Journal of Computing and Electronic Information Management

ISSN: 2413-1660

Generation of Common Distributed Random Numbers and Data Sampling Method Based on R Software

Shengbin Zheng, Qing Chen, Lianzhao Li, Ruiwen Shen, Xindan Zhang
School of Mathematics and Statistics, Shandong University of Technology, Shandong Province, 255049, China

Abstract: The purpose of this paper is to help readers use R software to generate random numbers and random sampling methods which obey various distribution laws. By introducing the fixed command of generating random numbers for each distribution in R software, readers can easily generate random numbers for testing. In addition, this paper gives the code of simple random sampling, stratified sampling, cluster sampling, unequal probability sampling and systematic sampling by using R software, which is clear and clear and convenient for readers to quickly solve the random sampling problem in experimental design.

Keywords: R software Common Distributed Random Numbers Multiple sampling methods.

1. Common Distributed Random Numbers and Statistical Applications

1.1 Summary

The so-called distributed random number is to generate a set of sample data based on a certain distribution, and there are uncertain rules in the number, size and sequence. For example: (0,1) the generation of uniformly distributed random numbers, the generation of the most common normal distribution random numbers, and so on, and regard these distributed random numbers as sample data for statistical analysis or related experiments. And the degree of fitting increases with the increase of sample size.

In statistical learning, we know that many practical models are based on the distribution function of random variables, such as the application of exponential distribution in queuing theory, the application of binomial distribution in gambling games, the application of normal distribution of students scores etc. In the absence of a given distribution of sample data, we need to generate the distribution of random numbers to obtain sample data. For example, when we test the law of majority and the central limit theorem, we need to use computer to generate multiple groups of random numbers with different distribution to judge the change of mean and the simulation of normal distribution image when the sample size is large enough.

1.2 Generating Common Distributed Random Numbers with R Software

Runif () function is often used in statistical tests to generate random numbers with uniform distribution. Its grammatical rules are: rnorm (n, mean, sd), n denotes the number of random numbers, mean denotes the mean value, Sd denotes the standard deviation. Similarly, the following table gives grammatical rules for generating common distributed random numbers:

Table 1 Syntax Rules for Generating Distributed Random Numbers

Distribution type	Grammatical Rules of Random Number Generation	Explain
binomial distribution	rbinom(n,size,prob)	Size: Test times，prob: Binomial distribution probability
Geometric Distribution	rgeom(n,prob)	Prob: Geometric Distribution Probability
Poisson distribution	rpos(n,k)	K: Poisson distribution parameters
Normal distribution	rnorm (n,mean,sd)	Mean: mean value,sd:standard deviation
t distribution	rt(n,f)	F: Degree of freedom of t distribution
F distribution	rf(n,k1,k2)	k1: First Degree of Freedom, k2: Second Degree of Freedom
chi-square distribution	rchisq(n,f)	F: Chi-Square Distribution Degree of Freedom
Gamma Distribution	rgamma(n,k1,k2)	K1: shape parameter, k2: Scale parameter

Through the grammatical rules given in the table, readers can quickly generate random numbers of each distribution through R software, which is very convenient.

2. Data sampling

2.1summary

The so-called data sampling refers to extracting a part of the unit from the whole as a sample and analyzing this part of the data. In statistical investigation, it is often difficult to judge and analyze the whole because of the large amount of data and high complexity. Sampling survey method can estimate the whole with a certain accuracy and get some characteristic data of the whole as soon as possible, which not only saves the cost of the survey greatly, but also helps to improve the quality of the survey data.

2.2 Sampling Method Realization

In practical application, we often use many sampling methods, such as simple random sampling, stratified sampling, unequal probability sampling, systematic sampling, etc. Different experiments need different sampling methods. The following code is given to implement common sampling methods using R software(Insurance in the code is the data set that comes with R software).

(1) Simple random sampling

```
library(MASS)# Input data package.
data(Insurance)# Select the total data set to be analyzed .
n=as.numeric(readline("print:"))# Input Sample Number.
data1=sample(Insurance,n,replace=F)# SRS without replacement.
data2=sample(Insurance,n,replace=T)# Play back simple random sampling.
Insurance[a,]# Output sample case information.
```

(2)Stratified random sampling

```
install.packages(sampling)
library(sampling)
data3=strata(data,stratnames,size,method)
#data: Sampled data.
#stratanames: The name of the variable on which the hierarchy is based.
#size: Number of observation samples to be extracted from each layer.
#method=srswor: Represents no playback, method=srswr: Put back
getdata(Insurance,data3)# Output sample case information.
```

(3)Cluster sampling

```
data4=cluster(Insurance,clustername="District",size=2,method)
#clustername: Group variable.
```

(4)Systematic sampling

```
library(sampling)# Loading Sampling Package .
pik=inclusionprobabilities(Insurance,d=10)# Computation of Inclusion Probability in System Sampling with Interval d=10.
s=UPrandomsystematic(pik)# Systematic Sampling for Random Arrangement of Population Units.
(1:length(pik))[s==1]# Sample Units Extracted
s1=UPsystematic(pik)# Systematic Sampling of Population Units Arranged in a Certain Order
(1:length(pik))[s1==1]# Sample Units Extracted
```

(5)Sampling with Unequal Probabilities

```
weigh# Input auxiliary variable
```

data5=sample(Insurance,n,replace=T,prob=weigh)# Play back unequal probabilistic sampling(PPS)

#replace=F:пPS of sample

Through the above code, readers can easily achieve data sampling, and statistical analysis of the sample data to estimate some of the total number of features.

References

[1] Fang KangNan. R Data Analysis [M]. Electronics Industr, 2015.

[2] Jin Yongjin. Sampling technique [M]. China Renmin University Press , 2002.

[3] Zhang Xiaojia, Hu Liangping. Random Sampling Based on R Software and Its Application [J]. Sichuan Mental Health,2016,29(06):497-502.

[4] Ma X , Li J J , Weiss D M . Prioritized Constraints with Data Sampling Scores for Automatic Test Data Generation[C]// Proceedings of the 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2007, July 30 - August 1, 2007, Qingdao, China. IEEE, 2007.

[5] Klekota J , Brauner E , Schreiber S L . Identifying Biologically Active Compound Classes Using Phenotypic Screening Data and Sampling Statistics[J]. Journal of Chemical Information and Modeling, 2005, 45(6):1824-1836.

[6] Elson E L , Fee J A , Wakatsuki T . Phenotypic Screening for Pharmaceuticals Using Tissue Constructs[J]. Current Pharmaceutical Biotechnology, 2004, 5(2):-.

[7]Kell, Douglas B . Finding novel pharmaceuticals in the systems biology era using multiple effective drug targets, phenotypic screening and knowledge of transporters: where drug discovery went wrong and how to fix it[J]. FEBS Journal, 2013, 280(23):5957-5980.