**Лабораторная работа №3**

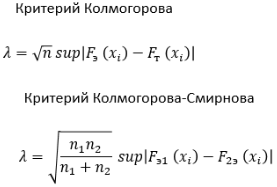
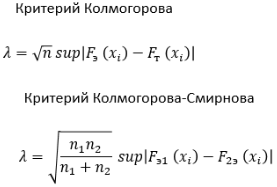
«Статистические гипотезы»

**Выполнил:** Николай Окуньков, 18ПИ-2.

**Цель работы:** проверка статистических гипотез с использованием встроенных в базовую версию пакета R функций, а также некоторого ряда критериев.

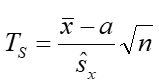
**Теоретическая часть:**

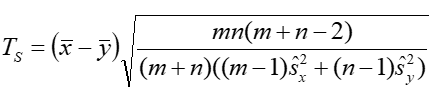
*Непрерывное равномерное распределение* – распределение случайной вещественной величины, принимающей значения, принадлежащие некоторому промежутку конечной длины. Характеризуется тем, что плотность вероятности на этом промежутке почти везде постоянна.

*Критерий однородности Смирнова* – используется для проверки гипотезы о принадлежности двух независимых выборок одному закону распределения, то есть о том, что два эмпирических распределения соответствуют одному и тому же закону.

*Критерий согласия Колмогорова* – используется для проверки гипотезы о принадлежности выборки некоторому закону распределения, то есть проверки того, что эмпирическое распределение соответствует предполагаемой модели.

*Распределение Стьюдента* – это однопараметрическое семейство абсолютно непрерывных распределений:

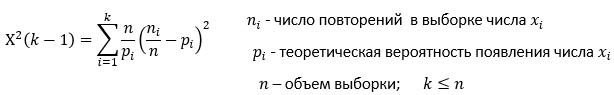
* *Одновыборочный t-критерий*– один из вариантов критерия Стьюдента, служащий для проверки нулевой гипотезы о равенстве среднего значения генеральной совокупности, из которой была взята выборка, некоторому известному значению. В общем виде проверка этой гипотезы выполняется при помощи t-критерия, рассчитывающийся как отношение разницы между выборочным средним и известным значением к стандартной ошибке выборочного среднего. Исходя из свойств t-распределения, рассчитанное значение критерия можно интерпретировать таким образом: если данное значение попадает в область отклонения нулевой гипотезы, то мы вправе отклонить проверяемую нулевую гипотезу. Область отклонения нулевой гипотезы для критерия Стьюдента определяется заранее принятым уровнем значимости и числом степеней свободы:
* *Двухвыборочный t-критерий* **–** при сравнении двух выборок проверяемая нулевая гипотеза состоит в том, что обе эти выборки происходят из нормально распределенных генеральных совокупностей с одинаковыми средними значениями. Данные генеральные средние оценивают при помощи выборочных средних значений. В знаменателе приведенной формулы находится стандартная ошибка разницы между выборочными средними.



*Значение 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.

*Выборочное среднее* – приближение теоретического среднего распределения, которое основывается на выборке из него.

*Экспоненциальное распределение* – описывает интервалы времени между независимыми событиями, которые происходят со средней интенсивностью. Количество наступлений такого события за некоторый отрезок времени описывается дискретным распределением Пуассона:

*Критерий согласия Пирсона* – непараметрический метод, позволяющий оценить значимость различий между фактическим (выявленным в результате исследования) количеством исходов или качественных характеристик выборки, попадающих в каждую категорию, и теоретическим количеством, которое можно ожидать в изучаемых группах при справедливости нулевой гипотезы:

где – число повторений в выборке числа **;** – теоретическая вероятность появления числа ; ***n*** – объём выборки; ***k ≤ n***.

*Уровень значимости статистического теста* – это допустимая для данной задачи вероятность ошибки первого рода, то есть вероятность отклонить нулевую гипотезу, когда на самом деле она верна.

*Используемые в данной работе функции:*

**lines(), mean(x), plot()**, **seq(from, to, by = ), runif(), ks.test(), sd(), t.test(), sapply(), abline(), cut()**, **hist( ), diff(), rexp(), chisq.test(), rnorm()**, **sort(х):**

**lines()** – функция служит для создания линий на графике;

**lines(x, col = , lwb = )**

(Параметры: **x** – значения; **col** – цвет; **lwd** – толщина линии.)

**mean(x)** – функция, находящая среднее арифметическое элементов объекта **х**;

**plot()** – функция, служащая для построения графиков;

(Параметры: **x**, **y** – координаты точек на графике; **type** – указывает на тип графикa (в данной работе используется тип “l”, lines); **xlab** – заголовок для оси х; **ylab** – заголовок для оси y; **main** – общее название графика.)

**seq(from, to, by = )** – функция, генерирующая последовательность чисел от **from** до **to** с шагом **by**;

**runif()** – функция для равномерного распределения;

**runif(n, min = , max = )**

(Параметры: **n** – объем выборки; **min** и **max** – нижний и верхний пределы распределения, которые должны быть представлены в виде целых чисел.)

**ks.test()** – функция для выполнения тестов Колмогорова-Смирнова. Возвращает значения статистики Колмогорова-Смирнова и строку с описанием гипотезы;

**ks.test(x, y, …, alternative = c(“two.sided”, “less”, “greater”), exact = NULL)**

(Параметры: **x** – вектор, содержащий выборку; **y** – вектор, содержащий вторую выборку, или символьная строка с именем распределения; **…** – параметры распределения; **alternative** – символьный аргумент, обозначающий тип альтернативной гипотезы (принимает одно из значений: **“two.sided”** (по умолчанию), **“less”** или **“greater”**); **exact** – **NULL** или логическое значение, обозначающее требуется ли точное вычисление **p-value** (не используется, если **alternative = “less”** или **alternative = “ greater ”.**)

**sd()** – функция для вычисления стандартного отклонения значений в **x**;

**sd(x, na.rm = FALSE)**

(Параметры: **х** – числовой вектор; **na.rm** **–** логический аргумент, указывающий на то, нужно ли исключать пропущенные значения.)

**t.test()** – функция, выполняющая тест Стьюдента;

**t.test(x, y=NULL, alternative = c("two.sided", "less", "greater"), conf.level = 0.95, mu=NULL, paired= FALSE, var.equal=FASLSE)**

(Параметры: **t.test(formula, data, subset, na.action)**, где **x, y** – числовые векторы, первой и второй выборок; **alternative** – альтернативная гипотез (может быть одна из **“two sided”** (по умолчанию) – двусторонняя критическая область, **“greater”** – правосторонняя критическая область или **“less”** – левосторонняя критическая область); **conf.level** – доверительная вероятность для возвращаемого доверительного интервала; **paired** – признак парного теста, проверяется гипотеза для **x** и **y**, оба вектора должны иметь одну и ту же длину; **formula** – формула вида **lhs ~ rhs** , где **lhs** - числовой вектор, **rhs** - фактор с двумя классами; **data** – матрица или таблица данных, из которых берутся данные для **formula; subset** – вектор, определяющий используемое подмножество наблюдений; **na.action –** функция, которая вызывается, как только в данных встретилось значение **NA**; **var.equal** –логическая переменная, указывающая, следует ли рассматривать две дисперсии как равные (если **TRUE**, то объединенная дисперсия используется для оценки дисперсии, в противном случае используется приближение Уэлча c приближенными степенями свободы).)

**sapply()** – функция, используемая в случаях, когда необходимо применить какую-либо функцию к каждому компоненту списка, но результат вывести в виде вектора;

**sapply(x, function, simplify = , USE.NAMES =** **)**

(Параметры: **x** – имя матрицы, массива или таблицы данных; **function** – имя применяемой функции; **simplify** – логический аргумент (нужно ли представлять выводимый результат в виде матрицы (**TRUE**) или вектора; **USE.NAMES** – логический аргумент (если данный аргумент принимает значение **TRUE** и **x** символьного типа, то в качестве названий для вывода используется **x**).)

**abline()** – функция добавляет одну или несколько прямых линий через текущий график;

**abline(h = , v = )**

(Параметры: **h** – значение для горизонтальной линии; **v** – значение для вертикальной линии.)

**cut()** – функция делит вектор на равные интервалы;

**cut(x = , breaks = )**

(Параметры: **x** – числовой вектор; **breaks** – либо количество интервалов, либо вектор точек, по которым нужно делить вектор **x**.)

**hist( )** – функция для создания гистограмм частот значений переменной **х**;

**hist(x, breaks = )**

(Параметры: **x** – переменная; **breaks** – количество столбцов.)

**diff()** – функция для вычисления различий между всеми последовательными значениями вектора;

**diff(x, lag = , differences = )**

(Параметры: **x** – вектор значений; **lag –** указывает задержку; **differences** - позволяет указать порядок различий.)

**rexp()** – функция экспоненциального распределения;

**rexp(N, rate = )**

(Параметры: **N** – объем выборки; **rate** – параметр **λ**.)

**chisq.test()** – функция критерия согласия Пирсона;

**chisq.test(x, y = NULL, p = rep(1/length(x), length(x)), rescale.p = FALSE)**

(Параметры: **x** – вектор или матрица; **y** – вектор (игнорируется, если **x** матрица); **p –** вектор, содержащий вероятности (должен иметь такую же длину, что и **x); rescale.p** – логическое значение (если TRUE, то **p** при необходимости нормируется так, чтобы сумма его компонент была равна 1.)

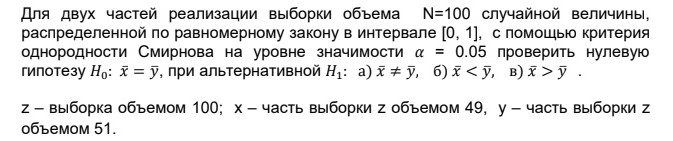
**rnorm()** – функция для случайной генерации совокупностей нормально распределенных чисел;

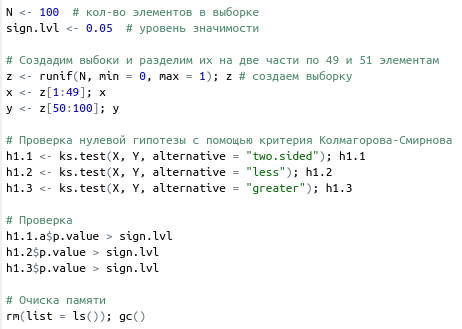
**rnorm(n, mean = , sd = )**

(Параметры: **n** **–** количество элементов выборки; **mean** **–** математическое ожидание; **sd** **–** среднеквадратическое отклонение.)

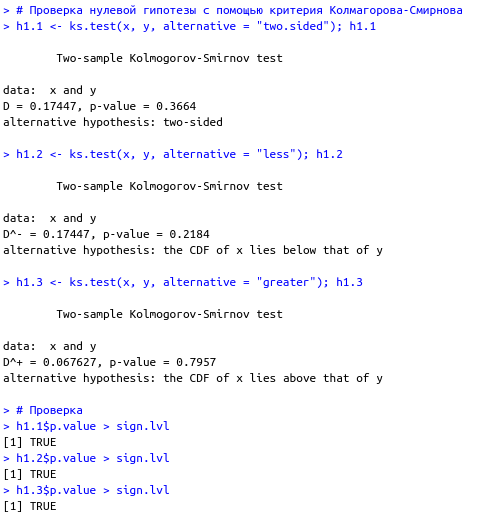
**sort(х)** – функция для сортировки элементов объекта х по возрастанию.

**Ход работы:**

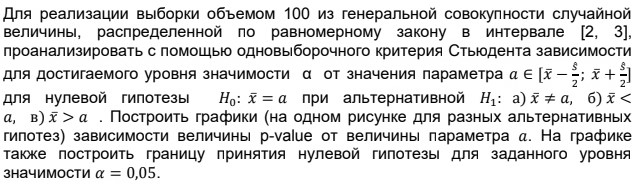
Задание 1:

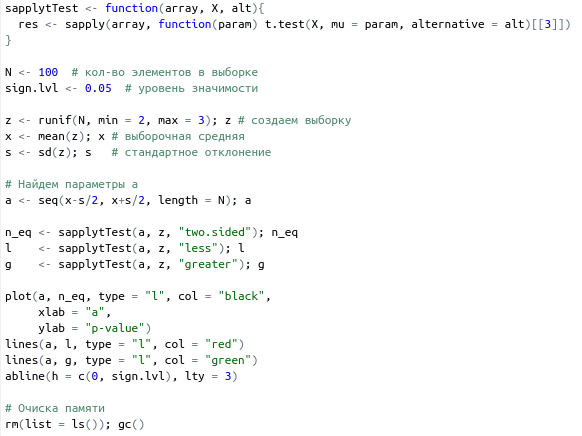


Результат:

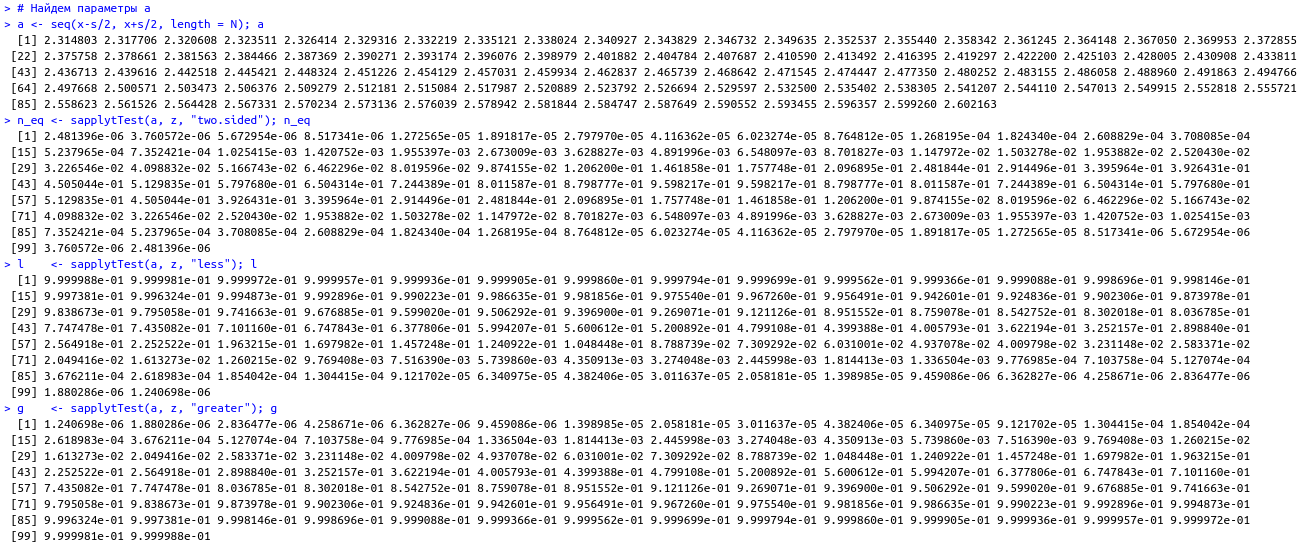


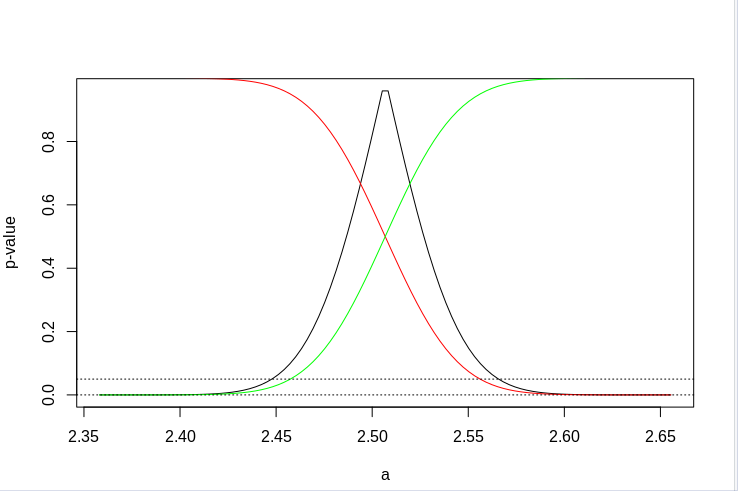
Задание 2:

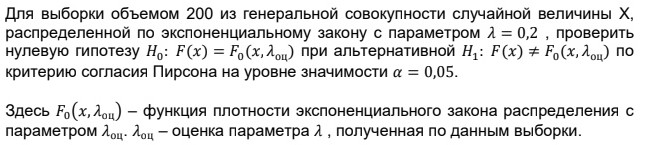


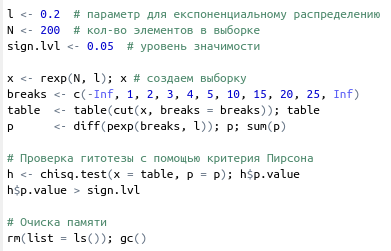


Результат:

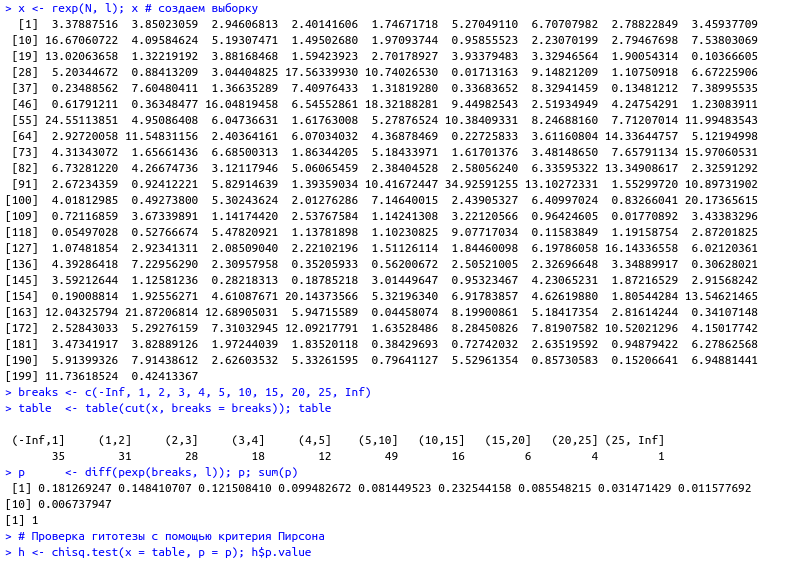




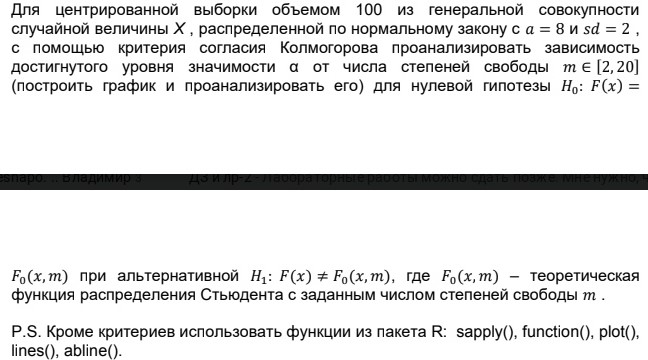
Задание 3:

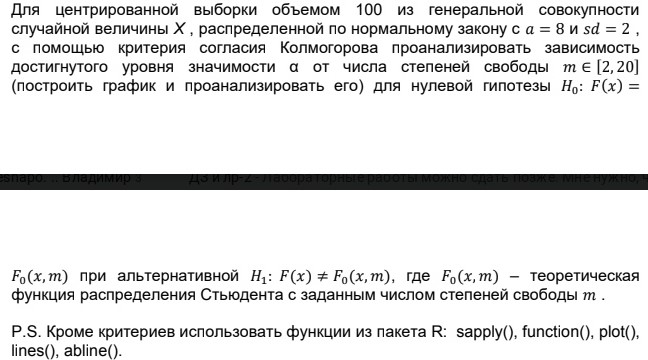


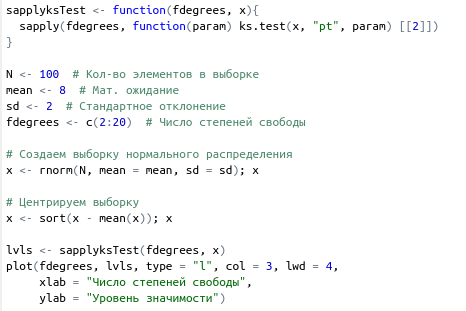
Результат:

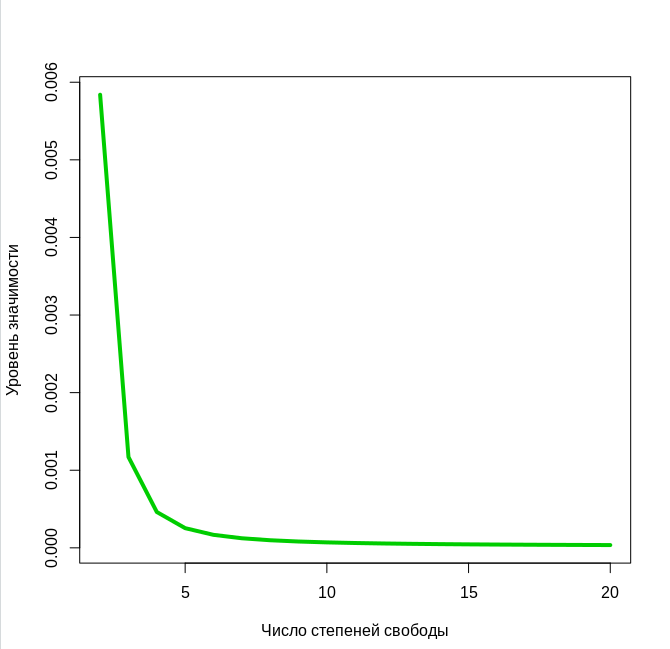


Задание 4:







****Результат:

**Вывод**

В ходе работы мной были проверены статистические гипотезы на разных наборах данных (выборках), что помогло мне усвоить теоретический материал, используемый в данной работе, узнать о новых функциях пакета R и ознакомиться с механизмом их работы.