Statistical Advisor

Bernoulli – describes all situations where a "trial" is made resulting in either "success" or "failure", such as when tossing a coin

Beta – arises from a transformation of the F distribution and is typically used to model the distribution of order statistics

Binomial – is useful for describing distributions of binomial events, such as the number of defective components in samples of 20 units taken from production process

Cauchy – is often used in statistics as canonical example of a pathological distribution since both its mean and variances are undefined

Chi-square – the sum of n independent squared random variables, each distributed following the standard normal distribution, is distributed as Chi-square with n degrees of freedom

Exponential – is frequently used to model the time interval between successive random events, example would be the gap length between cars crossing an intersection

Extreme Value – to model extreme events such as size of floods, gust velocities encountered by airplanes, maxima of stock indices over a given year, etc

F – is mostly used in tests of variance (e.g.., ANOVA)

Gamma – when modeling the distribution of the life-times of a product such as an electric light bulb, or the serving time taken at a ticket booth at a baseball game

Geometric – if the independent Bernoulli trials are made until a "success" occurs, then the total number of trials required is a geometric random variable

Gompertz – is theoretical distribution of survival times

Logistic – is used to model binary responses (e.g.., gender) and is commonly used in logistic regression

Log-normal – is often used in simulations of variables such as personal incomes, age at first marriage, or tolerance to poison in animals, etc

Normal – is a bell shaped curve which is symmetrical about mean, is a theoretical function commonly used in inferential statistics as an approximation to sampling distributions, it's a good model for random variable

Pareto – is commonly used in monitoring production processes, example- a machine which produces copper wire will occasionally generate a flaw at some point along the wire

Poisson – distribution of rare events, example – number of accidents per person, number of sweepstakes won per person, etc

Rayleigh – distance of darts from the target in a dart-throwing game

Rectangular – useful for describing random variables with a constant probability density over the defined range

Student's t – is symmetric about 0, its shape is similar to that of standard normal distribution. It is most commonly used in testing hypothesis about the mean of a particular population

Weibull – used when the failure probability varies over time, often used in reliability testing (e.g.., ball bearings, etc)

Statistical significance test - Compare the observed distribution of variables against several theoretical distributions and test the discrepancy of the observed data from the respective theoretical distributions

p-value – statistical significance of a result is an estimated measure of the degree to which it is "true"

Z-value – standardized value - value is expressed in terms of its difference from the mean, divided by sd

Confidence intervals – give us a range of values around the mean where we expect the "true" (population) mean is located

t-Test – to evaluate the differences in means between groups. f the resultant t-value is statistically significant then one can conclude that the means in the two variables are different

F-test – for comparison of the variances in two groups, if statistically significant, one can conclude that the variances (variability) in the two groups are different

General Linear Models (GLM) – assumes that the variables in the comparison are normally distributed within the groups

Generalized Linear Models (GLZ) – doesn't assume that the variables in the analysis follow normal distribution

Variance components & Mixed model ANOVA/MANOVA – techniques for analyzing research designs with random effects, including the estimation of variance components for such effects.
It is also well suited for analyzing large main effect designs, and designs with many factors where the higher order interactions are not of interest, and analysis involving case weights

Discriminant function analysis – how to identify the specific variables that show different means in different groups

Kolmogorov-Smirnov test & Shapiro - Wilks' W test – are used for test for normality

Monte-Carlo studies – to determine how sensitive they are to violations of the assumptions of normal distribution of the analyzed variables in population

X-bar chart – the sample means are plotted in order to control the mean value of a variable

R-chart – the sample ranges are plotted in order to control the variability of a variable

T2 chart** – simultaneously monitor several measurements

OC Curves – extremely useful for exploring the power of our quality control procedure

Independent variable – are those that are manipulated (male/female) / grouping variables

Dependent variable – are those that are measured / registered (WCC/height/weight/salary/etc)

Nominal /Categorical – gender, race, color, city, etc
Ordinal – middle class / rich, bachelor / masters, etc
Interval/Ratio – temperature, income, etc

Correlation – is a measure of the relation between two or more variables
Perfect Negative = -1, Perfect Positive = +1, No Correlation = 0
Residual = deviations from the regression line
Outliers = atypical / infrequent observations

Pearson r – determines the extent to which values of two variables or "proportional" to each other

Spearman R – similar to Pearson r except that it is computed from ranks

Process capability indices are ratios that summarize the extent to which a manufacturing process or supplier turns products or parts within specified engineering limits, example 6Sigma

Factor analysis – is to detect underlying dimensions that explain relations between multiple variables

Correspondence analysis – to analyze two-way and multi-way tables containing some measure of correspondence between rows and columns

Multidimensional scaling – is to detect underlying dimensions for a set of multiple input variables

PLS – allows you to extract factors (components) from a data set that includes one or more predictor variables and one ore more dependant variables

Association rules – is to detect relationships or associations between specific values of categorical variables in large data sets

Tree clustering – objects are linked in successive steps , yielding a tree that ultimately joins all objects

k-means – one specifies a priori how many clusters to expect, program will try to find the best division of objects into the requested number of clusters

Two-way joining – both cases and variables are clustered automatically, it will attempt to form clusters of similar data points (values)

EM clustering – by fitting a mixture of distributions to the data

Neural networks – analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations from other observations after executing a process so-called learning from existing data

ARIMA – is to detect seasonal patterns

Autocorrelation analysis – allows you to examine the lagged effect (correlation) of a variable with itself or with another variable

Fourier analysis – is to decompose a time series into simple underlying waves forms

Distributed lags – for analyzing the lagged effects of one independent variable with one dependant variable. This can simultaneously evaluate multiple lags

Neural networks – to predict lagged response variables

Log linear – techniques for the analysis of multi-way frequency (cross tabulation) tables, to select the best model for the data

Classification trees – for predicting the membership of cases or objects in the classes of a categorical dependant variable from their measurements on one or more predictor variables

MARSplines – nonparametric regression procedure which makes no assumption about the underlying functional relationship between the dependant and independent variables

Random Forest – consists of an arbitrary number of simple trees, which are used to determine the final outcome

Fourier – decompose a complex time series with cyclical components into a few underlying sinusoidal functions of particular wavelengths

ARIMA – allows us to uncover the hidden patterns in the data but also generate forecasts on patterns of the data are unclear and observations involve considerable error

Nonparametric – where we know nothing about the parameters of the variable of interest in the population, do not rely on mean or sd describing distribution of the variable of interest. Or data contains rankings than precise measurements

Life table analysis – the proportion surviving up to the respective interval, the proportion failing in the interval, the hazard rate for the interval, percentiles of the cumulative survival function

Kaplan-Meier product-limit estimates – life table analysis computed continuously

Specialized distributions – Weibull, Gompertz, linear hazard

Gage repeatability – to the extent to which repeated measurements of the same part by the same operator (of the gage) produce identical results

Gage reproducibility – to the extent to which different operators measuring the same parts with the same gage produce identical measurements

Mean – average or central tendency of a continuous distribution

Standard deviation & Variance – measures of dispersion, that is, the variability of data

Skewness – measure of the sidedness of the distribution – skewed towards right or left of the mean

Kurtosis – measure of pointedness or peakedness of the distribution - spread or centered closely around the mean

Histogram – examine frequency distributions of values of variables

Scatter plot – visualize relations between two or three variables

Probability plot – provide a quick way to visually inspect to what extent the pattern of data follows a distribution

Quantile-Quantile plot – is useful for finding the best fitting distribution within a family of distributions

Probability-Probability plot – is useful for determining how well a specific theoretical distribution fits the observed data

Line plot – individual data points are connected by a line, provides a simple way to visually represent a sequence values

Box plot – ranges of values of a selected variables are plotted separately for groups of cases

Pie chart – used for representing proportions of values of variables

Missing data point plot – to visualize the pattern or distribution of missing data

Ternary plot – to examine relations between three or more dimensions where three of those dimensions represent components of a mixture

Icon plots – represent cases or units of observation as multidimensional symbols, to spot complex relations

Chernoff faces – face is drawn with relative values of the selected variables for each case are assigned to shapes and sizes of individual face features (sun rays, stars, etc are variations of the same plot)

Contour plot – is the projection of a 3-dimesnsional surface onto a 2-dimensional plane

Ishikawa chart – to depict the factors or variables that make up a process (cause-and-effect diagram)

Gain /Lift chart – provides a visual summary of the usefulness of the information provided by one or more statistical models

Matrix plot – summarizes the relationships between several variables in a matrix of true x-y plots

Pareto chart – identify the causes of quality problems or loss

ROC curve – used to evaluate the goodness of fit for a classifier

Ternary plot – the triangular coordinate systems are used to plot three (or more) variables

Brushing – an interactive method that enables us to select on-screen specific data points and identify their characteristics

Regression – objective is to estimate the value of a continuous output variable from some input variables

Multiple regression analysis – in which one dependant variable is related to multiple independent variables

PLS – allow you to extract factors from a dataset that includes one or more predictor variables, and one or more dependent variables

Polynomial regression computes the relationship between a dependant variable with one or more independent variables, and those independent variables squared, cubed, etc.
Polynomial regression is to detect some curvilinearity in a relationship

Logit and probit regression are used for analyzing the relationship between one or more independent variables with a categorical dependant variable at two levels

A data set is **Censored** if some observations are incomplete but not missing. For example – a part or product may not fail within the time span covered by study, however we do not know how long it will function properly thereafter

ANOVA/MANOVA – is to test for significant differences between means by comparing variances

McNemar test for changes in proportions – example, if one wants to compare how many students in a class fail a particular test at the beginning of the semester, and at the end of semester

Canonical correlation – to investigate the simultaneous relationship between two set of variables

Process Analysis – to what extent does the long-term performance of the process comply with engineering requirements or managerial goals?

Pareto chart – identify the few that account for the majority of problems

Perform Pareto analysis

Regression control chart – monitor the relationship between two aspects of the production process

Regression control charts

Quality control charts – monitor the extent to which our products meet specifications

Compute quality control charts for on-line quality control (Shewhart control charts)

Compute quality control charts

Standard process capability histogram

Perform gage repeatability / reproducibility analysis

Perform process capability analysis

Factor Analysis, Correspondence analysis, Multidimensional scaling, Partial least squares

Survival analysis – Weibull, Gompertz, exponential, linear hazard

Analyze failure times

Power analysis – sample size estimation & confidence interval estimation

Develop/analyze sampling plans

Designing / analyzing experiments (including Taguchi)

Experimental designs – apply analysis of variance principles to product development

Factors or dimensions in multiple continuous variables

Compute statistics for industrial quality control / improvement?

Distributions – normal, exponential, gamma, log-normal, Chi-square, binomial, geometric, etc. Statistical significance tests – Chi-square test, Kolmogorov - Smirnov test
Survival analysis - Weibull, Gompertz, linear & exponential hazard
Various curve fitting routines
Kaplan-Meier product-limit method

The shape of or fitting distributions

Test hypothesis (predictions) about your data?

Do you want to

Explore data and search for structure / patterns / factors / clusters?

Describe / Summarize / Tabulate Data?

Differences between groups/samples

Differences between variables

Relationship between variables

Differences in relationship between variables in different groups

Association Rules Analysis

Associations between variables

Clusters or natural groups of variables or cases

Patterns or trends in observations over time

Relationships in multi-way cross tabulation tables

Relationships between predictors and responses on categorical dependant variables

Explore/Summarize a time series

Cluster analysis – tree clustering, k-means clustering, two-way joining, EM clustering

Autocorrelation analysis, ARIMA, Fourier analysis, Neural networks

Generalized linear models, Log linear

Decision trees, Classification trees, Discriminant analysis, cluster analysis, nonparametric statistics - MARSplines, nonlinear estimation, Naïve Bayes, Random Forest

Use Time Series, Autocorrelation analysis, Fourier decomposition, seasonal & non-seasonal ARIMA

Tabulate/Plot categorical data (such as gender, occupation) and compute frequencies, percentages, etc

Summarize/Plot the shape of the distribution of continuous variables/measurements

Simple descriptive statistics and frequency distributions

Basic statistics -> mean, sd, variance, skewness, kurtosis, frequency distribution tables,

histograms, pie charts, Chernoff faces, star, sun ray plots, etc

Detailed descriptive statistics Nonparametric

Specialized industrial descriptive stats and stats for survival/failure times

Descriptive statistics by groups, samples or variables

Breakdown table of means of a variable (income by gender, within gender by occupation, etc), compute box plots, statistics broken down by another variable.
Histograms, line graphs, categorized plots

Simple frequency tables and plot histograms or stem-and-leaf diagrams

Basic statistics -> mean, sd, variance, skewness, kurtosis, frequency distribution tables,

histograms, pie charts, Chernoff faces, star, sun ray plots, etc

Cross tabulate two or more variables to describe their joint distribution (cross tabulation tables, banner tables)

Basic statistics + Harmonic & geometric mean, mode, median, quartiles, percentiles, quartile range, minimum, maximum, sum, etc

Basic statistics + two & higher way cross tabulation tables, banner tables, log linear, 2d & 3d histograms

Censored data (failure times, survival times)

Use Survival analysis

Process capability (of an industrial/manufacturing process)

Use Process analysis

Gage repeatability & reproducibility

Use Process analysis

t-Test for independent samples, F-test for the comparison of variances in two groups, Nonparametric tests, histograms, line graphs, scatter plots, etc

Comparison of means/variances in two groups/samples

Comparison of means/variances in multiple groups

General (nonparametric) comparisons of distributions in two or more groups

Comparison of survival/failure times in two or more groups

General Linear Models – ANOVA / MANOVA
Generalized Linear Models – ANOVA like, maximum likelihood methods
Variance Components and mixed model ANOVA / ANCOVA,
Discriminant function Analysis,
Nonparametric tests,

Histograms, line graphs, scatter plots, etc

Rank order tests, Kolmogorov-Smirnov two sample test
Generalized Linear Models – maximum likelihood methods,

Histograms, line graphs, scatter plots, etc

Wald - Wolfowitz test, Mann-Whitney U test, Kolmogorov-Smirnov two-sample test, Kruskal -Wallis test, Friedman's two way analysis, Cochran Q test

Use Survival analysis – Gehan's generalized Wilcoxon test, the Cox-Mantel test, the Cox F-test, the log-rank test, Peto and Peto's generalized Wilcoxon test. Nonparametric tests.

Comparison of means for two variables or measurements

Comparison of means for several variables or repeated measurements

General (nonparametric) comparison of two variables

General (nonparametric) comparison of several variables

t-Test for dependant samples
Nonparametric tests,

Histograms, line graphs, scatter plots, etc

General Linear Model – repeated measures ANOVA / MANOVA
Nonparametric tests – Friedman analysis of variance,

Histograms, line graphs, scatter plots, etc

Nonparametric tests, rank order tests, McNemar test

Histograms, line graphs, scatter plots, etc

Nonparametric tests, rank order tests, McNemar test

Histograms, line graphs, scatter plots, etc

Simple linear relationships between two continuous variables

Simple relationships between two categorical variables

Multiple linear relationships between continuous variables

Multiple relationships between categorical variables

Relation between two sets of variables

Nonlinear relationships

Time-dependant (lagged) relationships

Analysis of covariance

Stratified linear or nonlinear regression

Standard Pearson r, Multiple regression, Nonparametric (Spearman R, Kendall tau, Gamma coefficient)
Survival analysis, Proportional hazard regression model

2D scatter plots, line graphs, etc

Two-way or higher-way cross tabulation tables, Nonparametric tests, Fisher exact test, McNemar's test

3D histograms

Log linear, Correspondence analysis, Factor analysis, predictive mapping, Visual generalized linear model, link functions

Canonical Correlation

Time series, Neural networks

ANCOVA

General linear models, Generalized linear models, Nonlinear estimation, Survival analysis, Scatter plots, surface plots etc

General regression models, Generalized linear models, multiple regression, partial least squares, survival analysis, proportional hazard regression model

Polynomial regression analysis

Probit / Logit regression for categorical dependant and continuous independent variables

General nonlinear regression

Nonlinear regression analysis for censored survival times

General regression models, general linear models, multiple regression

Generalized linear model, nonlinear estimation

Generalized linear model, nonlinear estimation
2D, 3D fitting of lines or surfaces to observed data

Use Survival analysis

Reference: StatSoft Online Textbook. Prepared By: Rajkumar BN (Draft version)