NAG C Library Chapter Introduction

g01 – Simple Calculations on Statistical Data

Contents

1 Scope of the Chapter

This chapter covers both the calculation of simple descriptive statistics such as means and the calculation of probabilities and deviates (percentage points) for standard distributions.

2 Background

2.1 Summary Statistics

Summary statistics consist of two groups. Firstly, there are those based on moments, for example mean, standard deviation, coefficient of skewness, and coefficient of kurtosis. Secondly, there are the summary statistics 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 and median values. A five-point summary of the data (minimum, maximum, median and hinges or quantiles) can be used in the construction of box and whisker plots which illustrate the location, spread and shape of the data, alternateively, a histogram can be drawn from a frequency table computed from the data.

2.2 Statistical Distribution Functions and their Inverses

Most of the functions available are concerned with statistical distribution functions. They are commonly used in three problem areas:

(i) evaluation of probabilities and expected frequencies for a distribution model

(ii) testing of hypotheses about the variables being observed

(iii) evaluation of confidence limits for parameters of a 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 for non-negative values. The beta is for variables in the range [0,1] and may take many different shapes. The assumption of the Normal distribution leads to procedures for testing and interval estimation based on the Normal, 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' is found. This probability (the significance) is usually then compared with a preassigned value (the significance level of the test), to decide whether, on the basis of the sample values, the null hypothesis can be rejected in favour of an alternate hypothesis. The probability that the null hypothesis will be rejected when an alternative 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 of observing a value not exceeding x is equal to α. For the parameter of interest a confidence interval of size $1 - 2\alpha$ can then be computed as a function of the values of x corresponding to probabilities α and $1 - \alpha$ and suitable statistics calculated from the sample values.

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.

3 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)* Griffin (3rd Edition)

Tukey J W (1977) *Exploratory Data Analysis* Addison-Wesley

4 Available Functions

4.1 Summary Statistics

g01aac Mean, variance, skewness, kurtosis etc, one variable, from raw data

g01alc Five-point summary (median, hinges and extremes)

g01aec Frequency table from raw data

4.2 Discrete Distributions

g01bjc Binomial distribution function

g01bkc Poisson distribution function

g01blc Hypergeometric distribution function

4.3 Statistical Distribution Functions

g01eac Probabilities for the standard Normal distribution

g01ebc Probabilities for Student's t-distribution

g01ecc Probabilities for χ^2 distribution

g01edc Probabilities for F-distribution

g01eec Upper and lower tail probabilities and probability density function for the beta distribution

g01efc Probabilities for the gamma distribution

g01hac Probability for the bivariate Normal distribution

g01hbc Computes probabilities for the multivariate Normal distribution

4.4 Inverses of Statistical Distribution Functions

g01fac Deviates for the Normal distribution

g01fbc Deviates for Student's t-distribution

g01fcc Deviates for the χ^2 distribution

g01fdc Deviates for the F-distribution

g01fec Deviates for the beta distribution

g01ffc Deviates for the gamma distribution

4.5 Non-central Distributions

g01gbc Computes probabilities for the non-central Student's t-distribution

g01gcc Computes probabilities for the non-central χ^2 distribution

g01gdc Computes probabilities for the non-central F-distribution

g01gec Computes probabilities for the non-central beta distribution

4.6 Testing for Normality and Other Distributions

g01ddc Shapiro and Wilk's W test for Normality

g01dhc Ranks, Normal scores, approximate Normal scores or exponential (Savage) scores