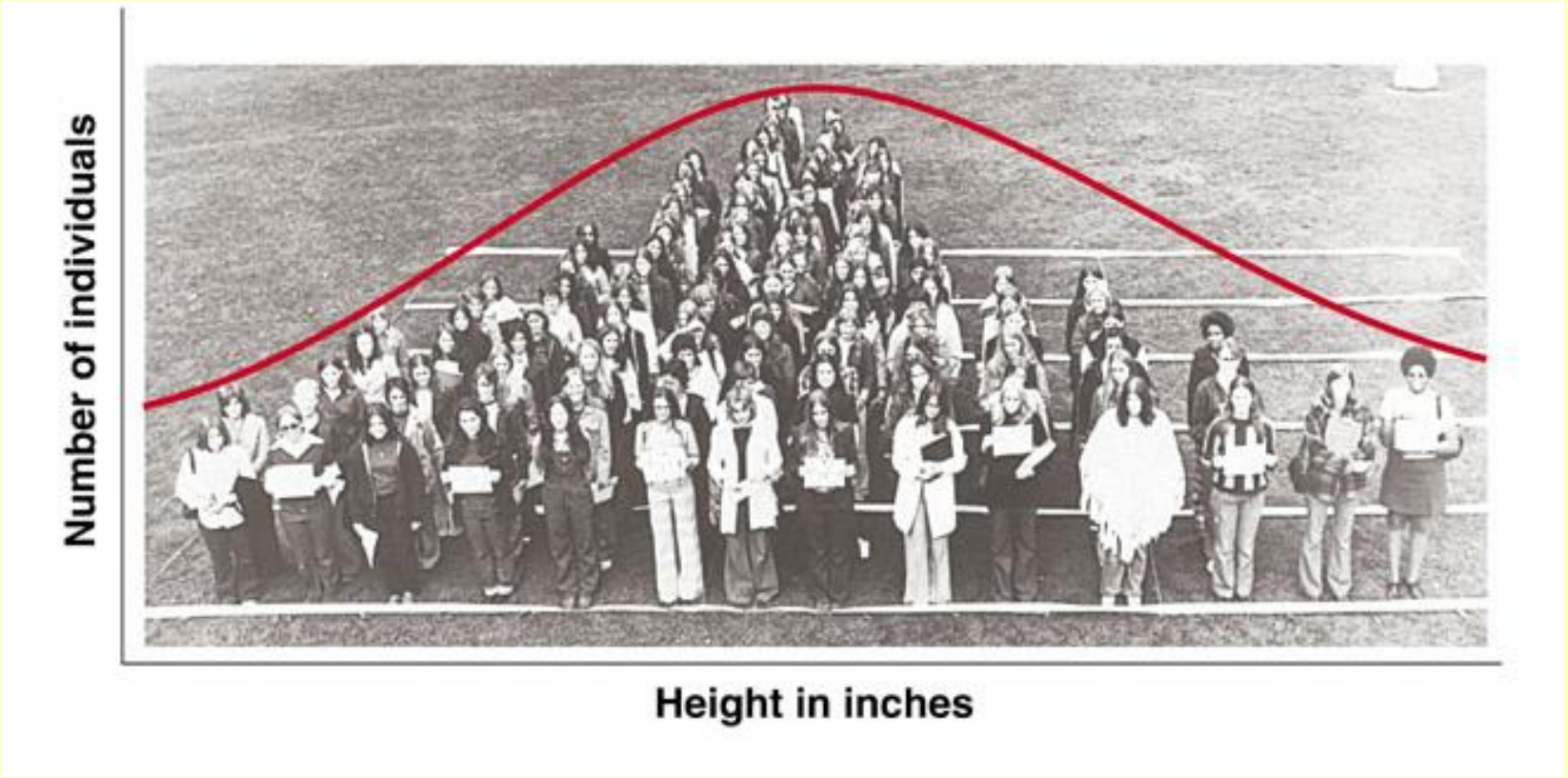
Data Distributions and Normality



Definition – (Non)Parametric

Parametric statistics assume that data come from a specific probability distribution (a normal distribution) and make inferences about parameters of the distribution.

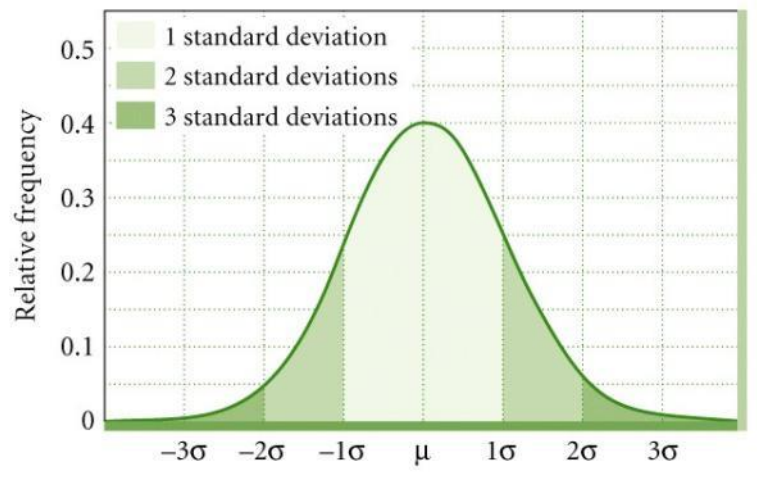
Non-parametric statistics involves:

- *distribution free techniques* do not rely on data belonging to a particular distribution. For instance, randomization tests, whereby observations are shuffled.

- *non-parametric statistics* whose interpretation does not depend on fitting any parameterized distribution. For instance, statistics based on ranks of observations are in the core of many non-parametric approaches.

The Normal Distribution

$X \sim N(\mu, \sigma)$



Every Normal Distribution can be described using only two parameters: Mean and S.D.

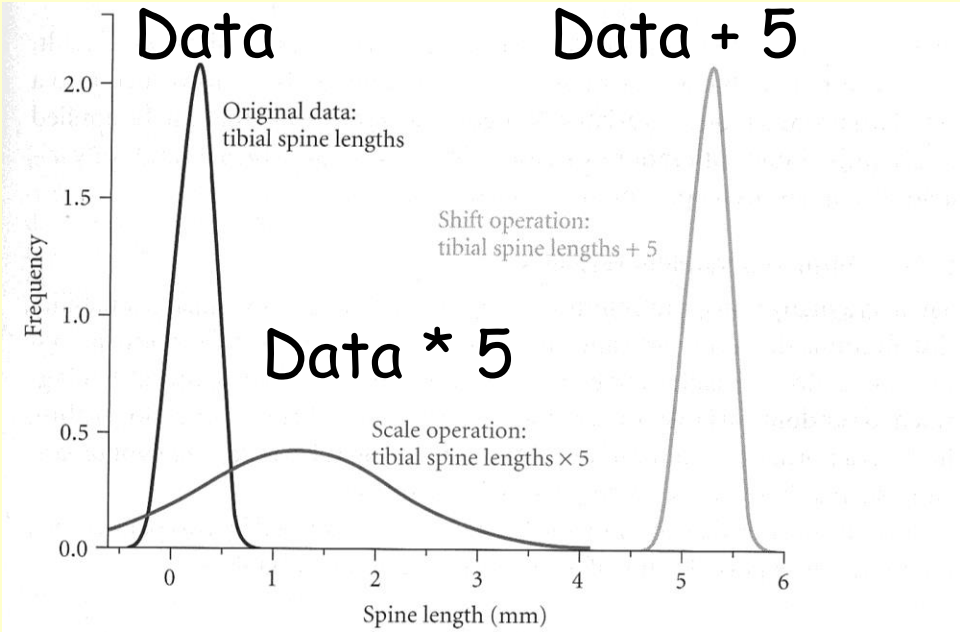


Figure 2.7 Shift and scale operations on a normal distribution. The normal distribution has two convenient algebraic properties. The first is a shift operation: if a constant b is added to a set of measurements with a mean μ, the mean of the new distribution is shifted to $\mu + b$, but the variance is unaffected. The black curve is the normal distribution fit to a set of 200 measurements of spider tibial spine length (Figure 2.6). The gray curve shows the shifted normal distribution after a value of 5 was added to each one of the original observations. The mean has shifted 5 units to the right, but the variance is unaltered. In a scale operation (green curve), multiplying each observation by a constant a causes the mean to be increased by a factor of a, but the variance to be increased by a factor of a^2. This curve is the normal distribution fit to the data after they have been multiplied by 5. The mean is shifted to a value of 5 times the original, and the variance has increased by a factor of $5^2 = 25$.

Quantifying Distributions

Distribution shapes categorized by symmetry (skew)

Skew: Measure of the symmetry of a distribution.

Symmetric distributions have a skew = 0.

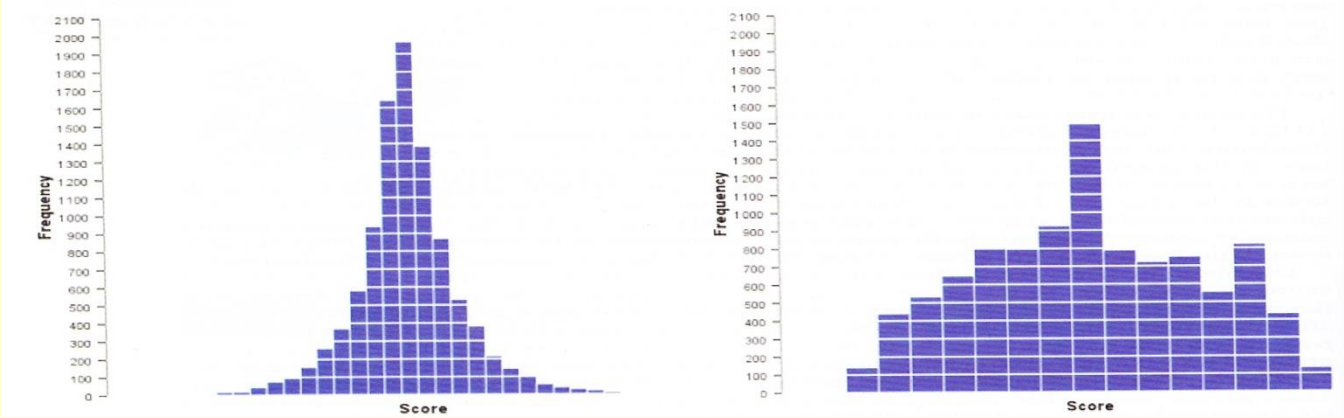
Positive skew:
the mean is larger than the median,
skewness > 0

Negative skew:
the mean is smaller than the median,
skewness < 0

Quantifying Distributions

Distribution shapes categorized by kurtosis

Kurtosis: Measure of the degree to which observations cluster in the tails or the center of the distribution.

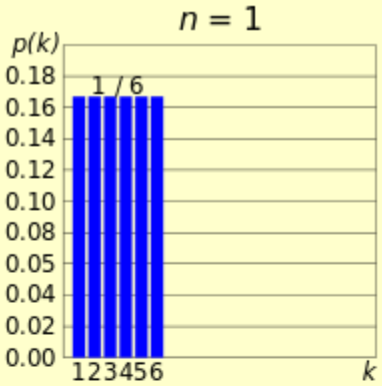


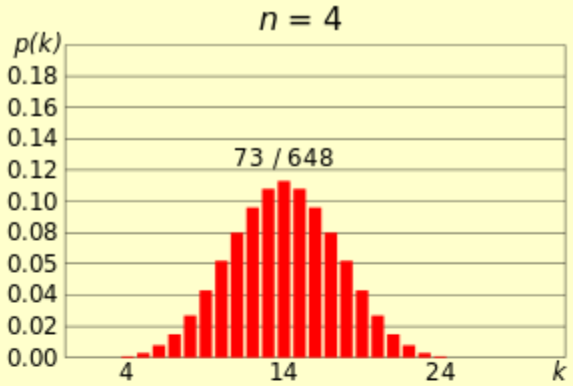
Positive kurtosis:
Less values in tails and more values close to mean.
Leptokurtic.

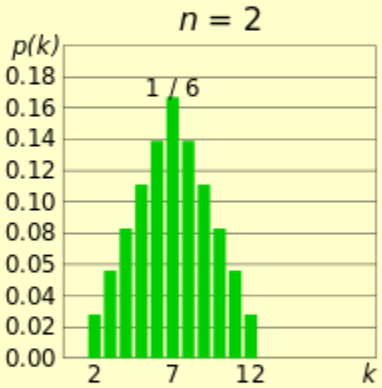
Negative kurtosis:
More values in tails and less values close to mean.
Platykurtic.

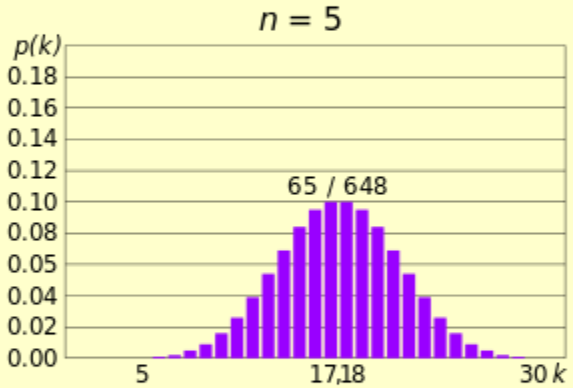
Central Limit Theorem

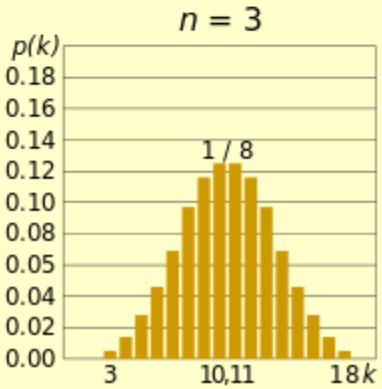
In probability theory, the **central limit theorem** states that the mean of a sufficiently large number of independent random variables, each with a finite mean and variance, will be normally distributed

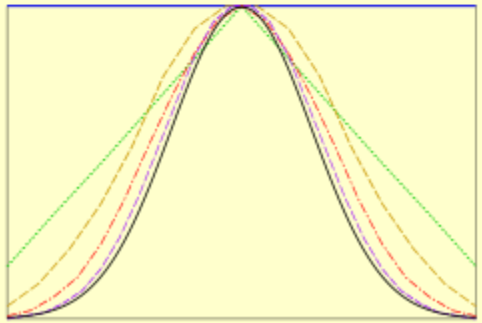












Assessing Normality

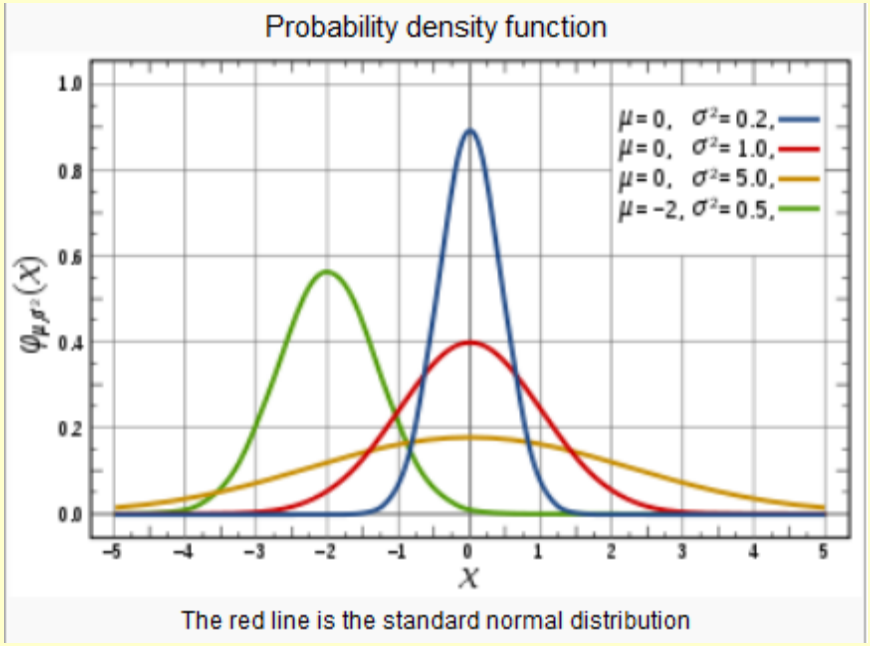
- We do not have access to sample the entire biological population, so we test observed data
- 1) Central Limit Theorem
  - If $N < 25$, sampling distribution rarely normal
- 2) Graphical Displays
  - Histogram
  - P-P Plot
- 3) Skewness / Kurtosis (point estimate +/- SE)
  - Do they overlap with 0 ?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

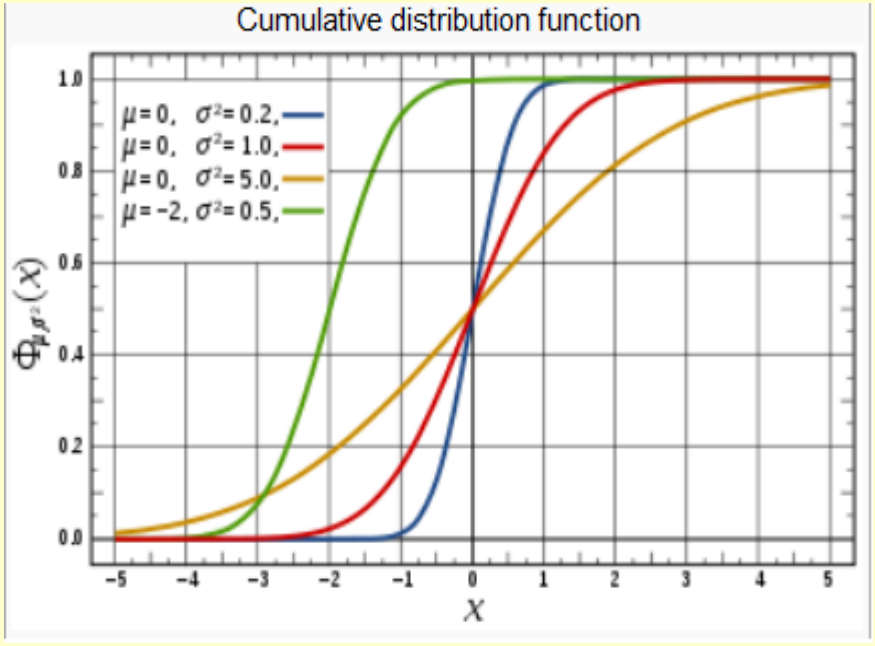
Assessing Normality

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





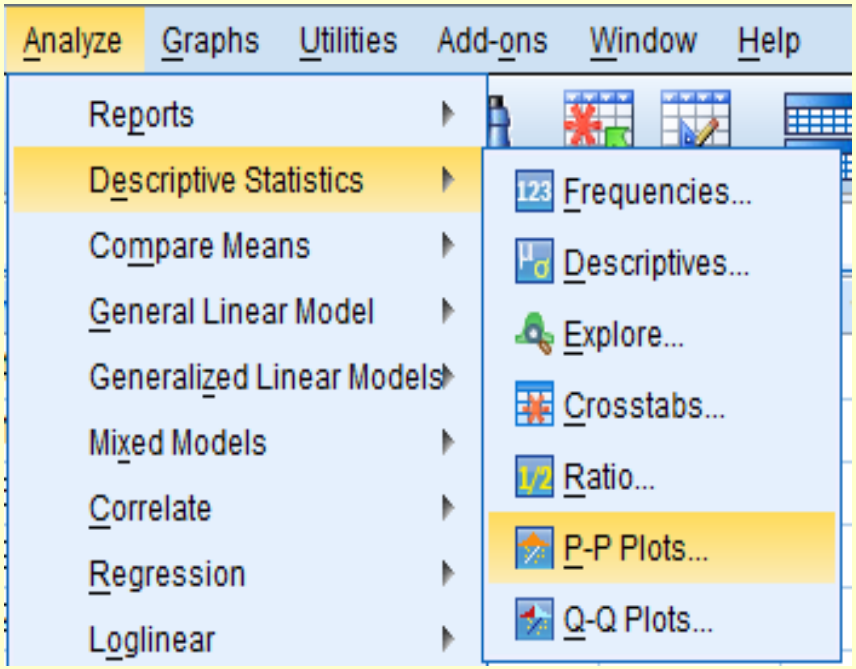
Comparing observations against a cumulative normal distribution (same mean and S.D.)

Assessing Normality - Graphically

Descriptive Statistics

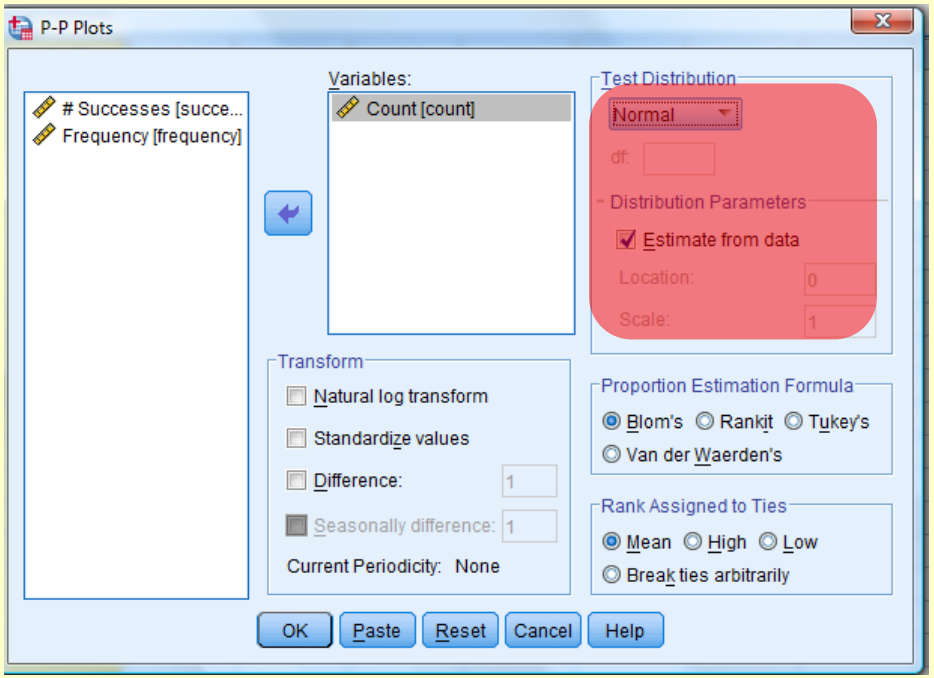
Tabular

Graphical



Test Distribution (df)

Select Parameters
(determined or estimated)

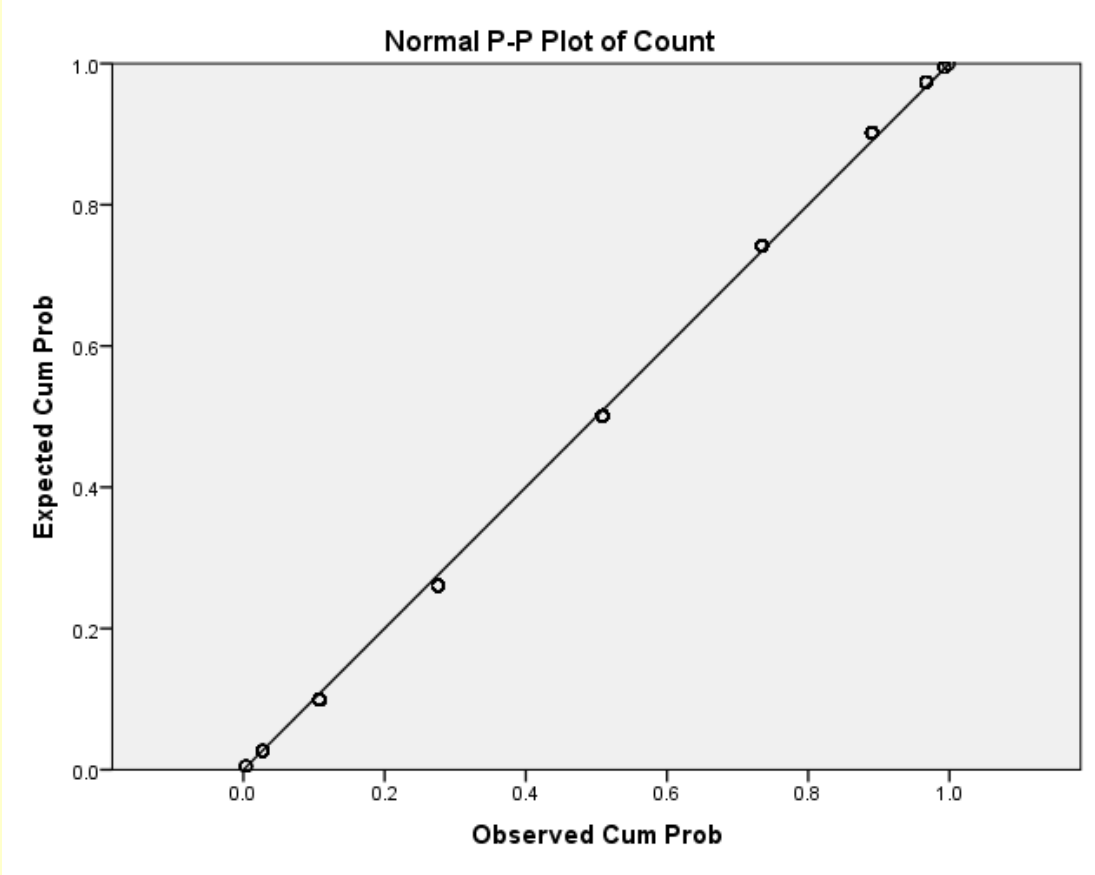


P – P Plots

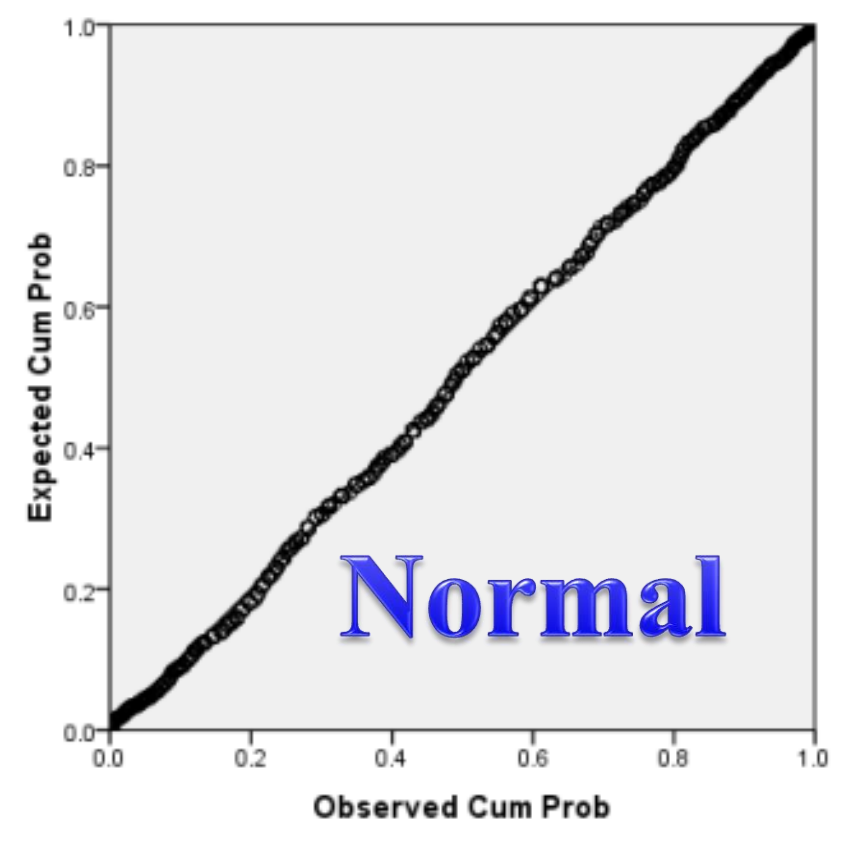
Visualizing deviations from cumulative frequency of the "normal"

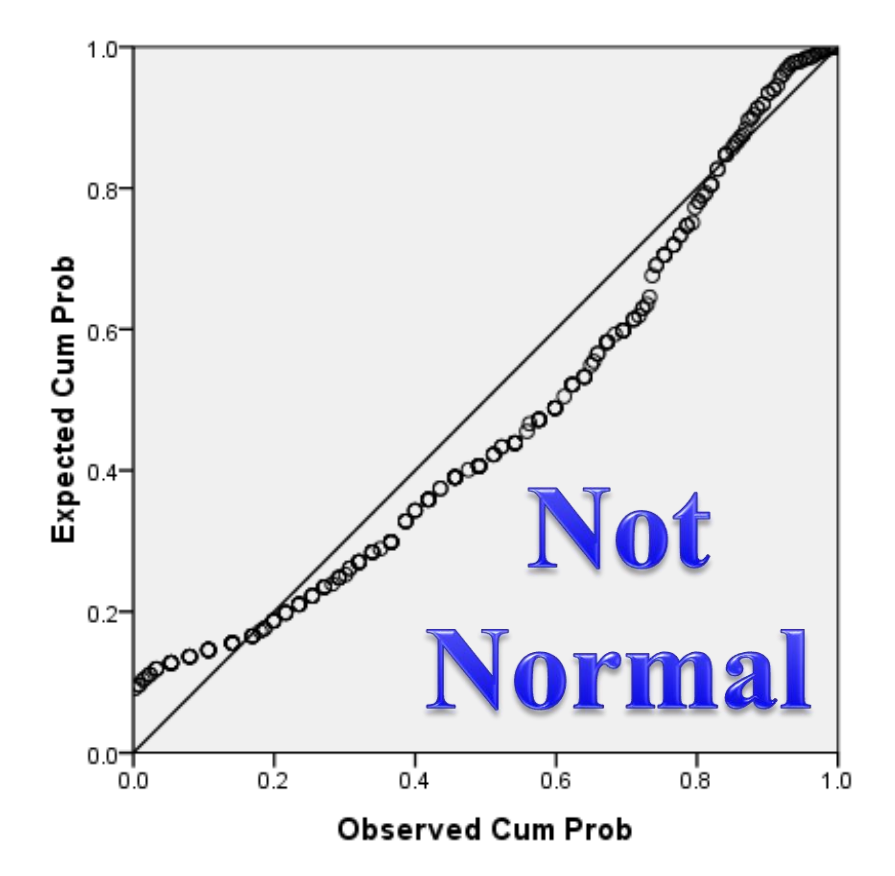
A **p - p plot** shows the cumulative frequencies of two datasets: the observed data and the expected data.

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



Assessing Normality - Graphically





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

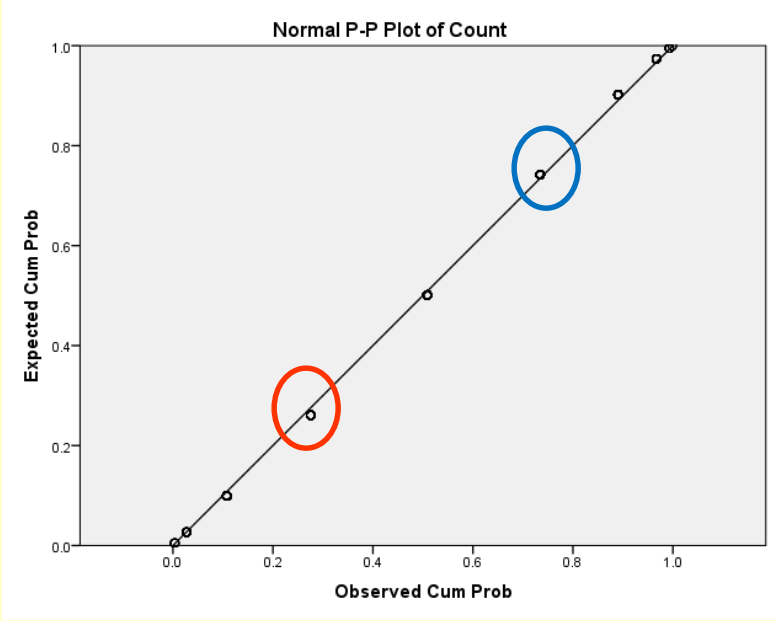
Assessing Normality - Graphically

Estimated Distribution Parameters

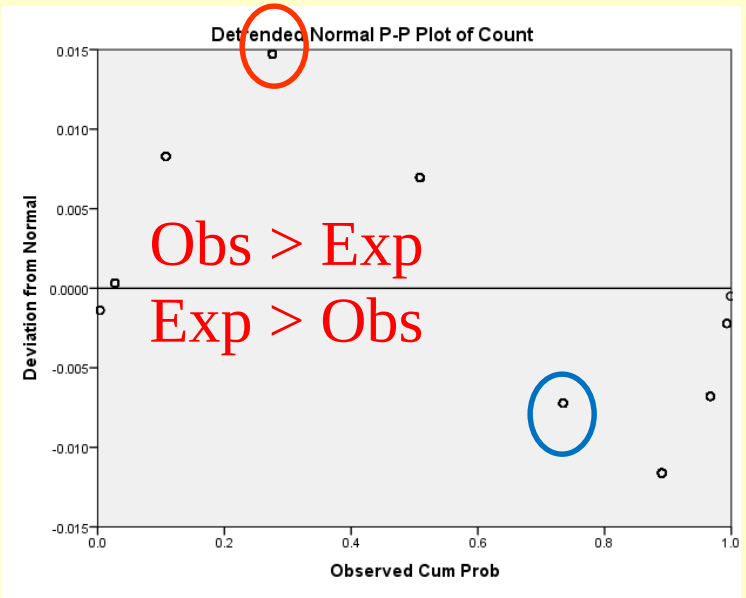
		Count
Normal Distribution	Location	3.9940
	Scale	1.55125

The cases are unweighted.

Mean
S.D.

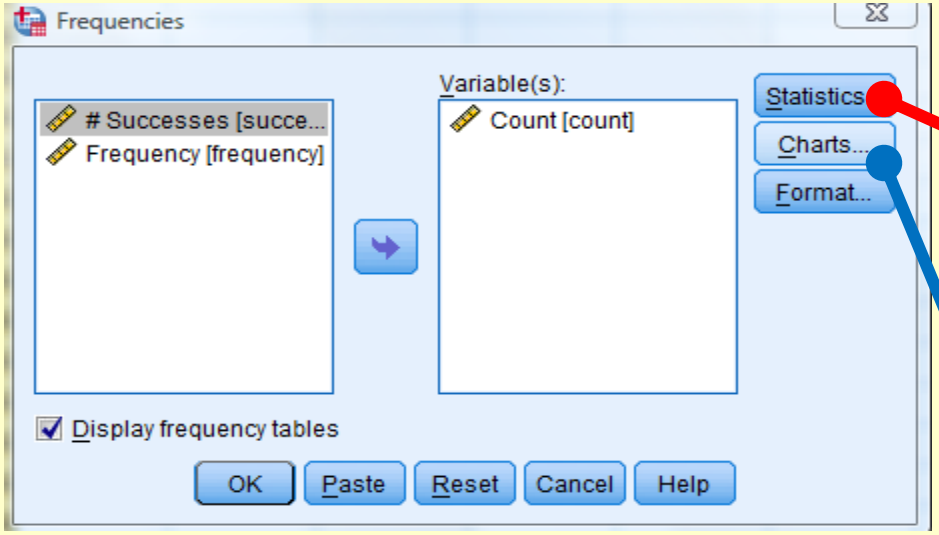


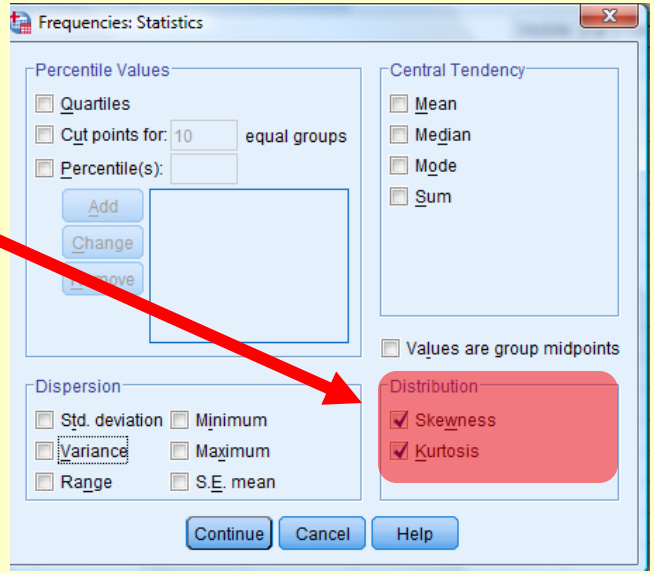
P-P plot (cumulative)

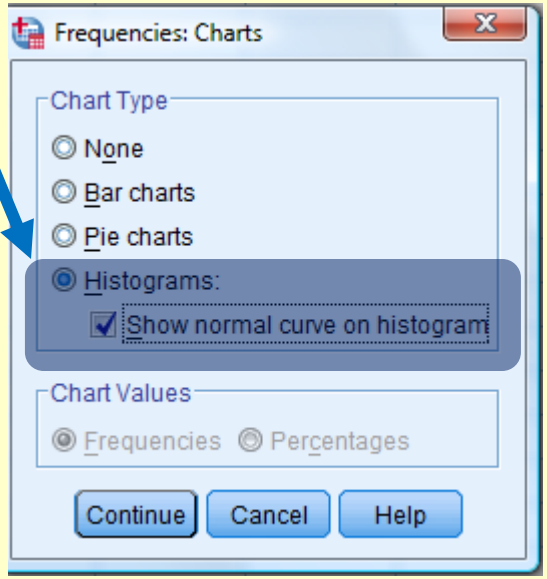


Deviations of P-P Plot

Assessing Normality – Skew & Kurtosis



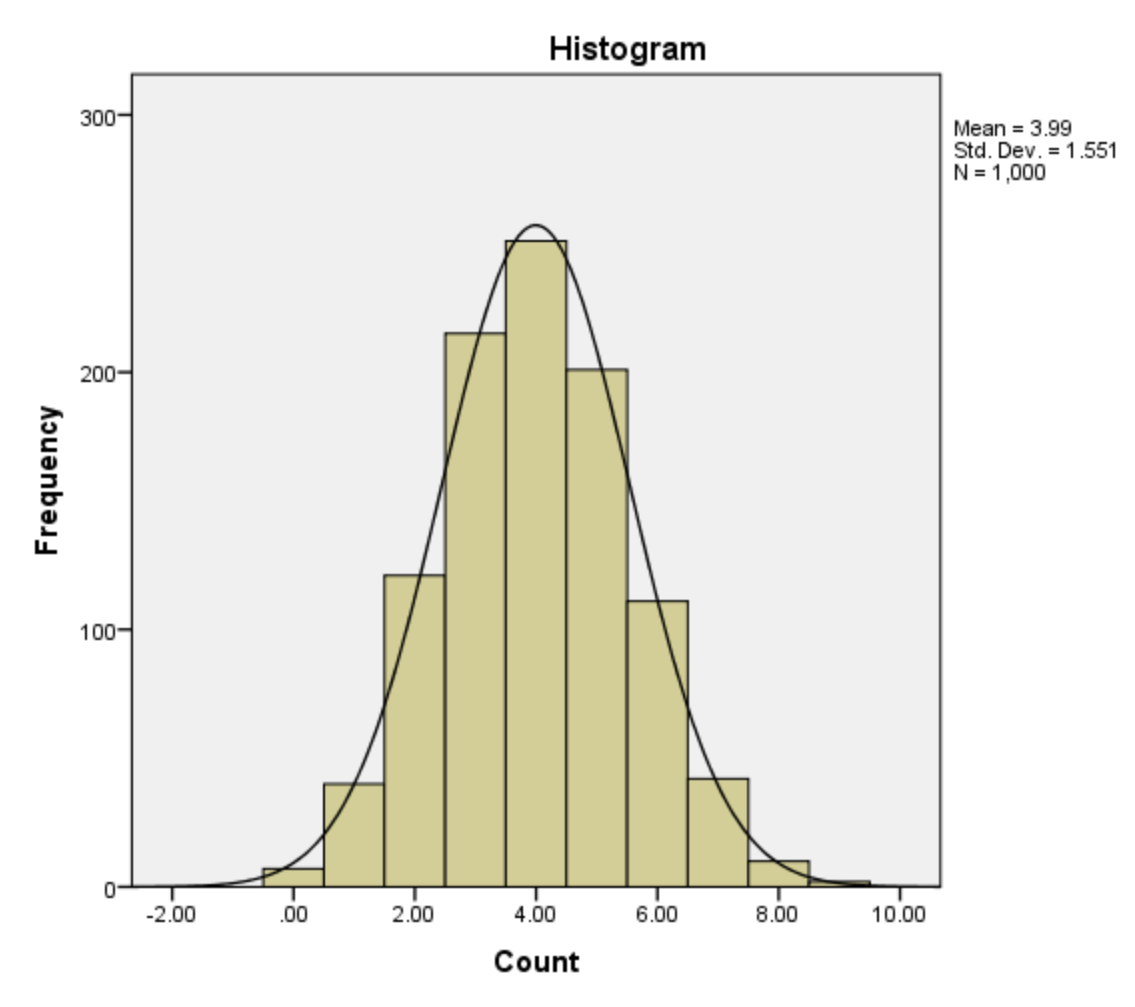




Statistics:
skewness / kurtosis

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

Assessing Normality - Tests

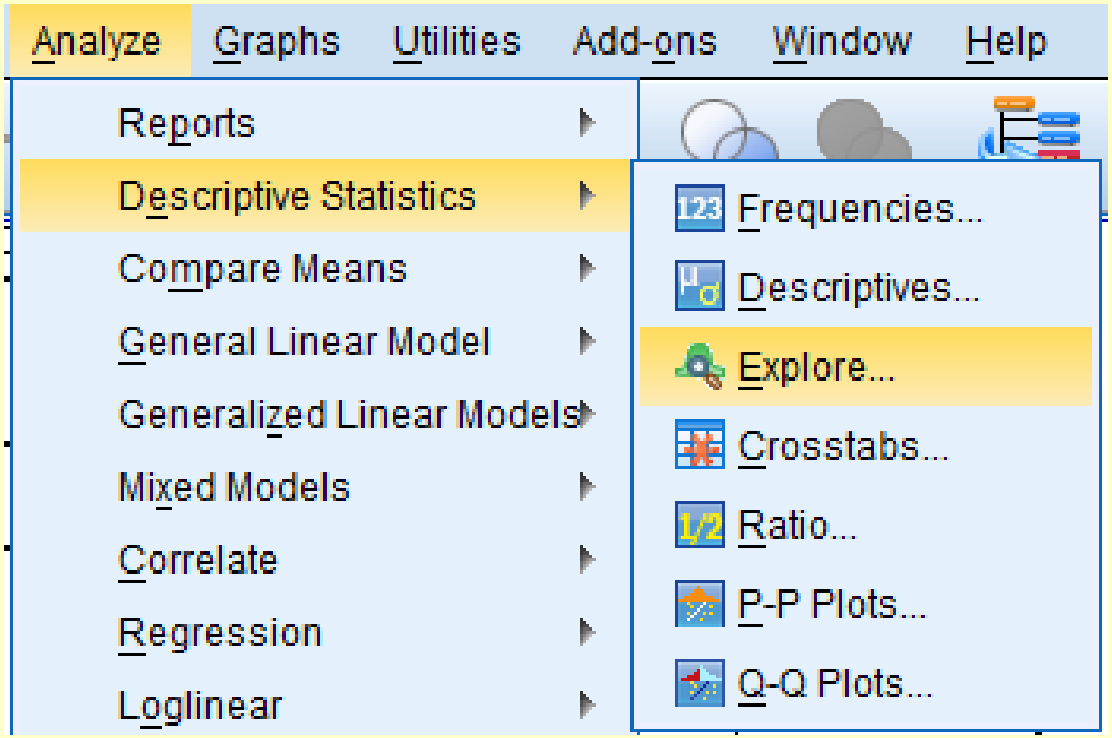
Performing Statistical Tests

- Kolmogorov-Smirnov & Shapiro – Wilk Tests
  - Test if data differ from a normal distribution

Assessing Normality - Tests

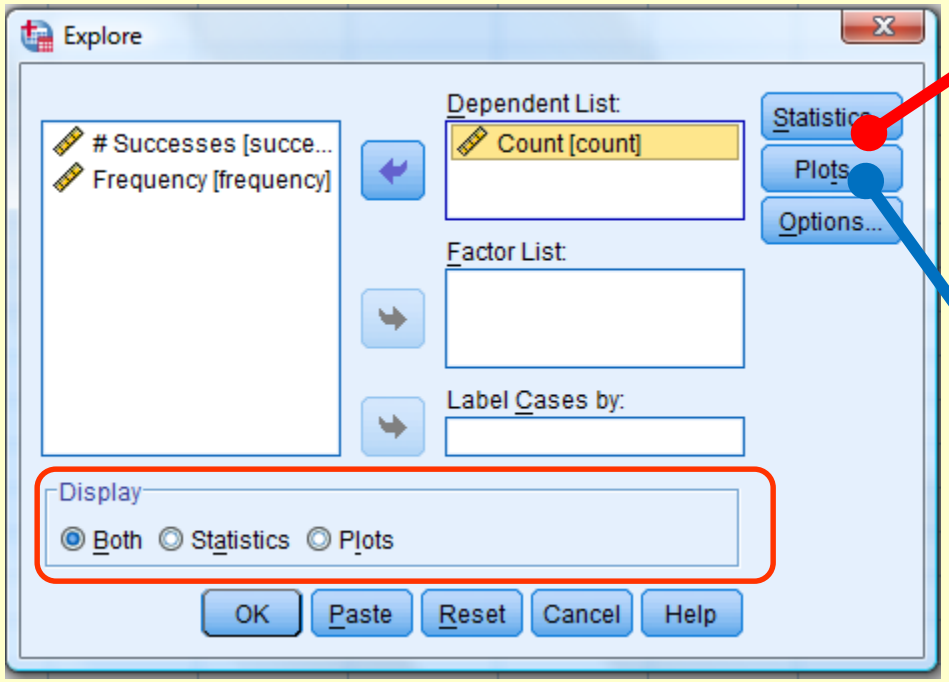
Part of Data Exploration :

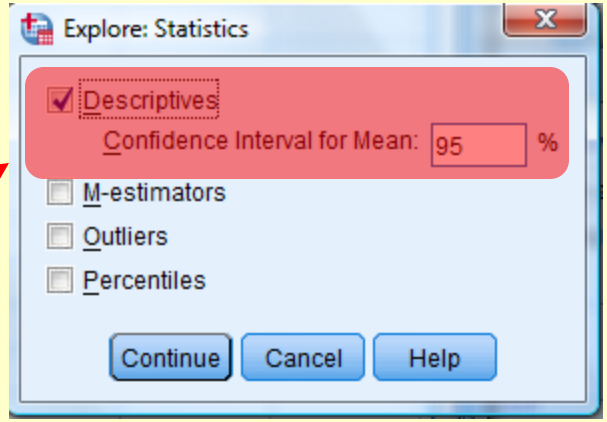
Under Descriptive Stats

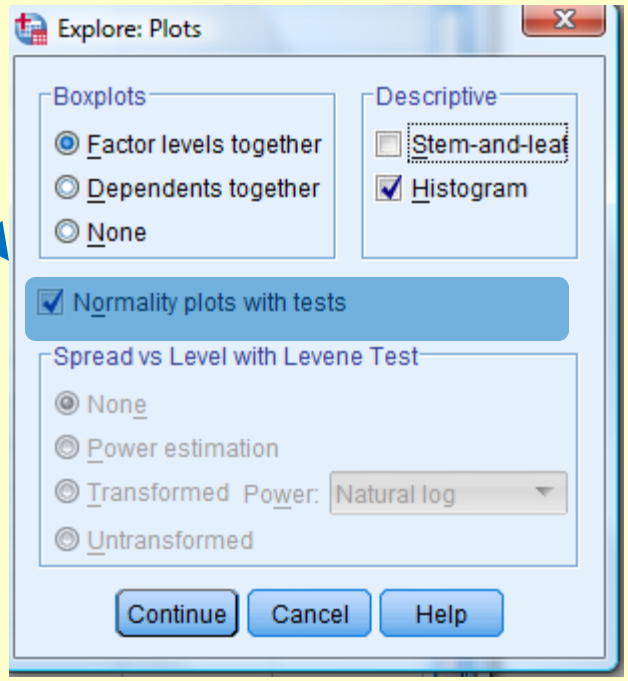


Assessing Normality - Tests

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 – Approach

Suggested Approach:

- Use parametric tests – whenever possible.

-Take care to examine diagnostic statistics and to determine if extra assumptions are met.

- Perform the matching non-parametric test and compare results. What causes disagreements?

- Remember: Because parametric statistics require a probability distribution, they are not distribution-free.

Summary - Parametric Statistics

Benefits and Costs:

- Parametric methods make more assumptions than non-parametric methods. If the extra assumptions are correct, parametric methods have more statistical power (produce more accurate and precise estimates.)

- However, if those assumptions are incorrect, parametric methods can be very misleading. They can cause false positives (type –I errors). Thus, they are often not considered robust.

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.

Summary – Assessing Normality

- Parametric tests based on normal distributions
- 4 ways of Checking the assumption of normality
  - Graphical displays: P-P
  - 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