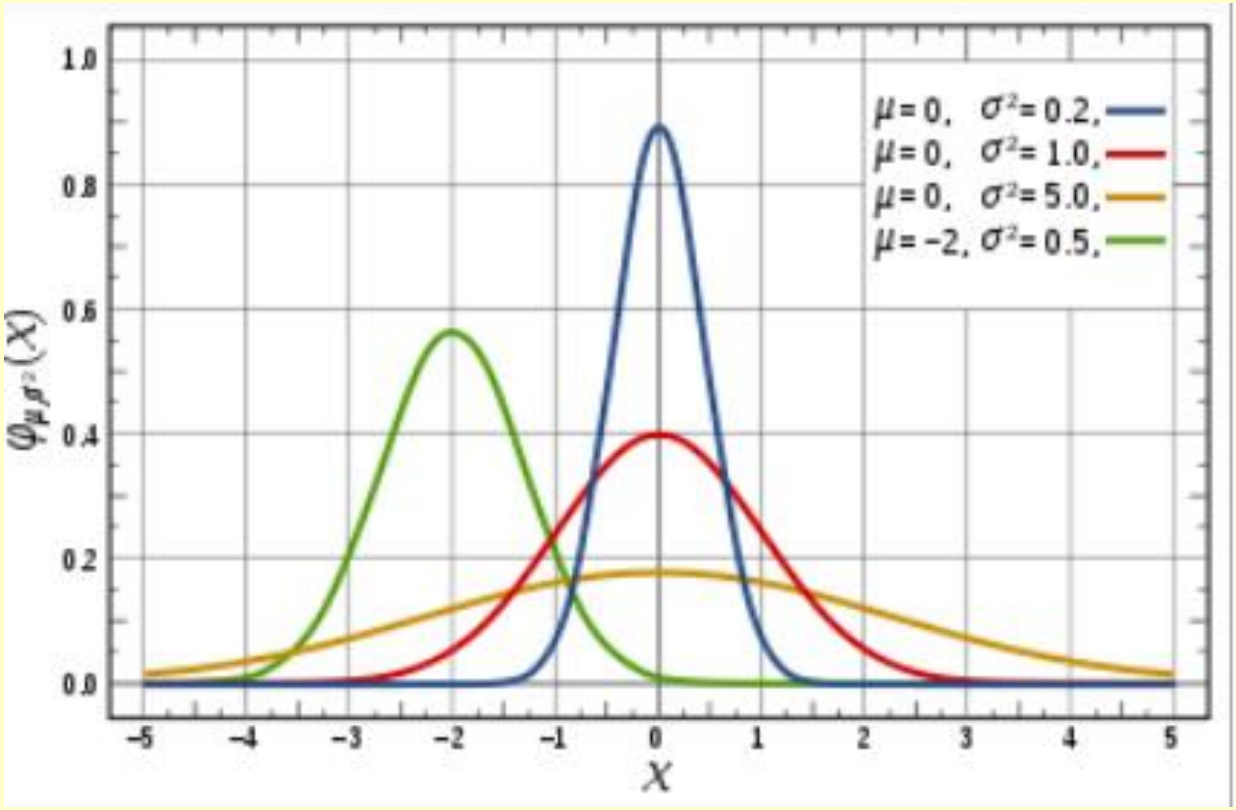
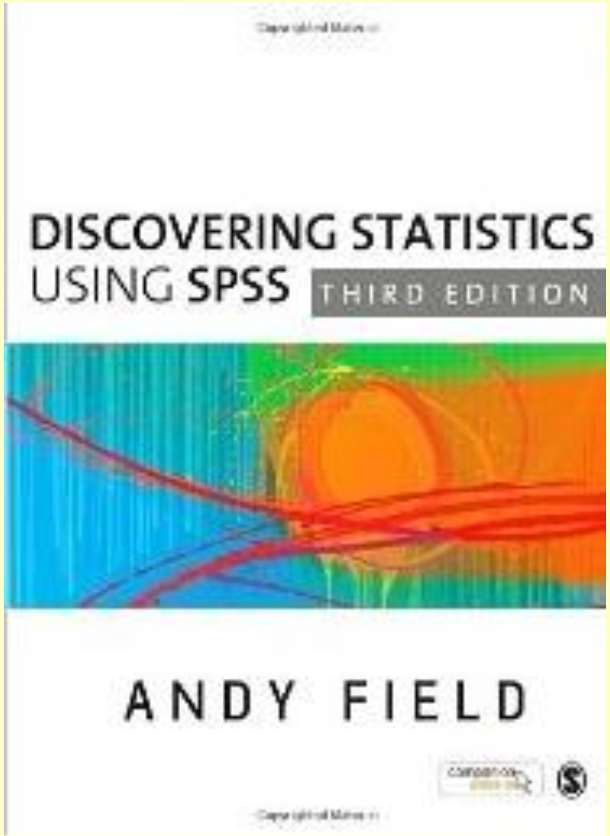
Parametric & Nonparametric Statistics: Assumptions



http://www.pelagicos.net/classes_biometry_fa16.htm

Reading - Field: Chapters 3, 4, 5



Assessing Normality

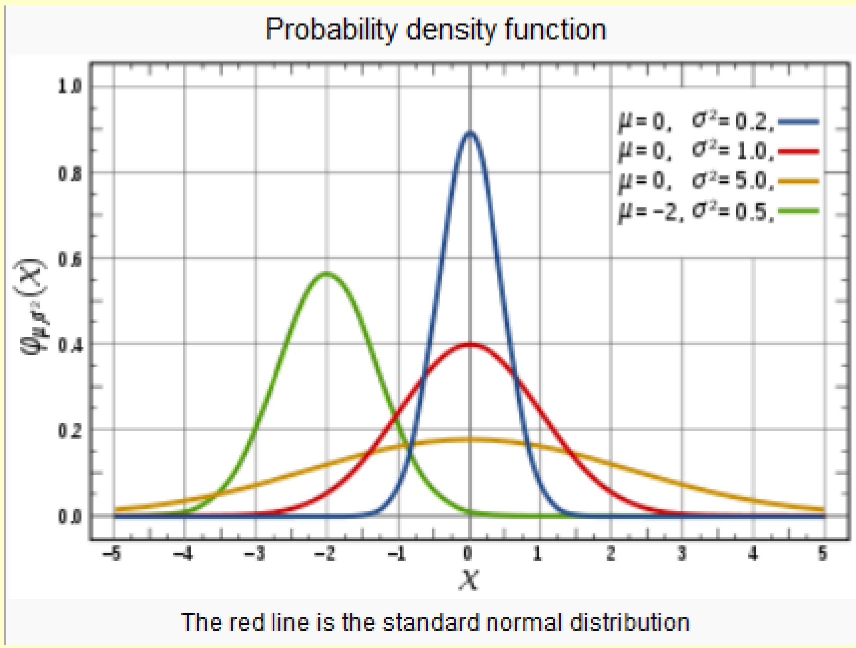
- We do not have access to sample the entire biological population, so we test observed data
- 1) Central Limit Theorem
  - If $N < 25$, sampling distribution rarely normal
- 2) Graphical Displays
  - Histogram
  - P-P Plot (or Q-Q plot)
- 3) Skewness / Kurtosis (point estimate +/- SE)
  - Do they overlap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

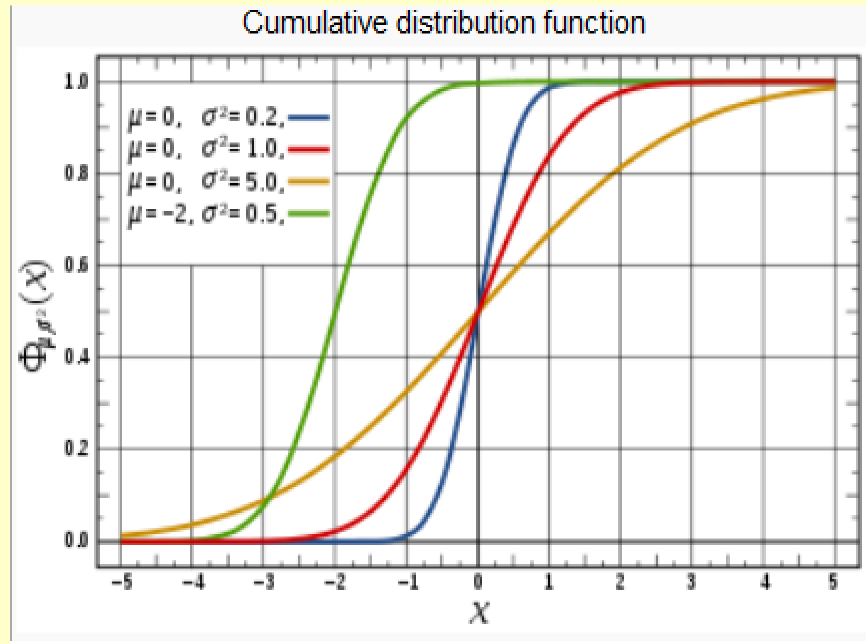
Assessing Normality

4) Performing Statistical Tests

- Skewness / Kurtosis
  - 0 in a normal distribution
  - Convert to *z* score (by dividing value by *SE*)
- Kolmogorov-Smirnov & Shapiro – Wilk Tests
  - Tests if data differ from a normal distribution
  - Significant = non-Normal data
  - Non-Significant = Normal data

Assessing Normality - Graphically



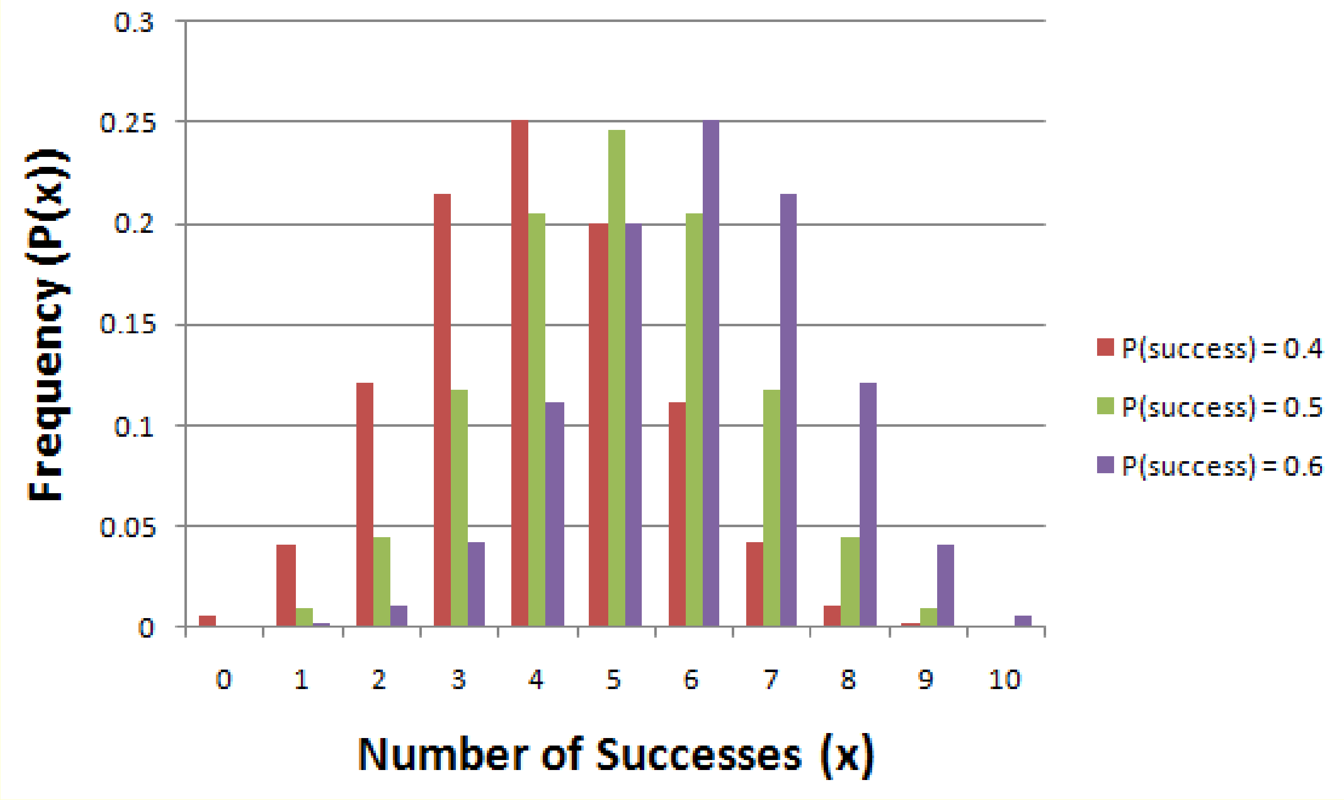


Comparing observations against a cumulative normal distribution (same mean and S.D.)

Assessing Normality - Graphically

Characteristics of Normal Distributions

Unimodal, Symmetrical, Bell-shaped



Assessing Normality - Graphically

Binomial distribution – from Hw #3

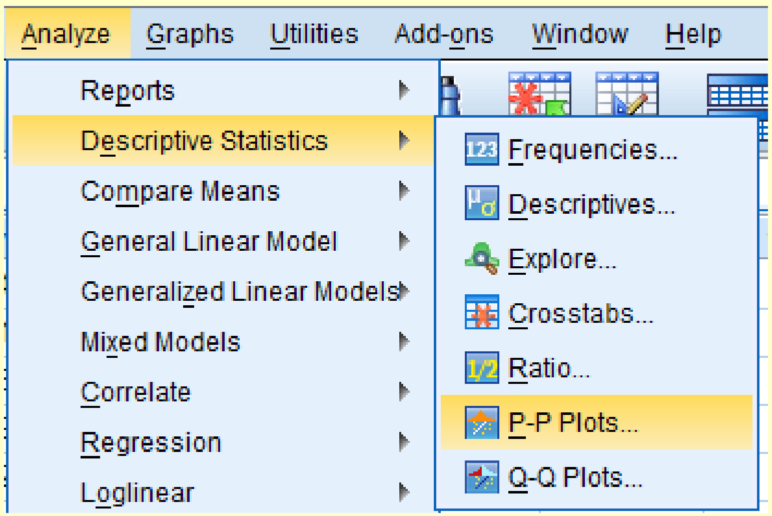
With 2 parameters:

p = 0.4
n = 10

# Successes	Frequency	Count (sample of 1000)
0	0.00605	6
1	0.04031	40
2	0.12093	121
3	0.21499	215
4	0.25082	241
5	0.20066	201
6	0.11148	111
7	0.04247	42
8	0.01061	11
9	0.00158	2
10	0.00010	0

Assessing Normality - Graphically

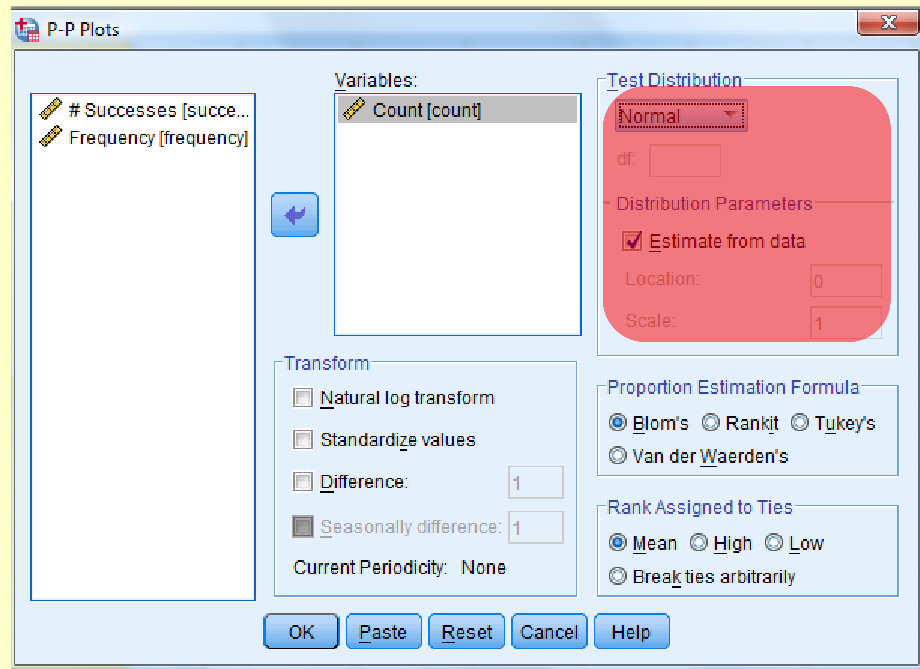
P-P Plot (or Q-Q plot)



Test Distribution (df)

Select Parameters

(determined or estimated)

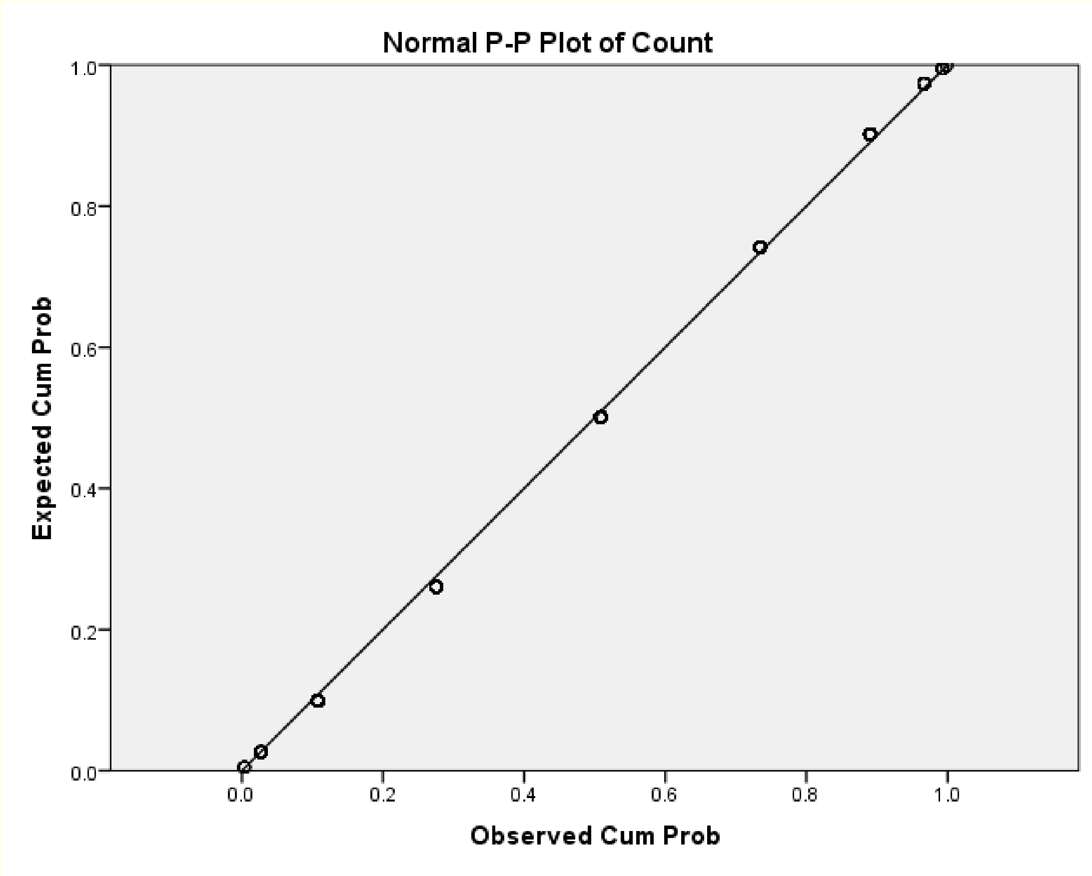


P – P Plots

Visualizing deviations from cumulative frequency of the "normal"

A **p - p plot** shows the cumulative frequencies of two datasets: the observed data and the expected data.

NOTE: Every value in the datasets is plotted with a different point.



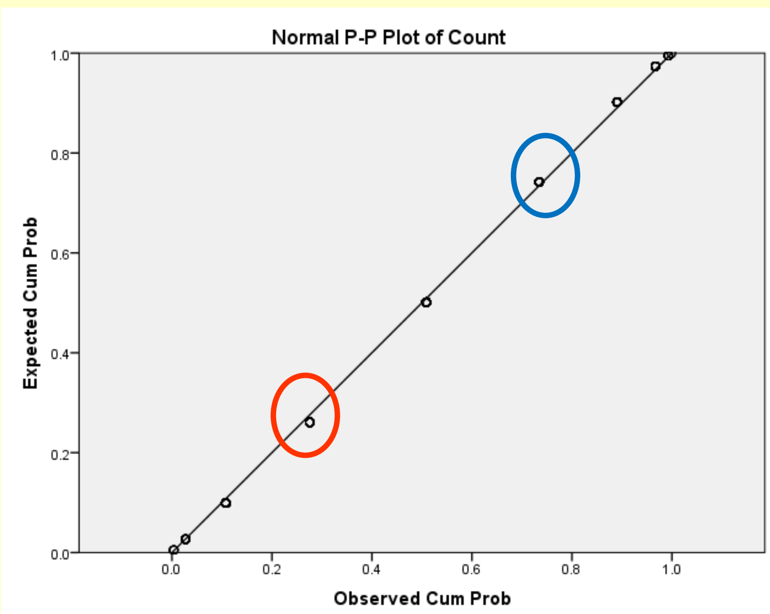
Assessing Normality - Graphically

Estimated Distribution Parameters

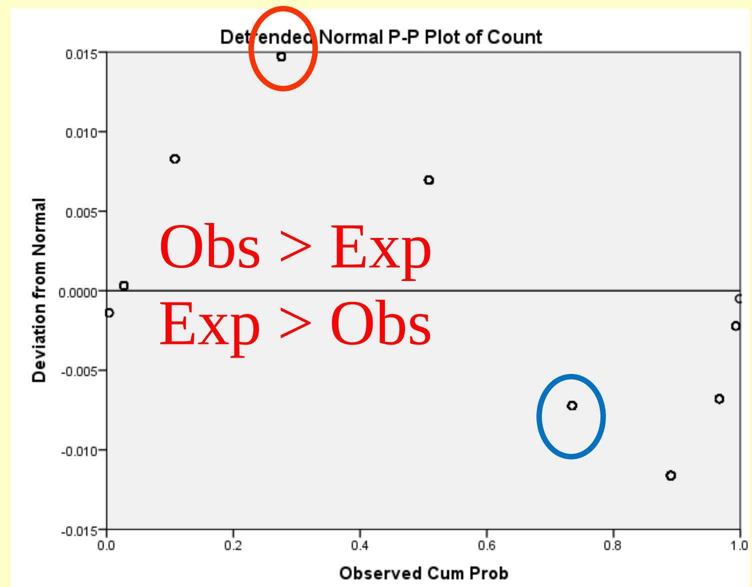
		Count
Normal Distribution	Location	3.9940
	Scale	1.55125

The cases are unweighted.

Mean
S.D.



P-P plot (cummulative)



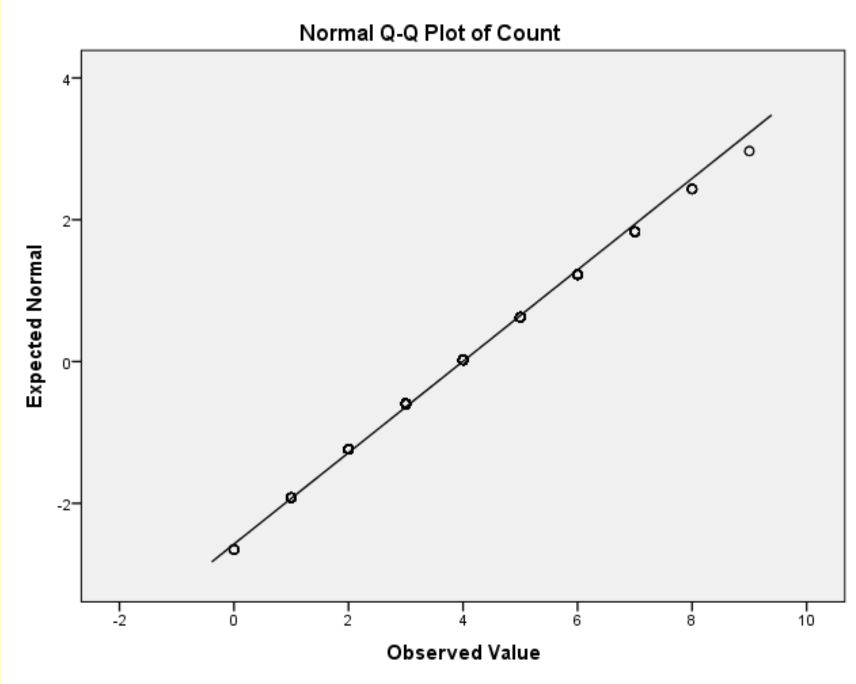
Deviations of P-P Plot

Q – Q Plots

Visualizing deviations from quantiles of "normal"

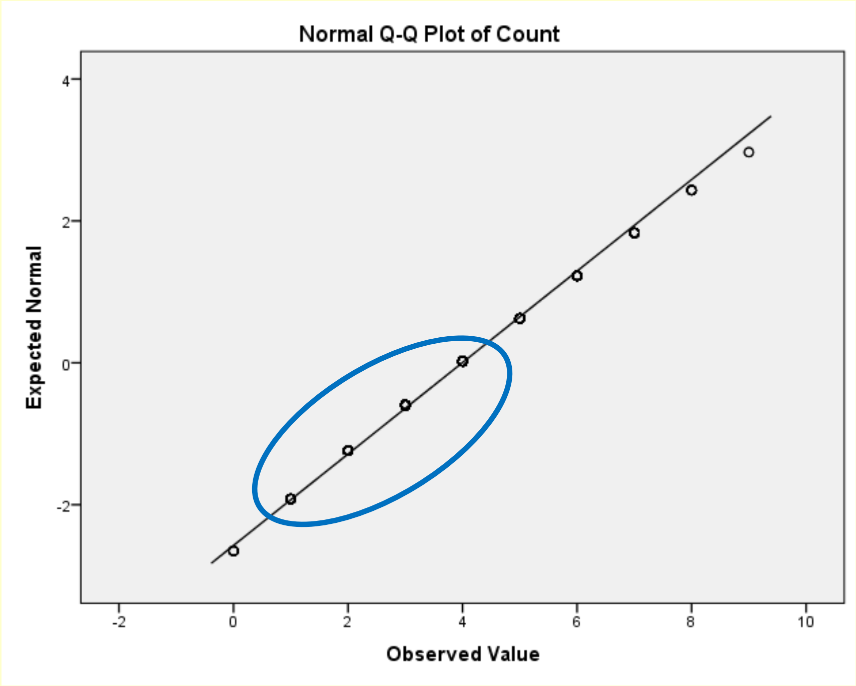
A **q-q plot** is a **plot** of the quantiles of the first data set against the quantiles of the second data set.

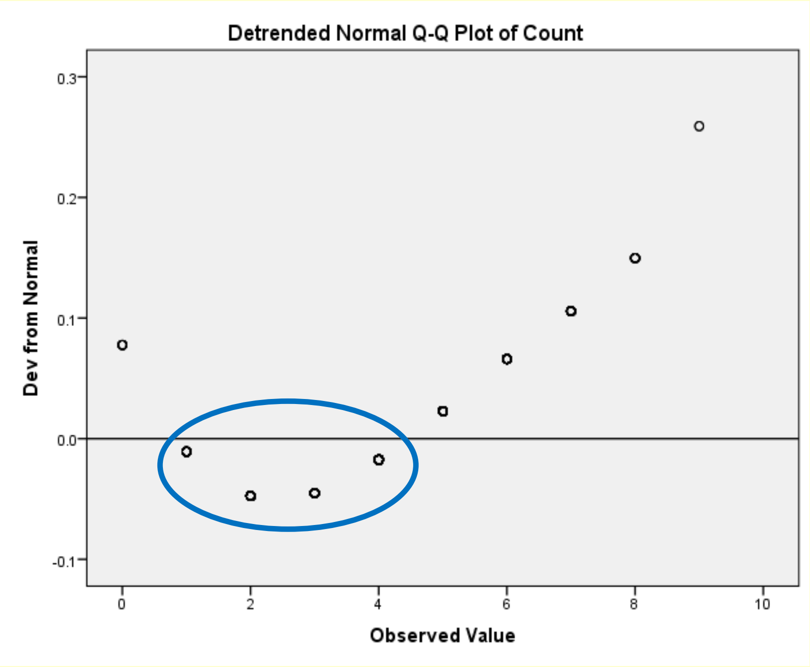
NOTE: Every value in the datasets is plotted with a different point.



Remember: By a quantile, we mean the fraction (or %) of data points below the given value.
NOTE: We use the center of mass of the distribution (mean = median) as the reference point for the normal.

Assessing Normality - Tests

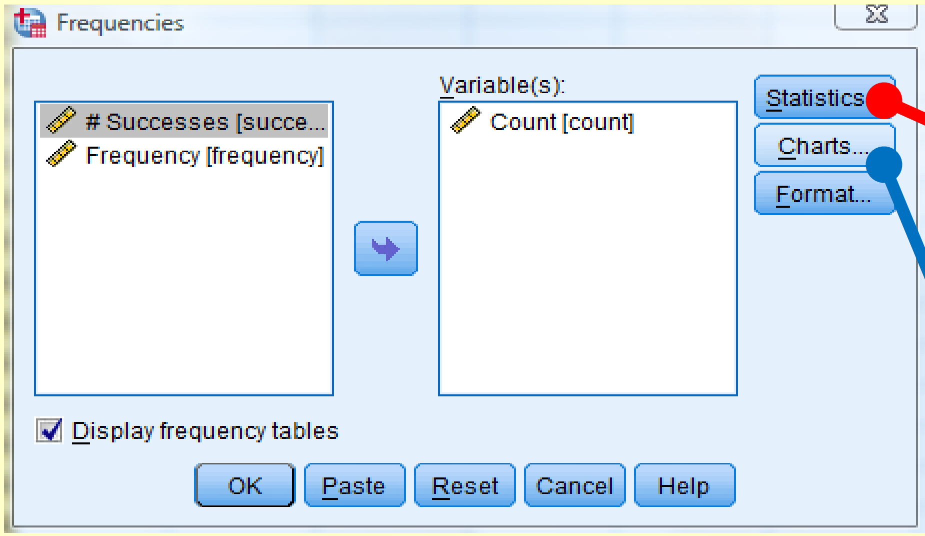


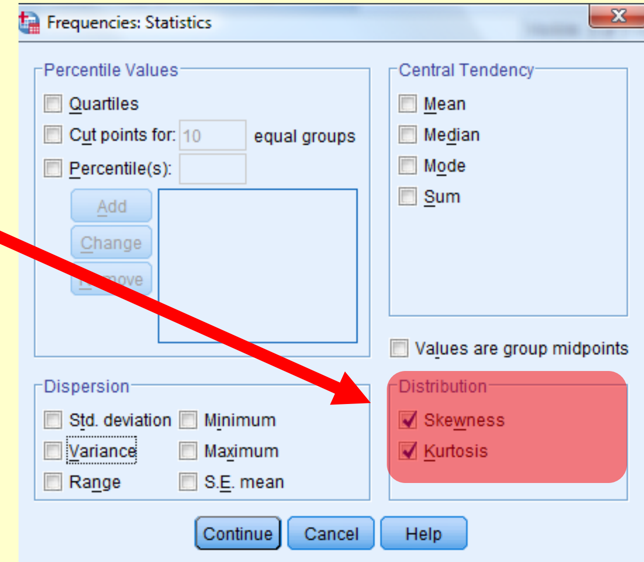


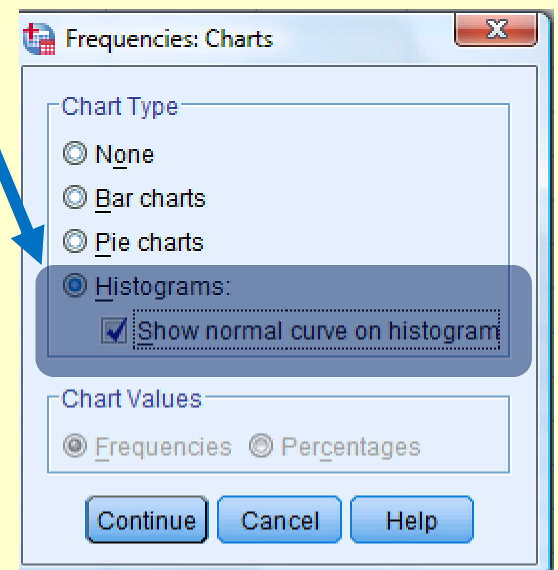
Q – Q Plots:

Visualizing deviations from the quantiles of the "normal expectation": Observed – Expected

Assessing Normality – Skew & Kurtosis



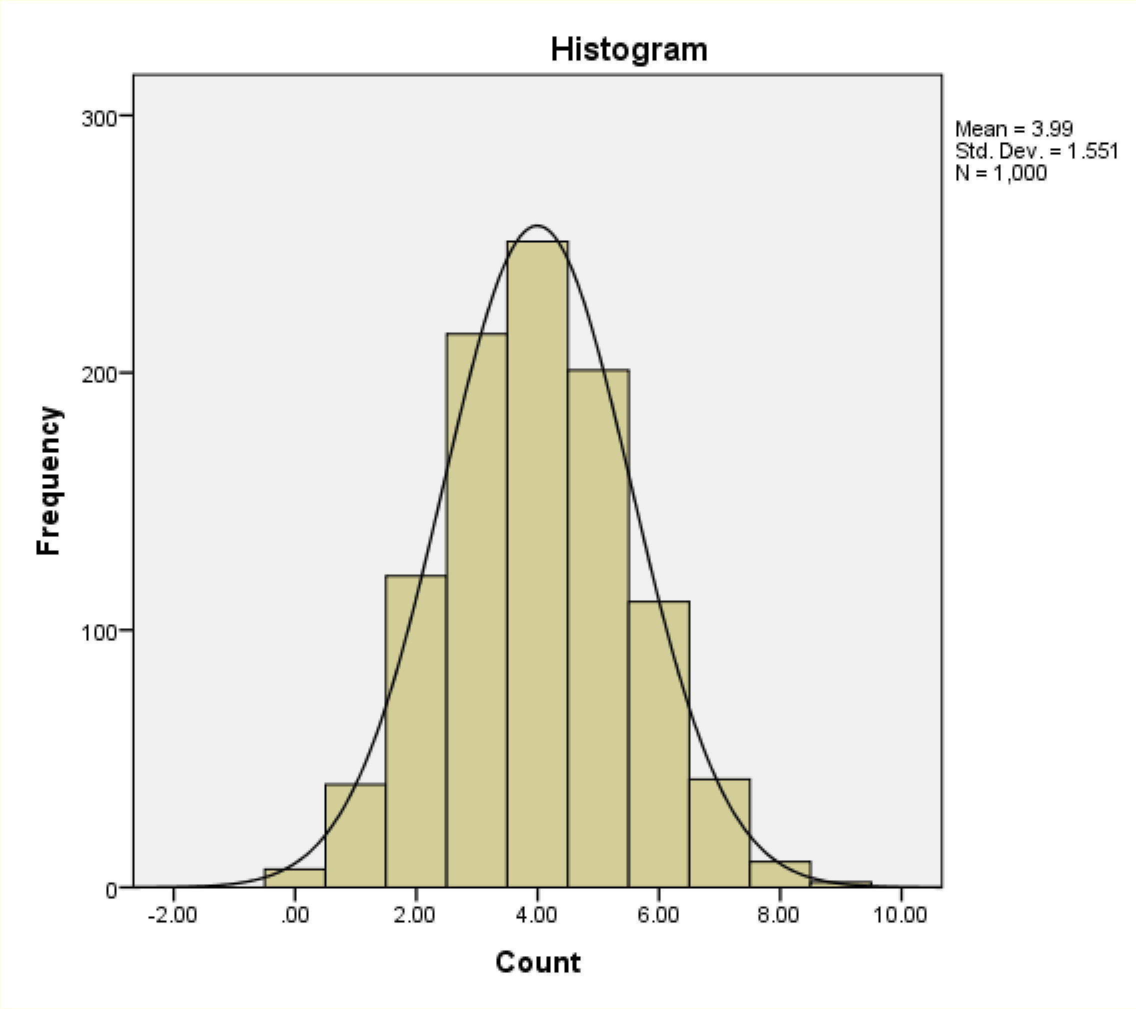




Statistics:
skewness / kurtosis

Charts: histograms
(show normal curve)

Assessing Skewness & Kurtosis



Median = 4

Mean = 3.99

If the Median & Mode are different then the Skewness … ?

Assessing Skewness & Kurtosis

Statistics

Count

N	Valid	1000
	Missing	0
Skewness		.120
Std. Error of Skewness		.077
Kurtosis		-.160
Std. Error of Kurtosis		.155

Limitations

These are not bounded metrics,

They can provide a measure of statistical significance, using the Z-score

Measuring Skewness

$$CM = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^r$$

Central Moment:
average of deviations
from the mean
(to the r power)

r = 1, first moment
r = 2, second moment
...

$$g_1 = \frac{1}{ns^3}\sum_{i=1}^{n}(Y_i - \bar{Y})^3$$

Third moment
(r = 3)

Measuring Kurtosis

$$CM = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^r$$

Central Moment:
average of deviations
from the mean
(to the r power)

r = 1, first moment
r = 2, second moment
…

$$g_2 = \left[\frac{1}{ns^4}\sum_{i=1}^{n}(Y_i - \overline{Y})^4\right] - 3$$

Fourth moment
(r = 4)

Assessing Normality - Tests

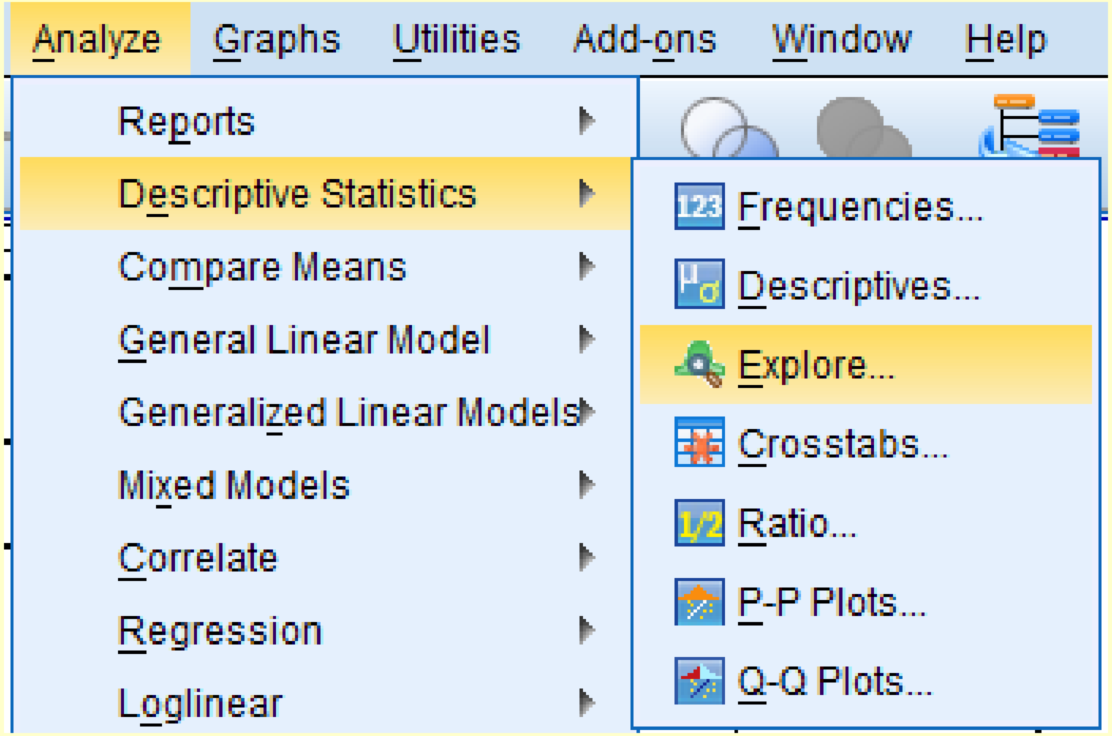
4) Performing Statistical Tests

- Calculate probabilities of observed data
  - Convert to z scores (subtract mean, dividing by *SE*)
- Kolmogorov-Smirnov & Shapiro – Wilk Tests
  - Test if data differ from a normal distribution

Assessing Normality - Tests

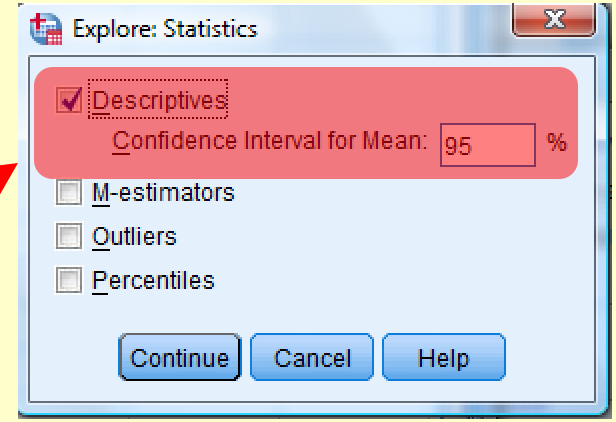
Part of Data Exploration :

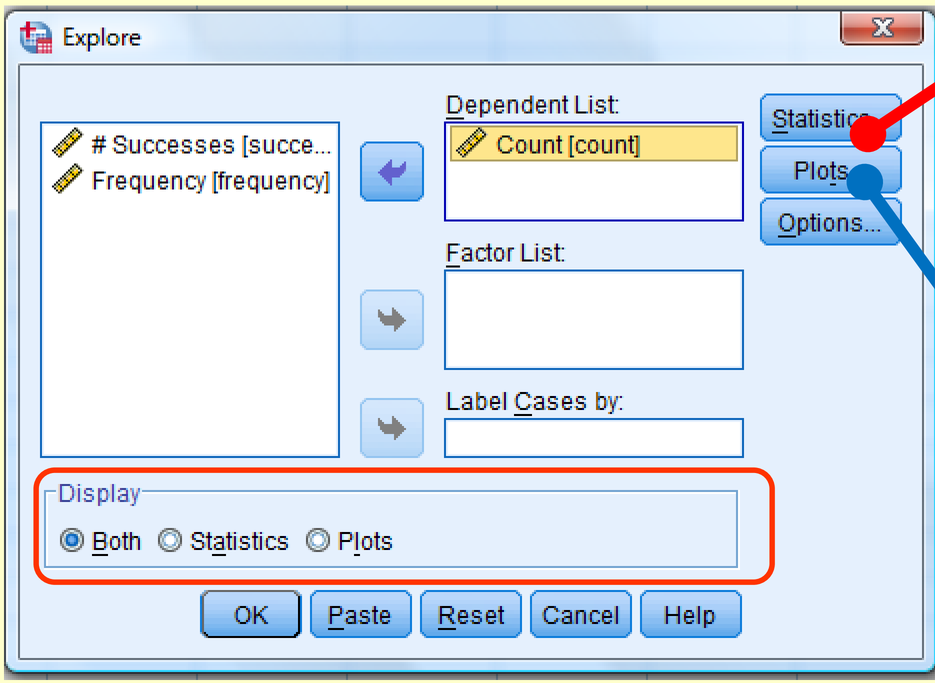
Under Descriptive Stats

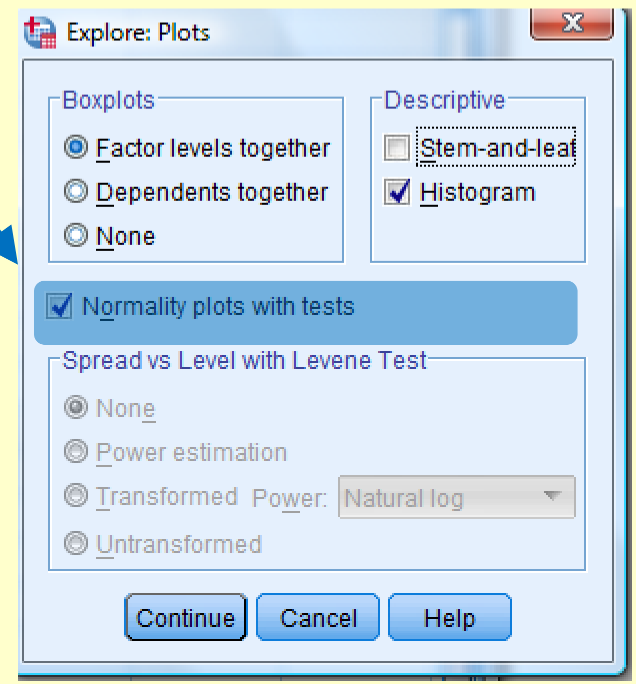


Assessing Normality - Tests

Descriptives (95% CI)







Normality plots with tests

Kolmogorov – Smirnov (K-S) Test

Nonparametric test for equality of continuous, one-dimensional probability distributions.

Used to compare a sample with a reference probability distribution (1-sample K–S test), or to compare two samples (2-sample K–S test).

The 2-sample K-S test is one of most useful nonparametric methods for comparing two samples, because it is sensitive to differences in the location (mean) and shape of the cumulative distribution functions of the two samples.

Kolmogorov – Smirnov (K-S) Test

The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution (in 2-sample case) or that the sample is drawn from the reference distribution (in 1-sample case).

In each case, the distributions considered under the null hypothesis are continuous distributions but are otherwise unrestricted.

Kolmogorov – Smirnov (K-S) Test

Note: K-S can be used to test any continuous distribution.

In special case of testing for normality of a distribution, samples are standardized and compared with a standard normal distribution.

This is equivalent to setting mean / S.D. of the reference distribution equal to sample estimates

Note: The K-S test is less powerful for testing normality than the Shapiro–Wilk test.

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.

Assessing Normality - Tests

Test if data differ from a normal distribution

Kolmogorov-Smirnov / Shapiro-Wilk

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Count	.132	1000	.000	.964	1000	.000

a. Lilliefors Significance Correction

Summary – Normality

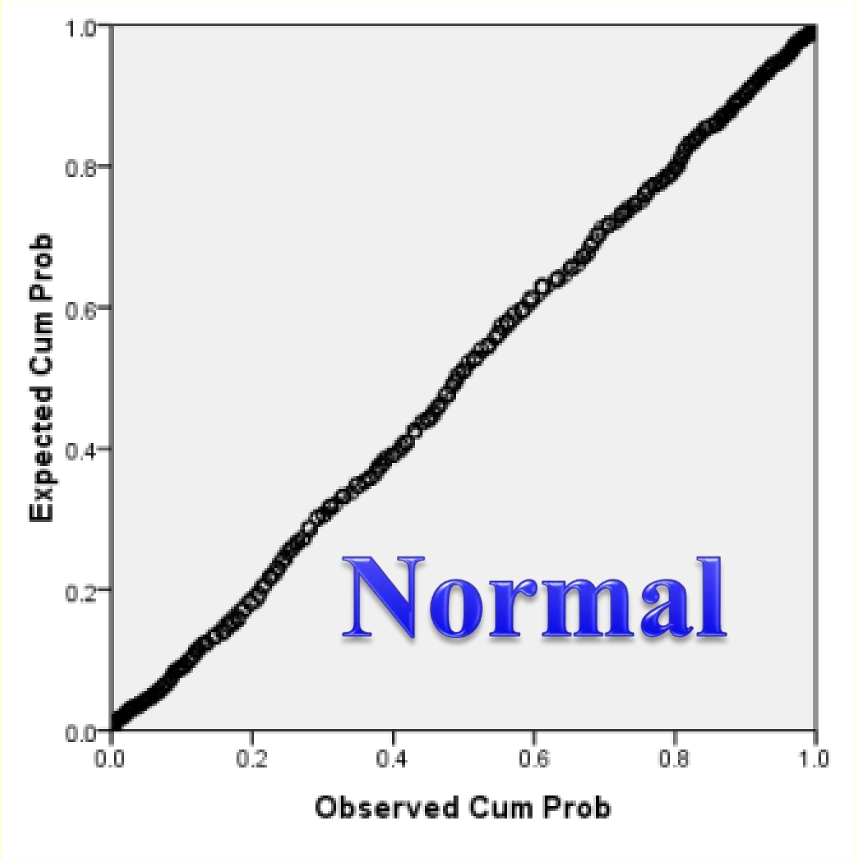
Indicators of a normal (Gaussian) distribution

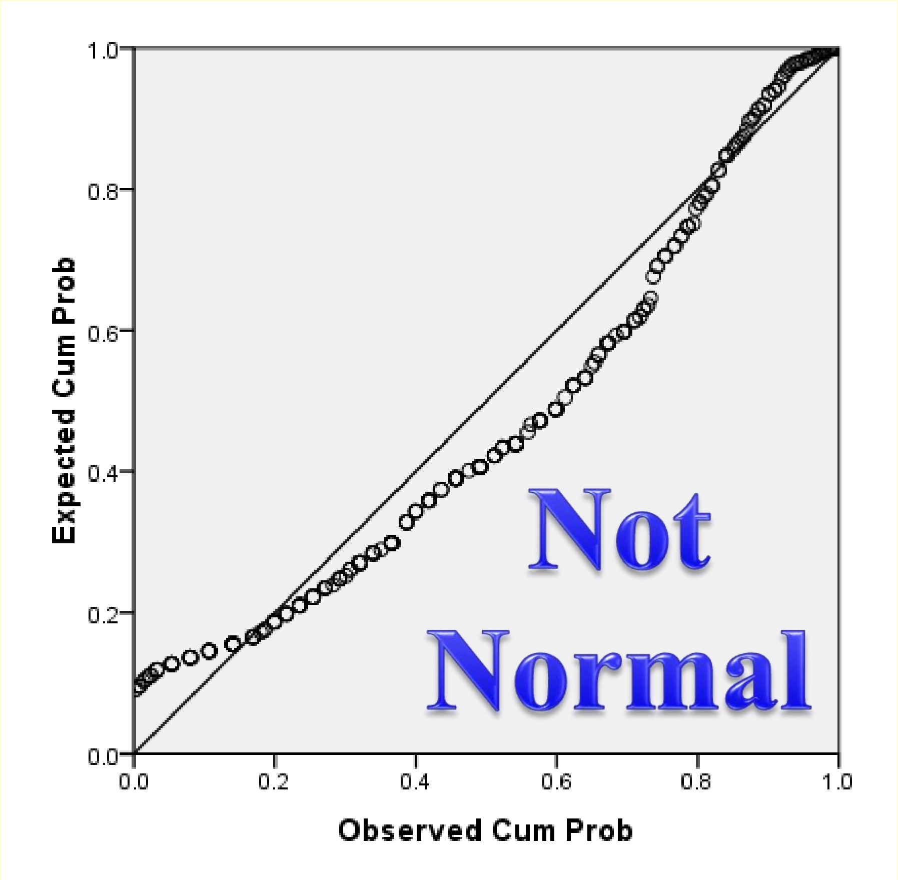
A. Mean = Median = Mode

B. Skewness: measures asymmetry of the distribution. A value of zero indicates symmetry. 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.

Assessing Normality - Graphically





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

Summary

- Parametric tests based on normal distributions
- 4 ways of Checking the assumption of normality
  - Graphical displays: P-P and Q-Q plots
  - Skew
  - Kurtosis
  - Normality tests: K-S and S-W tests
- Next Lecture: When and how to correct problems in the distribution of the data
  - Log, Square Root and Reciprocal Transformations
  - Pitfalls and alternatives