DAA PROJECT REPORT

STATISTICAL LIBRARY IN C

Team Members

1. SANKETH RANGREJI 01FB15ECS267

2. SHIVASAKTHI S.V 01FB15ECS281

3. SOURABH R.M 01FB15ECS300

ABSTRACT

The main goal of our project is to develop a Statistical library for C. The library provides basic functions to compute the Descriptive Measures of Statistics. We have also provided functions for Inferential Statistics. An implementation of a library that provides basic tools for descriptive and inferential statistical analysis. The objective is the design and/or implementation of algorithmically efficient methods for statistical computation. Some methods covered include finding the measures of central tendencies and measures of dispersion, interquartile ranges, calculation of pearson’s coefficient, least squares fit,etc.

Acknowledgement

We would like to thank **Prof. N S Kumar** for having given us this opportunity to do this project under him. We would like to express our gratitude to him for his guidance from selection of the project till its successful completion and for his assistance in overcoming the hurdles we faced in the process and in mentoring us throughout the project.

We would also like to thank PES University for providing this opportunity.

Features of the project:

- Descriptive statistics
  - Mean
  - Median
  - Mode
  - Variance
  - Standard deviation
  - Min and Max
  - Inter Quartile Range

- Inferential Statistics
  - Z test (Single and Double tailed)
  - Probability
  - Chi-squared
  - Pearson Co-efficient
  - Least square fit (Linear Regression)
  - T test
  - Chi-squared Goodness of Fit

Documentation

mean

Computes the average of all the given values.

Function signature:

double mean (double *, int)

Parameter 1 takes in an array of type Double.

Parameter 2 takes in the size of the array given into parameter 1.

median_sorted

Computes the median for a given sorted array.

Function signature:

double median_sorted (double *, int)

Parameter 1 takes in an array of type Double.

Parameter 2 takes in the size of the array given into parameter 1.

median_unsorted

Computes the median for a given unsorted array.

Function signature:

double median_unsorted (double *, int)

Parameter 1 takes in an array of type Double.

Parameter 2 takes in the size of the array given into parameter 1.

mode

Computes the value that appears most often in the data. Returns -1 if all the data is distinct.

Function signature:

double mode (double *, int)

Parameter 1 takes in an array of type Double.

Parameter 2 takes in the size of the array given into parameter 1.

variance

Computes how far a set of numbers is spread out from their mean.

Function signature:

double variance (double *, int)

Parameter 1 takes in an array of type Double.

Parameter 2 takes in the size of the array given into parameter 1.

standard_deviation_sample

Computes the root of the variance i.e. quantifies the amount of dispersion of the data.

Function signature:

double standard_deviation_sample (double *, int)

Parameter 1 takes in an array of type Double.

Parameter 2 takes in the size of the array given into parameter 1.

min

Computes the min of a given data set.

Function signature:

double min (double *, int)

Parameter 1 takes in an array of type Double.

Parameter 2 takes in the size of the array given into parameter 1.

max

Computes the max of a given data set.

Function signature:

double max (double *, int)

Parameter 1 takes in an array of type Double.

Parameter 2 takes in the size of the array given into parameter 1.

iqr

Computes the inter-quartile range which is the midspread of the data.

Function signature:

double iqr (double *, int)

Parameter 1 takes in an array of type Double.

Parameter 2 takes in the size of the array given into parameter 1.

z_test_onetail

Statistical test to determine whether 2 population means are different when variances are known and sample size is large enough.

Function signature:

int z_test_onetail(double, double, double, double, int)

Parameter 1 takes

Parameter 2 takes in the mean of the sample.

Parameter 3 takes in the standard deviation of the sample.

Parameter 4 takes in the alpha value of the sample.

Z_test_twotail

Has the same signature as that of one tail test.

Probability

Computes the probability of normal distribution for given z_score

Function signature :

double p_norm(const double);

Parameter takes in the z_score value

Chisquare

Computes the probability value of the chisquared statistic

Function signature:

double chisqr(int , double);

Parameter 1 takes in the degree of freedom

Parameter 2 takes in the critical value

Pearson_Co-efficient

Measures the linear linear correlation between two variables

Function prototype:

double pearson_r(double *,double *,int)

Parameter 1 takes in mean of x

Parameter 2 takes in mean of y

Least square fit

Gives the line for the best approximation of the given set

Function prototype:

int least_square_fit(double *, double *, int n, double *, double *)

Parameter 1 for the pointer to array with x values

Parameter 2 for the pointer to array with y values

Parameter 3 takes in n value

Parameter 4 to return the slope value

Parameter 5 to return the constant value

T test

Computes the test statistic follows a student's t-distribution under the null hypothesis

Function prototype:

Parameter 1 takes the t_value

Parameter 2 takes the alpha value

Parameter 3 takes the n value

Chi_square_goodness_of_fit

Comapre observed sample distribution with the expected probability distribution
Function prototype:
Parameter 1 takes observed frequencies
Parameter 2 takes expected frequencies
Parameter 3 takes alpha values
Parameter 4 takes n value

Description

The users can include our library in their program which provides common statistical tools for data analysis.

We have included the details of the functions for the user to know how the respective functions to be used and the order of parameters to be passed.

We are giving the statistical measures for floating point values which covers the larger group. We are comparing the floating point values by using an epsilon value which is defined to be a very small value.

We have tried to incorporate most of the functions. More functions can be included as an extension of this project .

Bibliography

1. Numerical recipes by William H Press
2. www.wikipedia.org
3. Introduction to Data Science notes
4. Algorithms notes
5. www.stackoverflow.com