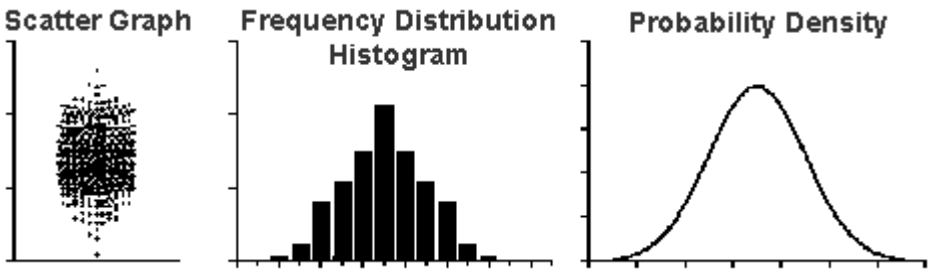
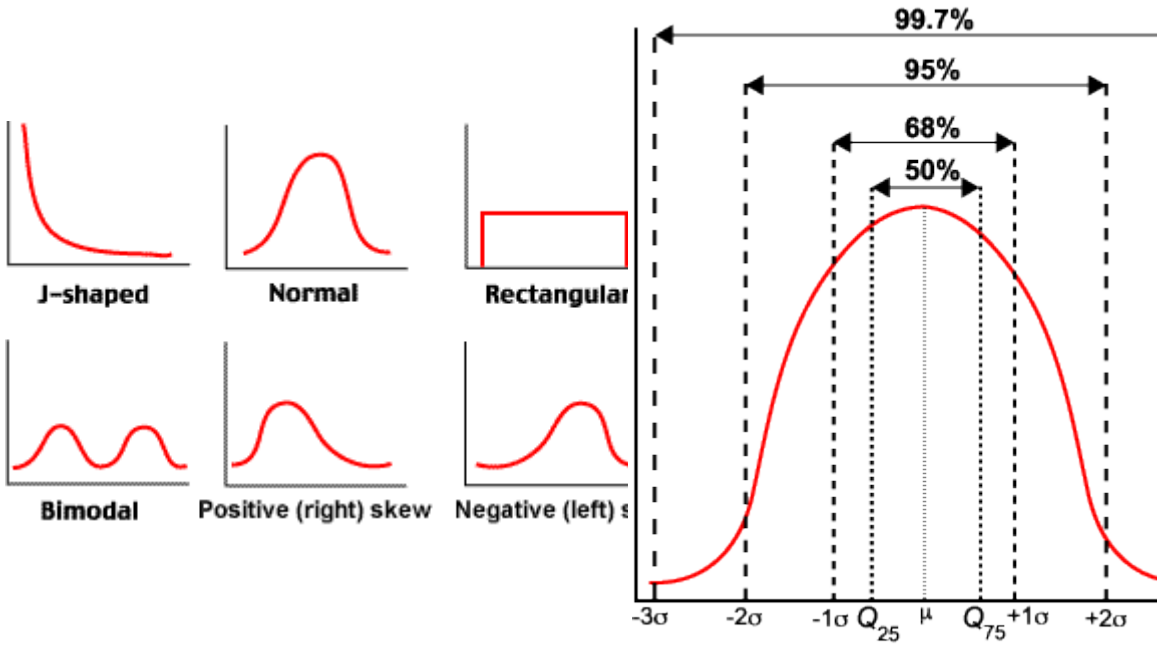
The Normal Frequency Distribution

When many independent random factors act in an additive manner to create variability, the dataset follows a bell-shaped distribution called the normal (or Gaussian distribution, after Carl Friedrich Gauss, 1777-1855):



The normal distribution has some **special mathematical properties** which form the basis of many statistical tests. Although no real datasets follow the normal distribution exactly, many kinds of data follow a distribution that is approximately Gaussian. A normal distribution can be defined by two parameters, the **mean** and the **standard deviation**. By definition, normal frequency distributions are continuous (not bimodal). Of course, not all datasets follow a normal distributions, e.g.

- Binomial distribution: A probability distribution of binary variables. A binary variable is one which can take one of two possible values, e.g. male or female, left or right-handed, heads or tails, etc - the two states of a binary variable are mutually exclusive. Binary variables are quite common in biological data.
- Poisson distribution: The Poisson Distribution is a discrete probability distribution that applies to occurrences of some event over a specific interval. The random variable X is the number of occurrences of the event in an interval. The interval can be time, distance, area, volume or some similar unit. This method can be used to model many events, e.g. the number of orchids in a field or the probability of your computer crashing.



How to recognize a normal (& non-normal) distribution:

1. In a perfect normal frequency distribution, the mean, median and mode are equal. The data is continuous and symmetrically distributed around the central point. This does not mean that there are no outliers, but the data is no bimodal (or multimodal).
2. In a perfect normal frequency distribution:
    - 68% of samples fall between ± 1 standard deviations from the mean
    - 95% of samples fall between ± 2 standard deviations from the mean
    - 99.7% of samples fall between ± 3 standard deviations from the mean
3. Kolmogorov-Smirnov & Shapiro-Wilk tests: statistical methods which determine whether one distribution is significantly different from another.
4. Normal Probability Plots: PP and QQ plots

Fortunately, SPSS contains powerful, easy to use tools which make it easy to assess frequency distributions and help you make decisions about which tests to use (see below).

Parametric & Nonparametric Methods:

Statistical methods which depend on the parameters of populations or probability distributions and are referred to as **parametric methods**. Parametric tests include:

- t-test
- ANOVA

- many others

These tests are only meaningful for numerical data which is sampled from a population with an underlying **normal distribution** or whose distribution can be **rendered normal by mathematical transformation**.

Nonparametric methods require fewer assumptions about a population or probability distribution and are applicable in a wider range of situations:

1. Methods used with **qualitative** data.
   or:
2. Methods used with **quantitative** data when **no assumption can be made about the population probability distribution**.

Nonparametric methods are useful in situations where the assumptions required by parametric methods appear questionable. A few of the more commonly used nonparametric methods include:

- Chi-squared test
- Spearman rank correlation coefficient
- Wilcoxon signed-rank test
- Mann-Whitney-Wilcoxon test

These tests are characterised as distribution free - i.e. neither the values obtained nor the population from which the sample was drawn need have a normal distribution. Unlike the parametric tests which can give erroneous results these can always be used safely regardless of the distribution of the data.

Unfortunately, they are less flexible in practice and less powerful than parametric tests. In cases where both parametric and nonparametric methods are applicable, **statisticians usually recommend using parametric methods** because they tend to provide better precision.

Exploratory Data Analysis (EDA)

Why perform EDA?

- To reveal possible errors in the data, e.g. outliers.
- To reveal features of the dataset, e.g. symmetry, skew, scatter.
- To test for a normal distribution.
- To determine whether parametric or non-parametric tests should be used.

Much of statistics is about detecting patterns - something which the human eye and brain are very good at. EDA shows you the patterns which are hidden when the data is in numerical form.

Never, ever, run any statistical test without performing EDA first!

EDA includes:

- **Descriptive statistics** (numerical summaries): mean, median, range, variance, standard deviation, etc. In SPSS choose **Analyze: Descriptive Statistics: Descriptives**.
- **Kolmogorov-Smirnov & Shapiro-Wilk tests**: These methods test whether one distribution (e.g. your dataset) is significantly different from another (e.g. a normal distribution) and produce a numerical answer, yes or no. Use the Shapiro-Wilk test if the sample size is between 3 and 2000 and the Kolmogorov-Smirnov test if the sample size is greater than 2000. Unfortunately, in some circumstances, both of these tests can produce misleading results, so "real" statisticians prefer graphical plots to tests such as these.

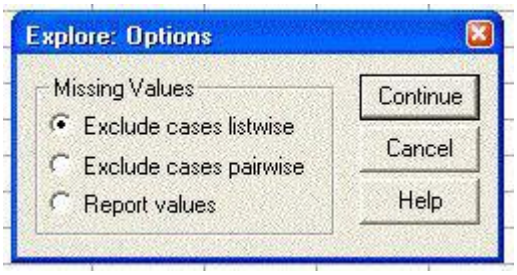
- **Graphical methods**:
    - frequency distribution histograms
    - stem & leaf plots
    - scatter plots
    - box & whisker plots
    - Normal probability plots: PP and QQ plots
    - Graphs with error bars (Graphs: Error Bar)

The good news is that SPSS makes EDA very easy:

- **Analyze: Descriptive Statistics: Explore** Select: **Both** (Statistics and Plots) and under Plots, select **Normality plots with tests.**
- **Graphs: Histogram**, select **Display normal curve**.
- **Graphs: P-P**, select **Test Distribution: Normal.**
    - Normal probability plots (P-P plots) show the expected statistics from a sample taken from the Normal distribution with mean 0 and variance 1 against the data values ordered from smallest to largest. If the data inspected have a Normal frequency distribution, a plot of the data against the expected statistics should produce a straight line. SPSS also produces a variant of the probability plot called a Q-Q plot, or quantile-quantile plot in which percentiles of the standard normal distribution are plotted against percentiles of the data. Both P-P and Q-Q plots highlight departures from Normality. The Q-Q plot is more sensitive to deviances from normality in the tails of the distribution, whereas the normal probability plot is more sensitive to deviances near the

mean of the distribution. If the observed points curve above or below the normal plot line, this indicates the kurtosis departs from a normal distribution, whereas if the observed plot is S-shaped, this shows the data is skewed.

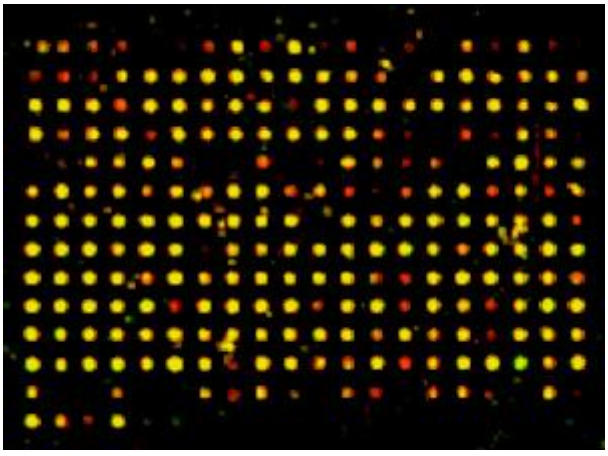
Important: Treatment of Missing Values:
If you run **Analyze: Descriptive Statistics: Explore**, SPSS will "Exclude cases listwise". What this means in practice is that all the variables (columns) will be treated as if they have the same number of cases (rows). If this is true then there's no problem, but if not, then some of the EDA statistics will be wrongly calculated. To avoid this problem, in the **Analyze: Descriptive Statistics: Explore** dialog, click the **Options** button and select: **Exclude cases pairwisewise**:



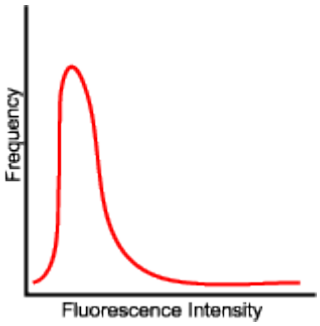
EDA on what?

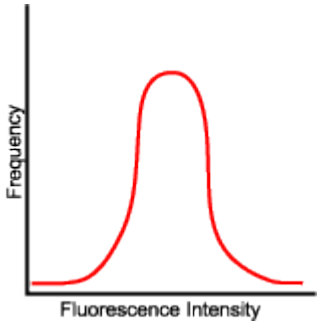
As a first approximation, it may be useful to perform EDA on an entire datset to get a quick feeling for what the data looks like. However, if you are going to be comparing groups (subsets of the data) using a parametric test, you need to ensure that each of the groups being compared has a normal frequency distribution, so you need to perform EDA on all the groups separately.

Transforming Data



In the results from DNA microarray experiments, many of the hybridized spots have very low fluorescence. However, a few of the spots have very strong fluorescence. This produces a strong positively-skewed distribution:





In this case, the relatively small number of high-intensity samples distorts the results. Since this is not a normal distribution, calculating the mean is misleading and statistical tests based on the normal distribution would give meaningless answers. The solution is to divide the fluorescence intensity of each sample by the **median** fluorescence intensity. Plotting the result of this gives a normal distribution.

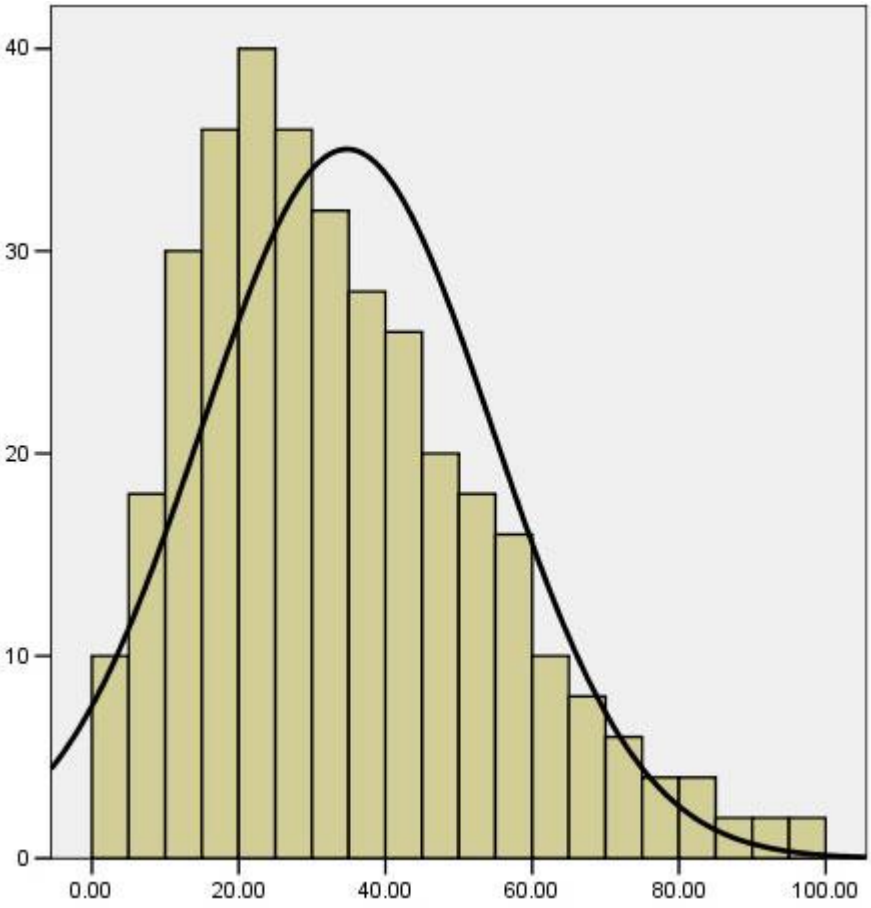
Different sorts of mathematical transformation work best for different datasets:

- Calculating the **square root** (Sqrt) of each datapoint may help to normalize weakly skewed datasets.
- Calculating the **logarithm** (Log10/Ln) of each datapoint helps to normalize more strongly skewed datasets ("log-normal" distribution). Any type of log will do this.
- Calculating the **reciprocal** (1/var) of of each datapoint may produce a normal distribution from an exponential dataset.

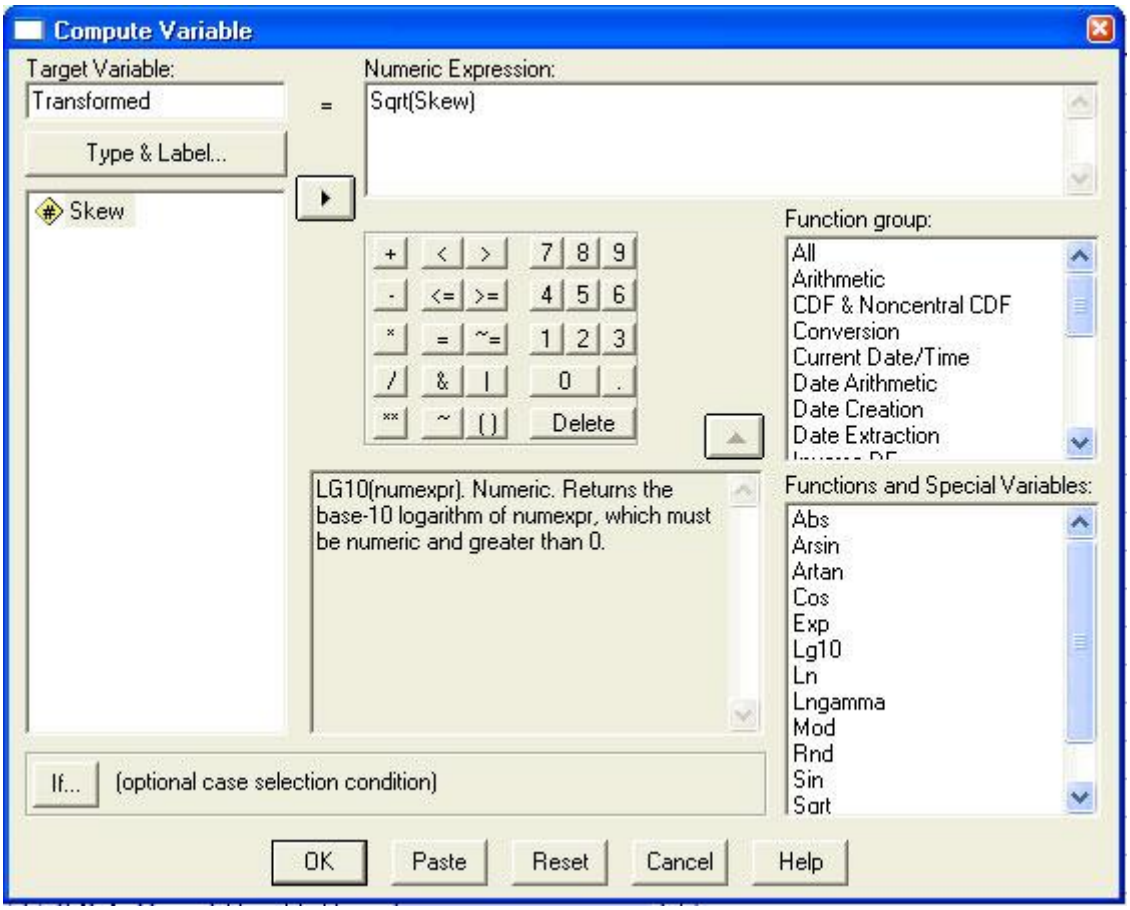
While it is worth remembering the above suggestions, transforming a dataset is an empirical exercise - perform several of the most likely transformations and test for a normal distribution by EDA. People get understandably concerned about transforming data. Transforming data to allow you to use parametric statistics is completely legitimate as log as:

- You clearly document what transformations you have performed.
- You do not forget that you are no longer working with raw data!
- You do not try to compare a transformed dataset with a raw dataset, e.g. if you want to do a t-test on two datasets and one is non-normal, you must apply the same transformation to both datasets!

Read this excellent article. To transform variables in SPSS:**Transform: Compute**, and select the options you want to construct a new variable, e.g. **Ln, Lg10, Sqrt, 1/[var]**, etc:



Transform: Compute:


Compute Variable
Target Variable:
Transformed
Type & Label...
Skew
=
Numeric Expression:
Sqrt(Skew)
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) Delete
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Date Extraction
LG10(numexpr). Numeric. Returns the base-10 logarithm of numexpr, which must be numeric and greater than 0.
Functions and Special Variables:
Abs
Arsin
Artan
Cos
Exp
Lg10
Ln
Lngamma
Mod
Rnd
Sin
Sqrt
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help

becomes: