«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Лабораторная работа №5 по теме: «Проверка гипотезы о законе распределения по критерию Хи-квадрат для смоделированных и реальных данных.»

Выполнил

Студент 1 курса

группы 09-115(3)

Зиновьев Е. А.

Преподаватель:

Шустова Е.П

Казань 2021

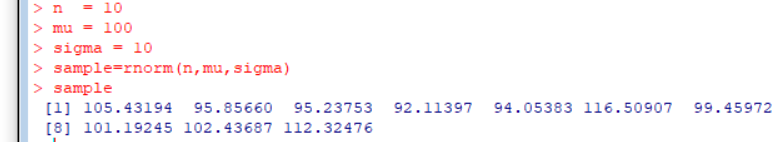
**Хи2 - 1 - проверка гипотезы о законе распределения по критерию Хи-квадрат для смоделированных данных.**

**План работы:**

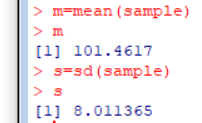
1. Смоделируйте выборку объёма 10 из нормального распределения (параметры распределения выберите произвольно);
2. Получите точечные оценки для параметров распределения;
3. Сформулируйте гипотезу о нормальности распределения с.в. (с параметрами, равными их выборочным оценкам) и проверьте её по критерию Хи-квадрат Пирсона;
4. Постройте диаграмму "Квантиль-Квантиль" (см. статью "Q-Q plot" в Глоссарии);
5. Смоделируйте выборку объёма 10000 из нормального распределения (с теми же параметрами, с которыми была получена первая выборка);
6. Получите для этой новой выборки точечные оценки для параметров распределения;
7. Сформулируйте гипотезу о нормальности распределения с.в. (с параметрами, равными их выборочным оценкам) и проверьте её по критерию Хи-квадрат Пирсона; постройте диаграмму "Квантиль-Квантиль" ;
8. Выберите любое другое распределение; зафиксируйте произвольно его параметр(ы);
9. Последовательно проверьте гипотезу о нормальности распределения для двух выборок (объёма 10 и 10000, соответственно) из этого закона распределения (в качестве параметров нормального распределения используйте их точечные оценки, найденные по выборкам); постройте диаграмму "Квантиль-Квантиль";
10. Полученные результаты оформите в виде таблицы следующего вида (в ячейки таблицы поместите диаграммы "Квантиль-Квантиль", значение p-value и сделанный Вами вывод о принятии или отклонении гипотезы о законе распределения).
11. Объясните полученные результаты.

**Ход работы:**

* 1. Смоделируем выборку sample объёма n=10 из нормального распределения с параметрами mu=100, sigma=10.



* 1. Получим точечные оценки для параметров распределения:

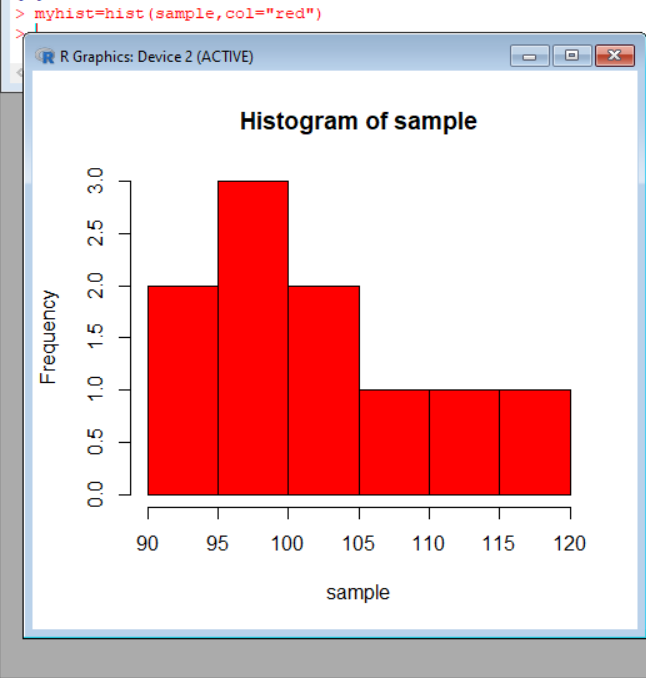


Как видим, среднее отклонилось на 1.5%, а среднеквадратическое отклонение на 20% от заданных значений.

* 1. Сформулируем гипотезу:

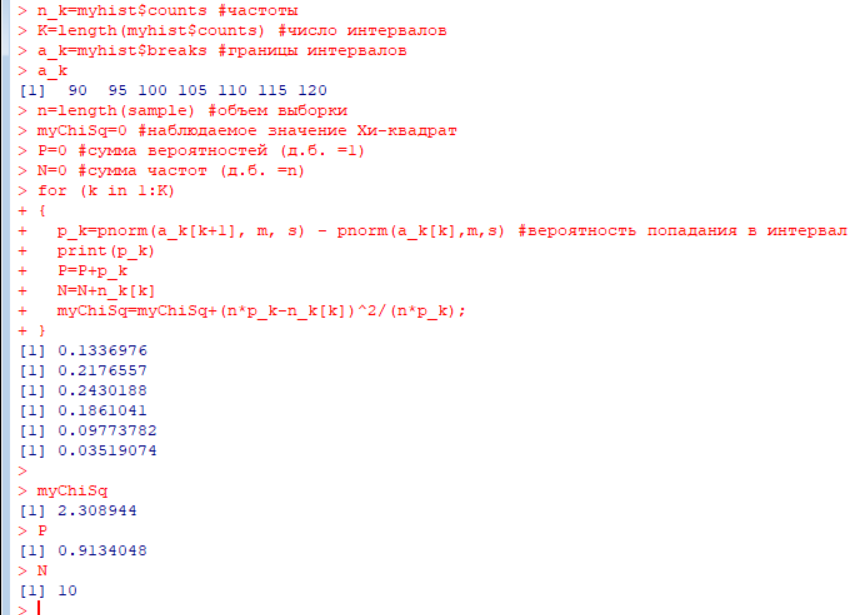
Наблюдаемая с.в. имеет нормальный закон распределения с параметрами:𝜇 = 101.4617; σ=8.0113. Проверим её по критерию Хи-квадрат Пирсона.

Строим гистограмму:

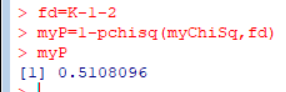


По ней видно, что ровно половина значений находится в половине стандартного отклонения.

Получаем наблюдаемое значение Хи-квадрат (myChiSq), сумму вероятностей(P) и сумму частот(N):



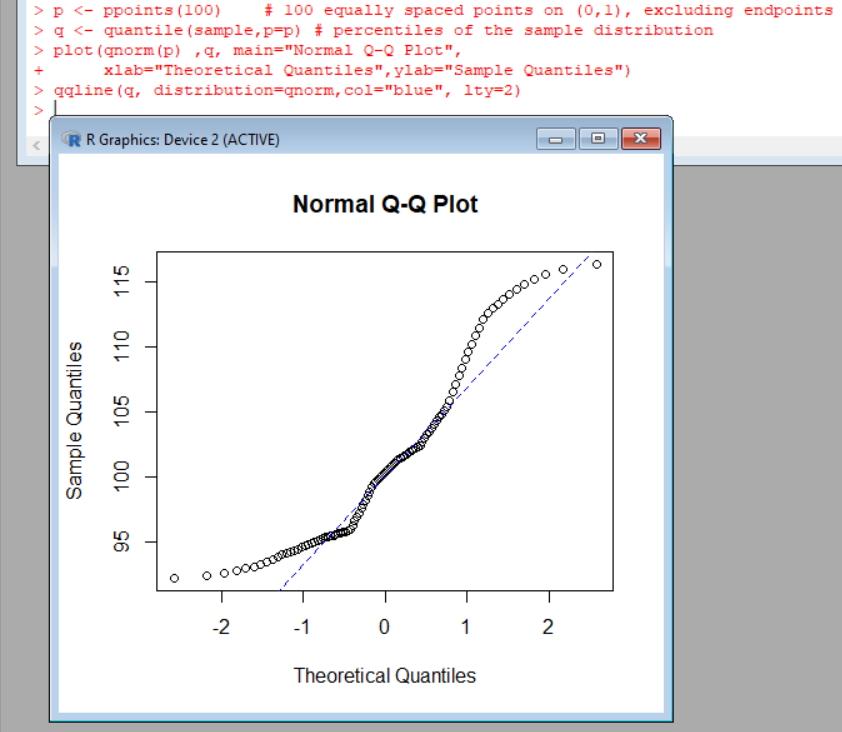
Как видим, сумма найденных вероятностей почти = 1 (0.91), как и должно быть; сумма эмпирических частот = n (10), а 𝜒набл = 2.309. Понятно, что чем меньше наблюдаемое значение критерия, тем лучше наша гипотеза согласуется с опытными данными. Учитывая, что критерий имеет распределение Хи-квадрат с числом степеней свободы, равным K – 1 – количество оцениваемых по выборке параметров, найдём вероятность получить значение критерия, большее, чем наблюдаемое:



Получили 0.51. Обычно пороговое значение этой вероятности полагают равным 0,05. Таким образом, вероятность превышения наблюдаемого значения отклонения эмпирических частот от теоретических, вычисленных в соответствии с гипотезой, (т.е. 0.51) сильно выше пороговой (0,05).

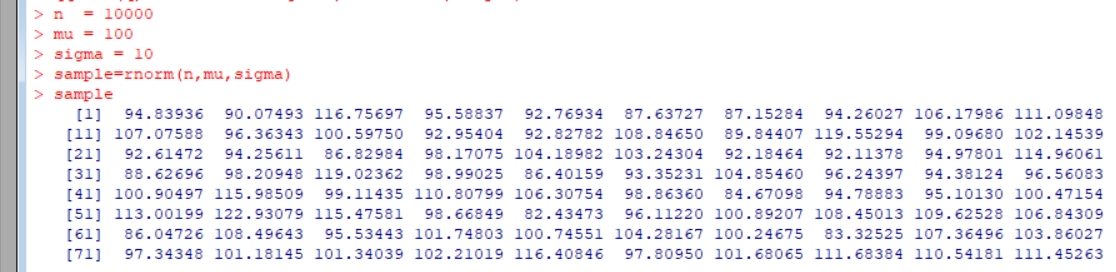
Вывод: выдвинутая гипотеза о законе распределения хорошо согласуется с опытными данными.

* 1. Построим диаграмму "Квантиль-Квантиль":

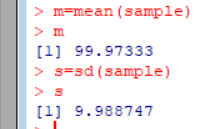


Построенная диаграмма также подтверждает, что выдвинутая гипотеза о законе распределения хорошо согласуется с опытными данными. Хотя и видно, что на хвостах данного графика есть отклонения нашей выборки от теоретического нормального распределения.

* 1. Смоделируем выборку sample объёма n=10000 из нормального распределения с параметрами mu=100, sigma=10.



* 1. Получим точечные оценки для параметров распределения:

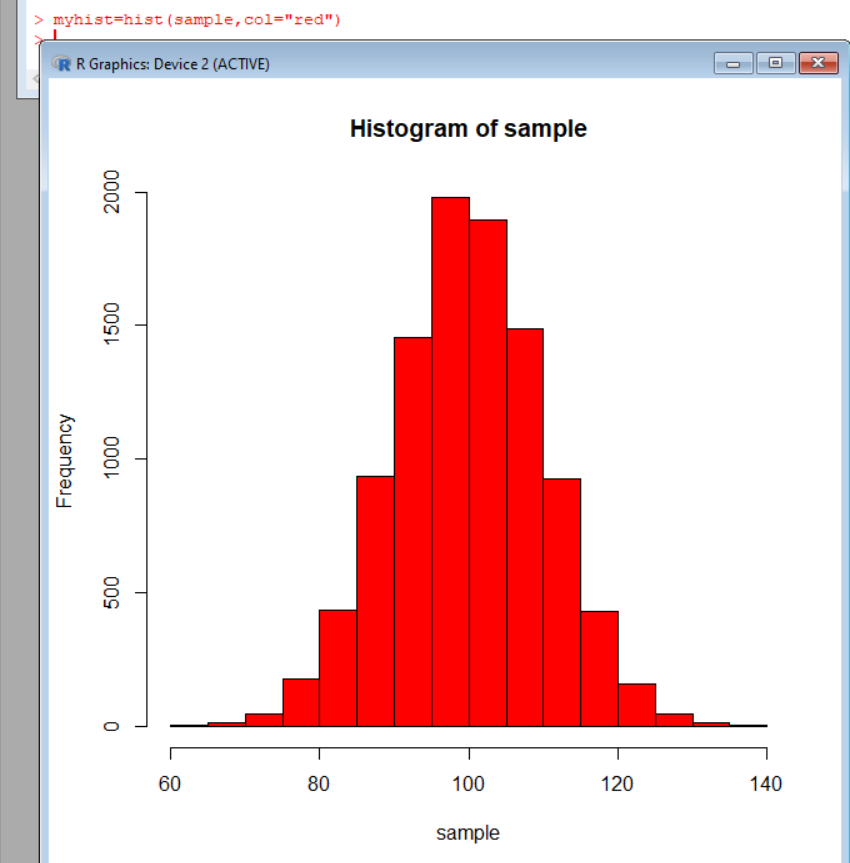


Как видим, по сравнению с выборкой из 10 значений, среднее значительно сильно стало ближе к заданному (0.03% отличие), как и среднеквадратическое отклонение (0.02%).

* 1. Сформулируем гипотезу:

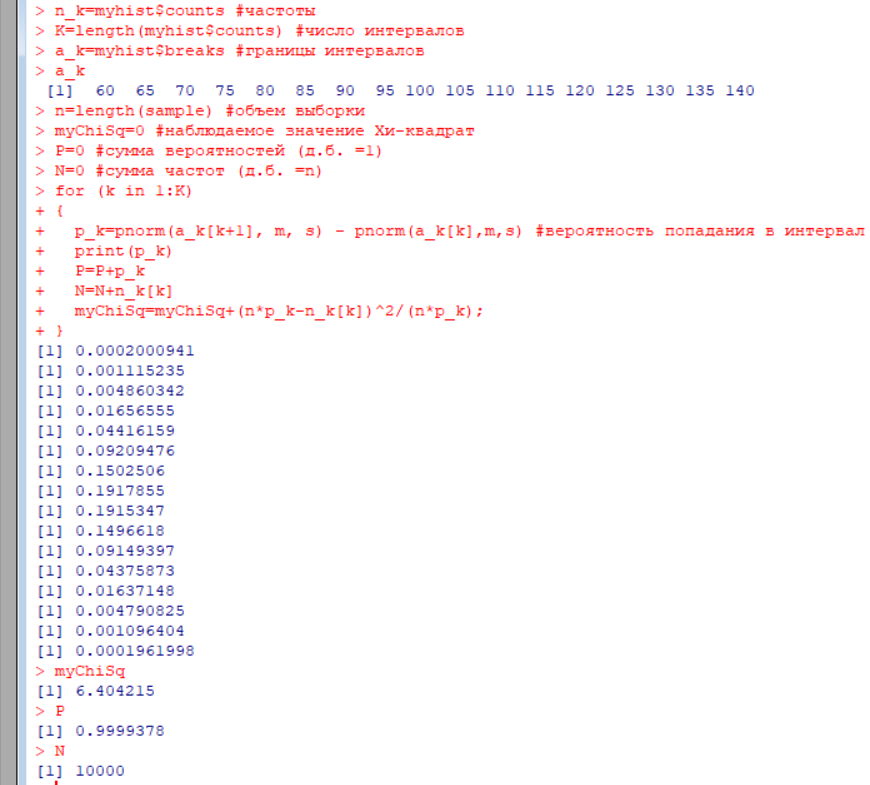
Наблюдаемая с.в. имеет нормальный закон распределения с параметрами:𝜇 = 99.973; σ=9.98. Проверим её по критерию Хи-квадрат Пирсона.

Строим гистограмму:

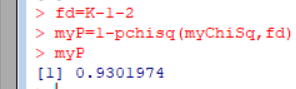


По сравнению с гистограммой для выборки из 10 значений данная гистограмма имеет более нормальную (куполообразную) форму.

Получаем наблюдаемое значение Хи-квадрат, сумму вероятностей и сумму частот:



Как видим, сумма найденных вероятностей теперь вообще приблизилась вплотную к 1, как и должно быть; сумма эмпирических частот = n (10000), а 𝜒набл. 2 = 6.404. Понятно, что чем меньше наблюдаемое значение критерия, тем лучше наша гипотеза согласуется с опытными данными. Учитывая, что критерий имеет распределение Хи-квадрат с числом степеней свободы, равным K – 1 – количество оцениваемых по выборке параметров, найдём вероятность получить значение критерия, большее, чем наблюдаемое:



Получили 0.93. Обычно пороговое значение этой вероятности полагают равным 0,05. Таким образом, вероятность превышения наблюдаемого значения отклонения эмпирических частот от теоретических, вычисленных в соответствии с гипотезой, (т.е. 0.93) выше пороговой (0,05).

Вывод: выдвинутая гипотеза о законе распределения хорошо согласуется с опытными данными.

Построим диаграмму "Квантиль-Квантиль":

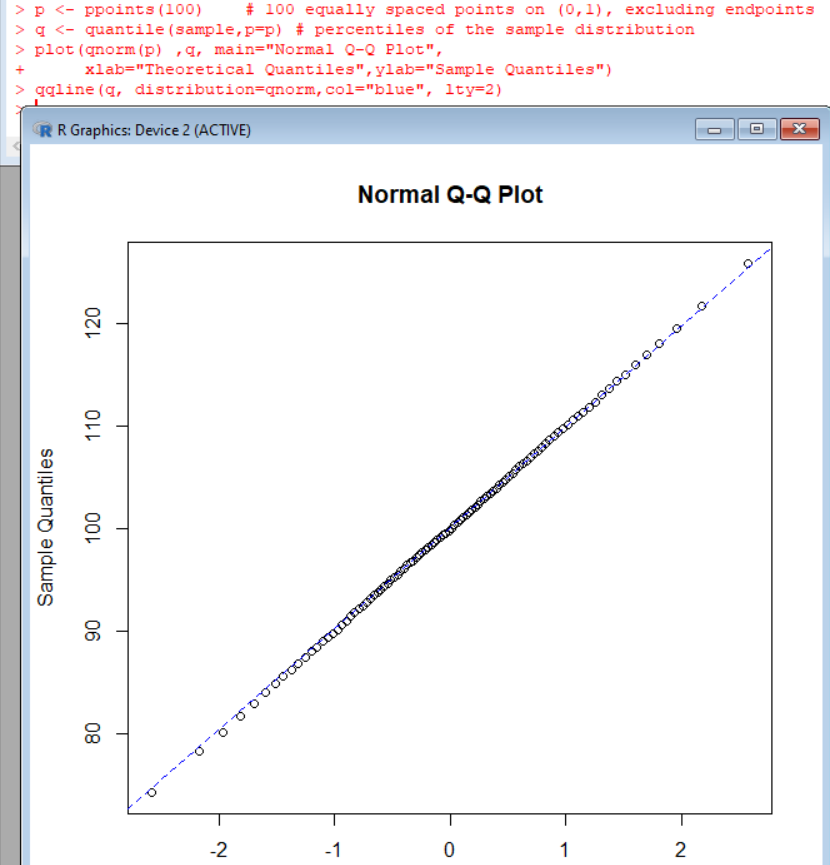
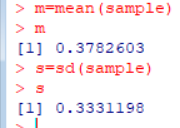


График нашей выборки идеально ложится на теоретическую нормальную кривую. Это говорит о том, что выдвинутая гипотеза о законе распределения хорошо согласуется с опытными данными.

* 1. Смоделируем выборку sample объёма n=10 из экспоненциального распределения с параметрам rate=2.



Получим точечные оценки для параметров распределения:

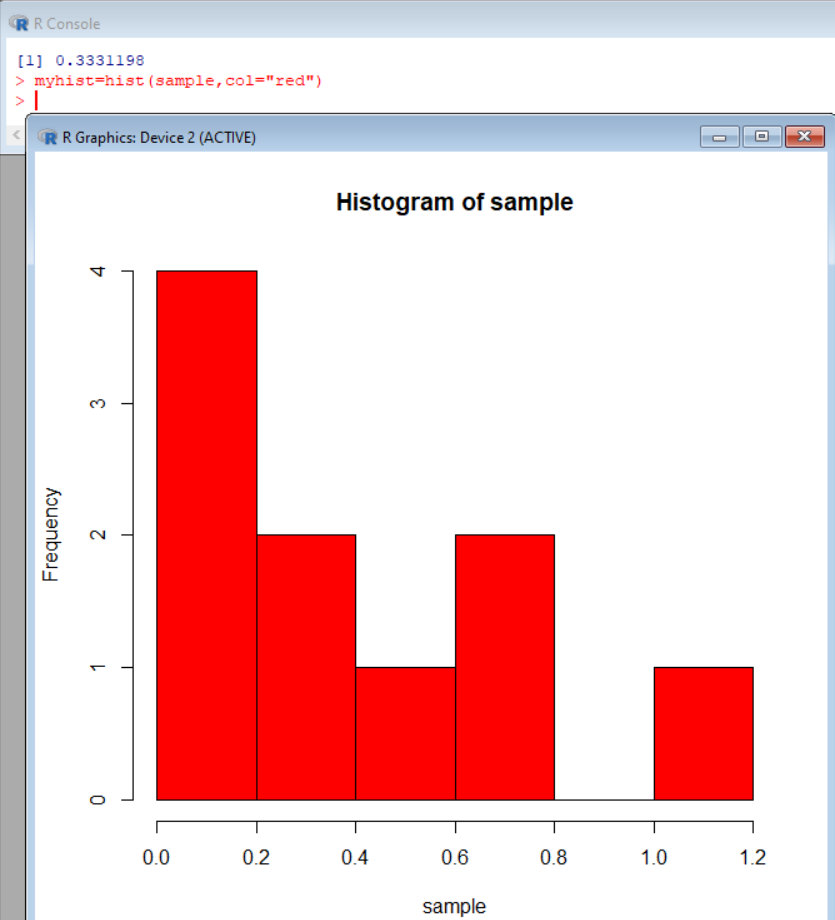


Видно, что у этого экспоненциального распределения среднее и стандартное отклонение очень близко друг к другу.

Сформулируем гипотезу:

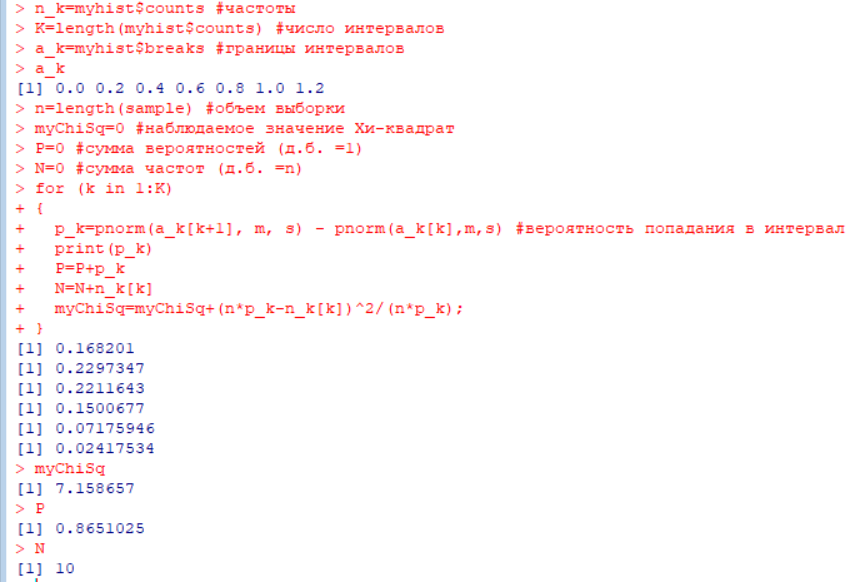
Наблюдаемая с.в. имеет нормальный закон распределения с параметрами:𝜇 = 0.378; σ=0.333. Проверим её по критерию Хи-квадрат Пирсона.

Строим гистограмму:

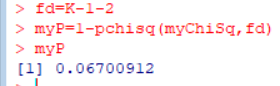


По графику распределение лишь отдаленно напоминает экспоненциальное.

Получаем наблюдаемое значение Хи-квадрат, сумму вероятностей и сумму частот:



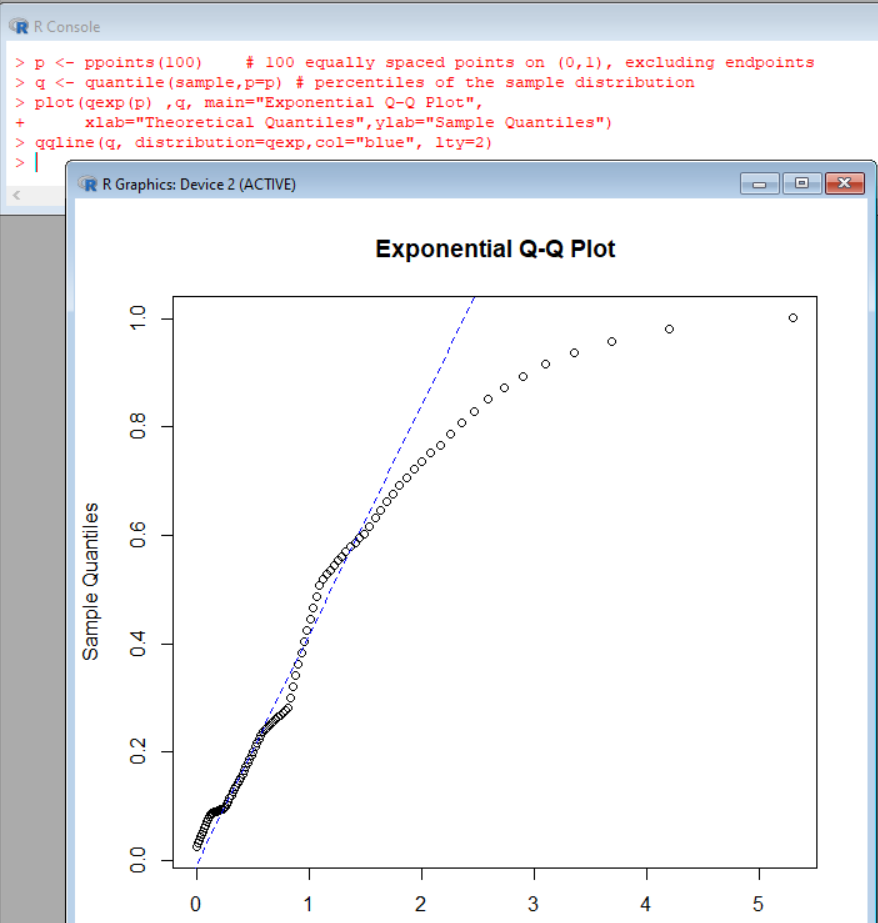
Как видим, сумма найденных вероятностей = 0.86, сумма эмпирических частот = n (10), а 𝜒набл. 2 = 7.16. Учитывая, что критерий имеет распределение Хи-квадрат с числом степеней свободы, равным K – 1 – количество оцениваемых по выборке параметров, найдём вероятность получить значение критерия, большее, чем наблюдаемое:



Получили 0.067, что очень близко к критическому значению. При 5% степени доверия нельзя сказать, что данное распределение не является нормальным.

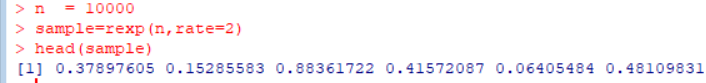
Вывод: выдвинутая гипотеза о законе распределения согласуется с опытными данными.

Построим диаграмму "Квантиль-Квантиль":

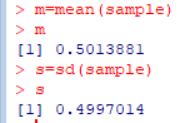


Как видим по графику, нижняя и средняя часть графика нашей выборки хорошо ложится на теоретическую кривую экспоненциального распределения, а верхняя часть отстает от теоретического. Поэтому сложно отрицать, что это не экспоненциальное распределение.

* 1. Смоделируем выборку sample объёма n=10000 из экспоненциального распределения с параметрам rate=2.



Получим точечные оценки для параметров распределения:

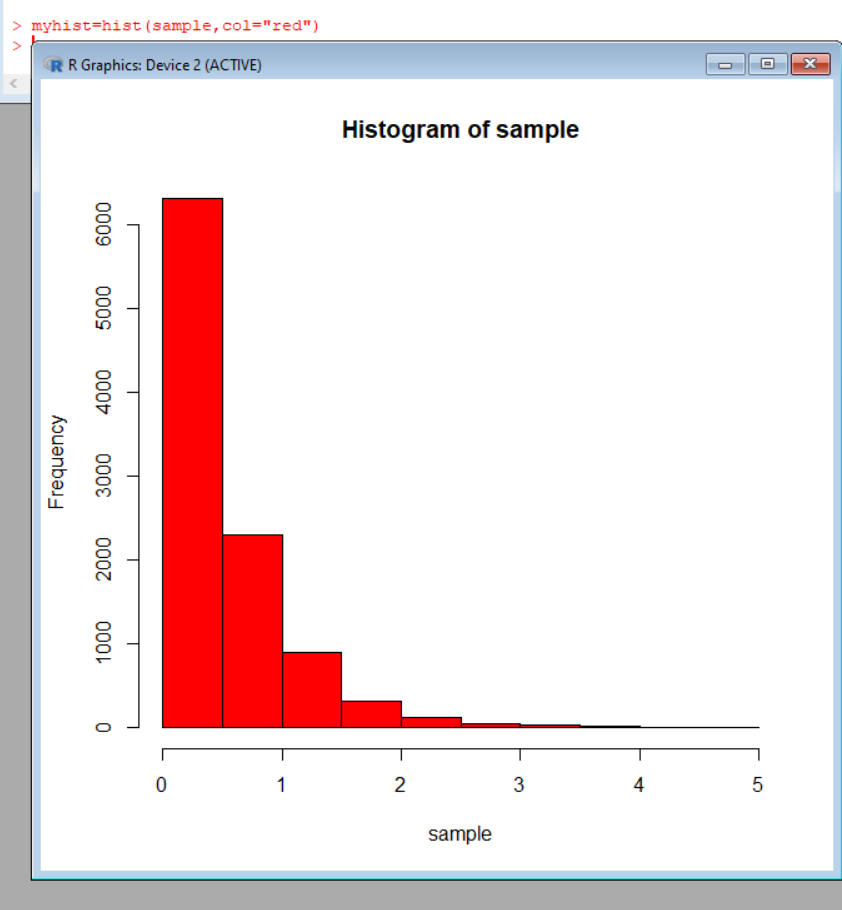


Видно, что у этого экспоненциального распределения среднее и стандартное отклонение стали еще ближе друг к другу.

Сформулируем гипотезу:

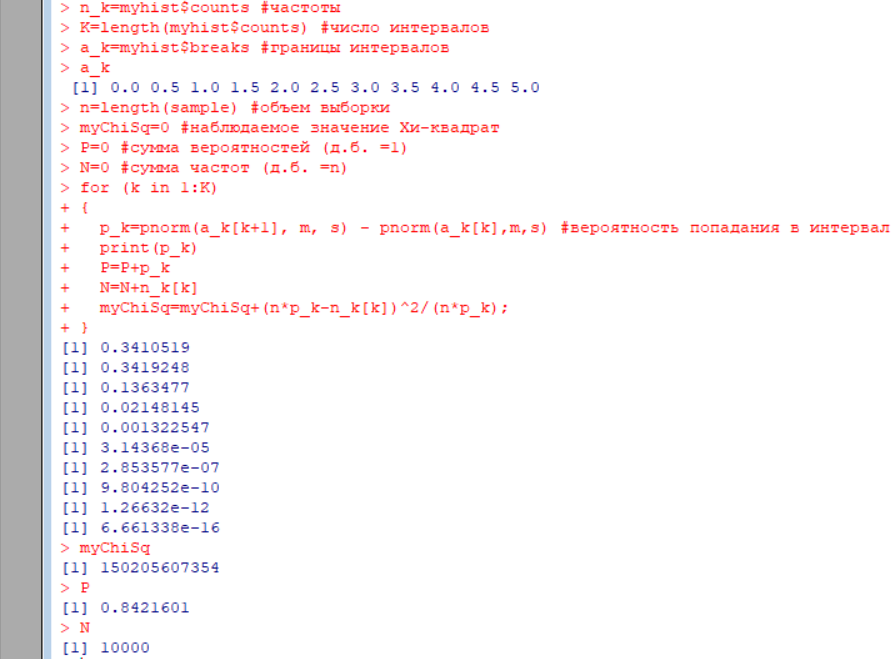
Наблюдаемая с.в. имеет нормальный закон распределения с параметрами:𝜇 = 0.501; σ=0.499. Проверим её по критерию Хи-квадрат Пирсона.

Строим гистограмму:

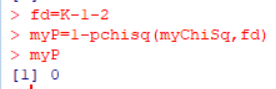


По графику четко видно экспоненциальное распределение.

Получаем наблюдаемое значение Хи-квадрат, сумму вероятностей и сумму частот:



Как видим, сумма найденных вероятностей = 0.84; сумма эмпирических частот = n (10000), а 𝜒набл. 2 = 150205607354. Найдем вероятность получить значение критерия, большее, чем наблюдаемое:



Получили 0. Обычно пороговое значение этой вероятности полагают равным 0,05. Таким образом, вероятность превышения наблюдаемого значения отклонения эмпирических частот от теоретических, вычисленных в соответствии с гипотезой, (т.е. 0) ниже пороговой (0,05).

Вывод: выдвинутая гипотеза о законе распределения совершенно не согласуется с опытными данными.

Построим диаграмму "Квантиль-Квантиль":

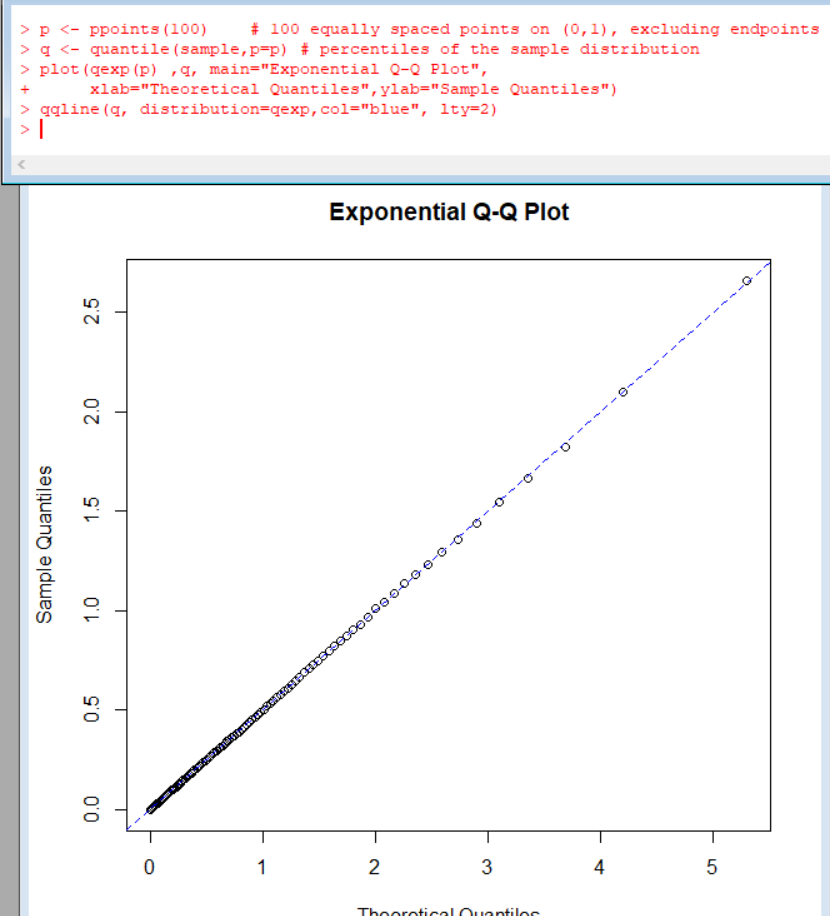


График нашей выборки идеально ложится на теоретическую экспоненциальную кривую, что говорит об экспоненциальном распределении данной выборки

10.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Моделированное распределение (columns)/Гипотетическое распределение (rows) | Нормальное, mu=100, sigma=10, объем выборки n=10 | Нормальное, mu=100, sigma=10, объем выборки n=10000 | Экспоненциальное, lambda=2, объем выборки n=10 | Экспоненциальное, lambda=2, объем выборки n=10000 |
| Нормальное, 𝜇 =101.461 σ=8.011. | p-value= 0.51.  Гипотеза принимается. |  |  |  |
| Нормальное, 𝜇 = 99.973; σ=9.98. |  | p-value= 0.93.  Гипотеза принимается. |  |  |
| Нормальное, 𝜇 = 0.378; σ=0.333. |  |  | p-value= 0.067.  Гипотеза принимается. |  |
| Нормальное, 𝜇 = 0.501; σ=0.499. |  |  |  | p-value= 0.  Гипотеза не принимается. |

Очевидно, что гипотеза о нормальном распределении принимается для первых 2 выборок, т.к. используемая выборка была смоделирована с помощью нормального распределения, к тому что же выборочные оценки практически совпадают с данными параметрами. Третья заданная выборка с 10 значениями, как экспоненциальная, также по коэффициенту р является нормальной, так как нельзя опровергнуть нулевую гипотезу о нормальности распределения при 5% степени доверия. Четвертая заданная выборка с 10 000 значениями, как экспоненциальная, по коэффициенту р=0 является явно ненормальной.

*Хи2 - 2 - проверка гипотезы о законе распределения по критерию Хи-квадрат для реальных данных.*

Выберите из репозитория 2 набора данных; для каждого из них:

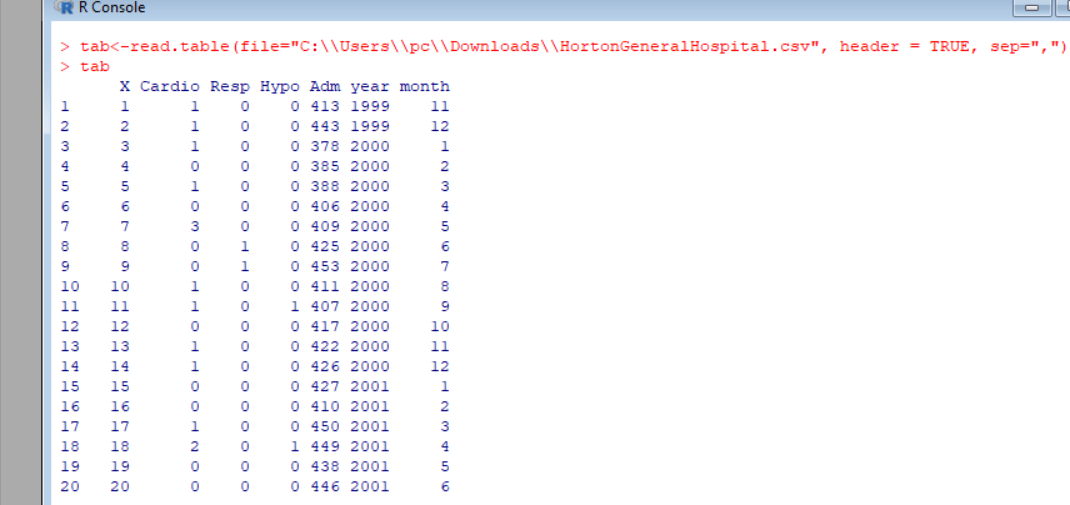
1. постройте гистограмму распределения;
2. сделайте предположение о возможном законе распределения;
3. найдите точечные оценки параметра(ов) предполагаемого распределения;
4. сформулируйте гипотезу о законе распределения (с параметрами, совпадающими с их выборочными оценками);
5. проверьте гипотезу по критерию Хи-квадрат;
6. постройте диаграмму "Квантиль-Квантиль";

**Ход работы:**

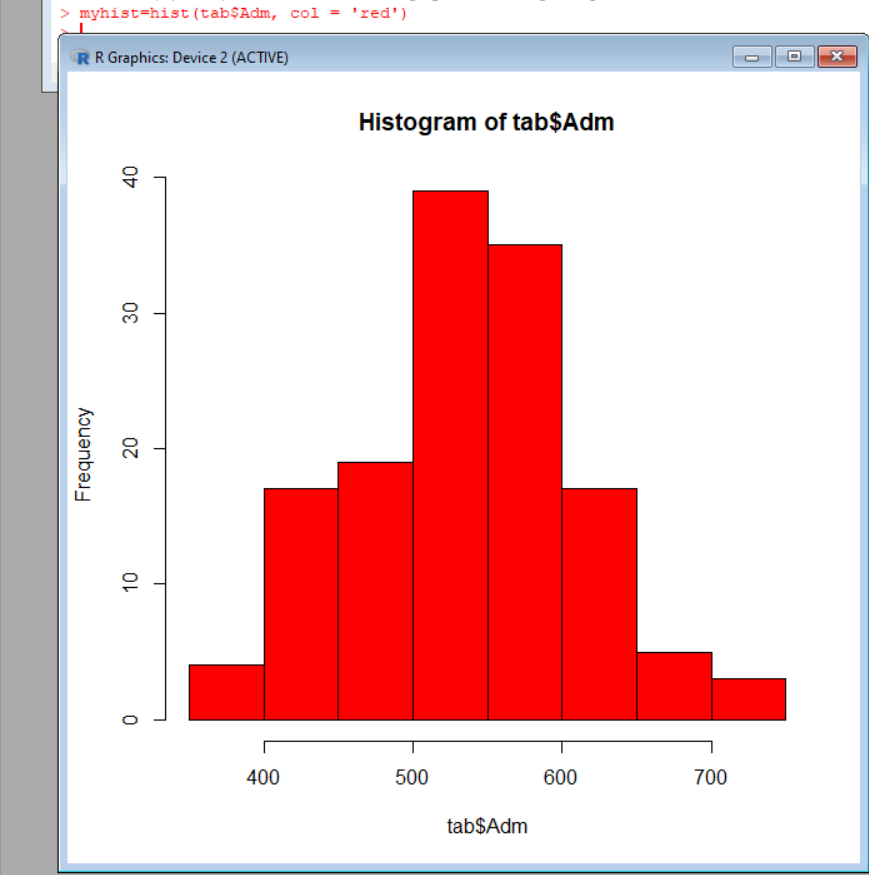
* + 1. Выбран набор данных:

<http://archive.ics.uci.edu/ml/datasets/Horton+General+Hospital>

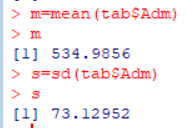
Выведем этот датасет:



Для анализа распределения используем столбец Adm. Построим гистограмму распределения для этой переменной:



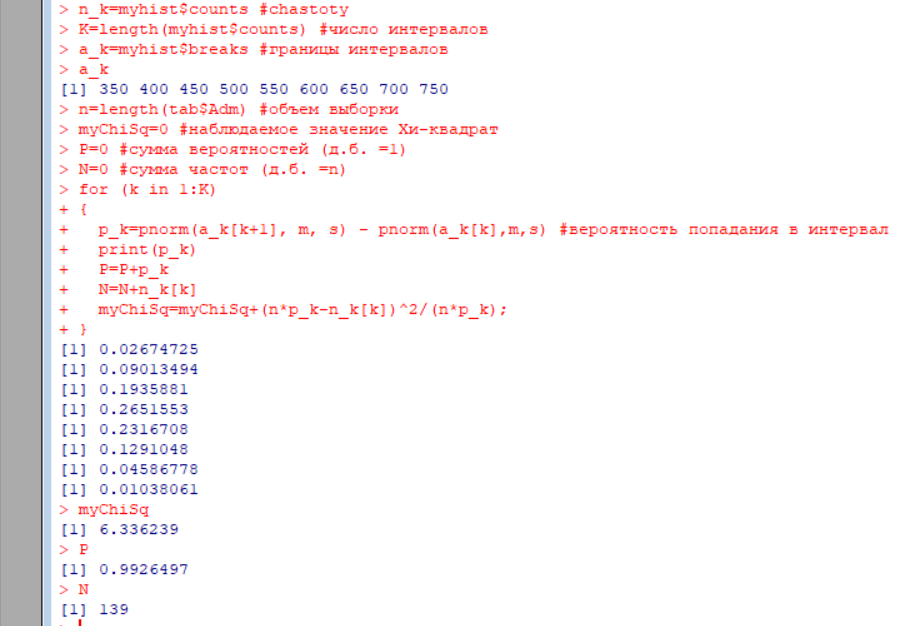
* + 1. По виду верхнего графика можно утверждать, что он имеет нормальное (из-за формы купола) распределение. При чем чуть более 70% значений лежит в промежутке между 500 и 600, в районе 1 стандартного отклонения (см.ниже).
    2. Получим точечные оценки для параметров распределения:



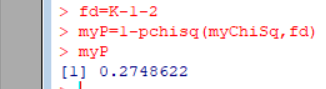
Среднее значение = 534.98, стандартное отклонение = 73.13, то есть составляет 15% от среднего, что может говорить о слишком крутом подъеме графика, который несвойственен теоретическому нормальному.

* + 1. Сформулируем гипотезу:

Наблюдаемая с.в. имеет нормальный закон распределения с параметрами:𝜇 = 534.98; σ=73.13. Проверим её по критерию Хи-квадрат Пирсона.



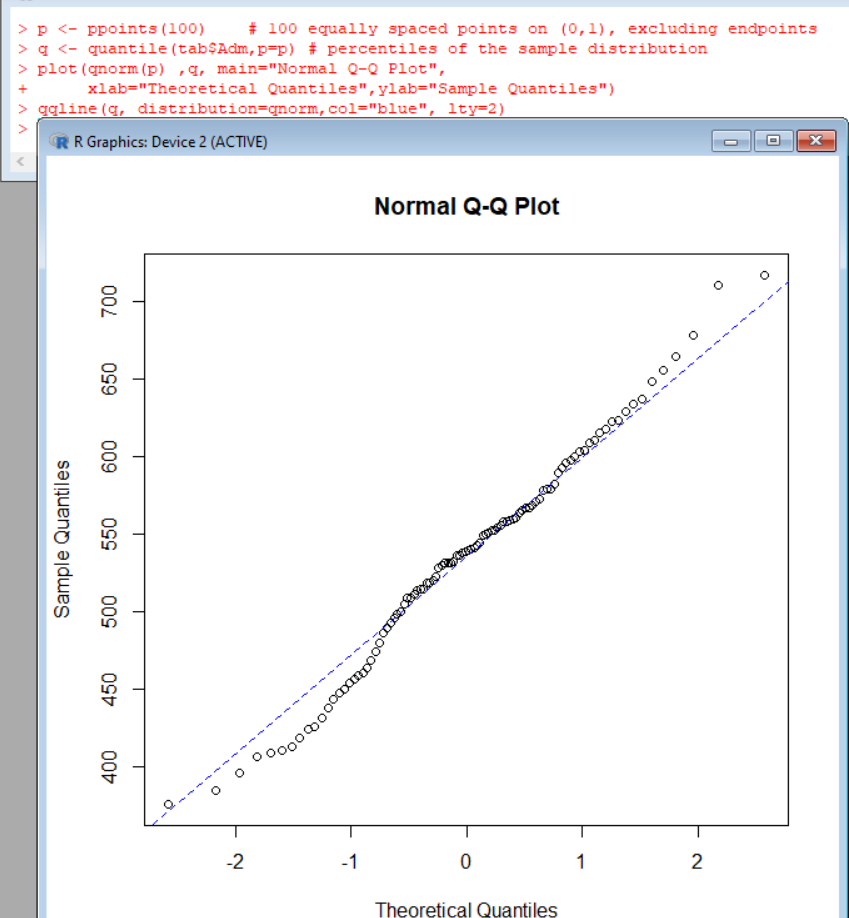
Как видим, сумма найденных вероятностей почти = 1, как и должно быть; сумма эмпирических частот = n (139), а 𝜒набл. 2 = 6.33. Учитывая, что критерий имеет распределение Хи-квадрат с числом степеней свободы, равным K – 1 – количество оцениваемых по выборке параметров, найдём вероятность получить значение критерия, большее, чем наблюдаемое:



Получили 0.27. Обычно пороговое значение этой вероятности полагают равным 0,05. Таким образом, вероятность превышения наблюдаемого значения отклонения эмпирических частот от теоретических, вычисленных в соответствии с гипотезой, (т.е. 0.2300673) выше пороговой (0,05).

Вывод: выдвинутая гипотеза о законе распределения хорошо согласуется с опытными данными.

* + 1. Построим диаграмму "Квантиль-Квантиль":

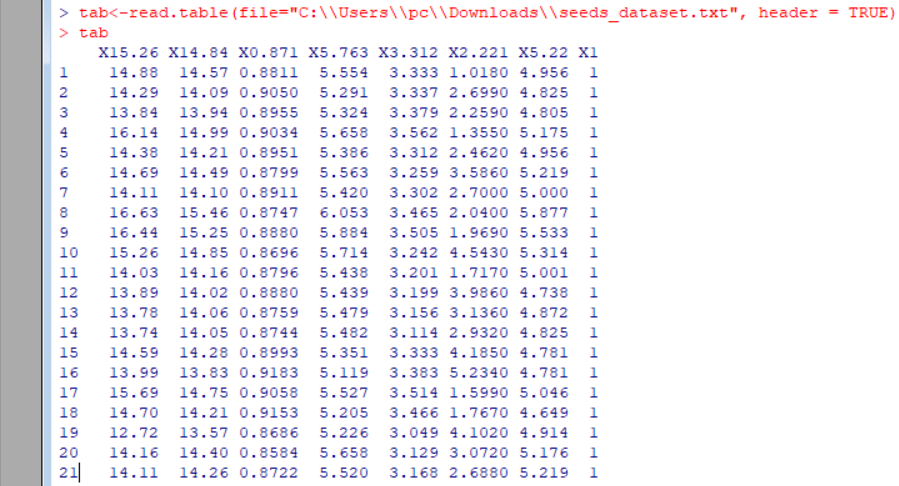


По графику также хорошо видно, что распределение является нормальным, несмотря на то, что нижняя часть нашей исследуемой лежит ниже теоретической нормальной кривой.

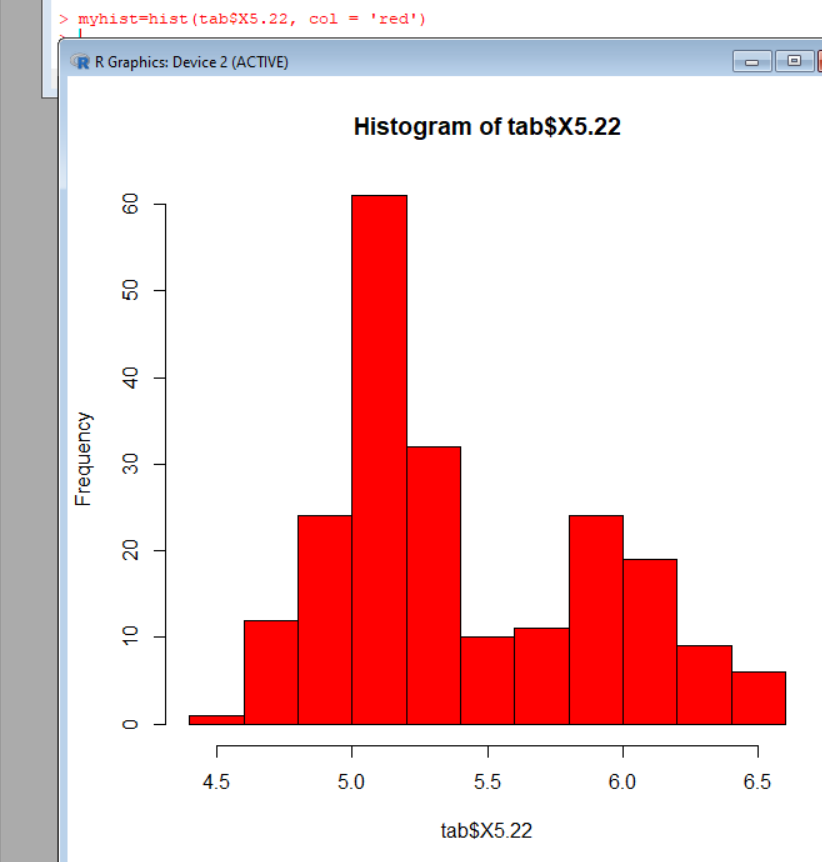
2-й набор данных:

<http://archive.ics.uci.edu/ml/datasets/seeds>

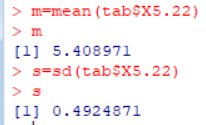
Выведем этот датасет:



* + 1. Для анализа распределения используем столбец Х5.22. Построим гистограмму распределения для этой переменной:



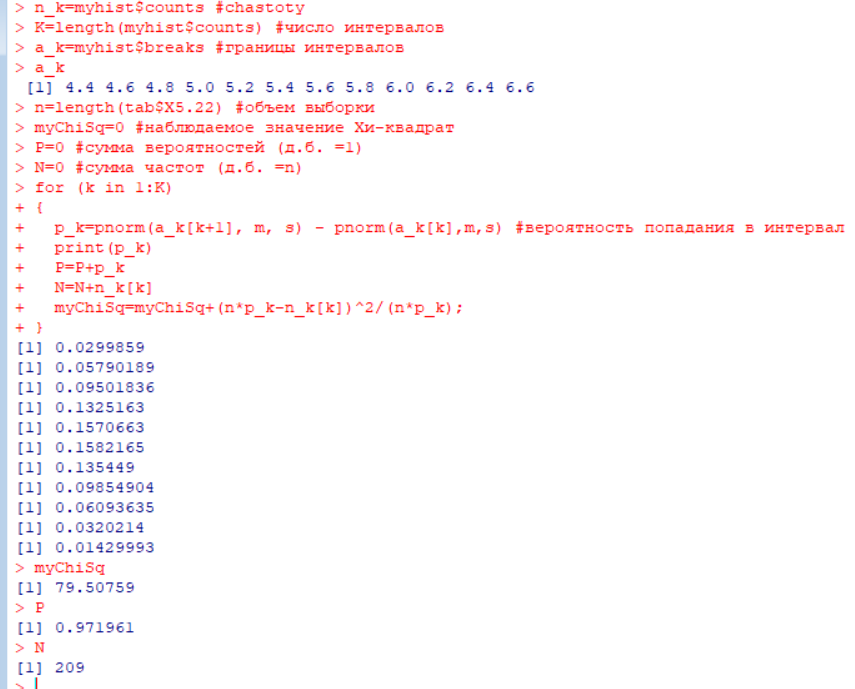
* + 1. По графику видно, что это похоже на нормальное распределение, у которого удалили часть значений на промежутке от 5.3 до 5.8 или их слишком много перекинуло в промежуток от 5 до 5.2(там 60% значений всех).
    2. Получим точечные оценки для параметров распределения:



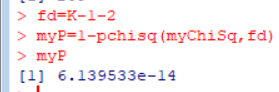
Стандартное отклонение составляет 10% от среднего значения.

* + 1. Сформулируем гипотезу:

Наблюдаемая с.в. имеет нормальный закон распределения с параметрами:𝜇 = 5.41; σ=0.49. Проверим её по критерию Хи-квадрат Пирсона.

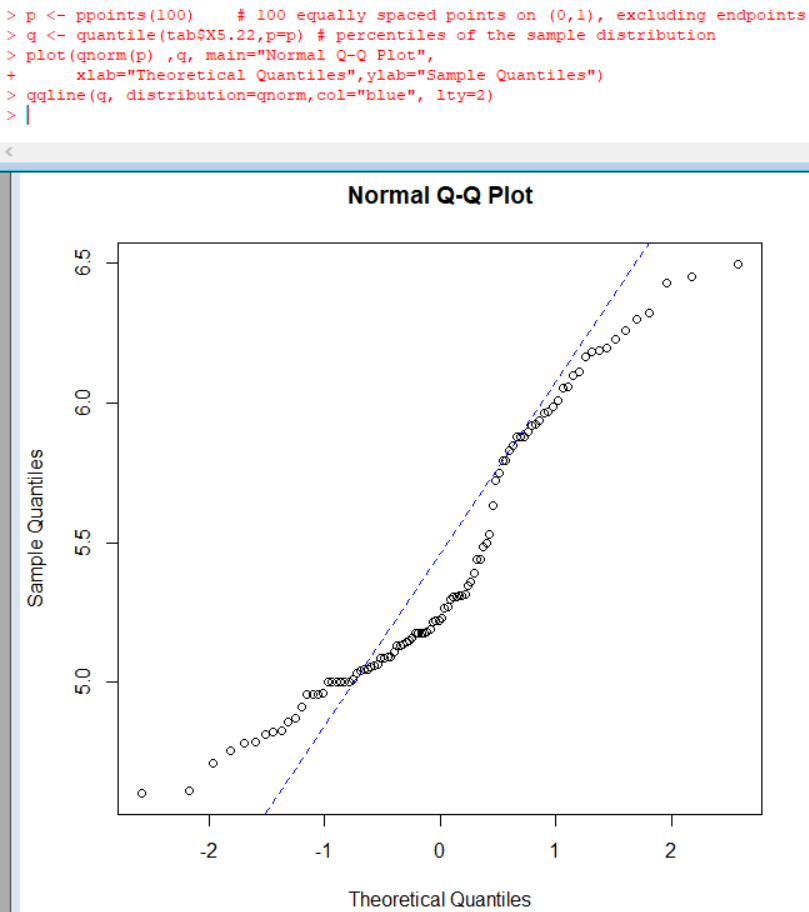


Как видим, сумма найденных вероятностей почти = 1, сумма эмпирических частот = n (209), а 𝜒набл. 2 = 79.5. Учитывая, что критерий имеет распределение Хи-квадрат с числом степеней свободы, равным K – 1 – количество оцениваемых по выборке параметров, найдём вероятность получить значение критерия, большее, чем наблюдаемое:



Таким образом, нулевая гипотеза сильно отрицается, что говорит об отсутствии нормального распределения данной выборки.

* + 1. Построим диаграмму "Квантиль-Квантиль":



По графику также видно, что исследуемая кривая практически вообще не ложится на теоретическую нормальную, что также подтверждает ненормальное распределение данной выборки.

**Общий вывод:** была изучена проверка гипотезы о законе распределения по критерию Хи-квадрат для смоделированных данных и реальных данных. По гистограмме был дан предварительный вывод о распределении выборки, а по значению р-коэффициента (если он меньше 0.05, то распределение ненормальное при 5% доверия) и графику q-q –plot (если точки исследуемой кривой хорошо ложатся на теоретическую нормальную, то распределение нормальное) уже точно можно его обобщить.

Код:

**n = 10**

**mu = 100**

**sigma = 10**

**sample=rnorm(n,mu,sigma)**

**sample**

**m=mean(sample)**

**m**

**s=sd(sample)**

**s**

**myhist=hist(sample,col="red")**

**n\_k=myhist$counts #частоты**

**K=length(myhist$counts) #число интервалов**

**a\_k=myhist$breaks #границы интервалов**

**a\_k**

**n=length(sample) #объем выборки**

**myChiSq=0 #наблюдаемое значение Хи-квадрат**

**P=0 #сумма вероятностей (д.б. =1)**

**N=0 #сумма частот (д.б. =n)**

**for (k in 1:K)**

**{**

**p\_k=pnorm(a\_k[k+1], m, s) - pnorm(a\_k[k],m,s) #вероятность попадания в интервал**

**print(p\_k)**

**P=P+p\_k**

**N=N+n\_k[k]**

**myChiSq=myChiSq+(n\*p\_k-n\_k[k])^2/(n\*p\_k);**

**}**

**fd=K-1-2**

**myP=1-pchisq(myChiSq,fd)**

**myP**

**p <- ppoints(100) # 100 equally spaced points on (0,1), excluding endpoints**

**q <- quantile(sample,p=p) # percentiles of the sample distribution**

**plot(qnorm(p) ,q, main="Normal Q-Q Plot",**

**xlab="Theoretical Quantiles",ylab="Sample Quantiles")**

**qqline(q, distribution=qnorm,col="blue", lty=2)**

**n = 10000**

**mu = 100**

**sigma = 10**

**sample=rnorm(n,mu,sigma)**

**sample**

**myhist=hist(sample,col="red")**

**n\_k=myhist$counts #частоты**

**K=length(myhist$counts) #число интервалов**

**a\_k=myhist$breaks #границы интервалов**

**a\_k**

**n=length(sample) #объем выборки**

**myChiSq=0 #наблюдаемое значение Хи-квадрат**

**P=0 #сумма вероятностей (д.б. =1)**

**N=0 #сумма частот (д.б. =n)**

**for (k in 1:K)**

**{**

**p\_k=pnorm(a\_k[k+1], m, s) - pnorm(a\_k[k],m,s) #вероятность попадания в интервал**

**print(p\_k)**

**P=P+p\_k**

**N=N+n\_k[k]**

**myChiSq=myChiSq+(n\*p\_k-n\_k[k])^2/(n\*p\_k);**

**}**

**myChiSq**

**P**

**N**

**fd=K-1-2**

**myP=1-pchisq(myChiSq,fd)**

**myP**

**p <- ppoints(100) # 100 equally spaced points on (0,1), excluding endpoints**

**q <- quantile(sample,p=p) # percentiles of the sample distribution**

**plot(qnorm(p) ,q, main="Normal Q-Q Plot",**

**xlab="Theoretical Quantiles",ylab="Sample Quantiles")**

**qqline(q, distribution=qnorm,col="blue", lty=2)**

**n = 10**

**sample=rexp(n,rate=2)**

**sample**

**m=mean(sample)**

**m**

**s=sd(sample)**

**s**

**myhist=hist(sample,col="red")**

**n\_k=myhist$counts #частоты**

**K=length(myhist$counts) #число интервалов**

**a\_k=myhist$breaks #границы интервалов**

**a\_k**

**n=length(sample) #объем выборки**

**myChiSq=0 #наблюдаемое значение Хи-квадрат**

**P=0 #сумма вероятностей (д.б. =1)**

**N=0 #сумма частот (д.б. =n)**

**for (k in 1:K)**

**{**

**p\_k=pnorm(a\_k[k+1], m, s) - pnorm(a\_k[k],m,s) #вероятность попадания в интервал**

**print(p\_k)**

**P=P+p\_k**

**N=N+n\_k[k]**

**myChiSq=myChiSq+(n\*p\_k-n\_k[k])^2/(n\*p\_k);**

**}**

**myChiSq**

**P**

**N**

**fd=K-1-2**

**myP=1-pchisq(myChiSq,fd)**

**myP**

**p <- ppoints(100) # 100 equally spaced points on (0,1), excluding endpoints**

**q <- quantile(sample,p=p) # percentiles of the sample distribution**

**plot(qexp(p) ,q, main="Exponential Q-Q Plot",**

**xlab="Theoretical Quantiles",ylab="Sample Quantiles")**

**qqline(q, distribution=qexp,col="blue", lty=2)**

**n = 10000**

**sample=rexp(n,rate=2)**

**head(sample)**

**myhist=hist(sample,col="red")**

**n\_k=myhist$counts #частоты**

**K=length(myhist$counts) #число интервалов**

**a\_k=myhist$breaks #границы интервалов**

**a\_k**

**n=length(sample) #объем выборки**

**myChiSq=0 #наблюдаемое значение Хи-квадрат**

**P=0 #сумма вероятностей (д.б. =1)**

**N=0 #сумма частот (д.б. =n)**

**for (k in 1:K)**

**{**

**p\_k=pnorm(a\_k[k+1], m, s) - pnorm(a\_k[k],m,s) #вероятность попадания в интервал**

**print(p\_k)**

**P=P+p\_k**

**N=N+n\_k[k]**

**myChiSq=myChiSq+(n\*p\_k-n\_k[k])^2/(n\*p\_k);**

**}**

**myChiSq**

**P**

**N**

**fd=K-1-2**

**myP=1-pchisq(myChiSq,fd)**

**myP**

**p <- ppoints(100) # 100 equally spaced points on (0,1), excluding endpoints**

**q <- quantile(sample,p=p) # percentiles of the sample distribution**

**plot(qexp(p) ,q, main="Exponential Q-Q Plot",**

**xlab="Theoretical Quantiles",ylab="Sample Quantiles")**

**qqline(q, distribution=qexp,col="blue", lty=2)**

**tab<-read.table(file="C:\\Users\\pc\\Downloads\\HortonGeneralHospital.csv", header = TRUE, sep=",")**

**tab**

**myhist=hist(tab$Adm, col = 'red')**

**m=mean(tab$Adm)**

**m**

**s=sd(tab$Adm)**

**s**

**n\_k=myhist$counts #chastoty**

**K=length(myhist$counts) #число интервалов**

**a\_k=myhist$breaks #границы интервалов**

**a\_k**

**n=length(tab$Adm) #объем выборки**

**myChiSq=0 #наблюдаемое значение Хи-квадрат**

**P=0 #сумма вероятностей (д.б. =1)**

**N=0 #сумма частот (д.б. =n)**

**for (k in 1:K)**

**{**

**p\_k=pnorm(a\_k[k+1], m, s) - pnorm(a\_k[k],m,s) #вероятность попадания в интервал**

**print(p\_k)**

**P=P+p\_k**

**N=N+n\_k[k]**

**myChiSq=myChiSq+(n\*p\_k-n\_k[k])^2/(n\*p\_k);**

**}**

**myChiSq**

**P**

**N**

**fd=K-1-2**

**myP=1-pchisq(myChiSq,fd)**

**myP**

**p <- ppoints(100) # 100 equally spaced points on (0,1), excluding endpoints**

**q <- quantile(tab$Adm,p=p) # percentiles of the sample distribution**

**plot(qnorm(p) ,q, main="Normal Q-Q Plot",**

**xlab="Theoretical Quantiles",ylab="Sample Quantiles")**

**qqline(q, distribution=qnorm,col="blue", lty=2)**

**tab<-read.table(file="C:\\Users\\pc\\Downloads\\seeds\_dataset.txt", header = TRUE)**

**tab**

**myhist=hist(tab$X5.22, col = 'red')**

**m=mean(tab$X5.22)**

**m**

**s=sd(tab$X5.22)**

**s**

**n\_k=myhist$counts #chastoty**

**K=length(myhist$counts) #число интервалов**

**a\_k=myhist$breaks #границы интервалов**

**a\_k**

**n=length(tab$X5.22) #объем выборки**

**myChiSq=0 #наблюдаемое значение Хи-квадрат**

**P=0 #сумма вероятностей (д.б. =1)**

**N=0 #сумма частот (д.б. =n)**

**for (k in 1:K)**

**{**

**p\_k=pnorm(a\_k[k+1], m, s) - pnorm(a\_k[k],m,s) #вероятность попадания в интервал**

**print(p\_k)**

**P=P+p\_k**

**N=N+n\_k[k]**

**myChiSq=myChiSq+(n\*p\_k-n\_k[k])^2/(n\*p\_k);**

**}**

**myChiSq**

**P**

**N**

**fd=K-1-2**

**myP=1-pchisq(myChiSq,fd)**

**myP**

**p <- ppoints(100) # 100 equally spaced points on (0,1), excluding endpoints**

**q <- quantile(tab$X5.22,p=p) # percentiles of the sample distribution**

**plot(qnorm(p) ,q, main="Normal Q-Q Plot",**

**xlab="Theoretical Quantiles",ylab="Sample Quantiles")**

**qqline(q, distribution=qnorm,col="blue", lty=2)**