NAG Library Chapter Introduction

G01 – Simple Calculations on Statistical Data

Contents

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.

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 ith order statistic in a sample is the ith smallest observation in that sample. Examples of such statistics are minimum, maximum, median, hinges and quantiles.

In addition to summarising 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 non-negative 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. Nonparametric 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 routines in this chapter, and Chapter G02 (for regression), Chapter G04 (for analysis of designed experiments), Chapter G13 (for time eries), and Chapter E04 (for nonlinear least squares problems).

Pseudorandom numbers from many statistical distributions can be generated by routines 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 ith smallest observation from a sample of size n (i.e., the ith 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 Wilk's 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

$$Q = X^{\mathrm{T}} A X.$$

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 when the Vavilov parameter κ is less than 0.01 and greater than 10.0 respectively.

2.6 Vectorized Routines

A number of vectorized routines are included in this chapter. Unlike their scalar counterparts, which take a single set of parameters and perform a single function evaluation, these routines take vectors of parameters and perform multiple function evaluations in a single call. The input arrays to these vectorized routines are designed to allow maximum flexibility in the supply of the parameters by reusing, in a cyclic manner, elements of any arrays that are shorter than the number of functions to be evaluated, where the total number of functions evaluated is the size of the largest array.

To illustrate this we will consider G01SFF, a vectorized version of G01EFF, which calculates the probabilities for a gamma distribution. The gamma distribution has two parameters α and β therefore G01SFF has four input arrays, one indicating the tail required (TAIL), one giving the value of the gamma variate, g, whose probability is required (G), one for α (A) and one for β (B). The lengths of these arrays are LTAIL, LG, LA and LB respectively.

For sake of argument, lets assume that $\mathrm{LTAIL} = 1$, $\mathrm{LG} = 2$, $\mathrm{LA} = 3$ and $\mathrm{LB} = 4$, then $\max(\mathrm{LTAIL}, \mathrm{LG}, \mathrm{LA}, \mathrm{LB}) = 4$ values will be returned. These four probabilities would be calculated using the following parameters:

i	**Tail**	g	α	β
1	TAIL(1)	G(1)	A(1)	B(1)
2	TAIL(1)	G(2)	A(2)	B(2)
3	TAIL(1)	G(1)	A(3)	B(3)
4	TAIL(1)	G(2)	A(1)	B(4)

3 Recommendations on Choice and Use of Available Routines

Note: the Student's t, χ^2, and F routines 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 routines.

Note: G01DHF computes either ranks, approximations to the Normal scores, Normal, or Savage scores for a given sample. G01DHF also gives you control over how it handles tied observations. G01DAF computes the Normal scores for a given sample size to a requested accuracy; the scores are returned in ascending order. G01DAF 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.

3.1 Working with Streamed or Extremely Large Datasets

The majority of the routines in this chapter are 'in-core', that is all the data required must be held in memory prior to calling the routine. In some situations this might not be possible, for example, when working with extremely large datasets or where all of the data is not available at once (i.e., the data is being streamed).

There are five routines in this chapter applicable to datasets of this form:

G01ATF computes the mean, variance and the coefficients of skewness and kurtosis for a single variable.

G01AUF, takes the results from two calls to G01ATF and combines them, returning the mean, variance and the coefficients of skewness and kurtosis for the combined dataset. This routine allows the easy utilization of more than one processor to spread the computational burden inherent in summarising a very large dataset.

G01ANF and G01APF compute the approximate quantiles for a dataset of known and unknown size respectively.

G01WAF computes the mean and standard deviation in a rolling window.

In addition, see G02BUF and G02BZF for routines to summarise two or more variables.

4 Auxiliary Routines Associated with Library Routine Parameters

None.

5 Routines Withdrawn or Scheduled for Withdrawal

The following lists all those routines that have been withdrawn since Mark 17 of the Library or are scheduled for withdrawal at one of the next two marks.

Withdrawn Routine	**Mark of Withdrawal**	**Replacement Routine(s)**
G01AAF	26	G01ATF
G01CEF	18	G01FAF

6 References

Hastings N A J and Peacock J B (1975) *Statistical Distributions* Butterworth

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