

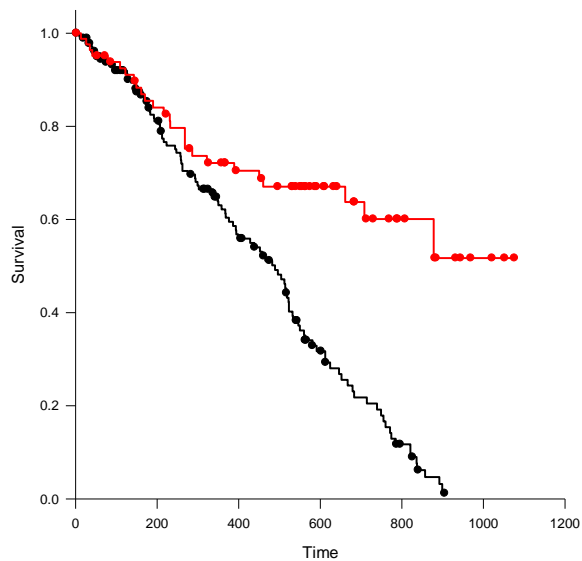
# New Statistics and Transform Language Features in SigmaPlot 11

## Major Statistical Test

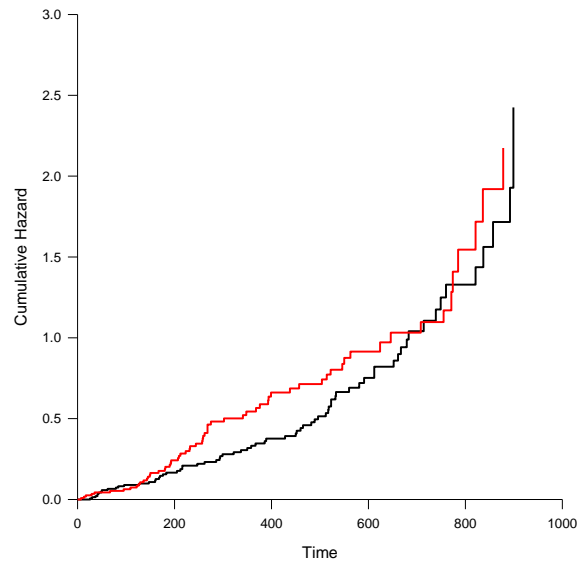
Cox Regression – Includes the *proportional-hazards* model with stratification to study the impact of potential risk factors on the survival time of a population. Input data can be categorical.

Examples of Cox Regression Result Graphs:

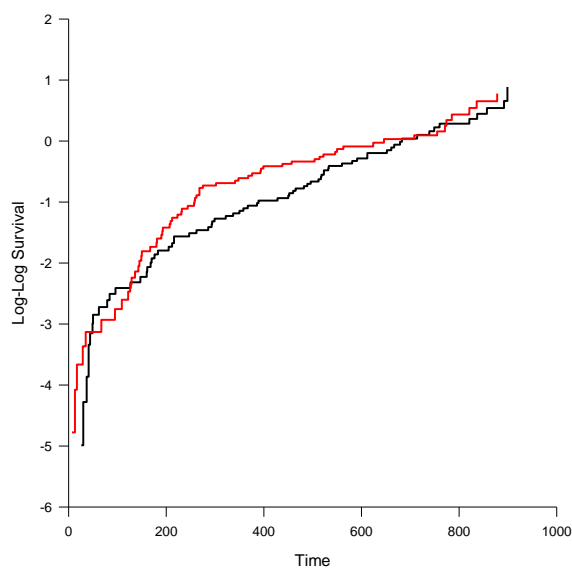
**Adjusted Survival Curves**



**Cumulative Hazard Curves**



**Log-Log Survival Curves**



## Minor Statistical Tests

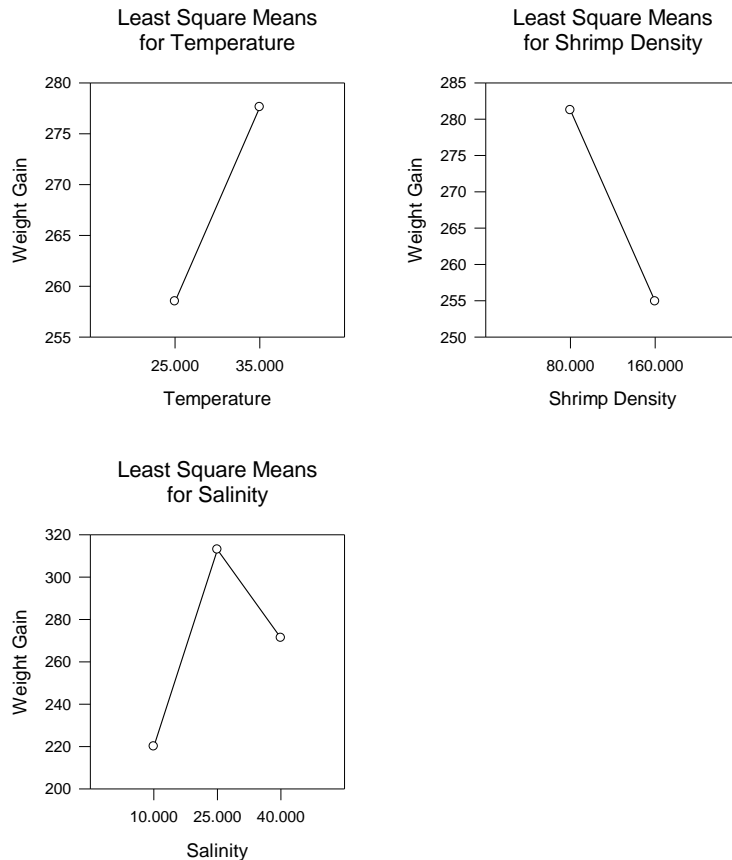
- One-Sample T-test – Tests the hypothesis that the mean of a population equals a specified value.
- Odds Ratio and Relative Risk tests – Both tests the hypothesis that a treatment has no effect on the rate of occurrence of some specified event in a population. Odds Ratio is used in *retrospective* studies to determine the treatment effect after the event has been observed. Relative Risk is used in *prospective* studies where the treatment and control groups have been chosen before the event occurs.
- Shapiro-Wilk Normality test – A more accurate test than Kolmogorov-Smirnov for assessing the normality of sampled data. Used in assumption checking for many statistical tests, but can also be used directly on worksheet data.

## Result Graphs

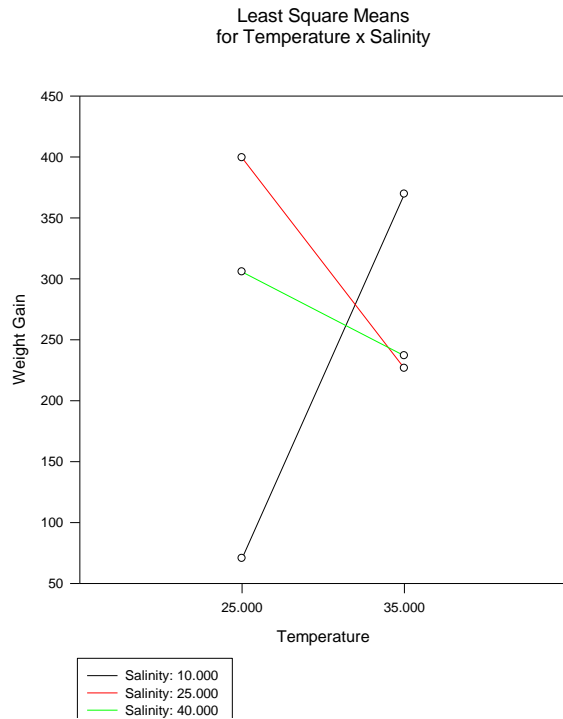
ANOVA Profile Plots– Used to analyze the main effects and higher-order interactions of factors in a multi-factor ANOVA design by comparing averages of the least square means.

Examples of Profile Plots for a 3-Way ANOVA design:

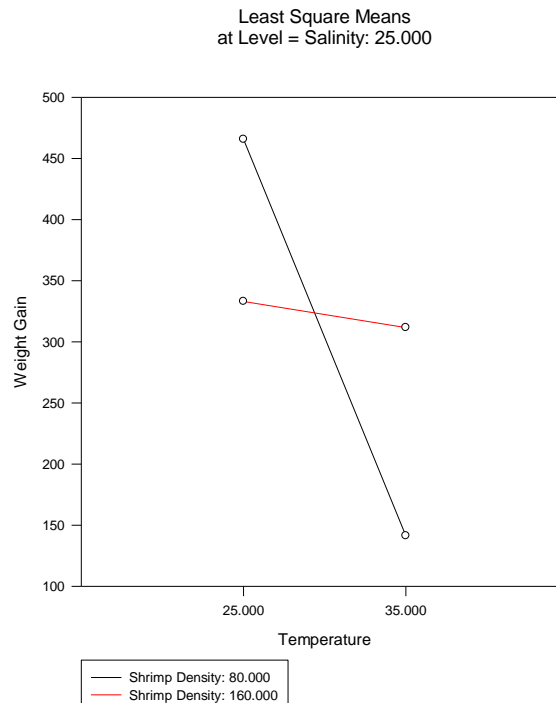
### 1. Main Effects Plots



## 2. Two-Way Effects Plot



## 3. Three-Way Effects Plot



## Probability Transforms

The following functions have been added to SigmaStat's Transform language for calculating probabilities and scores associated with distributions that arise in many fields of study. The abbreviation CDF means cumulative distribution function.

Transform language name	Description
<i>gammadist</i>	Gamma CDF
<i>gammainv</i>	Inverse Gamma CDF
<i>gammaden</i>	Gamma Density
<i>weibulldist</i>	Weibull CDF
<i>weibullinv</i>	Inverse Weibull CDF
<i>weibullden</i>	Weibull Density
<i>cauchydists</i>	Cauchy CDF
<i>cauchyinv</i>	Inverse Cauchy CDF
<i>cauchyden</i>	Cauchy Density
<i>erf</i>	Error Function
<i>erfc</i>	Complementary Error Function
<i>lognormdist</i>	Lognormal CDF
<i>lognorminv</i>	Inverse Lognormal CDF
<i>lognormden</i>	Lognormal Density
<i>expdist</i>	Exponential CDF
<i>expinv</i>	Inverse Exponential CDF

<i>expden</i>	Exponential Denstiy
<i>logisdist</i>	Logistic CDF
<i>logisinv</i>	Inverse Logistic CDF
<i>logisden</i>	Logistic Density
<i>loglogisdist</i>	Loglogistic CDF
<i>loglogisinv</i>	Inverse Loglogistic CDF
<i>loglogisden</i>	Loglogistic Density
<i>median</i>	A number that is both less than or equal to and greater than or equal to half of the values in the data set.

### Minor Enhancements to Existing Features

- Added Yates correction to Mann-Whitney test.
- Improved messages with added information when assumption checking for ANOVA fails.

### More In-Depth Discussion of New Features

The following information external specifications are the sources for the following information. The specs also contain screen shots of the dialog interfaces for the features and example reports and graphs.

#### Cox Regression

**Background and Terminology** – Cox Regression is a part of Survival Analysis that studies the impact of potential risk factors on the survival time of a population. The risk factors are often called *covariates*, *predictors*, or *explanatory variables*. We will use the term *covariates* in our applications. As an example, consider the possible effects of gender, age, and two types of drug therapy on the survival of a population suffering from some form of cancer. The survival time may decrease as age increases. Death rates among males may be higher than for females. Finally, drug A may increase survival time more than drug B. In this study, *Gender*, *Age*, and *Drug Therapy* are the covariates that affect the survival experience. In Cox Regression, a model is defined that describes the relationship between the covariates and survival time. This model is then used to predict the likelihood of survival at each point in time for any values of the covariates. It also allows us to determine the significant effect of each covariate.

There are two types of covariates. The above covariates, *Gender* and *Drug Therapy*, each have two categories of non-numeric values and are called *categorical covariates*. Since the covariate *Age* can assume a continuous range of numeric values, it is called a *continuous* or *nominal covariate*. Frequently, a categorical covariate has numeric values assigned to its categories, but these values are only used for naming purposes and are not used to indicate a measurement.

The simplest way to visualize the effect of covariates on survival time is to construct a *survival curve*. A survival curve plots the relationship between each value of time and the probability of surviving beyond that value. This relationship is called the *survival function* (or *survivorship function*). In Kaplan-Meier survival analysis, one survival function is defined that is independent of any covariates. In Cox survival

analysis, specific values for each of the covariates lead to one estimated survival function for the population. The graph of such a function is called a *covariate-adjusted survival curve*.

In Cox Regression, the primary object of study is the *hazard function* of the population, as estimated from the sampled survival data. This function is closely related to the survival function. The hazard function (sometimes known as the *conditional failure rate*, *hazard rate*, or just the *hazard*) is defined as the instantaneous rate of change in the likelihood of failure at each point in time, given survival up to that point. As an example, suppose  $h$  is the hazard function and suppose  $h(t) = .1$  at some time  $t$ , then an interpretation of this value is that there is approximately a 10% chance that a subject will fail within the next unit time period, given the subject has survived up to time  $t$ . Another function, the *cumulative hazard function*, is defined at each value of time as the integral of the hazard over all previous values of time. It provides a smoothed alternative to the hazard function as estimates of the hazard function itself can be too “noisy” for practical use. If  $H$  denotes the cumulative hazard function, then the above definitions can be used to show that the survival function  $S$  is defined at each time  $t$  by:

$$S(t) = \exp(-H(t)).$$

All of the functions discussed above are not only functions of time, but also depend upon the covariates in the survival study. In the Cox model, the hazard function assumes a specific form given by:

$$h(t, X_1, X_2, \dots, X_n) = h_0(t) \cdot \exp(b_1 X_1 + b_2 X_2 + \dots + b_n X_n)$$

where  $X_1, X_2, \dots, X_n$  are the covariates in the study. The function  $h_0$  is called the *baseline hazard function* and only depends upon time. The exponential factor on the right-hand side of the equation involves the covariates, but does not depend on time. In our implementation of Cox Regression, we are assuming that every covariate is time-independent and so its value for each subject remains constant over time (it is possible, however, to extend Cox Regression to include time-dependent covariates).

The coefficients  $b_1, b_2, b_n$  in our model are constants, independent of both time and the covariates, and their values are determined from the regression analysis by maximizing a quantity known as the *partial likelihood function*. The resulting values of the coefficients are called the *best-fit coefficients* or, sometimes, the *maximum likelihood estimates*. Once the coefficients are determined, there is a procedure that estimates the values of the *baseline survival function* at the sampled event times. The baseline survival function is defined by setting all covariates to zero. Denoting this function by  $S_0$ , the covariate-adjusted survival functions and cumulative hazard functions are determined for each event time  $t$  by:

$$H_0(t) = -\log(S_0(t))$$

$$H(t, X_1, \dots, X_n) = H_0(t) \exp(b_1 X_1 + \dots + b_n X_n)$$

$$S(t, X_1, \dots, X_n) = S_0(t)^{\exp(b_1 X_1 + \dots + b_n X_n)}$$

Our model of the hazard function shows that if there are two specifications for the values of the covariates, then the corresponding values of the hazards are proportional over time. This is the reason

the Cox model is called a *proportional hazards model*. It is possible that a potential covariate for the model does not satisfy this assumption. For example, suppose we have the covariate *Gender* in a survival study. If males are dying at twice the rate of females during the first month of a study, and both genders die at the same rate during the next month of the study, then the ratio of the hazards, or the *hazard ratio*, for males to females is not constant over time and the proportionality assumption fails. Such a covariate cannot be included in the hazard model.

A covariate may also be omitted from the model because its value is based on the design of the study and has secondary importance as a risk factor for survival. For example, when a study is performed at two different clinics to determine the impact of age and drug therapy on patient recovery, then the variable *Clinic* is such a covariate.

Any variable whose values have been included in the survival data but is not included as a covariate in the hazard model for the reasons described above is called a *stratification variable*. Each value or level of such a variable is called a *stratum*; collectively, the levels are the *strata*. When a stratification variable is present, then the survival study is partitioned into groups, one for each stratum, where each group has its own survival function that is determined from the regression analysis. The best-fit coefficients are the same for each stratum, but the baseline time-dependent factors in the model are different.

**Feature Description** – The Cox Regression feature in SigmaStat and SigmaPlot consists of two separate analyses, the *Proportional Hazards* model, with no stratification variable, and the *Stratified Model*, where the user selects a worksheet column containing the strata. Each test is accessed from the Statistics menu on a submenu under the item *Survival – Cox Regression*. For each test, the user selects a time column, status column, and any number of covariate columns from the worksheet. The user can subsequently designate which selected covariates should be interpreted as categorical. When the analysis is completed, a report will be generated to provide the numeric results. Result graphs will also be available for obtaining covariate-adjusted survival curves, cumulative hazard functions, and log-log survival functions (discussed below). Various options for controlling the regression process, displaying report results, and for setting result graph attributes can be set in the Test Options dialog box. The Advisor wizard has also been modified to suggest the usage of these tests.

### **Related documents and texts**

Hosmer, D.W., Jr. and Lemeshow, S. (1999). *Applied Survival Analysis – Regression modeling of time to event data*. New York: John Wiley & Sons.

Kleinbaum, David G. (1996). *Survival Analysis – A Self-Learning Text*. Statistics in the Health Sciences series, Springer-Verlag, New York.

*SurvivalGuide.pdf* – internal document.

## One-Sample t-test

**Feature Description** – The *one-sample t-test* is used to test the hypothesis that the mean of a sampled normally-distributed population equals a value specified by the user. SigmaStat currently has no one-sample testing except for normality.

The menu and test combo box will be modified to include a command for this test. The menu command will be on a submenu under a new test category called *Single Group*. The unpaired t-test that is currently in SigmaStat is simply called *t-test*, and this name will be kept. The one-sample t-test will have its own options that are set from the Test Options dialog. The first panel will provide an edit control option for entering the value of the hypothesized population mean. The remaining options in the dialog will be a subset of the options available for the two-sample test. The Test Wizard for the one-sample case will provide the same data format options as the two-sample case except for the Indexed format, which makes no sense when you have one sample. In addition to producing a report, there will be three result graphs that are a subset of those produced for the two-sample t-test.

### Computational Results –

#### *Hypothesis testing:*

The null hypothesis is that the mean of the sampled population equals the user supplied value. The sample mean of the selected data is compared with the hypothesized population mean supplied by the user by computing:

$$t = \frac{\bar{x} - \mu}{\sigma} \sqrt{n}$$

where

$\bar{x}$  = sample mean

$\mu$  = hypothesized population mean

$\sigma$  = sample standard deviation

n = sample size

By random sampling of the population, assuming the null hypothesis is true, this quantity defines a random variable  $T$ , whose distribution is Student's central T-distribution with  $n - 1$  degrees of freedom. The (two-sided) P-value for this test is computed as  $P(|T| > |t|)$ , where  $P$  denotes the probability distribution for  $T$ . This P-value is then compared to the significance level  $\alpha$  that is set by the user. If the value is less than  $\alpha$ , there is a significant difference between the mean of the sampled population and  $\mu$ .

#### *Confidence Interval for the Population Mean:*

The  $(1 - \alpha)100\%$  confidence interval for the true population mean is

$$\bar{x} \pm \sigma t_{\alpha, n-1}$$

where  $\bar{x}$  and  $\sigma$  are defined above, and  $t_{\alpha, n-1}$  is the value that satisfies  $P(|T| > t_{\alpha, n-1}) = \alpha$ .

### Retrospective Power

It may be of interest to know the power of the test based upon the difference in means that was observed from the sample. This will not assist, however, in reaching conclusions about the significance results since you already know the test is not powerful enough if you failed to reject the null hypothesis. For retrospective, or observed, power one must keep in mind that it is not simply defined as the probability of rejecting a hypothesis that is known to be false, but as the probability of detecting the difference that was actually observed in the data.

To compute the power, the critical value  $t_c$  of Student's central  $T$  distribution is computed for the given significance level  $\alpha$ . This value is found by solving the equation  $P(|T| > t_c) = \alpha$ . Now, let  $T_{\delta, n-1}$  be a random variable whose distribution is Student's non-central distribution with  $n-1$  degrees of freedom and with non-centrality parameter equal to  $\delta = \sqrt{n}/\sigma |\bar{x} - \mu|$ . Then

$$Power = P(|T_{\delta, n-1}| > t_c).$$

## Odds Ratio and Relative Risk

**Feature Description** – The *odds ratio* and *relative risk* are values that measure the strength of association between a treatment or risk factor and a specified event that occurs in members of a population. In a study for which these values are computed, you have a control group and a treatment group, each of whose members are randomly selected, and you have an event, like a disease, whose frequency in the population may be affected by the treatment administered. The total number of subjects in each group can be different.

A study using relative risk assumes the control group and the treatment group have been selected in advance. Observations are then made to determine how many from each group experience the event. This is an example of a *prospective study*. The relative risk  $RR$  is defined as the probability of the event in the treatment group divided by the probability of the event in the control group, where each probability is estimated as the relative frequency of the event in the group.

$$RR = \frac{\text{Probability of event in treatment group}}{\text{Probability of event in control group}}$$

Odds ratio is frequently used in *case-control* studies. This type of study is done *retrospectively*, in which the investigator samples two groups of subjects from the population according to whether a subject did or did not experience the event. The two groups are called the Cases and Controls, respectively. The number of subjects from each group who were exposed to the treatment or risk factor is then noted. The odds ratio  $OR$  is defined by:

$$OR = \frac{\text{Odds of event in treatment group}}{\text{Odds of event in control group}}$$

The odds ratio is an estimate of how much more likely the event occurs for an individual in the population exposed to the risk factor as compared to an individual not exposed to the risk factor.

In summary, the main computational difference between Relative Risk and Odds Ratio is that the former is computed as a ratio of probabilities whereas the latter is computed as a ratio of odds.

The null hypothesis for both the relative risk and the odds ratio is that its value equals 1. This means that the treatment or risk factor does not affect the event rate. A value significantly different from 1 indicates that the treatment either significantly increases or decreases the risk of the event in the population.

The data that is used for computing either quantity can be represented in a 2x2 contingency table. The probability of significance calculation for the test uses the chi-square statistic for this table. If the expected number of observations for any cell of the table is less than 5, then the Fisher-Exact test is used to compute the probability.

In SigmaStat, we implement relative risk and odds ratio as two separate tests since they are used with different assumptions. The menu and test combo box will be modified to include a command for each test. The menu commands will be on a submenu under the test category *Rates and Proportions*. Each test will have its own options in the Test Options dialog. These options will include settings for the Yates continuity correction factor for the chi-square statistic, the confidence interval for the ratio, the control group row selection, and the power. The Test Wizard for each test will have two data formats to select from: Tabulated and Raw. The Tabulated format assumes the data is in the form of a 2x2 contingency table where the two column selections represent event counts and non-event counts. The Raw data assumes the selected data is in two columns, one for the risk factor/control group labels and one for the event/non-event labels. After finishing the Test Wizard, a report will be produced. There are no result graphs for either test.

**Example –**

Suppose we are given a 2x2 contingency table of observations for studying the association of some risk factor to an event. Suppose the risk factor is radiation and the event is cancer.

	<i>Cancer</i>	<i>No Cancer</i>
<i>Radiation</i>	50	43
<i>Control</i>	14	35

The relative risk for the above table is  $RR = (50/93)/(14/49) = 1.88$ , so that the risk of developing cancer in the population is estimated to be 1.88 times higher for those receiving the radiation. The chi-square

probability for this table is .007 so that risk of developing cancer is significantly greater for those exposed to the radiation.

The odds ratio the above table is  $OR = (50/43)/(14/35) = 2.91$ , so that exposure to radiation increases the odds of developing cancer by an estimated 2.91 times among the population. With the same probability value as above, it is clear the effect of the radiation is significant.

### Computations –

Results in the report for relative risk and odds ratio use the computations below. It is assumed the input data can be put into the form of the 2x2 contingency table below:

	<i>Event</i>	<i>No Event</i>
<i>Treatment</i>	<i>a</i>	<i>b</i>
<i>Control</i>	<i>c</i>	<i>d</i>

#### Relative Risk:

$$RR = \frac{a/a+b}{c/c+d}$$

$$\text{Standard Error for RR} = s_{RR} = \sqrt{1/a + 1/c - 1/(a+b) - 1/(c+d)}$$

$$(1 - \alpha) * 100\% \text{ Confidence Interval for RR} = RR * \exp(\pm z_{\alpha} * s_{RR})$$

where  $z_{\alpha}$  is the two-sided critical value of the standard normal distribution with probability =  $\alpha$ .

#### Odds Ratio:

$$OR = \frac{a/b}{c/d}$$

$$\text{Standard Error for OR} = s_{OR} = \sqrt{1/a + 1/b + 1/c + 1/d}$$

$$(1 - \alpha) * 100\% \text{ Confidence Interval for OR} = OR * \exp(\pm z_{\alpha} * s_{OR})$$

where  $z_{\alpha}$  is the two-sided critical value of the standard normal distribution with probability =  $\alpha$ .

## Shapiro-Wilk Normality Test

**Feature Description** – The Shapiro-Wilk’s normality test will be added to both SigmaStat and SigmaPlot to determine if either worksheet data or the residuals that result from curve fitting are consistent with data drawn from a normal distribution. Normally distributed data is a principal assumption when using many of the tests in SigmaStat and in the non-linear regression analysis of SigmaPlot.

The current normality test in both programs is based upon the Kolmogorov-Smirnov (KS) statistic that computes the maximum difference between the sample cumulative distribution function of the data and the theoretical normal distribution having the same population parameters as the data.

The main advantage of this method is that the distribution of the statistic is independent of the underlying theoretical distribution (so long as it is continuous). For example, the KS –statistic could be used to test data against a gamma or Weibull distribution. The primary disadvantage of the method is the assumption that the population parameters are known. In practice, these parameters are only estimated from the data by using the sample mean and the (unbiased) sample variance. To compensate for this lack of information, simulation studies by Lilliefors and Wilkinson have yielded a correction for normal distributions that is used by both programs. Another disadvantage is that for small to moderate sample sizes, the test has difficulty discriminating between distributions that are roughly similar (low power).

The Shapiro-Wilk’s so-called W-statistic is specifically designed for the normal distribution and has higher power than Kolmogorov-Smirnov. The statistic is a ratio of two estimates of the variance of a normal distribution based on a random sample. The numerator of W is proportional to the square of the best linear estimator of the standard deviation, and the denominator is the sum of squares of the observations about the sample mean. The main limitation of the method is that the sample size is restricted to values between 3 and 5000, inclusive.

Given a set of observations  $x_1, x_2, \dots, x_n$  sorted into either ascending or descending order, the W-statistic is defined by:

$$W = \left( \sum_{k=1}^n a_k x_k \right)^2 / \sum_{k=1}^n (x_k - \bar{x})^2$$

where  $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$  is the sample mean and  $a_k, k = 1, \dots, n$  are weights whose values depend only on the sample size  $n$ .

## Probability Transforms

**Feature Description** – Seven new sets of probability functions will be added to the Transform language for computing the values of cumulative distribution functions and their inverses, and for computing probability density functions. In addition, we will add two other functions frequently used in statistical calculations: the error function and the complementary error function. Finally, a function for computing the median of a column of data will be added.

Like all transform language functions, these functions can be accessed from the User-defined Transform dialog, the Quick Transform dialog, SigmaPlot's automation interface (macros), the Nonlinear Regression Wizard, and SigmaPlot's Plot Equation dialog.

*General Definitions of Probability Functions:*

- Cumulative Distribution Function (CDF) – If  $X$  is a random variable with respect to a particular probability measure  $P$ , the  $CDF(x)$  is the probability that the values of  $X$  are less than  $x$ , i.e.  $CDF(x) = P(X < x)$ . For a continuous random variable,  $CDF(x)$  can be computed as the indefinite integral of the probability density function (if it exists).
- Inverse Cumulative Distribution Function – This function computes the value  $x$  of the random variable  $X$  that yields a specified probability value  $p$  for the CDF. That is, Inverse  $CDF(p) = x$  such that  $P(X < x) = p$ .
- Probability Density Function (PDF) – For a continuous random variable, the derivative of the CDF, if it exists. In this case, the probability that the values of the random variable lie within a small interval can be estimated by the product of the density at some point in the interval and the size of the interval. For a discrete random variable,  $PDF(x) = P(X=x)$ . For sampled data, the density function is approximated by a histogram.

*Some applications of the functions being added:*

A cumulative distribution function, an inverse distribution function, and a density function will be added for each of the three families of probability distributions below.

Function Family	Use	Parameters
Gamma	Describes the distribution of time until the $n^{\text{th}}$ occurrence in a Poisson process.	Two positive parameters – a shape parameter and a scale parameter. Setting the shape parameter to 1 yields the <i>exponential distribution</i> .
Weibull	Describes the failure time distributions when the failure rate is assumed to increase as some power.	Two positive parameters – a shape parameter and a scale parameter.
Cauchy	Gives the distribution of the ratio of two standard normal random variables.  Also gives the distribution of the random variable $Y = \tan(X)$ , where $X$ has a uniform distribution.	Two parameters - a location parameter and a positive scale parameter.
Lognormal	Gives the distribution of the random variable $Y = \exp(X)$ , where $X$ has a normal distribution.	Two parameters - a location parameter and a positive scale parameter. They are the mean and standard deviation of the underlying normal distribution.
Exponential	A special case of the	One positive scale parameter.

	Gamma distribution. Gives the distribution of time until the first occurrence in a Poisson process.	
Logistic	Similar in shape to the normal distribution, but with wider tails and is easier to compute	Two parameters - a location parameter and a positive scale parameter.
LogLogistic	Gives the distribution of the random variable $Y = \exp(X)$ , where $X$ has a logistic distribution.	Two parameters - a location parameter and a positive scale parameter.

*Mathematical Descriptions*

Some of the functions below are expressed in terms of the gamma function:

$$\Gamma(t) = \int_0^{\infty} e^{-x} x^{t-1} dx \text{ for positive } t$$

<p><b>Gamma Cumulative Distribution Function</b></p> $\text{GammaDist}(x, a, b) = \frac{1}{b^a \Gamma(a)} \int_0^x u^{a-1} e^{-u/b} du$	$x \geq 0$ $a > 0$ $b > 0$
<p><b>Weibull Cumulative Distribution Function</b></p> $\text{WeibullDist}(x, a, b) = 1 - \exp(-(x/b)^a)$	$x \geq 0$ $a > 0$ $b > 0$
<p><b>Cauchy Cumulative Distribution Function</b></p> $\text{CauchyDist}(x, a, b) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-a}{b}\right)$	$-\infty < x < \infty$ $-\infty < a < \infty$ $b > 0$

<p><b>Error Function</b></p> $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$	$-\infty < x < \infty$
<p><b>Complementary Error Function</b></p> $\text{erfc}(x) = 1 - \text{erf}(x)$	$-\infty < x < \infty$
<p><b>Lognormal Cumulative Distribution Function</b></p> $\text{LognormalDist}(x, \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\ln(x)} e^{-z-\mu^2/2\sigma^2} dz$	$x \geq 0$ $-\infty < \mu < \infty$ $0 < \sigma < \infty$
<p><b>Exponential Cumulative Distribution Function</b></p> $\text{ExponentialDist}(x, \lambda) = 1 - \exp(-\lambda x)$	$x \geq 0$ $\lambda > 0$

<p><b>Logistic Cumulative Distribution Function</b></p> $\text{LogisticDist}(x, \mu, \sigma) = \frac{1}{1 + \exp(-(x - \mu) / \sigma)}$	$-\infty < x < \infty$ $-\infty < \mu < \infty$ $0 < \sigma < \infty$
<p><b>LogLogistic Cumulative Distribution Function</b></p> $\text{LogLogisticDist}(x, \mu, \sigma) = \frac{1}{1 + \exp(-(\ln(x) - \mu) / \sigma)}$	$x \geq 0$ $-\infty < \mu < \infty$ $0 < \sigma < \infty$

## ANOVA Profile Plots

**Feature Description** – Profile plots are useful for comparing the *least square means*, also called *estimated marginal means*, in a multifactor ANOVA model. Differences in the means, or *effects*, among the levels of a specified factor, when computed over a range of levels of the remaining factors, determine how the data is affected by that factor and its interaction with other factors. Profile plots provide a quick qualitative assessment of the various treatment effects so that the investigator can determine the impact of each factor on the data. The hypothesis testing in ANOVA reports quantifies these effects to determine if any of the differences are statistically significant.

In ANOVA analysis, the least square means are first computed for the individual cells. A cell is defined as the collection of observations made for a particular combination of levels, where one level is selected from each factor. Generally, the cell means are obtained as the predicted values in a regression model that is associated with the ANOVA model. The cells means determine the *two-way interaction effects* in a Two-Way ANOVA and the *three-way interaction effects* in a Three-Way ANOVA. If the cell means are averaged over all levels of one factor while fixing the levels of the remaining factors, you obtain lower-order effects. This is how the *main effects* are computed in Two-Way ANOVA and the two-way interaction effects are computed in Three-Way ANOVA. Finally, the main effects for a given factor in a Three-Way ANOVA are determined by averaging the cell means over all levels of the remaining two factors while fixing each level of the given factor.

Profile plots are line plots with the levels of one factor represented on the horizontal axis of the graph and the experiment's data (and the least square means of that data) represented on the vertical axis. The least square means have the same scale as the data and so are positioned relative to the data axis for each factor level on the horizontal axis.

We will use the following design for presenting profile plots:

- For Main Effects, there is one plot per graph and the number of graphs equals the number of factors.

- For 2-Way Effects, we have one graph for each distinct pairwise-combination of factors (so one graph for Two-Way ANOVA and three graphs for Three-Way ANOVA). Each of these graphs contains multiple profile plots, one for each level of one of the factors.
- For 3-Way Effects in Three-Way ANOVA, the number of graphs equals the number of levels of the third factor (this factor is the last factor that was selected for running the test). Each graph for 3-Way Effects contains multiple profile plots, one for each level of one of the factors.

All of the data that is graphed for Profile plots is listed in the Summary table of the report.