NAG Fortran Library 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 Routines 4

4 Routines Withdrawn or Scheduled for Withdrawal 6

5 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 the user, 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 routines 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 ith order statistic in a sample is the ith 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 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 routines 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 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 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 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.

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 then 0.01 and greater than 10.0 repectively.

3 Recommendations on Choice and Use of Available Routines

Descriptive statistics / Exploratory analysis:
- plots:
  - Box and Whisker G01ASF
  - histogram G01AJF
  - Normal probability ($Q - Q$) plot G01AHF
  - scatter plot G01AGF
  - stem and leaf G01ARF
- summaries:
  - frequency / contigency table,
    - one variable G01AEF
    - two variables, with χ^2 and Fisher's exact test G01AFF
  - mean, variance, skewness, kurtosis (one variable),
    - from frequency table G01ADF
  - mean, variance, sums of squares and products (two variables) G01ABF
  - median, hinges / quartiles, minimum, maximum G01ALF

Descriptive statistics / Exporatory analysis:
- summaries:
  - mean, variance, skewness, kurtosis (one variable),
    - from raw data G01AAF

Distributions:
- Beta:
  - central:
    - deviates G01FEF
    - probabilities and probability density function G01EEF
  - non-central:
    - probabilities G01GEF
- Binomial:
  - distribution function G01BJF
- Durbin–Watson statistic:
  - probabilities G01EPF
- Energy loss distributions:
  - Landau:
    - density G01MTF
    - derivative of density G01RTF
    - distribution G01ETF
    - first moment G01PTF
    - inverse distribution G01FTF
    - second moment G01QTF
  - Vavilov:
    - density G01MUF
    - distribution G01EUF
    - initialization G01ZUF

- F:
  - central:
    - deviates G01FDF
    - probabilities G01EDF
  - non-central:
    - probabilities G01GDF
- Gamma:
  - deviates G01FFF
  - probabilities G01EFF
- Hypergeometeric:
  - distribution function G01BLF
- Kolomogorov–Smirnov:
  - probabilities:
    - one-sample G01EYF
    - two-sample G01EZF
- Normal:
  - bivariate:
    - probabilities G01HAF
  - multivariate:
    - probabilities G01HBF
    - quadratic forms:
      - cumulants and moments G01NAF
      - moments of ratios G01NBF
  - univariate:
    - deviates G01FAF
    - probabilities G01EAF
    - reciprocal of Mill's Ratio G01MBF
    - Shapiro and Wilks test for Normality G01DDF
- Poisson:
  - distribution function G01BKF
- Student's t:
  - central:
    - deviates G01FBF
    - probabilities G01EBF
  - non-central:
    - probabilities G01GBF
- Studentized range statistic:
  - deviates G01FMF
  - probabilities G01EMF
- von Mises:
  - probabilities G01ERF
- χ^2:
  - central:
    - deviates G01FCF
    - probabilities G01ECF
    - probability of linear combination G01JDF
  - non-central:
    - probabilities G01GCF
    - probability of linear combination G01JCF

Scores:

- Normal scores, ranks or exponential (Savage) scores G01DHF
- Normal scores:
  - accurate G01DAF
  - approximate G01DBF
  - variance-covariance matrix G01DCF

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

Note: G01DHF computes either ranks, approximations to the Normal scores, Normal, or Savage scores for a given sample. G01DHF also gives the user 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 the user will have to sort the data or scores.

4 Routines Withdrawn or Scheduled for Withdrawal

Withdrawn Routine	Mark of Withdrawal	Replacement Routine(s)
G01ACF	9	G04BBF
G01BAF	16	G01EBF
G01BBF	16	G01EDF
G01BCF	16	G01ECF
G01BDF	16	G01EEF
G01CAF	16	G01FBF
G01CBF	16	G01FDF
G01CCF	16	G01FCF
G01CDF	16	G01FEF
G01CEF	18	G01FAF

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