Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Институт Информационных Технологий и Анализа Данных

**Название работы** – “Программные средства статистического анализа”

Отчет по лабораторной работе “Лабораторная работа №1”

по дисциплине Моделирование систем и процессов

Вариант 21

Выполнил

Студент, номер группы ИСМб-19-1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_ Д.Д.Солопов

Дата: \_\_\_\_\_\_\_\_\_\_

Принял

Должность \_\_\_\_\_\_\_\_\_\_\_\_\_\_ О.С.Бучнев

Дата: \_\_\_\_\_\_\_\_\_\_

Иркутск 2021 г.

Содержание

[Цель работы 3](#_Toc66528846)

[Постановка задач 4](#_Toc66528847)

[Решение задач 5](#_Toc66528848)

[2.1. График и гистограмма частот 6](#_Toc66528849)

[2.2. Таблица и график накопленных частот (оценка функции распределения вероятностей) 8](#_Toc66528850)

[2.3. Оценка математического ожидания, дисперсии и среднеквадратического отклонения. 12](#_Toc66528851)

[2.4. Доверительные интервалы для оценок математического ожидания и среднеквадратического отклонения. 15](#_Toc66528852)

[2.5. Выдвинуть гипотезу о теоретическом распределении генеральной совокупности, из которой извлечена выборка, и с помощью критерия χ^2-квадрат (критерия Пирсона) проверить эту гипотезу при уровне значимости ∝= 0,1. 18](#_Toc66528853)

[Вывод 20](#_Toc66528854)

Цель работы

Целью лабораторной работы является ознакомление с программными средствами статистического анализа и обработки наблюдений, а так же повторение базовых понятий теории вероятностей и математической статистки.

Постановка задач

Задание

1. Получив от преподавателя номер варианта (от 1 до 30) получить выборку (файл Варианты\_R.xls).

2. Используя возможности языка R, провести анализ сгенерированной выборки. При этом необходимо получить:

2.1. таблицу и график гистограммы частот,

2.2. таблицу и график накопленных частот (оценку функции распределения вероятностей),

2.3. оценки математического ожидания, дисперсии и среднеквадратического отклонения,

2.4. доверительные интервалы для оценок математического ожи­дания и среднеквадратического отклонения,

2.5. выдвинуть гипотезу о теоретическом распределении генеральной совокупности, из которой извлечена выборка, и с помощью критерия -квадрат (критерия Пирсона) проверить эту гипотезу при уровне значимости 0,1.

3. Составить отчет по лабораторной работе, в который включить все результаты статистического анализа, снабдив их развернутыми пояснениями и иллюстрациями. Пояснения должны содержать описание всех данных, полученных при выполнении пунктов 2.1 – 2.5.

Решение задач

**Задание 1**

Выборка согласно индивидуальному варианту 21:

x<-c(0.055142,0.013344,0.122560,0.238551,0.212595,0.350787,0.019994,0.259986,0.075240,0.124642,0.068656,0.024851,0.070090,0.068509,0.009442,0.123500,0.038525,0.070715,0.478898,0.299773,0.214555,0.065536,0.063025,0.120898,0.002092,0.156874,0.028269,0.056316,0.131069,0.213805,0.032344,0.140175,0.011813,0.095309,0.158439,0.031042,0.010544,0.047412,0.021714,0.020424,0.010584,0.019542,0.093444,0.046933,0.019223,0.258758,0.181235,0.310185,0.110045,0.078190,0.248625,0.106835,0.200646,0.071645,0.057742,0.013889,0.002268,0.153634,0.020300,0.181236,0.084836,0.004219,0.029782,0.000013,0.142415,0.015641,0.074098,0.048807,0.031320,0.043216,0.044386,0.118830,0.078281,0.038214,0.247450,0.010158,0.016772,0.015630,0.064190,0.011424,0.110394,0.179514,0.019766,0.020522,0.285728,0.031544,0.030696,0.039059,0.136562,0.007522,0.005928,0.277302,0.065102,0.106700,0.083071,0.059818,0.026216,0.033623,0.259605,0.000024)

Случайной величиной называется величина, которая в результате опыта может принять то или иное значение, причем заранее неизвестное.

Генеральная совокупность — совокупность всех объектов относительно которых предполагается делать выводы при изучении конкретной задачи.

Генеральная совокупность состоит из всех объектов, которые имеют качества, свойства, интересующие исследователя.

Совокупность, из которой производится выборка, носит в статистике название генеральной совокупности.

Выборку образуют значения случайной величины, которые она уже приняла в результате серии опытов.

Выборка или выборочная совокупность — часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом).

2.1. График и гистограмма частот

**Код:**

x<-c(0.055142,0.013344,0.122560,0.238551,0.212595,0.350787,0.019994,0.259986,0.075240,0.124642,0.068656,0.024851,0.070090,0.068509,0.009442,0.123500,0.038525,0.070715,0.478898,0.299773,0.214555,0.065536,0.063025,0.120898,0.002092,0.156874,0.028269,0.056316,0.131069,0.213805,0.032344,0.140175,0.011813,0.095309,0.158439,0.031042,0.010544,0.047412,0.021714,0.020424,0.010584,0.019542,0.093444,0.046933,0.019223,0.258758,0.181235,0.310185,0.110045,0.078190,0.248625,0.106835,0.200646,0.071645,0.057742,0.013889,0.002268,0.153634,0.020300,0.181236,0.084836,0.004219,0.029782,0.000013,0.142415,0.015641,0.074098,0.048807,0.031320,0.043216,0.044386,0.118830,0.078281,0.038214,0.247450,0.010158,0.016772,0.015630,0.064190,0.011424,0.110394,0.179514,0.019766,0.020522,0.285728,0.031544,0.030696,0.039059,0.136562,0.007522,0.005928,0.277302,0.065102,0.106700,0.083071,0.059818,0.026216,0.033623,0.259605,0.000024)

par(mfrow=c(1, 2))

p1<-hist(x, breaks = 30, freq = FALSE, col = "lightgreen")

lines(density(x),col="red")

p1

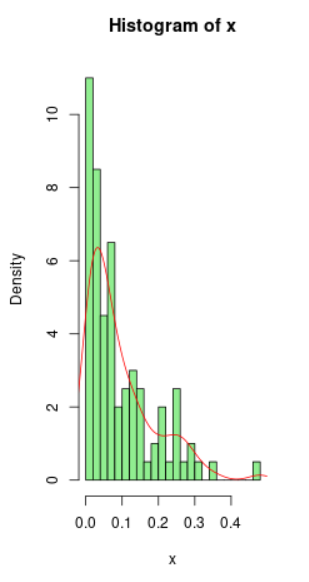


Рисунок 1 – График и гистограмма частот

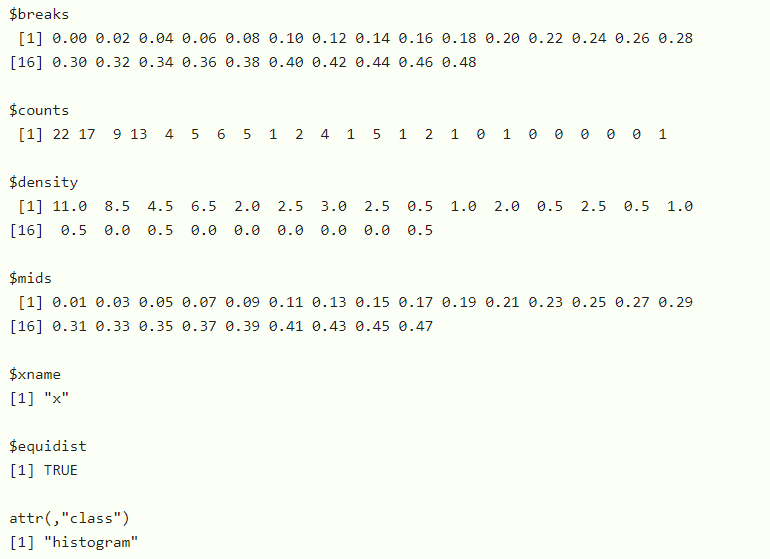


Рисунок 2 – Выходные данные после выполнения программного кода на языке R

Функция density(x) вычисляет ядерные плотности вероятностей, которые продемонстрированы на рисунке 2 в виде выходного значения работы программного кода, на языке R.

С помощью функции hist() происходит рисование графика, в соответствии с его параметрами.

Гистограмма представляет собой диаграмму (обычно столбиковую), которая используется в статистике для графического представления распределения вероятностей значений случайной величины.

По горизонтальной оси гистограммы откладывается диапазон наблюдаемых значений величины, разбитый на определенное число (обычно 10-15) интервалов, а по вертикальной — вероятность или частота ее попадания в каждый интервал. Тогда столбик будет отражать значения этих показателей для интервала, на который он опирается.

По форме гистограммы аналитик может в первом приближении оценить, какому статистическому закону распределения подчиняется случайная величина.

Плотность распределения вероятностей (плотность распределения) случайной величины — первая производная от функции распределения. Другими словами, плотность вероятности — это некоторая средняя вероятность, приходящая на бесконечно малый отрезок. Функция распределения — это первообразная плотности распределения вероятностей.

2.2. Таблица и график накопленных частот (оценка функции распределения вероятностей)

**Код:**

x<-c(0.055142,0.013344,0.122560,0.238551,0.212595,0.350787,0.019994,0.259986,0.075240,0.124642,0.068656,0.024851,0.070090,0.068509,0.009442,0.123500,0.038525,0.070715,0.478898,0.299773,0.214555,0.065536,0.063025,0.120898,0.002092,0.156874,0.028269,0.056316,0.131069,0.213805,0.032344,0.140175,0.011813,0.095309,0.158439,0.031042,0.010544,0.047412,0.021714,0.020424,0.010584,0.019542,0.093444,0.046933,0.019223,0.258758,0.181235,0.310185,0.110045,0.078190,0.248625,0.106835,0.200646,0.071645,0.057742,0.013889,0.002268,0.153634,0.020300,0.181236,0.084836,0.004219,0.029782,0.000013,0.142415,0.015641,0.074098,0.048807,0.031320,0.043216,0.044386,0.118830,0.078281,0.038214,0.247450,0.010158,0.016772,0.015630,0.064190,0.011424,0.110394,0.179514,0.019766,0.020522,0.285728,0.031544,0.030696,0.039059,0.136562,0.007522,0.005928,0.277302,0.065102,0.106700,0.083071,0.059818,0.026216,0.033623,0.259605,0.000024)

par(mfrow=c(1, 2))

p1<-hist(x, breaks = 30, freq = FALSE, col = "lightgreen")

lines(density(x),col="red")

p1

p1$counts <- cumsum(p1$counts)

plot(p1,col = "red")

p1

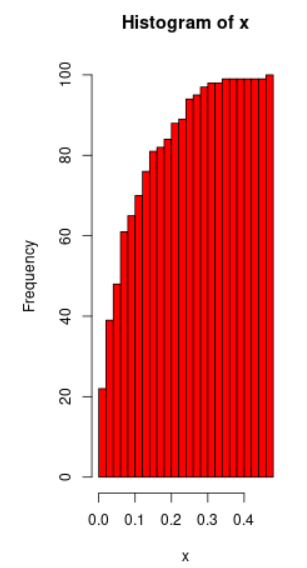


Рисунок 3 – Гистограмма накопленных частот

Гистограмма накопленных частот отличается от гистограммы распределения тем, что высота каждого столбика пропорциональна частоте, накопленной к данному значению (интервалу).

Площадь каждого последующего столбца в гистограмме накопленных частот равна сумме площадей предыдущих столбцов.

Законом распределения случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

Значение функции распределения это оценка вероятности события, заключающегося в том, что случайная величина примет значение, меньшее, чем наблюдаемое (которое является аргументом функции).

Функция f(x) – производная функции распределения, характеризует, как бы, плотность, с которой распределяются значения случайной величины в данной точке. Эта функция называется плотностью распределения непрерывной случайной величины X.

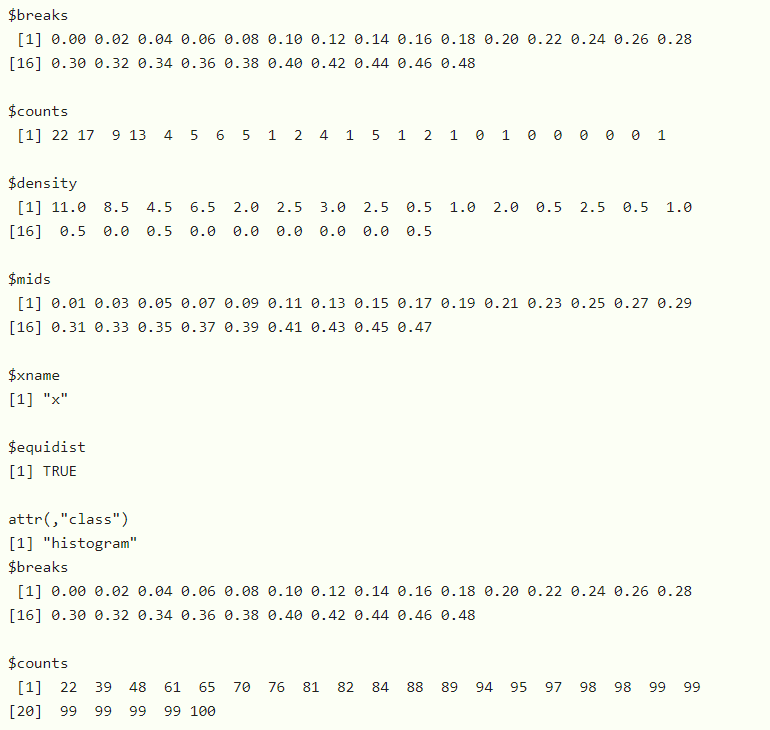


Рисунок 4 – Результат после выполнения программного кода на языке R (часть 1)

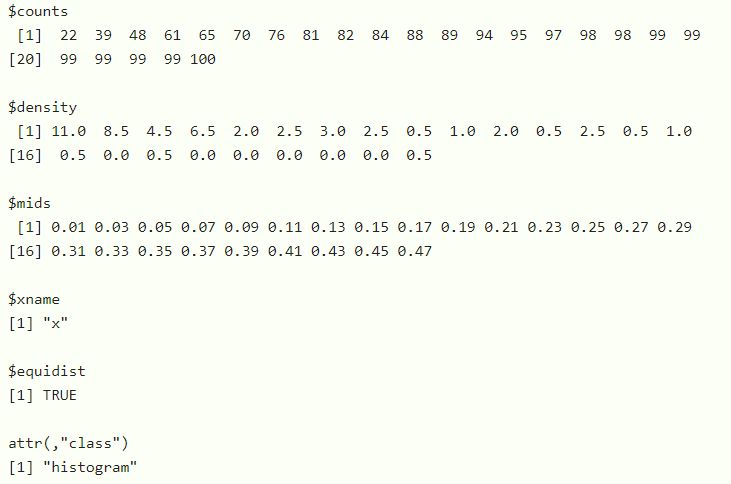


Рисунок 5 – Результат после выполнения программного кода на языке R (часть 2)

Как видно из вывода программы на рисунке 5, программа рисует ещё одну гистограмму, однако уже она является гистограммой накопленных частот (см. рис. 3).

2.3. Оценка математического ожидания, дисперсии и среднеквадратического отклонения.

**Код:**

m\_x=mean(x)

m\_x

disp=var(x)

disp

sqo=sd(x)

sqo

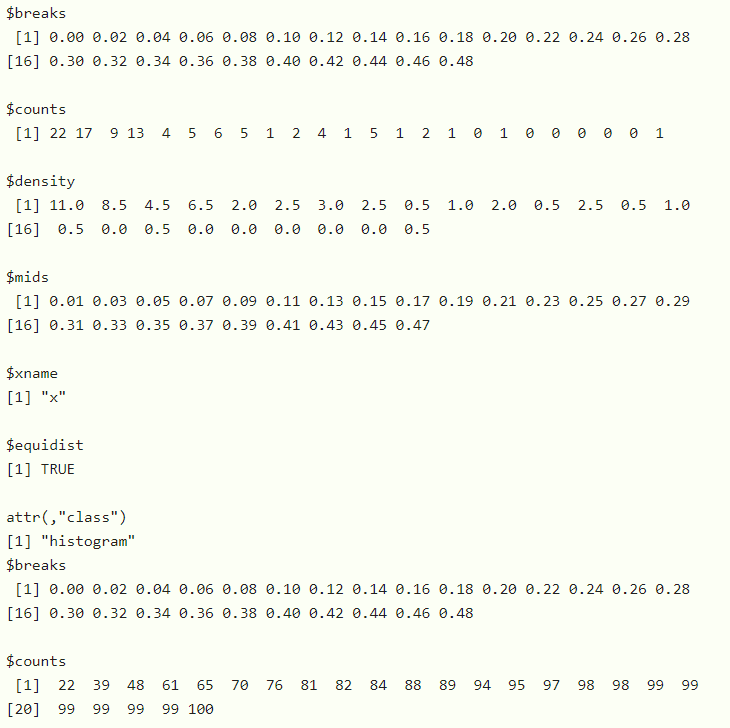


Рисунок 6 – Результат выполнения кода на языке R (часть 1)

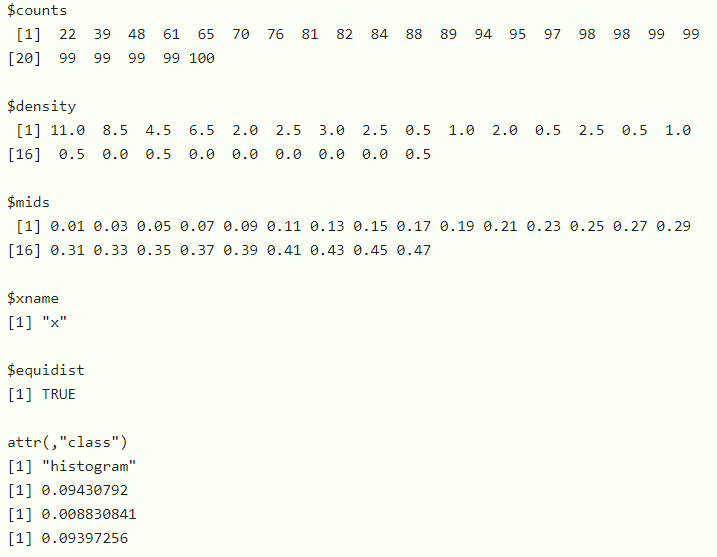


Рисунок 7 – Результат выполнения кода на языке R (часть 2)

Следует обратить внимание на рисунок 7, на последний вывод данных:

Attr(, “class”)

[1] “histogram”

[1] 0.09430792 - математическое ожидание

[1] 0.008830841 - дисперсия

[1] 0.09397256 - среднеквадратическое отклонение

Любое значение искомого параметра, вычисленное на основе ограниченного числа опытов, всегда будет содержать элемент случайности. Такое приближенное, случайное значение называется оценкой параметра.

Математическое ожидание — понятие, означающее среднее (взвешенное по вероятностям возможных значений) значение случайной величины. В случае непрерывной случайной величины подразумевается взвешивание по плотности распределения.

Значение среднего арифметического некоторого множества при стремлении его членов к бесконечности называется математическим ожиданием (М).

Абсолютной погрешностью называется разность измеряемой величины со средним значением.

Дисперсией называется среднее арифметическое квадратов всех абсолютных погрешностей.

Дисперсия случайной величины — мера разброса значений случайной величины относительно её математического ожидания.

Среднеквадратическое отклонение— в теории вероятностей и статистике наиболее распространённый показатель рассеивания значений случайной величины относительно её математического ожидания. Обычно указанные термины означают квадратный корень из дисперсии случайной величины, но иногда могут означать тот или иной вариант оценки этого значения.

Среднеквадратическое отклонение определяется как квадратный корень из дисперсии случайной величины.

2.4. Доверительные интервалы для оценок математического ожидания и среднеквадратического отклонения.

**Код:**

library(DescTools)

MeanCI( x,conf.level = 0.95)

VarCI(x,conf.level = 0.95)

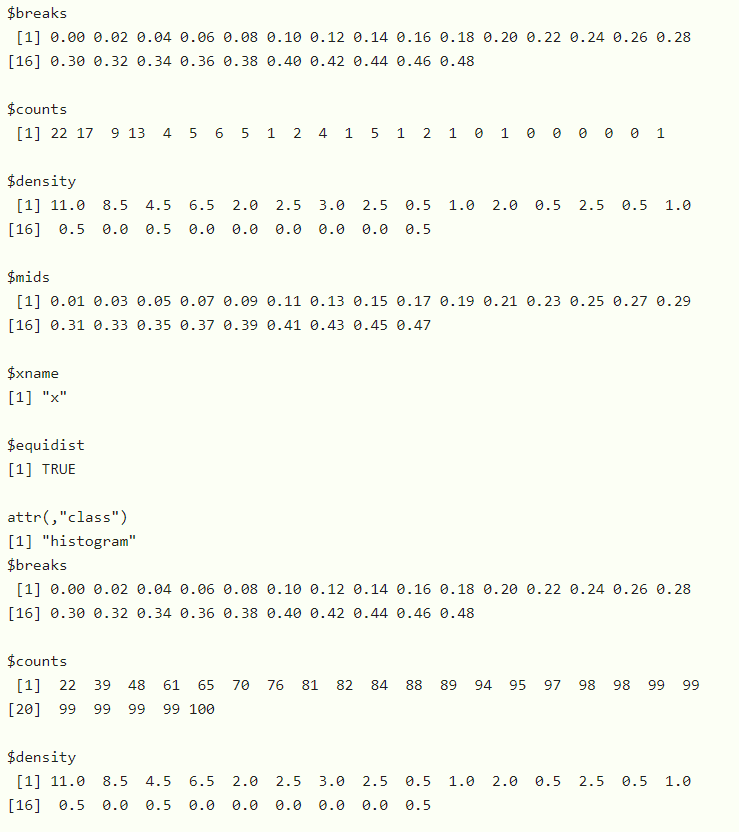


Рисунок 8 – Результат выполнения кода на языке R (часть 1)

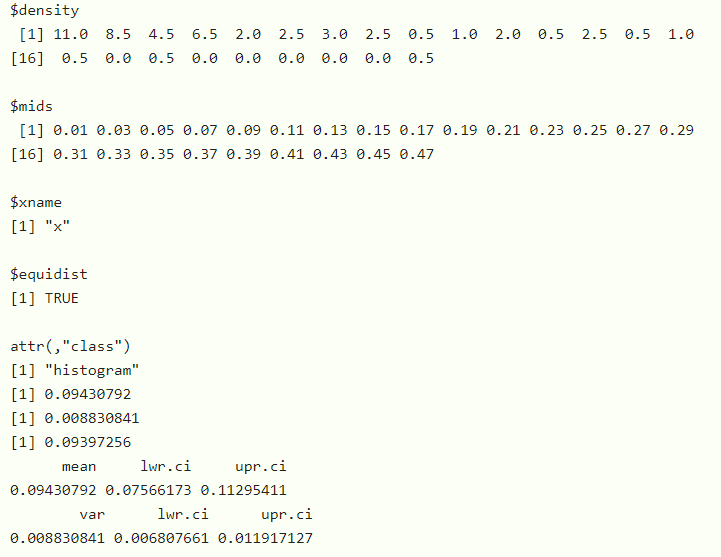


Рисунок 9 – Результат выполнения кода на языке R (часть 2)

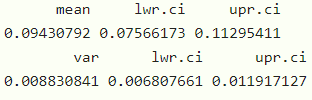


Рисунок 10 – Доверительные интервалы для оценок математического ожидания и среднеквадратического отклонения

Доверительным называется интервал, в который попадают измеренные в эксперименте значения, соответствующие доверительной вероятности

Доверительным называют интервал, который покрывает неизвестный параметр с заданной надёжностью.

Доверительные интервалы могут быть полезны при проведении опытов, для основания доказательной базы о состоятельности гипотезы распределения.

2.5. Выдвинуть гипотезу о теоретическом распределении генеральной совокупности, из которой извлечена выборка, и с помощью критерия χ^2-квадрат (критерия Пирсона) проверить эту гипотезу при уровне значимости ∝= 0,1.

Рассмотрев рисунок 1, можно предположить, что теоретическое распределение имеет экспоненциальную форму. Для проверки данной гипотезы будет использован следующий код:

**Код:**

library(DescTools)

MeanCI( x,conf.level = 0.95)

VarCI(x,conf.level = 0.95)

k<-50

int<-seq(min(x),max(x),(max(x)-min(x))/k)

x.fact<-hist(x,breaks=int,plot=FALSE)

int[1]<-(0)

int[k+1]<-(Inf)

x.theor<-pexp(int,rate=1/mean(x))

x.theor<-(x.theor[2:(k+1)]-x.theor[1:k])

chisq.test(x.fact$counts,p=x.theor)

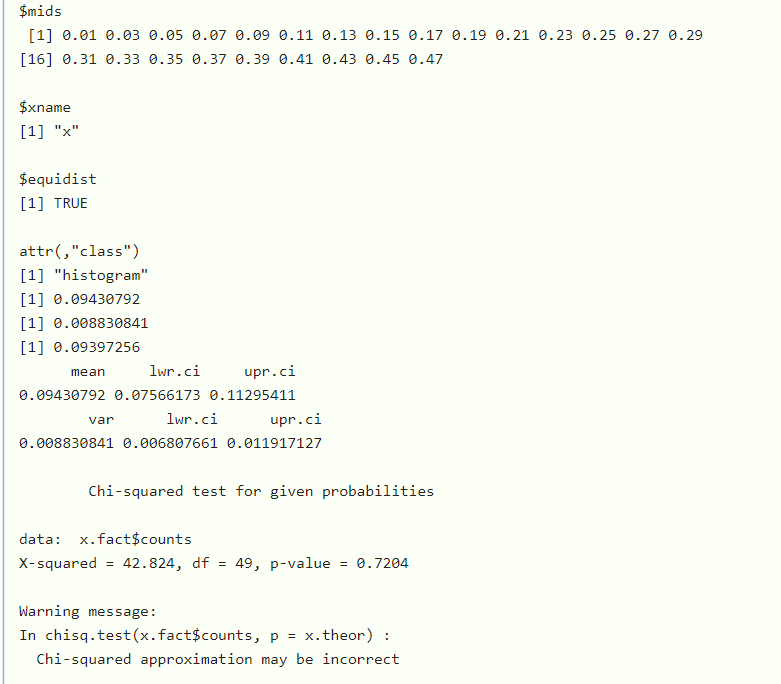


Рисунок 11 – Результат выполнения кода на языке R

Рассмотрев рисунок 11 можно убедиться в том, что гипотеза о экспоненциальном теоретическом распределении генеральной совокупности, из которой извлечена выборка, верна, поскольку значение p-value равно 0.7204, а данное числовое значение превышает уровень значимости, характеризуемый числом 0.1 (∝= 0.1).

P-value — вероятность получить для данной вероятностной модели распределения значений случайной величины такое же или более экстремальное значение статистики (среднего арифметического, медианы и др.), по сравнению с ранее наблюдаемым, при условии, что нулевая гипотеза верна.

С помощью p-value можно оценить состоятельность предположенной гипотезы.

Уровень значимости – это пороговая (критическая) вероятность ошибки 1-го рода, т.е. непринятия гипотезы H0, когда она верна («ложная тревога»). Требуемый уровень значимости α задает критическое значение для статистического теста.

Вероятность совершить ошибку первого рода принято обозначать через , ее называют уровнем значимости.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза.

Ошибка второго рода состоит в том, что будет принята неправильная.

В случае, если p-value имеет меньшее значение, чем уровень значимости, тогда происходит ошибка первого рода и непринятие гипотезы Н0, когда она верна.

Вывод

Было проведено ознакомление с программными средствами статистического анализа и обработки наблюдений, а также повторение базовых понятий теории вероятностей и математической статистики.

Используя возможности языка R был проведён анализ сгенерированной выборки. Была получена таблица и график гистограммы частот, на основе сгенерированной выборки. Была получена таблица и график накопленных частот. Были получены оценки математического ожидания, дисперсии и среднеквадратического отклонения, которые программа-интерпретатор вывела на экран, после завершения выполнения программы. Были найдены доверительные интервалы для оценок математического ожидания и среднеквадратического отклонения. Была выдвинута гипотеза о теоретическом распределении генеральной совокупности, гипотеза заключается в том, что теоретическое распределение генеральной совокупности является экспоненциальной, что можно наблюдать из рисунков, представленных в отчёте. Гипотеза была подтверждена значением p-value равным 0.7274, что является большим, чем значение уровня значимости 0,1.