NAG C Library Chapter Introduction

g01 – Simple Calculations on Statistical Data

Contents

1 Scope of the Chapter

This chapter covers three topics:

summary statistics

statistical distribution functions and their inverses;

testing for Normality and other distributions.

2 Background to the Problems

2.1 Summary Statistics

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 and hinges.

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. 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 non-central 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 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 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

$$Q = X^T AX.$$

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 non-central 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.

Bot 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 then 0.01 and greater than 10.0 repectively.

3 Recommendations on Choice and Use of Available Functions

The following functions are recommended for the tasks described.

Distribution Functions and their Inverses
- Continuous distributions:
  - χ^2 distribution ..... nag_prob_chi_sq (g01ecc)
  - beta distribution ..... nag_prob_beta_dist (g01eec)
  - bounds for the significance of the Durbin–Watson statistic nag_prob_durbin_watson (g01epc)
  - distribution of the Studentized range statistic ..... nag_prob_studentized_range (g01emc)
  - F (variance-ratio) distribution ..... nag_prob_f_dist (g01edc)
  - gamma distribution ..... nag_gamma_dist (g01efc)
  - Normal distribution ..... nag_prob_normal (g01eac)
  - one sample Kolmogorov–Smirnov distribution ..... nag_prob_1_sample_ks (g01eyc)
  - Student's t-distribution ..... nag_prob_students_t (g01ebc)
  - two sample Kolmogorov–Smirnov distribution ..... nag_prob_2_sample_ks (g01ezc)
  - von Mises distribution ..... nag_prob_von_mises (g01erc)
- Discrete distributions:
  - binomial ..... nag_binomial_dist (g01bjc)
  - hypergeometric ..... nag_hypergeom_dist (g01blc)
  - Poisson ..... nag_poisson_dist (g01bkc)
- Distribution of functions of quadratic forms of Normal variables:
  - linear combination of (central) χ^2 variables ..... nag_prob_lin_chi_sq (g01jdc)
  - moments of quadratic forms ..... nag_moments_quad_form (g01nac)
  - moments of ratios of quadratic forms ..... nag_moments_ratio_quad_forms (g01nbc)
  - positive linear combination of (non-central) χ^2 variables nag_prob_lin_non_central_chi_sq (g01jcc)
- Inverses of distribution functions:
  - χ^2 ..... nag_deviates_chi_sq (g01fcc)
  - beta ..... nag_deviates_beta (g01fec)
  - distribution of the Studentized range statistic .... nag_deviates_studentized_range (g01fmc)
  - F (variance-ratio) ..... nag_deviates_f_dist (g01fdc)
  - gamma ..... nag_deviates_gamma_dist (g01ffc)
  - Normal distribution ..... nag_deviates_normal (g01fac)
  - Student's t ..... nag_deviates_students_t (g01fbc)
- Multivariate distributions:
  - bivariate Normal distribution ..... nag_bivariate_normal_dist (g01hac)
  - multivariate Normal distribution ..... nag_multi_normal (g01hbc)
- Non-central distributions:
  - non-central χ^2 distribution ..... nag_prob_non_central_chi_sq (g01gcc)
  - non-central beta distribution ..... nag_prob_non_central_beta_dist (g01gec)
  - non-central F (variance-ratio) distribution ..... nag_prob_non_central_f_dist (g01gdc)
  - non-central Student's t-distribution ..... nag_prob_non_central_students_t (g01gbc)
- Other related functions:
  - reciprocal of Mills' ratio ..... nag_mills_ratio (g01mbc)

Plots, Descriptive Statistics, Summary Statistics and Exploratory Data Analysis
- Data displays:
  - frequency table produced from a set of data ..... nag_frequency_table (g01aec)
- Descriptive statistics:
  - five-point summary ..... nag_5pt_summary_stats (g01alc)
  - mean,
    - standard deviation etc. from a frequency table ..... nag_summary_stats_freq (g01adc)
    - standard deviation etc. from raw data ..... nag_summary_stats_1var (g01aac)

Testing for Normality and other distributions
- Calculation of Normal Scores,
  - the expected value of the order statistics from a standard Normal sample nag_normal_scores_exact (g01dac)
- Calculation of ranks and scores ..... nag_ranks_and_scores (g01dhc)

Calculation of the variance-covariance matrix of the order statistics from a standard Normal sample ..... nag_normal_scores_var (g01dcc)
Calculation of the W test for Normality ..... nag_shapiro_wilk_test (g01ddc)
Energy loss distributions
  Landau distribution
    density ..... nag_prob_density_landau (g01mtc)
    derivative of density ..... nag_prob_der_landau (g01rtc)
    distribution ..... nag_prob_landau (g01etc)
    first moment ..... nag_moment_1_landau (g01ptc)
    inverse distribution ..... nag_deviates_landau (g01ftc)
    second moment ..... nag_moment_2_landau (g01qtc)
  Vavilov distribution
    density ..... nag_prob_density_vavilov (g01muc)
    distribution ..... nag_prob_vavilov (g01euc)
    initialisation ..... nag_init_vavilov (g01zuc)

Note: the Student's t, χ^2, and F functions do not aim to achieve a high degree of accuracy, only about 4 or 5 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: nag_ranks_and_scores (g01dhc) computes either ranks, approximations to the Normal scores, Normal, or Savage scores for a given sample. nag_ranks_and_scores (g01dhc) also gives the user control over how it handles tied observations. nag_normal_scores_exact (g01dac) computes the Normal scores for a given sample size to a requested accuracy; the scores are returned in ascending order. nag_normal_scores_exact (g01dac) 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 the user will have to sort the data or scores.

4 Functions Withdrawn or Scheduled for Withdrawal

None.

5 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