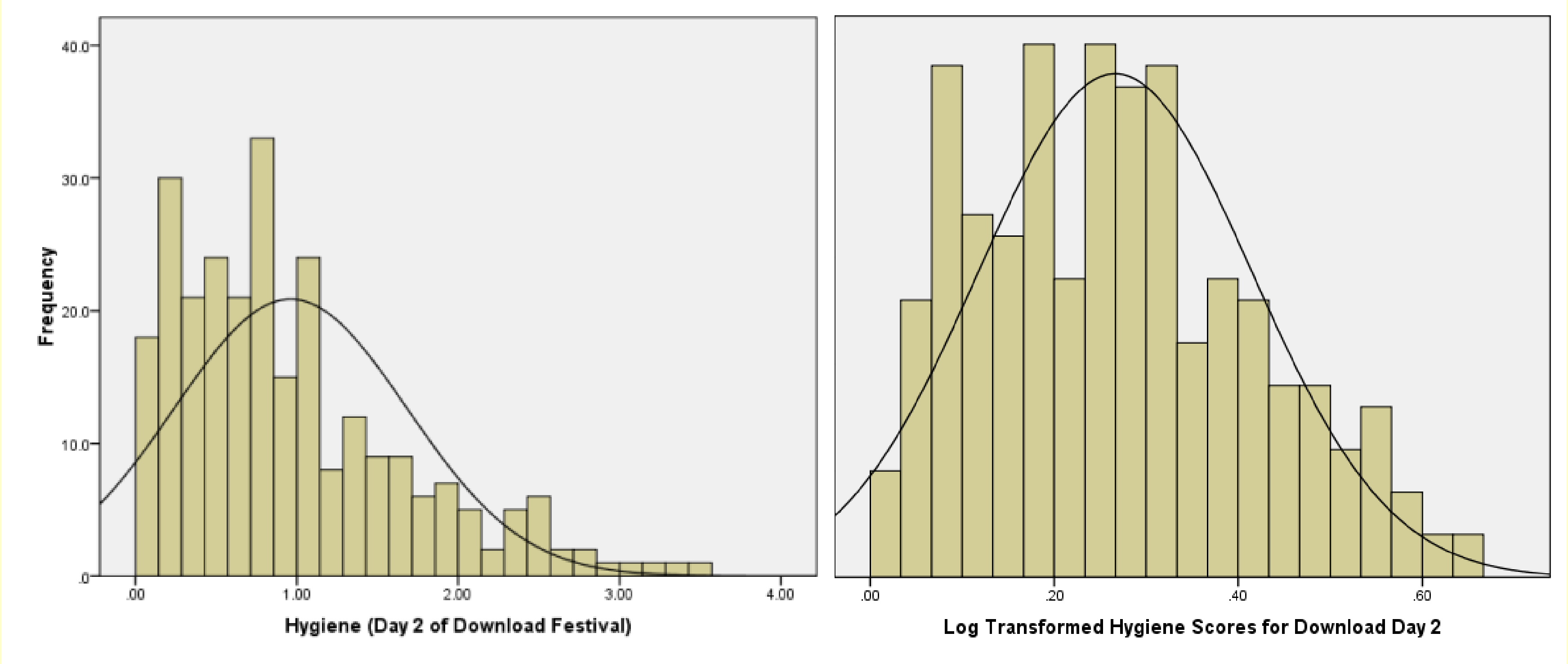
- Note: to log-transform data containing zeros, a small number should be added to all data points.

- With count data, add one, so that: fx = log(0+1) =0
- With density data, add constant smaller than smallest possible sample, so that: fx = log(0+0.001) = -3

Log Transformation

Log Transformation ($\log(X_i)$): Reduces positive skew

Before | After



Monotonic Data Transformations

MONOTONIC TRANSFORMATIONS	Reasonable and acceptable domain of x	Range of $f(x)$
$x^{½}$ (power)		

(x) $\longrightarrow$ **f(x)**

Power exponents:

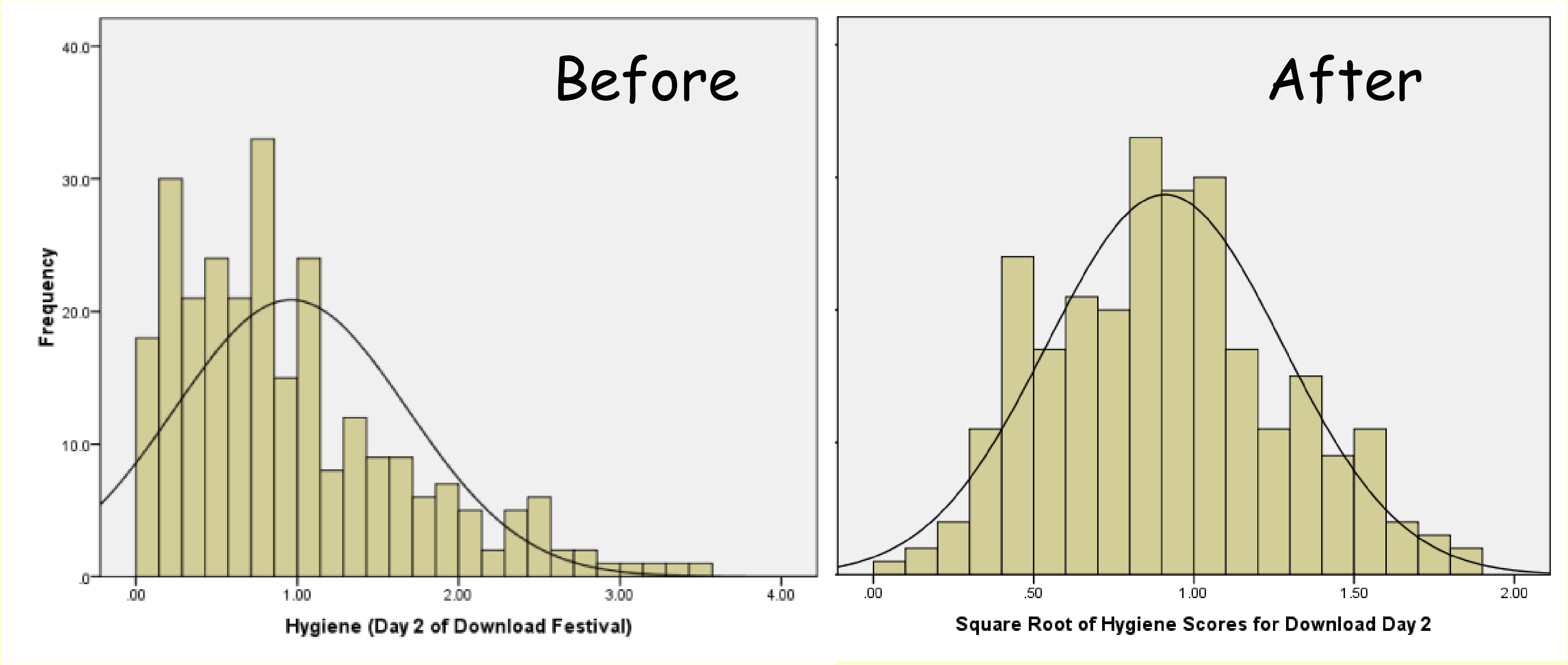
½ power (square root)

Square root transform deals with positive skew, by bringing in large tails.
Special treatment of zeros not necessary.

Square Root Transformation

Square Root Transformation ($\sqrt{X_i}$):

Reduces positive skew. Useful for stabilizing variance



Data Transformations – For Proportions

Arcsine / Arcsine-squareroot transformation

TRANSFORMATION	Reasonable and acceptable domain of x	Range of $f(x)$
$(2/\pi)\cdot\arcsin(x)$	$0 \leq x \leq 1$	0 to 1 inclusive
$(2/\pi)\cdot\arcsin(x^{½})$	$0 \leq x \leq 1$	0 to 1 inclusive

- This transformation is useful when dealing with proportional data (e.g., Percent Cover)

- Note: data must range between 0 and 1, inclusive.

The constant 2 / pi scales the result of arcsin(x) [in radians] to range from 0 to 1, assuming that $0 \leq x \leq 1$.

Non-monotonic Data Transformations

NONMONOTONIC TRANSFORMATIONS	Reasonable and acceptable domain of x	Range of $f(x)$	
x^0 (power)	all	0 or 1 only	**P / A**

(x) → **f(x)**

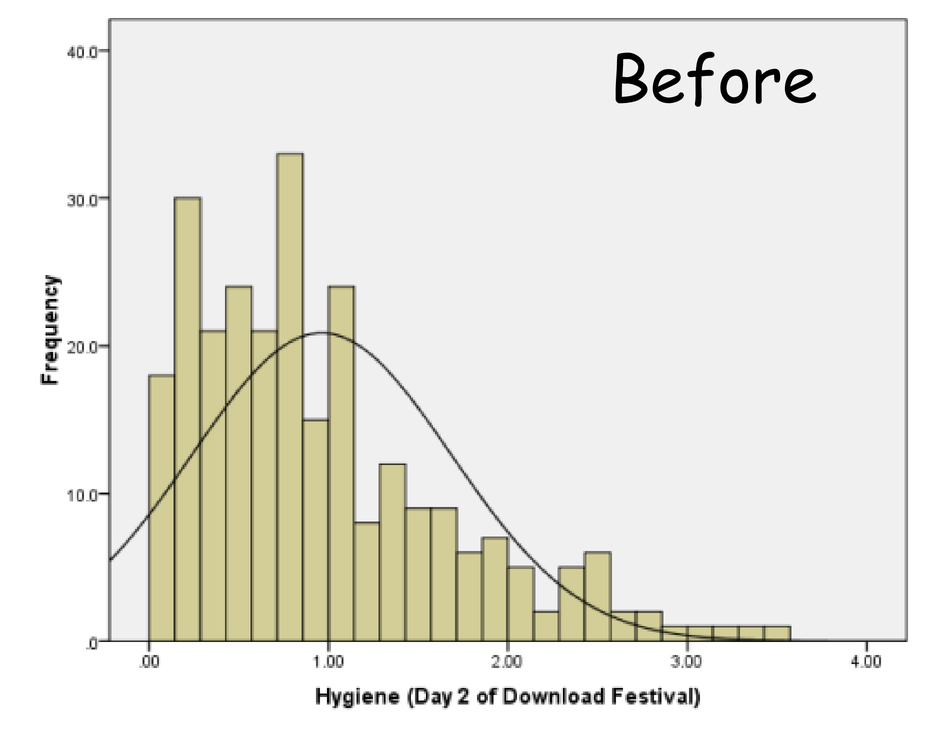
Note: 0 power transformation is NOT monotonic

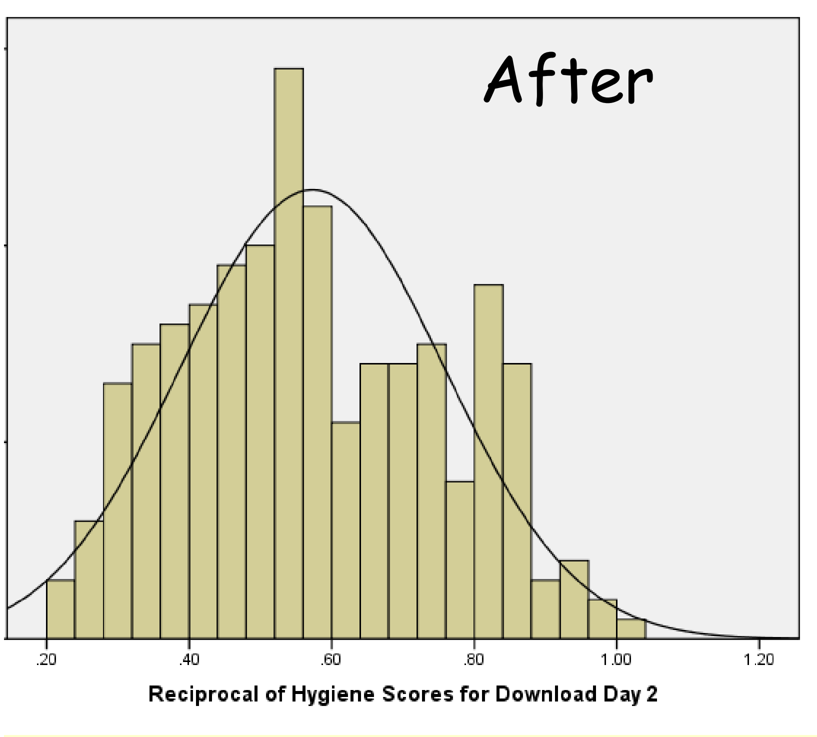
It recodes data as Presence / Absence (0 / 1)

Reciprocal Transformation

$(1 / X_i)$: Dividing 1 by each value reduces the impact of large scores.

Beware: This transformation is non-monotonic, because it reverses the scores.

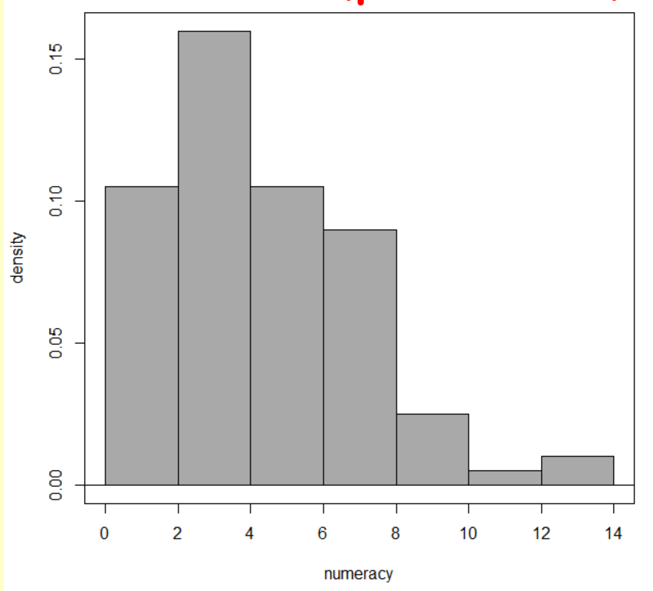




Numeracy:

skewness	skew.2SE
0.93271513942	1.93204903727
kurtosis	**kurt.2SE**
0.76349270501	0.79807966944

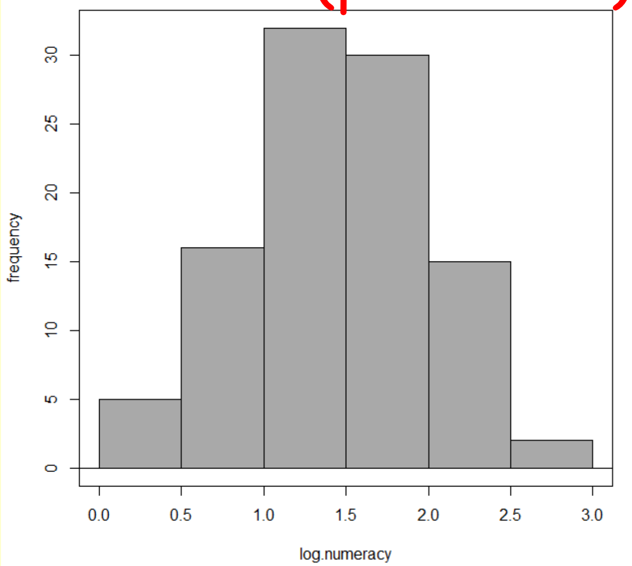
S-W test ($p < 0.001$)



Log (Numeracy):

skewness	skew.2SE
0.401220988	-0.831099004
kurtosis	**kurt.2SE**
-0.275254548	-0.287723848

S-W test ($p = 0.003$)



Take-home Lessons

Data transformations are one of the most difficult issues in parametric statistics:

- Conflicting advice: transform or not

-Conflicting results: various normality tests

Recommendation:

Select one approach that provides multiple evidence and come up with criteria before starting analysis

Be as strict as you wish: one or more criteria

But, if a test significant… cannot back-track

Take-home Lessons

- Parametric tests are more powerful, but are based on assumption of normally distributed data
- Determine normality criteria and undertake data transformations, if needed
- If you are unsure, data transformations can always be attempted to compare the same test results, using transformed and un-transformed data
- Test normality before / after data transformations
- If transformations do not work…

  use non-parametric tests

To Transform or Not ?

- Tricky to achieve normality with small samples (often impossible when $n < 20$)

- Transforming data does not always work (e.g., fix skew / kurtosis)

- Transforming data affects hypothesis being tested

  E.g. when using a log transformation and comparing means you are switching from comparing arithmetic means to comparing geometric means

Summary

Rules for Data Transformations

Most Important Rule: Do not Reverse the Order of the Values (larger remains larger… smaller remains smaller)

Monotonic: change values but retain ranks

Non-monotonic: change values and ranks

(For example: Add random number, Multiply by (-1))