Advanced Analyses Calculations

Introduction

This document is a collection of the calculations used in the advanced analyses in STATGRAPHICS *Plus* Quality and Design, and STATGRAPHICS *Plus* Professional.

The calculations are arranged by analysis in the order that they are listed on the Special menu:

Quality Control Calculations

Acceptance Chart

The Acceptance Chart shows the measurements or subgroup averages with modified control limits. If Sigma Multiple is selected:

$$UCL = (USL - Z_{\delta}\sigma) + \frac{3\sigma}{\sqrt{n}}$$

$$LCL = (LSL + Z_{\delta}\sigma) - \frac{3\sigma}{\sqrt{n}}$$

where

δ is the maximum allowable proportion of non-conforming items.

If Beta Risk is specified, then the control limits are set at:

$$UCL = (USL - Z_{\delta}\sigma) - \frac{Z_{\beta}\sigma}{\sqrt{n}}$$

$$LCL = (LSL + Z_{\delta}\sigma) + \frac{Z_{\beta}\sigma}{\sqrt{n}}$$

where

β is the probability of being within the control limits when the proportion of non-conforming items equals δ.

Acceptance Sampling

Acceptance Sampling is the process for determining sample sizes and decision rules for incoming or outgoing products. STATGRAPHICS *Plus* has procedures for generating two types of acceptance sampling plans: attributes and variables. For each of the two basic plan types there are three variations: OC Plans, AOQL PLans and LTPD Plans.

Notation:

N = the number of items in the lot

n = the number of items sampled from the lot and inspected

AQL = Acceptable Quality Level, the poorest level of quality which the consumer finds acceptable on average

LTPD = Lot Tolerance Percent Defective, the poorest level of quality that the consumer is willing to tolerate in any given lot

AOQL = Averrage Outgoing Quality Limit, the maximum percent of defective items accepted by a given sampling plan

AOQL Plans

The focus is on the maximum percent of non-conforming items shipped after verification.

$$AOQL = \max_{\theta}\left[\theta P\,(\text{accept} \mid \theta)\left(\frac{N-n}{N}\right)\right]$$

LTPD Plans

The LTPD Plans minimize the number of inspections while controlling the probability of accepting a "bad" lot. The consumer's risk is calculated according to:

$$\beta = P\,(\text{accept} \mid \text{LTPD})$$

OC Plans

The OC Plan controls the probability of accepting a lot at both the the AQL and the LTPD. All compute acceptance probabilities according to:

$$P(\text{accept} \mid \theta) = \sum_{x=0}^{c} P_H(x \mid \theta, N) = \sum_{x=0}^{c} \frac{\binom{N\theta}{x}\binom{N(1-\theta)}{n-x}}{\binom{N}{n}}$$

ARIMA Charts

Underlying the ARIMA control charts is the general AutoRegressive Integrated Moving Average Model:

$$z_j = \theta_0 + \Phi_1 z_{t-1} + \Phi_2 z_{t-2} + \ldots + \Phi_p z_{t-p} + a - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

where

$z_j = \nabla^d \bar{x}_j$

and

$a_j \sim \text{NID}(0, \sigma_a)$

Parameters which define this model:

θ_0 = constant

$\Phi_1, \Phi_2, \ldots \Phi_p$ = autoregressive parameters

$\theta_1, \theta_2, \ldots \theta_q$ = moving average parameters

Chart Type

Data with Long-Term Limits - plots the subgroup averages with limits defined by:

$\hat{\mu} \pm 3\hat{\sigma}_z$

Data with One-Step Limits - plots the subgroup averages with limits defined by:

$\hat{z}_j(j-1) \pm 3\sigma_a$

where

$\hat{z}_j\,(j-1)$ is the conditional expectation of the subgroup average at period j given all information through period j-1.

Residuals - plots the residuals defined by:

$$\hat{a}_j = z_j - \hat{z}_j\,(j-1)$$

Normalized Residuals - plots standardized residuals defined by:

$$\frac{\hat{a}_j}{\hat{\sigma}_a}$$

Process Capability

Normal Capability Indices

$$C_p = \frac{USL - LSL}{6\hat{\sigma}}$$

$$C_{pk} = \text{Min}\left\{\frac{(USL - \bar{\bar{X}})}{3\hat{\sigma}}, \frac{(\bar{\bar{X}} - LSL)}{3\hat{\sigma}}\right\}$$

$$C_{pk(upper)} = \frac{USL - \bar{\bar{X}}}{3\hat{\sigma}}$$

$$C_{pk(lower)} = \frac{\bar{\bar{X}} - LSL}{3\hat{\sigma}}$$

$$C_r = \frac{1}{C_p}$$

$$C_{pm} = \frac{USL - LSL}{6\hat{\sigma}'}$$

where

$$\hat{\sigma}' = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \text{Nominal})^2}{n-1}}$$

$$K = \frac{\bar{X} - \text{Nominal}}{(USL - LSL)/2}$$

Note: For variables control charts, STATGRAPHICS uses the appropriate $\hat{\sigma}$ to calculate the capability indices. For example, with X-bar and R Charts,

$$\hat{\sigma} = \frac{\bar{R}}{d_2}$$

Confidence Limits for Capability Indices

$$C_{pk} \pm Z_{\alpha/2} \sqrt{\frac{1}{9n} + \frac{C_{pk}^2}{2n-2}}$$

$$\left[\sqrt{\frac{C_1}{\nu}}\, C_{pm} \;,\; \sqrt{\frac{C_2}{\nu}}\, C_{pm} \right]$$

where

$$\nu = \frac{(n+\lambda)^2}{n+2\lambda}$$

and

$$\lambda = n\left(\frac{\bar{X} - \text{Target}}{s}\right)^2$$

and

$$C_1 = \frac{\alpha}{2} \text{ lower tail of the central } \chi^2_v$$

and

$$C_2 = 1 - \frac{\alpha}{2} \text{ of the central } \chi^2_v$$

$$\left[C_p \sqrt{\frac{\chi^2_{1-\alpha/2}}{n-1}} , C_p \sqrt{\frac{\chi^2_{\alpha/2}}{n-1}} \right]$$

Non-normal Capability Indices

$$C_p = \frac{USL - LSL}{U_p - L_p}$$

$$C_r = \frac{U_p - L_p}{USL - LSL} = \frac{1}{C_p}$$

$$C_{pk(upper)} = \frac{USL - Median_{est}}{U_p - Median_{est}}$$

$$C_{pk(lower)} = \frac{Median_{est} - LSL}{Median_{est} - L_p}$$

$$C_{pk} = \min [C_{pk(upper)} , C_{pk(lower)}]$$

$$K = \frac{Median_{est} - Nominal}{\left(\frac{USL - LSL}{2}\right)}$$

$$U_p = \bar{x} + s\, U_p^*$$

$$L_p = \bar{x} - s\, L_p^*$$

$$Median_{est} = \bar{x} + s\, Median^*$$

The system uses the kurtosis and skewness statistics to locate the L_p, Median, and U_p values in the Pearson Curve tables that Gruska, Mirkhani, and Lamberson display in their work.

Pareto

For information on the calculations STATGRAPHICS uses in Pareto Analysis, see Duncan (1974).

All Variables Control Charts

In the following formulas, coefficients like A_2, c_4, d_2, d_3, and so on have their usual meaning in the quality control literature. You can look up their values in tables found in almost any quality control textbook.

N = number of subgroups

n_j = number of observations in subgroup j

j = 1,2,...,N

x_{ij} = ith observation in subgroup j

Subgroup

Means:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

Standard Deviations:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{(n_j - 1)}}$$

Ranges:

$$R_j = \max\{x_{ij} \mid 1 \le i \le n_j\} - \min\{x_{ij} \mid 1 \le i \le n_j\}$$

Grand Mean

Mean Range: $$\bar{R} = \frac{1}{N}\sum_{j=1}^{N} R_j$$

Let

CL = centerline

UCL = upper control limit

LCL = lower control limit

N = number of subgroups

n_j = size for the jth subgroup

$\bar{\bar{X}}$ = grand average of all subgroups

$\bar{R}$ = average (mean) range

μ = standard process mean

σ = standard process sigma

$\hat{\sigma}$ = estimated process sigma

k_u = multiple of sigma for upper control limit

k_l = multiple of sigma for lower control limit

d_2 = expected value of R/σ (depends on n)

d_3 = standard deviation of R/σ (depends on n)

$\bar{s}$ = average of subgroup standard deviations

c_4 = expected value of s/σ (depends on n)

pooled s = $\sqrt{\text{pooled variance}}$

$$\text{pooled variance} = \frac{\sum_{j=1}^{N} (n_j - 1)\ s_j^2}{\sum_{j=1}^{N} (n_j - 1)}$$

α_u = alpha for upper control limit
(probability of a Type I error beyond limit)

α_l = alpha for lower control limit
(probability of a Type I error beyond limit)

$\chi^2_{1-\alpha,\nu}$ = $1 - \alpha$ probability point of chi-square distribution with ν degrees of freedom

X-bar and R Charts

Estimated:

process mean = $\bar{\bar{X}}$

process sigma = $\frac{\bar{R}}{d_2} = \hat{\sigma}$

mean range = $\bar{R}$

Standard:

process mean = μ

process sigma = σ

mean range = $d_2 \sigma$

Initial Study

X-bar Chart:

$CL = \bar{\bar{X}}$

$UCL = \bar{\bar{X}} + k_u \frac{\hat{\sigma}}{\sqrt{n}}$

$$LCL = \bar{\bar{X}} - k_l \frac{\hat{\sigma}}{\sqrt{n}}$$

Range Chart:

$$CL = \bar{R}$$

$$UCL = \bar{R} + k_u d_3 \frac{\bar{R}}{d_2}$$

$$LCL = Max\left\{0, \bar{R} - k_l d_3 \frac{\bar{R}}{d_2}\right\}$$

Control-to-Standard

X-bar Chart:

$$CL = \mu$$

$$UCL = \mu + k_u \frac{\sigma}{\sqrt{n}}$$

$$LCL = \mu - k_l \frac{\sigma}{\sqrt{n}}$$

Range Chart:

$$CL = d_2 \sigma$$

$$UCL = (d_2 \sigma) + k_u d_3 \sigma$$

$$LCL = Max\{0, (d_2 \sigma) - k_l d_3 \sigma\}$$

Normalization: X-bar (i = 1,...,N)

Initial Study:

$$\frac{(\bar{x}_j - \bar{\bar{X}})}{\hat{\sigma}}$$

Control-to-Standard:

$$\frac{(\bar{x}_j - \mu)}{\sigma}$$

X-bar and S Charts

Estimated:

process mean $= \bar{\bar{X}}$

process sigma $= \frac{\bar{s}}{c_4} = \hat{\sigma}$

mean sigma $= \bar{s}$

Standard:

process mean $= \mu$

process sigma $= \sigma$

mean sigma $= c_4 \sigma$

Initial Study

X-bar Chart:

$$CL = \bar{\bar{X}}$$

$$UCL = \bar{\bar{X}} + k_u \frac{\hat{\sigma}}{\sqrt{n}}$$

$$LCL = \bar{\bar{X}} - k_l \frac{\hat{\sigma}}{\sqrt{n}}$$

S Chart:

$$CL = \bar{s}$$

$$UCL = \bar{s} + k_u \bar{s} \frac{\sqrt{1 - (c_4)^2}}{c_4}$$

$$LCL = \text{Max} \left\{ 0, \bar{s} - k_l \bar{s} \frac{\sqrt{(1 - (c_4)^2}}{c_4} \right\}$$

Control-to-Standard

X-bar Chart:

$$CL = \mu$$

$$UCL = \mu + k_u \frac{\sigma}{\sqrt{n}}$$

$$LCL = \mu - k_l \frac{\sigma}{\sqrt{n}}$$

X-bar and S-squared Charts

Estimated:

$$\text{process mean} = \bar{\bar{X}}$$

$$\text{process sigma} = \sqrt{\text{pooled } s^2} = \hat{\sigma}$$

Standard:

$$\text{process mean} = \mu$$

$$\text{process sigma} = \sigma$$

$$\text{variance} = \sigma^2$$

Initial Study:

X-bar Chart

$$CL = \bar{\bar{X}}$$

$$UCL = \bar{\bar{X}} + k_u \frac{\hat{\sigma}}{\sqrt{n}}$$

$$LCL = \bar{\bar{X}} - k_l \frac{\hat{\sigma}}{\sqrt{n}}$$

S-squared Chart:

$$CL = \text{pooled } s^2$$

$$UCL = \frac{\text{pooled } s^2}{n-1} \chi^2_{\alpha_u, n-1}$$

$$LCL = \frac{\text{pooled } s^2}{n-1}\ \chi^2_{1-\alpha_l,\,n-1}$$

Control-to-Standard:

X-bar Chart:

$$CL = \mu$$

$$UCL = \mu + k_u \frac{\sigma}{\sqrt{n}}$$

$$LCL = \mu - k_l \frac{\sigma}{\sqrt{n}}$$

S-squared Chart:

$$CL = \sigma^2$$

$$UCL = \frac{\sigma^2}{n-1}\ \chi^2_{\alpha_u,\,n-1}$$

$$LCL = \frac{\sigma^2}{n-1}\ \chi^2_{\alpha_l,\,n-1}$$

Individuals Charts

Ranges are estimated by:

$$MR(2)_j = |X_j - X_{j-1}|\ ,\ j = 2,...,N$$

$$MR(2)_1 = \text{missing}$$

STATGRAPHICS calculates d_2 and d_3 on a subgroup size of 2.

Estimated:

$$\text{process mean} = \bar{X}$$

$$\text{process sigma} = \frac{\overline{MR(2)}}{d_2} = \hat{\sigma}$$

$$\text{mean MR(2)} = \overline{MR(2)}$$

Standard:

$$\text{process mean} = \mu$$

$$\text{process sigma} = \sigma$$

$$\text{mean MR(2)} = d_2 \sigma$$

Initial Study

X Chart:

$$CL = \overline{X}$$

$$UCL = \overline{X} + k_u \hat{\sigma}$$

$$LCL = \overline{X} - k_l \hat{\sigma}$$

MR(2) Chart:

$$CL = \overline{MR(2)}$$

$$UCL = \overline{MR(2)} + k_u d_3 \frac{\overline{MR(2)}}{d_2}$$

$$LCL = \text{Max} \left\{ 0, \overline{MR(2)} - k_l d_3 \frac{\overline{MR(2)}}{d_2} \right\}$$

Control-to-Standard

X Chart:

$$CL = \mu$$

$$UCL = \mu + k_u \sigma$$

$$LCL = \mu - k_l \sigma$$

Range Chart:

$$CL = d_2 \sigma$$

$$UCL = (d_2 \sigma) + k_u d_3 \sigma$$

$$LCL = \text{Max} \{0, (d_2 \sigma) - k_l d_3 \sigma\}$$

Normalization: X

Initial Study: $\frac{(x_j - \bar{x})}{\hat{\sigma}}$

Control to Standard: $\frac{(x_j - \mu)}{\sigma}$

Normalization: Range

Initial Study: $(MR(2)_j - \overline{MR(2)}) \frac{d_2}{(MR(2)\, d_3)}$

Control to Standard: $\frac{(MR(2)_j - d_2\sigma)}{(\sigma\, d_3)}$

Moving Average Chart

Moving averages of order q, q = 1,2,...,n

Display Limits:

$$CL = \mu$$

$$UCL = \mu + \frac{k_u \sigma}{\sqrt{nq}}$$

$$LCL = \mu - \frac{k_l \sigma}{\sqrt{nq}}$$

Chart Limits:

$$CL = \mu$$

$$UCL[t] = \mu + \frac{k_u \sigma}{\sqrt{nq[t]}}$$

$$LCL[t] = \mu - \frac{k_l \sigma}{\sqrt{nq[t]}}$$

where

$$q[t] = \begin{cases} n\,t & \text{for } t < 8 \\ n\,q & \text{for } t \geq q \end{cases} \quad \text{for } t = 1,\ldots,n$$

Chart Data:

$$Z[t] = \frac{(x[t] + x[t-1] + \ldots + x[t-q+1])}{q} \quad \text{for } t \geq q$$

$$= \frac{(x[t] + x[t-1] + \ldots + x[1])}{t} \quad \text{for } t < q$$

Exponentially Weighted Moving Average (EWMA) Chart

Display Limits:

$$CL = \mu$$

$$UCL = \mu + k_u \sigma \sqrt{\frac{\lambda}{(2-\lambda)n}}$$

$$LCL = \mu - k_l \sigma \sqrt{\frac{\lambda}{(2-\lambda)n}}$$

Chart Limits:

$$CL = \mu$$

$$UCL = \mu + k_u \sigma \sqrt{(\frac{\lambda}{2-\lambda})\,(1-(1-\lambda)^{2t})\,\frac{1}{n}}$$

$$LCL = \mu - k_l \sigma \sqrt{(\frac{\lambda}{2-\lambda})\,(1-(1-\lambda)^{2t})\,\frac{1}{n}}$$

Chart Data:

$$z[0] = \mu$$

$$z[t] = (\lambda\, x[t]\,) + (1-\lambda)\, z[t-1] \quad \text{for } t=1,2,\ldots,N$$

Moving Range: Moving Range of order q for q=1,2,...,N

Control Limits:

$$CL = d_2 \sigma$$

$$UCL = (d_2 \sigma) + k_u d_3 \sigma$$

$$LCL = (d_2 \sigma) - k_l d_3 \sigma$$

where d_2 and d_3 are based on a subgroup size of q

Chart Data:

$$R[t] = \text{Max}(x[t] + x[t-1] + \ldots + x[t-q+1]) - \text{Min}(x[t] + x[t-1] + \ldots + x[t-q+1]) \quad \text{for } t \geq q$$

$$= \text{missing value for } t < q$$

Cumulative Sum (CuSum) Chart

μ = control mean
σ = standard deviation
Δ = difference to detect
α = Type I error
β = Type II error
n = number of data points (subgroups)
$\bar{x}$ = mean of subgroup i

Plot cumulative sums C_t versus t where:

$$C_t = \sum_{i=1}^{t} (\bar{x}_i - \mu) \quad \text{for } t = 1,2,\ldots,n$$

The V-mask is located at distance

$$d = \frac{2}{\Delta}\left[\frac{\frac{\sigma^2}{\bar{n}}}{\Delta} \ln \frac{1-\beta}{\frac{\alpha}{2}}\right]$$

in front of the last data point.

$$\text{Angle of mask} = 2\tan^{-1}\left(\frac{\Delta}{2}\right)$$

$$\text{Slope of the line} = \pm\frac{\Delta}{2}$$

Let

μ = process mean

σ = process sigma

x[i] = ith observation (if individual data are input)
ith $\bar{x}$ (if subgroup statistics)

R[i] = ith range (if subgroup statistics)

n = 1 (if individual data are input)
average subgroup size (if subgroup statistics)

N = number of subgroups

k_u = sigma multiple for upper control

k_l = sigma multiple for lower control

Multivariate Control Chart

S = sample covariance matrix

$\underset{\sim}{X}_t$ = observation vector at time t

$\underset{\sim}{\bar{X}}$ = vector of column averages

Then Hotelling's T-squared is

$$T_t^2 = (\underset{\sim}{X}_t - \underset{\sim}{\bar{X}})' \; S^{-1} (\underset{\sim}{X}_t - \underset{\sim}{\bar{X}})$$

and the upper control limit is

$$UCL = \left(\frac{k(n-1)}{(n-k)} \right) F_{k,n-k,\alpha}$$

Points are plotted at T_t^2; glyphs are centered at T_t^2.

All Attributes Control Charts

N = number of subgroups

n_j = number of observations in subgroup j

j = 1,2,...,N

$$\bar{n} = \frac{1}{N} \sum_{j=1}^{N} n_j$$

u_j = number of defects in sample j

$$\bar{u} = \frac{\text{number of defects in all samples}}{\text{number of units in all samples}} = \frac{\sum u_j}{\sum n_j}$$

u' = standard u

$$p_j = \frac{\text{number of defective units in subgroup j}}{\text{number of units inspected in subgroup j}}$$

$$\bar{p} = \frac{\text{number of defectives in all samples}}{\text{number of units in all samples}} = \frac{\sum p_j n_j}{\sum n_j}$$

p' = standard p

$$\bar{f} = \bar{p} N$$

$$f' = p' N$$

c_j = number of defects in item j

$\bar{c} = \frac{\text{number of defects in all items}}{\text{number of items}} = \frac{\sum c_j}{n}$

c' = standard c = standard number of defects

CL = centerline

UCL = upper control limit

LCL = lower control limit

k_u = multiplier for upper control limit

k_l = multiplier for lower control limit

n[j] = $\bar{n}$ (when "Avg. subgroup size" is selected)
n_j (when "Avg. subgroup size" is not selected)

p Chart

Estimated:

mean p = $\bar{p}$

sigma = $\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$

Standard:

mean p = p'

sigma = $\sqrt{\frac{p'(1-p')}{\bar{n}}}$

Initial:

CL = $\bar{p}$

UCL = $\bar{p} + k_u \sqrt{\frac{\bar{p}(1-\bar{p})}{n[j]}}$

$$LCL = \bar{p} - k_l \sqrt{\frac{\bar{p}(1-\bar{p})}{n[j]}}$$

Control-to-Standard: $CL = p'$

$$UCL = p' + k_u \sqrt{\frac{p'(1-p')}{n[j]}}$$

$$LCL = p' - k_l \sqrt{\frac{p'(1-p')}{n[j]}}$$

Normalization

Initial:
$$\frac{(p_j - \bar{p})}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n[j]}}}$$

Control-to-Standard
$$\frac{(p_j - p')}{\sqrt{\frac{p'(1-p')}{n[j]}}}$$

np Chart

Estimated:

$$\text{mean np} = \bar{f}$$

$$\text{sigma} = \sqrt{\bar{f}(1-\bar{f})}$$

Standard: $\text{mean np} = f'$

$$\text{sigma} = \sqrt{f'(1-f')}$$

Initial:

$$CL = \bar{f}$$

$$UCL = \bar{f} + k_u \sqrt{n[j]\,\bar{p}\,(1-\bar{p})}$$

$$LCL = \bar{f} - k_l \sqrt{n[j]\,\bar{p}\,(1-\bar{p})}$$

Control-to-Standard:

$$CL = f'$$

$$UCL = f' + k_u \sqrt{n[j]\,p'\,(1-p')}$$

$$LCL = f' - k_l \sqrt{n[j]\,p'\,(1-p')}$$

Normalization

Initial:

$$\frac{(f_j - \bar{f})}{\sqrt{n[j]\,\bar{p}\,(1-\bar{p})}}$$

Control-to-Standard:

$$\frac{(f_j - f')}{\sqrt{n[j]\,p'\,(1-p')}}$$

c Chart

Estimated:

$$\text{mean c} = \bar{c}$$

$$\text{sigma} = \sqrt{\bar{c}}$$

Standard:

$$\text{mean } c = c'$$

$$\text{sigma} = \sqrt{c'}$$

Initial:

$$CL = \bar{c}$$

$$UCL = \bar{c} + k_u \sqrt{\bar{c}}$$

$$LCL = \bar{c} - k_l \sqrt{\bar{c}}$$

Control-to-Standard:

$$CL = c'$$

$$UCL = c' + k_u \sqrt{c'}$$

$$LCL = c' - k_l \sqrt{c'}$$

Normalization

Initial:

$$\frac{(c_j - \bar{c})}{\sqrt{\bar{c}}}$$

Control-to-Standard:

$$\frac{(c_j - c')}{\sqrt{c'}}$$

u Chart

Estimated:

$$\text{mean } u = \bar{u}$$

$$\text{sigma} = \sqrt{\frac{\bar{u}}{\bar{n}}}$$

Standard:

$$\text{mean } u = u'$$

$$\text{sigma} = \sqrt{\frac{u'}{\bar{n}}}$$

Initial:

$$CL = \bar{u}$$

$$UCL = \bar{u} + k_u \sqrt{\frac{\bar{u}}{n[j]}}$$

$$LCL = \bar{u} - k_l \sqrt{\frac{\bar{u}}{n[j]}}$$

Control-to-Standard:

$$CL = u'$$

$$UCL = u' + k_u \sqrt{\frac{u'}{n[j]}}$$

$$LCL = u' - k_l \sqrt{\frac{u'}{n[j]}}$$

Normalization

Initial:

$$\frac{(u_j - \bar{u})}{\sqrt{\frac{\bar{u}}{n[j])}}}$$

Control-to-Standard:

$$\frac{(u_j - u')}{\sqrt{\frac{u'}{n[j]}}}$$

Gage R&R

For information on the calculations STATGRAPHICS uses in the Gage R&R Analysis, see AIAG (1990), *Measurement Systems Analysis Reference Manual.*

Toolwear Chart

The toolwear chart was originally developed to monitor the wear of tools where the the measurements were expected to follow a natural trend. That linear trend is defined by:

$\mu_j = A + B_j$

where

μ_j = the mean wear at period j

A = the y-axis intercept

B = the slope

Control limits are placed by default at:

$$\hat{A} + \hat{B}j \pm 3\frac{\hat{\sigma}}{\sqrt{\bar{n}}}$$

For initial studies the model estimation error can be included:

$$\hat{A} + \hat{B}j \pm 3\hat{\sigma}\sqrt{\frac{1}{\bar{n}}\left(1 + \frac{(j-\bar{j})^2}{\sum_{i=1}^{m} n_i(i-\bar{j})^2}\right)}$$

where

$\bar{n}$ = average subgroup size ($\bar{n}$ = 1 for individuals data)

$\bar{j}$ = average of the period numbers used to fit the line

Design of Experiments Calculations

D-Optimal Design

The optimize design procedure seeks to find the design which maximizes the determinant of the M matrix defined by:

$$|M| = \frac{|X'X|}{N^p}$$

where

N = total number of runs
p = the number of estimated coefficients
X = the normalized design matrix.

Optimize design also shows various measures of efficiency:

D-efficiency - compare the determinant of M to the value of the best possible design according to:

$$\text{D–efficiency} = 100\left(\frac{1}{N}|X'X|^{\frac{1}{p}}\right)\%$$

A-efficiency - compares the sum of the variances of the estimated regression coefficients without considering their covariances:

$$\text{A–efficiency} = \left(\frac{p}{\text{trace}\,(N\,(X'X)^{-1})}\right)\%$$

G-efficiency - compares the maximum prediction standard error over the design points σ_m through:

$$\text{G–efficiency} = 100\left(\frac{\sqrt{p/N}}{\sigma_m}\right)\%$$

Inner/Outer Arrays

For information on the STATGRAPHICS *Plus* implementation of arrays, see Taguchi (1987), *The System of Experimental Designs: Engineering Methods to Optimize Quality and Minimize Costs.*

Mixture Designs

The Scheffe polynomial formulas for the Linear, Quadratic, Special Cubic, or Cubic models have the following forms.

Linear Model:

$$Y = b_1X_1 + b_2X_2 + \ldots + b_NX_N$$

Quadratic Model:

$$Y = \text{Linear Model} + b_{12}X_1X_2 + b_{13}X_1X_3 + \ldots + b_{N-1,\,N}X_{N-1}X_N$$

Special Cubic Model:

$$Y = \text{Quadratic Model} + b_{123}X_1X_2X_3 + \ldots + b_{N-2,\,N-1,\,N}X_{N-2}X_{N-1}X_N$$

Cubic Model:

$$Y = \text{Special Cubic Model} + d_{12}X_1X_2(X_1 - X_2) + d_{13}X_1X_3(X_1 - X_3) + \ldots + d_{N-1,\,N}X_{N-1}X_N(X_{N-1} - X_N)$$

Multi-Factor Categorical Designs

Multi-Factor Categorical Designs are analyzed using the Multifactor ANOVA procedure. Additional information on the calculations can be found in Neter, et al (1996), *Applied Linear Statistical Models*, fourth edition.

Multilevel Factorial Designs

This design fits the general second order model such as the following for three factors from Polhemus (1999):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \varepsilon$$

Response Surface Designs

Axial Distance

A design is orthogonally blocked if the axial distance is

$$\alpha = \left[k \times \frac{(1 + n_{so}/n_s)}{(1 + n_{co}/n_c)} \right]^{1/2}$$

where

k = the number of factors

n_{so} = the number of centerpoints in the star portion of the design

n_{co} = the number of centerpoints in the cube portion of the design

n_s = the number of base points in the star portion

n_c = the number of base points in the cube portion

In a Rotatable design, the variance of the predicted response is constant on a hyperspherical surface centered at the design origin. The axial distance is

$$\alpha = (n_c)^{1/4}$$

where n_c is the number of points in the cube portion of the design, excluding centerpoints.

An orthogonal design is one in which the estimates of all terms in the second-order model are uncorrelated. To produce an orthogonal design, the axial distance is set to

$$\alpha = \{[(n_c + n_s + n_o)^{1/2} - n_c^{1/2}]^2 \times \frac{n_c}{4}\}^{1/4}$$

where

n_c = the number of cube points

n_s = the number of star points

n_o = the number of centerpoints

You can make a design both rotatable and approximately orthogonal if you set the axial distance to the value used for rotatable designs, and then set the number of centerpoints to the integer closest to

$$n_o = 4 \times n_c^{1/2} + 4 - 2k$$

where

n_c = the number of cube points

k = the number of factors

Screening Designs

See Chapters 10 and 12 in Box, Hunter, and Hunter (1978) for the calculations used to estimate main effects, interactions, variances, and normal plots.

The design generators and block generators for the factorial designs are listed in Box, Hunter, and Hunter (1978) in Tables 10.B.1 and 12.15.

The error degrees of freedom shown on the Design Selection List screen is computed using the formula

$$\text{d.f.} = n - (k + f + (b - 1)) - 1$$

where

n = the number of runs
k = the number of factors
f = the number of two-factor interactions
b = the number of blocks.

If this number is negative, the system displays a 0. The equation is exactly correct if main effects, two-factor interactions, and block effects are clear of each other. If confounding exists, the degrees of

freedom are increased by adding back the number of nonestimable effects. For Plackett-Burman designs, the system considers only main factor effects to be estimable when computing the degrees of freedom for error.

Single Factor Categorical Designs

The Single Factor Catergorical Designs are analyzed using the One-Way ANOVA analysis. The calculations are in Appendix E of the *Standard Edition Manual.*

Variance Components Designs

The general class of hierarchical experiments in which factors are nested one within the other. Frequently these experiments are performed to determine what level of a process needs further investigation. The model is of the form

$$Y = \mu + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_k$$

where

Y = response variable

μ = process mean

ε_j = experimental error due to component j

k = number of variance components

The process standard deviation is generally:

$$\sigma_y = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_k^2}$$

where

σ_y = process standard deviation

σ_j = standard deviation of error component j

Time-Series Calculations

ARIMA Model

The basic form of the model to be fitted is:

$$W_t = \mu + \frac{\theta(B)\,\Theta_s(B)}{\varphi(B)\,\Phi_s(B)}\,a_t$$

This model expresses the data as a combination of the series' past values and the past values of the random input, where:

t = time

B = the backshift operation; that is, $B^l W(t) = W(t-l)$

W_t = the original data or a difference of that data

μ = the mean

$\theta(B)$ = the nonseasonal moving-average operator, $1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$

$\varphi(B)$ = the nonseasonal autoregressive operator, $1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_q B^p$

$\Theta_s(B)$ = the seasonal moving-average operator, $1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs}$

$\Phi_s(B)$ = the seasonal autoregressive operator, $1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}$

a_t = the random error.

This is typically denoted as a $(p,d,q) = (P,D,Q)^s$ model, where:

p = the order of the nonseasonal autoregressive term
d = the order of nonseasonal differencing
q = the order of the nonseasonal moving-average term
P = the order of the seasonal autoregressive term
D = the order of seasonal differencing
Q = the order of the seasonal moving-average term
s = the length of seasonality

Autocorrelations

The autocorrelation r_k at lag k is calculated as follows:

$$r_k = \frac{c_k}{c_0}$$

where

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} (y_t - \bar{y})\,(y_{t+k} - \bar{y})$$

and

$$\bar{y} = \frac{\left(\sum_{t=1}^{n} y_t \right)}{n}$$

and

$$c_0 = \sum_{t=1}^{n} (y_t - \bar{y})^2$$

$$\text{standard error at lag k} = \sqrt{\frac{1}{n}\left\{1 + 2\sum_{v=1}^{k-1} r_v^2\right\}}$$

where

y_t = observation at time t
n = number of observations

Box-Cox Transformations

For the original value Z, the transformation Z_T is given by:

$$Z_T = \frac{(Z + \lambda_2)^{\lambda_1 - 1}}{\lambda_1 g^{(\lambda_1 - 1)}} \qquad \text{if } \lambda_1 \neq 0$$

$$Z_T = g \ln (Z + \lambda_2) \qquad \text{if } \lambda_1 = 0$$

where g is the sample geometric mean of $Z + \lambda_2$. The first parameter λ_1 governs the strength of the transformation.

$\lambda_1 = 1$ corresponds to the original data; $\lambda_1 = 0$ to a logarithm. The system adds λ_2 to the data before it applies λ_1.

Brown's Linear and Quadratic Exponential Smoothing

Let

α = the smoothing constant

Y_t = the observed value at time t

S_t = the smoothed value at time t

Then

$$S_t = \alpha Y_t + (1 - \alpha) S_{t-1}$$

Crosscorrelations

Crosscorrelation at lag k: x = input time series

y = output time series

$$r_{xy}(k) = \frac{c_{xy}(k)}{s_x s_y} \qquad k = 0,\pm1,\pm2,...$$

where

$$c_{xy}(k) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(y_{t+k} - \bar{y}) & k = 0,1,2,... \\ \frac{1}{n}\sum_{t=1}^{n+k}(y_t - \bar{y})(x_{t-k} - \bar{x}) & k = 0,-1,-2,... \end{cases}$$

and

$$s_x = \sqrt{c_{xx}(0)}$$

$$s_y = \sqrt{c_{yy}(0)}$$

Holt's Linear Exponential Smoothing

N = the number of observations

X_t = the tth observation in the time series

S_t = the smoothed value of the time series at time t

b_t = the estimated trend at time t

α = the smoothing constant for the level of the time series

β = the smoothing constant for the trend of the time series

F_{t+m} = the forecast for the time period t+m

The initial values for the estimates are:

$S_1 = X_1$

$b_1 = X_2 - X_1$

The forecasted values of the time series are given by:

F_{t+m} for t+m {>= ^ 3 }

$$S_t = \alpha X_t + (1-\alpha)(S_{t-1} + b_{t-1})$$

$$b_t = \beta(S_t - S_{t-1}) + (1-\beta)b_{t-1}$$

$$F_{t+m} = S_t + b_t\, m$$

Partial Autocorrelation at lag k

The partial autocorrelation at lag k, $\hat{\Theta}_{kk}$, is obtained by solving the Yule-Walker equations:

$$3r_j = \hat{\Theta}_{kl}{}^r j{-}1 + \hat{\Theta}_{k2}{}^r j{-}2 + \ldots + \hat{\Theta}_{k(k-1)}{}^r j{-}k{+}1 + \hat{\Theta}_{kk}{}^r j{-}k$$

$$j = 1,2,\ldots,k$$

$$\text{standard error} = \sqrt{\frac{1}{n}}$$

Periodogram and Integrated Periodogram

Periodograms are computed using Fourier transforms. The value of the ordinate at each frequency f_i is given by the following formulas.

If n is odd, where n is the number of observations

$$I(f_i) = \frac{n}{2}\left(a_i{}^2 + b_i{}^2\right) \quad i=1,2,\ldots,\left[\frac{n-1}{2}\right]$$

where

$$a_i = \frac{2}{n}\sum_{t=1}^{n} y_t \cos(2\pi f_i\, t)$$

$$b_i = \frac{2}{n}\sum_{t=1}^{n} y_t \sin(2\pi f_i t)$$

$$f_i = \frac{i}{n}$$

If n is even, an additional term is added:

$$I(0.5) = n\left(\frac{1}{n}\sum_{t=1}^{n} (1)^t Y_t\right)^2$$

Resistant Nonlinear Smoothing

For the calculations used in resistant nonlinear smoothing, see Tukey (1977).

Seasonal Decomposition

For information on the calculations used in seasonal decomposition, see Makridakis, Wheelwright, and McGee (1983).

Smoothing

For information on the calculations used in resistant nonlinear smoothing, see Tukey (1977).

For information on the calculations used in weighted moving averages, see Kendall and Stuart (1961).

Tests for Randomness

Tests are variants of the basic runs test. The calculation formulas are the same, but the value of the parameters the formulas use are different.

For the test of the number of runs above and below the median:

n_1 = number of observations above the median
n_2 = number of observations below the median

For the test of the number of runs up and down:

n_1 = number of times the value rises
n_2 = number of times the value decreases

The system ignores adjacent pairs (two consecutive numbers that are equal).

For both tests, the expected number of runs is

$$E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1$$

The estimated variance for the runs is

$$V(R) = \frac{2n_1n_2\,(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}$$

The test statistic for the runs test is

$$Z = \frac{|R - E(R)| - 0.5}{\sqrt{V(R)}}$$

Vertical Time Sequence Plot

For information on the calculations the system uses to produce a vertical time sequence plot, see Box and Jenkins (1976).

Winter's Exponential Smoothing

x_t = actual demand at time period t
L = length of seasonality
I_t = seasonal factor for period t
α = smoothing constant for level
β = smoothing constant for trend
γ = smoothing constant for seasonality
k = number of periods into the future (k=1,2, ...)

$$\text{level}_t = \alpha \frac{x_t}{I_{t-L}} + (1 - \alpha)\,(\text{level}_{t-1} + \text{trend}_{t-1})$$

$$\text{trend}_t = \beta(\text{level}_t - \text{level}_{t-1}) + (1 - \beta)\,\text{trend}_{t-1}$$

$$I_t = \gamma\left(\frac{X_t}{\text{level}_t}\right) + (1 - \gamma)\,I_{t-L}$$

$$\text{Forecast}_{t+k} = (\text{level}_t + k \times \text{trend}_k)I_{t-L+k}$$

Multivariate Methods Calculations

Cluster Analysis

Lance and William Flexible Method

The formula for the distance measures between groups is:

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \mid d_{ki} - d_{kj} \mid$$

where

d_{ij} = the distance between groups i and j
α = parameter1
β = parameter2
γ = parameter3

Hierarchical Clustering Scheme

Quadruple Constraint

$$\alpha_i + \alpha_j + \beta = 1$$

$$\alpha_i = \alpha_j$$

$$\beta < 1$$

$$\gamma = 0$$

Nearest Neighbor

$$\alpha_i = \alpha_j = \frac{1}{2}$$

$$\beta = 0$$

$$\gamma = -\frac{1}{2}$$

Furthest Neighbor

$$\alpha_i = \alpha_j = \frac{1}{2}$$

$$\beta = 0$$

$$\gamma = \frac{1}{2}$$

Centroid

$$\alpha_i = \frac{n_i}{n_i + n_J}$$

$$\alpha_j = \frac{n_j}{n_i + n_J}$$

$$\beta = -\alpha_i \alpha_j$$

$$\gamma = 0$$

Median

$$\alpha_i = \alpha_j = \frac{1}{2}$$

$$\beta = -\frac{1}{4}$$

$$\gamma = 0$$

Group Average

$$\alpha_i = \frac{n_i}{n_i + n_J}$$

$$\alpha_j = \frac{n_j}{n_i + n_J}$$

$$\beta = \gamma = 0$$

Ward's Method

$$\alpha_i = \frac{n_k + n_i}{n_k + n_i + n_J}$$

$$\alpha_j = \frac{n_k + n_j}{n_k + n_i + n_J}$$

$$\beta = \frac{-n_k}{n_k + n_i + n_J}$$

$$\gamma = 0$$

Factor Analysis

Factor Scores for Unstandardized Variables

$$F_{jk} = \sum_{i=1}^{p} W_{ji} X_i = W_{j1} X_1 + W_{j2} X_2 + \ldots + W_{jp} X_p$$

where

p = number of variables

X_i = unstandardized variables

W_j = factor weights

Factor Scores for Standardized Variables

$$F_{jk} = \sum_{i=1}^{p} W_{ji} Z_{ik} = W_{j1} Z_{1k} + W_{j2} Z_{2k} + ... + W_{jp} Z_{pk}$$

where

p = number of variables

W_j = factor weights

Z_i = standardized variables

k = observations

$$\text{Percent of Variance} = \frac{\text{Eigenvalue}}{p} \times 100$$

Discriminant Analysis

Discriminant Function

$$D_i = d_{i1} Z_1 + d_{i2} Z_2 + \ldots + d_{ip} Z_p$$

where

p = number of variables

$Z_1, Z_2, ... Z_p$ = standardized variables

d_i = standardized classification function coefficients

$$\text{Eigenvalues} = \frac{\text{Between–groups sums of squares}}{\text{Within–groups sums of squares}}$$

$$\text{Wilks' Lambda} = \frac{\text{Within–groups sums of squares}}{\text{Total sums of squares}}$$

$$\text{Percent of Variance} = \frac{\text{Eigenvalue}}{p} \times 100$$

Principal Components

Principal Components for Unstandardized Variables

$$P_{jk} = \sum_{i=1}^{p} W_{ji} X_i = W_{j1} X_1 + W_{j2} X_2 + ... + W_{jp} X_p$$

where

p = number of variables

X_i = unstandardized variables

W_j = component weights

Principal Components for Standardized Variables

$$P_{jk} = \sum_{i=1}^{p} W_{ji} Z_i = W_{j1} Z_1 + W_{j2} Z_2 + ... + W_{jp} Z_p$$

where

p = number of variables

Z_i = standardized variables

W_j = component weights

$$\text{Percent of Variance} = \frac{\text{Eigenvalue}}{p} \times 100$$

Advanced Regression Calculations

Notation Common to All Analyses

$\underset{\sim}{Y}$ = vector of n observations for the dependent variable

$\underset{\sim}{X}$ = n-by-p matrix of observations for p-1 independent variables and the constant term, if any

$\underset{\sim}{W}$ = vector of weights (default = $\underset{\sim}{1}$)

$\underset{\sim}{\beta}$ = $p \times 1$ vector of unknown model coefficients

$\underset{\sim}{\varepsilon}$ = $n \times 1$ vector of random errors

$\underset{\sim}{b}$ = $p \times 1$ vector of estimated model coefficients

$s^2\{b\}$ = estimated variance-covariance matrix

MSE = mean squared error

$\underset{\sim}{\hat{Y}}$ = vector of predicted values for $\underset{\sim}{Y}$

$\underset{\sim}{e}$ = vector of residuals

$\underset{\sim}{h}$ = vector of leverages

$\underset{\sim}{d}$ = vector of studentized residuals

Mean:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i w_i}{\sum_{i=1}^{n} w_i}$$

$\underset{\sim}{\bar{X}}$ = vector of column means for $\underset{\sim}{X}$

General Linear Model:

$$\underset{\sim}{Y} = \underset{\sim}{X}\ \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

Least Squares Estimates:

$$\underset{\sim}{b} = (\underset{\sim}{X}'\ \underset{\sim}{W}\ \underset{\sim}{X}\ )^{-1}\ \underset{\sim}{X}'\ \underset{\sim}{W}\ \underset{\sim}{Y}$$

Estimated Variance-Covariance Matrix:

$$s^2\{b\} = MSE\,(X'WX)^{-1}$$

where

$$MSE = \frac{\sum_{i=1}^{n} w_i (Y_i - \hat{Y}_i)^2}{n-p} = \frac{SSE}{n-p}$$

Predicted Values:

$$\underset{\sim}{\hat{Y}} = \underset{\sim}{X}\ \underset{\sim}{b}$$

Residuals:

$$\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}}$$

R-Squared:

$$R^2 = \frac{SSTO - SSE}{SSTO}$$

where

$$SSTO = \begin{cases} \sum_{i=1}^{n} w_i (Y_i - \bar{Y})^2 & \text{if constant in model} \\ \sum_{i=1}^{n} w_i Y_i^2 & \text{if no constant} \end{cases}$$

Adjusted R-Squared:

$$1 - \left(\frac{n-1}{n-p}\right)\left(1 - R^2\right)$$

Standard Error of Estimate:

$$SE = \sqrt{MSE}$$

Durbin-Watson Statistic:

$$D = \frac{\sum_{i=1}^{n-1} \left(e_{i+1} - e_i\right)^2}{\sum_{i=1}^{n} e_i^2}$$

Mean Absolute Error:

$$MAE = \frac{\sum_{i=1}^{n} |e_i| w_i}{\sum_{i=1}^{n} w_i}$$

Mean Error:

$$ME = \frac{\sum_{i=1}^{n} e_i w_i}{\sum_{i=1}^{n} w_i}$$

Mean Percentage Error:

$$MPE = \frac{100 \sum_{i=1}^{n} \frac{e_i}{Y_i} w_i}{\sum_{i=1}^{n} w_i}$$

Mean Absolute Percentage Error:

$$MAPE = \frac{100 \sum_{i=1}^{n} \left| \frac{e_i}{Y_i} \right| w_i}{\sum_{i=1}^{n} w_i}$$

Forecasts:

$\underset{\sim}{X}_h$ = m-by-p matrix of independent variables
for m predictions

Predicted Value:

$$\hat{\underset{\sim}{Y}}_h = \underset{\sim}{X}_h \underset{\sim}{b}$$

Standard Error of Prediction:

$$S(\hat{\underset{\sim}{Y}}_{h(new)}) = \sqrt{\text{diagonal elements of MSE } (1 + \underset{\sim}{X}_h (\underset{\sim}{X}' \underset{\sim}{W} \underset{\sim}{X})^{-1} \underset{\sim}{X}_h')}$$

Standard Error of Mean Response:

$$S(\underset{\sim}{\hat{Y}}_h) = \sqrt{\text{diagonal elements of MSE } (\underset{\sim}{X}_h (\underset{\sim}{X}' \underset{\sim}{W} \underset{\sim}{X})^{-1} \underset{\sim}{X}_h')}$$

Variance Inflation Factors:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where

R_i^2 = coefficient of multiple determination when regressing x_i against other regressor variables

Leverage:

$$\underset{\sim}{h} = \text{diagonal elements of } \underset{\sim}{W} \underset{\sim}{X} (\underset{\sim}{X}' \underset{\sim}{W} \underset{\sim}{X})^{-1} \underset{\sim}{X}'$$

Studentized Residuals:

$$d_i = e_i \sqrt{w_i} \left[\frac{n - p - 1}{SSE(1 - h_i) - e_i^2 w_i} \right]^{\frac{1}{2}}$$

where

$$SSE = \sum_{i=1}^{n} w_i (Y_i - \hat{Y})^2$$

DFITS:

$$DFITS_i = d_i \sqrt{\frac{h_i}{1 - h_i}}$$

Mahalanobis Distance:

$$MD_i = \frac{n \left(n - 2 \right) \left(h_i - \frac{w_i}{\sum_{i-1} w_i} \right)}{(n - 1)(1 - h_i)}$$

Calibration Models

A $100\ (1-\alpha)\%$ **Prediction Interval for X given Y:**

$$\frac{Y-a}{b} \pm \frac{t\left(1-\frac{\alpha}{2}, n-2\right)\sqrt{MSE}}{b}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{\left(\frac{Y-\bar{Y}_W}{b}\right)^2}{S_{WXX}}}$$

where

S_{WXX} is the weighted sum of squares $= \sum w_i (X-\bar{X}_W)^2$,

$\bar{X}_W$, $\bar{Y}_W$ denote weighted averages,

and m is the Mean Size or Weight for the new observation Y.

Refer to Caulcutt & Boddy (1983) for details.

General Linear Models

Indicator Variables for Categorical Factors:

k = number of levels

Construct k - 1 indicator variable where

$$I_1 = \begin{cases} 1 \text{ if level} = 1 \\ -1 \text{ if level} = k \\ 0 \text{ otherwise} \end{cases}$$

$$I_2 = \begin{cases} 1 \text{ if level} = 2 \\ -1 \text{ if level} = k \\ 0 \text{ otherwise} \end{cases}$$

. . .

Multiple Comparison Intervals:

ith contrast

$$L_i = \sum_{j=1}^{k} c_j \mu_j$$

k = number of factor levels

$$\hat{L}_i = \sum_{j=1}^{k} c_{ij}\hat{\mu}_j$$

where

$\hat{\mu}_j$ = predicted value of Y for level j when all quantitative factors are set at their mean levels

$s(\hat{L}_i)$ = standard error of estimated contrast

ν = number of degrees of freedom for error

w = number of contrasts

Least Significant Differences (LSD) Intervals:

$$\hat{L}_i \pm t\, s(\hat{L}_i)$$

where

t is the upper $\alpha/2$ critical value of Student's t distribution with ν degrees of freedom.

Tukey Intervals:

$$\hat{L}_i \pm \frac{Q}{\sqrt{2}} s(\hat{L}_i)$$

where

Q is the upper α critical value of studentized range distribution with parameters k and ν.

Dunnett Intervals:

$$\hat{L}_i \pm |d|\, s(\hat{L}_i)$$

where

|d| is the upper α critical value of (k - 1) - variate Student's t distribution with ν degrees of freedom

Scheffé Intervals:

$\hat{L} \pm S\, s(\hat{L})$

where

$S^2 = (k - 1)\, F$ and F is the upper α critical value of Snedecor's F distribution with k - 1 and ν degrees of freedom

Bonferroni Intervals:

$\hat{L} \pm B\, s(\hat{L})$

where

B = upper $\alpha/2w$ critical value of student's t distribution with ν degrees of freedom

Multivariate t Intervals:

$\hat{L} \pm P\, s(\hat{L})$

where

P = upper $\alpha/2$ critical value of w-variate t distribution with ν degrees of freedom.

MANOVA Tests:

H = the hypothesis matrix

E = the error matrix

p = the number of dependent variables

df_h = the degrees of freedom for H

df_e = the degrees of freedom for E

S = min (df_h, p)

$$m = \frac{|\, df_h - p\,| - 1}{2}$$

$$n = \frac{df_e - p - 1}{2}$$

$$r = df_e - \frac{p - df_h + 1}{2}$$

$$u = \frac{p \times df_h - 2}{4}$$

$$t = \sqrt{\frac{p^2 df_h^2 - 4}{p^2 + df_h^2 - 5}}, \qquad \text{if } p^2 + df_h^2 - 5 > 0$$

$t = 1$, otherwise,

eigenvalues of $E^{-1} H$ are $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p$

Wilk's Lambda:

$$A = \frac{1}{1+\lambda_1} \times \frac{1}{1+\lambda_2} \times \ldots \times \frac{1}{1+\lambda_p}$$

$$F = \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \times \frac{rt - 2u}{p \times df_h}, \text{ with } p \times df_h \text{ and } (rt - 2u) \text{ df}$$

Hotelling-Lawley Trace:

$$U = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

$$F = 2\,(Sn + 1)\,\frac{U}{S^2\,(2m + S + 1)}, \text{ with } S(2m + S + 1) \text{ and } 2(Sn + 1) \text{ df}$$

Pillai's Trace:

$$V = \frac{\lambda_1}{1+\lambda_1} \times \frac{\lambda_2}{1+\lambda_2} \times \ldots \times \frac{\lambda_p}{1+\lambda_p}$$

$$F = \frac{2n + S + 1}{2m + S + 1}\,\frac{V}{S - V}, \text{ with } S(2m + S + 1) \text{ and } S(2n + S + 1) \text{ df}$$

Roy's Greatest Root Test:

$\max(\lambda_1, \lambda_2, \ldots, \lambda_p)$

Logistic Regression

Model:

$$E(\underset{\sim}{Y}) = \frac{1}{1 + \exp(-\underset{\sim}{X}\ \underset{\sim}{\beta})}$$

Logistic Transformation:

$$\log\left(\frac{E(\underset{\sim}{Y})}{1 - E(\underset{\sim}{Y})}\right) = \underset{\sim}{X}\ \underset{\sim}{\beta}$$

Weighted Least Squares:

$Y_i = p_i$ = (sample proportions)

n_i = samples sizes

$$w_i = \frac{n_i p_i}{1 - p_i}$$

by default, all p_i restricted to the range (user controlled)

$$\left(\frac{1}{2n_i}, 1 - \frac{1}{2n_i}\right)$$

n = number of samples

Maximum Likelihood (Proportions):

Maximize

$$L(\hat{\beta}) = \prod_{i=1}^{n}\left(\frac{1}{1 + e^{-x_i \underset{\sim}{\beta}}}\right)^{r_i}\left(\frac{e^{-x_i\beta}}{1 + e^{-x_i\beta}}\right)^{n_i - r_i}$$

where

$r_i = n_i p_i$

Maximum Likelihood (0's and 1's):

Maximize

$$L(\hat{\beta}) = \left(\frac{\prod_{i=1}^{n_i} e^{x_i\beta}}{\prod_{i=1}^{n} (1 + e^{x_i\beta})} \right)$$

where

n_i = number of successes

Likelihood Ratio Test:

$$\lambda(\beta) = -2 \ln \left[\frac{L(\hat{\beta})}{L(\hat{p})} \right]$$

where

$$L(\hat{p}) = \prod_{i=1}^{n} (Y_i)^{Y_i} (1 - Y_i)^{(1 - Y_i)}$$

Percentage of Deviance:

$$R^2 = \frac{\lambda(\beta_1, \beta_2, ..., \beta_{p-1} \mid \beta_o)}{\lambda(\beta_o)}$$

Adjusted Percentage of Deviance:

$$\text{Adjusted } R^2 = \frac{\lambda(\beta_1, \beta_2, ..., \beta_{p-1} \mid \beta_o) - 2p}{\lambda(\beta_o)}$$

Nonlinear Regression

Notation:

$F(X,\hat{\beta})$ = values of nonlinear function using parameter estimates

Marquardt Algorithm:

Estimated coefficients are obtained by minimizing the residual sum of squares using a search procedure suggested by Marquardt. This is a

compromise between the Gauss-Newton linearization and steepest descent methods.

The user specifies:

1. Initial value of Marquardt parameter λ.
2. Multiple p used to modify Marquardt parameter after each interation, where

   $$\lambda_i = \frac{\lambda_{i-1}}{p}$$

   provided SSE decreases, or

   $\lambda_i = p\lambda_{i-1}$ if SSE increases.
3. Convergence Criterion #1 — estimation stops if relative change in SSE is less than this criterion.
4. Convergence Criterion # 2 — estimation stops if relative change in all parameter estimates is less than this criterion.
5. Initial estimates β_0

   Find revised estimates where

   $\beta_i = \beta_{i-1} + \Delta$

   such that

   $SSE(\beta_i) < SSE\left(\beta_{i-1}\right)$

   where

   $$\Delta = (X'X + \lambda \, \text{diag}(X'X))^{-1} X'e$$

   An extended discussion of nonlinear estimation can be found in Draper and Smith, 1981.

Regression Model Selection

$$C_p = \frac{SSE_p}{MSE_F} - (n - 2_p)$$

where

F = the full model

n = the number of observations

p = 1 + the number of independent variables

Ridge Regression

Standardized Coefficient Estimates:

If $\underset{\sim}{Z}$ = n × p matrix of independent variables standardized so that $\underset{\sim}{Z}' \underset{\sim}{Z}$ equals the correlation matrix

and Θ = value of the ridge parameter

then $\underset{\sim}{b}'(\Theta) = (\underset{\sim}{Z}' \underset{\sim}{Z} + \Theta I_p)^{-1} \underset{\sim}{Z}' \underset{\sim}{Y}$

where

I_p is a p × p identity matrix.

Natural Coefficients:

For the ith independent variable,

$$b_i = \frac{b'_i}{S_i} \text{ for i = 1 ... number of independent variables}$$

for the constant,

$$b_o = b'_o - \frac{b'_1 \bar{x}_1}{S_1} - \frac{b'_2 \bar{x}_2}{S_2} - \ldots - \frac{b'_k \bar{x}_k}{S_k}$$

Variance Inflation Factors:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where

R_i^2 = the coefficient of multiple determination R^2 when regressing X_i against the other independent variables