

NAG Toolbox for MATLAB**Chapter Introduction****G01 – Simple Calculations on Statistical Data****Contents**

1	Scope of the Chapter	2
2	Background to the Problems	2
2.1	Plots, Descriptive Statistics and Exploratory Data Analysis	2
2.2	Statistical Distribution Functions and Their Inverses	2
2.3	Testing for Normality and Other Distributions	3
2.4	Distribution of Quadratic Forms	3
2.5	Energy Loss Distributions	4
3	Recommendations on Choice and Use of Available Functions	4
4	References	6

1 Scope of the Chapter

This chapter covers three topics:

- plots, descriptive statistics, and exploratory data analysis;
- statistical distribution functions and their inverses;
- testing for Normality and other distributions.

2 Background to the Problems

2.1 Plots, Descriptive Statistics and Exploratory Data Analysis

Plots and simple descriptive statistics are generally used for one of two purposes:

- the presentation of data;
- exploratory data analysis.

Exploratory data analysis (EDA) is used to pick out the important features of the data in order to guide the choice of appropriate models. EDA makes use of simple displays and summary statistics. These may suggest models or transformations of the data which can then be confirmed by further plots. The process is interactive between you, the data, and the program producing the EDA displays. In a formal presentation of data, selected features of the data are displayed for others to examine. In this situation high-quality graphics are often needed (for example, the NAG Graphics Library) but the character plots produced by functions in this chapter are usually adequate for EDA work.

The summary statistics consist of two groups. The first group are those based on moments; for example mean, standard deviation, coefficient of skewness, and coefficient of kurtosis (sometimes called the ‘excess of kurtosis’, which has the value 0 for the Normal distribution). These statistics may be sensitive to extreme observations and some robust versions are available in Chapter G07. The second group of summary statistics are based on the order statistics, where the i th order statistic in a sample is the i th smallest observation in that sample. Examples of such statistics are minimum, maximum, median and hinges.

In addition to summarizing the data by using suitable statistics the data can be displayed using tables and diagrams. Such data displays include frequency tables, stem and leaf displays, box and whisker plots, histograms and scatter plots.

2.2 Statistical Distribution Functions and Their Inverses

Statistical distributions are commonly used in three problems:

- evaluation of probabilities and expected frequencies for a distribution model;
- testing of hypotheses about the variables being observed;
- evaluation of confidence limits for parameters of fitted model, for example the mean of a Normal distribution.

Random variables can be either discrete (i.e., they can take only a limited number of values) or continuous (i.e., can take any value in a given range). However, for a large sample from a discrete distribution an approximation by a continuous distribution, usually the Normal distribution, can be used. Distributions commonly used as a model for discrete random variables are the binomial, hypergeometric, and Poisson distributions. The binomial distribution arises when there is a fixed probability of a selected outcome as in sampling with replacement, the hypergeometric distribution is used in sampling from a finite population without replacement, and the Poisson distribution is often used to model counts.

Distributions commonly used as a model for continuous random variables are the Normal, gamma, and beta distributions. The Normal is a symmetric distribution whereas the gamma is skewed and only appropriate for nonnegative values. The beta is for variables in the range $[0, 1]$ and may take many different shapes. For circular data, the ‘equivalent’ to the Normal distribution is the von Mises distribution. The assumption of the Normal distribution leads to procedures for testing and interval estimation based on the χ^2 , F (variance ratio), and Student’s t -distributions.

In the hypothesis testing situation, a statistic X with known distribution under the null hypothesis is evaluated, and the probability α of observing such a value or one more ‘extreme’ value is found. This probability (the significance) is usually then compared with a preassigned value (the significance level of the test), to decide whether the null hypothesis can be rejected in favour of an alternate hypothesis on the basis of the sample values. Many tests make use of those distributions derived from the Normal distribution as listed above, but for some tests specific distributions such as the Studentized range distribution and the distribution of the Durbin–Watson test have been derived. Non-parametric tests as given in Chapter G08, such as the Kolmogorov–Smirnov test, often use statistics with distributions specific to the test. The probability that the null hypothesis will be rejected when the simple alternate hypothesis is true (the power of the test) can be found from the noncentral distribution.

The confidence interval problem requires the inverse calculation. In other words, given a probability α , the value x is to be found, such that the probability that a value not exceeding x is observed is equal to α . A confidence interval of size $1 - 2\alpha$, for the quantity of interest, can then be computed as a function of x and the sample values.

The required statistics for either testing hypotheses or constructing confidence intervals can be computed with the aid of functions in this chapter, and Chapter G02 (Regression), Chapter G04 (Analysis of Designed Experiments), Chapter G13 (Time Series), and Chapter E04 (Non-linear Least-squares Problems).

Pseudo-random numbers from many statistical distributions can be generated by functions in Chapter G05.

2.3 Testing for Normality and Other Distributions

Methods of checking that observations (or residuals from a model) come from a specified distribution, for example, the Normal distribution, are often based on order statistics. Graphical methods include the use of **probability plots**. These can be either $P - P$ plots (probability–probability plots), in which the empirical probabilities are plotted against the theoretical probabilities for the distribution, or $Q - Q$ plots (quantile–quantile plots), in which the sample points are plotted against the theoretical quantiles. $Q - Q$ plots are more common, partly because they are invariant to differences in scale and location. In either case if the observations come from the specified distribution then the plotted points should roughly lie on a straight line.

If y_i is the i th smallest observation from a sample of size n (i.e., the i th order statistic) then in a $Q - Q$ plot for a distribution with cumulative distribution function F , the value y_i is plotted against x_i , where $F(x_i) = (i - \alpha)/(n - 2\alpha + 1)$, a common value of α being $\frac{1}{2}$. For the Normal distribution, the $Q - Q$ plot is known as a Normal probability plot.

The values x_i used in $Q - Q$ plots can be regarded as approximations to the expected values of the order statistics. For a sample from a Normal distribution the expected values of the order statistics are known as **Normal scores** and for an exponential distribution they are known as **Savage scores**.

An alternative approach to probability plots are the more formal tests. A test for Normality is the Shapiro and Wilks W Test, which uses Normal scores. Other tests are the χ^2 goodness of fit test and the Kolmogorov–Smirnov test; both can be found in Chapter G08.

2.4 Distribution of Quadratic Forms

Many test statistics for Normally distributed data lead to quadratic forms in Normal variables. If X is a n -dimensional Normal variable with mean μ and variance-covariance matrix Σ then for an n by n matrix A the quadratic form is

$$Q = X^T A X.$$

The distribution of Q depends on the relationship between A and Σ : if $A\Sigma$ is idempotent then the distribution of Q will be central or noncentral χ^2 depending on whether μ is zero.

The distribution of other statistics may be derived as the distribution of linear combinations of quadratic forms, for example the Durbin–Watson test statistic, or as ratios of quadratic forms. In some cases rather than the distribution of these functions of quadratic forms the values of the moments may be all that is required.

2.5 Energy Loss Distributions

An application of distributions in the field of high-energy physics where there is a requirement to model fluctuations in energy loss experienced by a particle passing through a layer of material. Three models are commonly used:

- (i) Gaussian (Normal) distribution;
- (ii) the Landau distribution;
- (iii) the Vavilov distribution.

Both the Landau and the Vavilov density functions can be defined in terms of a complex integral. The Vavilov distribution is the more general energy loss distribution with the Landau and Gaussian being suitable for when the Vavilov parameter κ is less than 0.01 and greater than 10.0 respectively.

3 Recommendations on Choice and Use of Available Functions

Descriptive statistics / Exploratory analysis:

plots:

box and whisker	g01as
histogram	g01aj
Normal probability ($Q - Q$) plot	g01ah
scatter plot	g01ag
stem and leaf	g01ar

summaries:

frequency / contingency table,	
one variable	g01ae
two variables, with χ^2 and Fisher's exact test	g01af
mean, variance, skewness, kurtosis (one variable),	
from frequency table	g01ad
mean, variance, sums of squares and products (two variables)	g01ab
median, hinges / quartiles, minimum, maximum	g01al

Descriptive statistics / Exploratory analysis:

summaries:

mean, variance, skewness, kurtosis (one variable),	
from raw data	g01aa

Distributions:

Beta:

central:

deviates	g01fe
probabilities and probability density function	g01ee

non-central:

probabilities	g01ge
---------------------	-------

Binomial:

distribution function	g01bj
-----------------------------	-------

Durbin–Watson statistic:

probabilities	g01ep
---------------------	-------

Energy loss distributions:

Landau:

density	g01mt
derivative of density	g01rt
distribution	g01et
first moment	g01pt
inverse distribution	g01ft
second moment	g01qt

Vavilov:

density	g01mu
distribution	g01eu
initialization	g01zu

<i>F</i> :	
central:	
deviates	g01fd
probabilities	g01ed
non-central:	
probabilities	g01gd
Gamma:	
deviates	g01ff
probabilities	g01ef
Hypergeometric:	
distribution function	g01bl
Kolomogorov–Smirnov:	
probabilities:	
one-sample	g01ey
two-sample	g01ez
Normal:	
bivariate:	
probabilities	g01ha
multivariate:	
probabilities	g01hb
quadratic forms:	
cumulants and moments	g01na
moments of ratios	g01nb
univariate:	
deviates	g01fa
probabilities	g01ea
reciprocal of Mill's Ratio	g01mb
Shapiro and Wilks test for Normality	g01dd
Poisson:	
distribution function	g01bk
Student's <i>t</i> :	
central:	
deviates	g01fb
probabilities	g01eb
non-central:	
probabilities	g01gb
Studentized range statistic:	
deviates	g01fm
probabilities	g01em
von Mises:	
probabilities	g01er
χ^2 :	
central:	
deviates	g01fc
probabilities	g01ec
probability of linear combination	g01jd
non-central:	
probabilities	g01gc
probability of linear combination	g01jc
Scores:	
Normal scores, ranks or exponential (Savage) scores	g01dh
Normal scores:	
accurate	g01da
approximate	g01db
variance-covariance matrix	g01dc

Note: the Student's *t*, χ^2 , and *F* functions do not aim to achieve a high degree of accuracy, only about four or five significant figures, but this should be quite sufficient for hypothesis-testing. However, both the Student's *t* and the *F* distributions can be transformed to a beta distribution and the χ^2 -distribution can be

transformed to a gamma distribution, so a higher accuracy can be obtained by calls to the gamma or beta functions.

Note: g01dh computes either ranks, approximations to the Normal scores, Normal, or Savage scores for a given sample. g01dh also gives you control over how it handles tied observations. g01da computes the Normal scores for a given sample size to a requested accuracy; the scores are returned in ascending order. g01da can be used if either high accuracy is required or if Normal scores are required for many samples of the same size, in which case you will have to sort the data or scores.

4 References

Hastings N A J and Peacock J B 1975 *Statistical Distributions* Butterworths

Kendall M G and Stuart A 1969 *The Advanced Theory of Statistics (Volume 1)* (3rd Edition) Griffin

Tukey J W 1977 *Exploratory Data Analysis* Addison–Wesley
