Анализ статистических данных в R Studio

Исходные данные:	Файл с данными
Данные приведены в столбцах. Первый столбец (без имени) номер измерения, второй и третий — значения с.в. X и Y. Данные в строках разделены точкой с запятой.	;X;Y 0;15.04;12.0 1;16.24;14.18 2;7.96;12.12 3;17.4;7.98 ...

> **Код на языке программирования R будет приведён таким шрифтом**

Вывод полученный после выполнения команд таким шрифтом.

В некоторых случаях результат выполнения команд будет записан в переменную и на экран значение не будет выведено. Для просмотра значения переменной достаточно вместо команды написать её имя. Однако, можно выполнять команды без записи результатов в переменные.

Для выполнения некоторых действий может понадобится установка дополнительных модулей (пакетов). R Studio предложить сделать это автоматически, при попытке использования функции из нового модуля.

Справку по любой функции можно получить выполнив `?функция`, например:

`? mean`.

1. Загрузка данных из CSV файла

File → Import Dataset → From CSV…

- Указать разделитель, которым отделены друг от друга значения. В поле предпросмотра данные должны разделится на несколько столбцов.
- Указать название данных в поле Name (напр. **mydata**)
- При необходимости отменить загрузку некоторых столбцов нажав на его заголовок и выбрав skip.
- Импортировать данные — кнопка Import

Альтернативой будет использование команды read.delim:

```
mydata = read.delim('stud-lab.csv',';', header=TRUE)
```

2. Показать сводку по данным

summary(mydata)

```
X1               X                Y
Min.   : 0.00   Min.   : 6.26   Min.   : 3.74
1st Qu.:24.75   1st Qu.:10.54   1st Qu.: 9.33
Median :49.50   Median :12.27   Median :11.14
```

```
      Mean   :49.50   Mean   :12.34   Mean   :10.91
      3rd Qu.:74.25   3rd Qu.:14.12   3rd Qu.:12.55
      Max.   :99.00   Max.   :20.65   Max.   :17.94
```

Выше, первый столбец (X1) — номера строк с данными в файле.
Запишем выборки X и Y в соответствующие переменные для удобства

```
x = mydata$X
y = mydata$Y
```

3. Таблица частот

Число интервалов рекомендуется выбирать по формуле[1]:

$$n = 1 + \lfloor \log_2 N \rfloor ,$$

где N — объём выборки, $\lfloor x \rfloor$ — обозначает целую часть числа.
Тогда ширина интервала:

$$h = \frac{max(X) - min(X)}{n} \qquad (1)$$

Число строк в загруженном файле (соответствует объёму выборки)

```
N = nrow(mydata)
n = 1 + trunc( log2(N) )
h = ( max(x) - min(x) ) / n
```

Выислим границы интервалов

```
x_breaks = seq( min(x) - h/2, max(x)+h/2, h)
```

```
5.232  7.288  9.344 11.399 13.455 15.511 17.566 19.622 21.678
```

середины интервалов

```
x_mids = seq( min(x), max(x), h)
```

```
   6.260  8.316 10.371 12.427 14.483 16.539 18.594 20.650
```

интервалы с частотами

```
x_ints = table( cut( x, breaks=x_breaks ) )
```

```
Создадим отдельную таблицу (DataFrame)
df = data.frame( x_ints )
```

Куда добавим необходимые поля указывая после имени талицы знак доллара, а за ним имя нового столбца

```
df$mids = x_mids
```

Добавим относительные частоты

```
df$w = df$Freq / sum( df$Freq )
```

---

1 https://ru.wikipedia.org/wiki/Правило_Стёрджеса

Относительные накопленные частоты:

```
df$cw = cumsum(df$w)
```

Ранги, где вариантой будем считать середину интервала:

```
df$r = rank(x_mid)
```

В результате будем иметь следующее содержимое таблицы df:

```
          Var1 Freq   mids    w   cw r
1 (5.23,7.29]    5  6.260 0.05 0.05 1
2 (7.29,9.34]    7  8.316 0.07 0.12 2
3 (9.34,11.4]   25 10.371 0.25 0.37 3
4 (11.4,13.5]   27 12.427 0.27 0.64 4
5 (13.5,15.5]   24 14.483 0.24 0.88 5
6 (15.5,17.6]    8 16.539 0.08 0.96 6
7 (17.6,19.6]    3 18.594 0.03 0.99 7
8 (19.6,21.7]    1 20.650 0.01 1.00 8
```

4. Построить гистограмму:

```
hist(mydata$X, breaks = n)
```

или

```
qplot(x, geom='histogram', binwidth=h, xlab='X', ylab='частота',
fill=I("grey32"), col=I("white"))[2]
```

5. Числовые характеристики

Средние выборочное

```
mX = mean(x)
```

```
12.3393
```

Стандартное отклонение по выборке

```
sdx = sd(x)
```

```
2.87531
```

```
minx = min(x)
```

```
6.26
```

```
maxx = max(x)
```

```
20.65
```

```
medx = median(x)
```

```
12.27
```

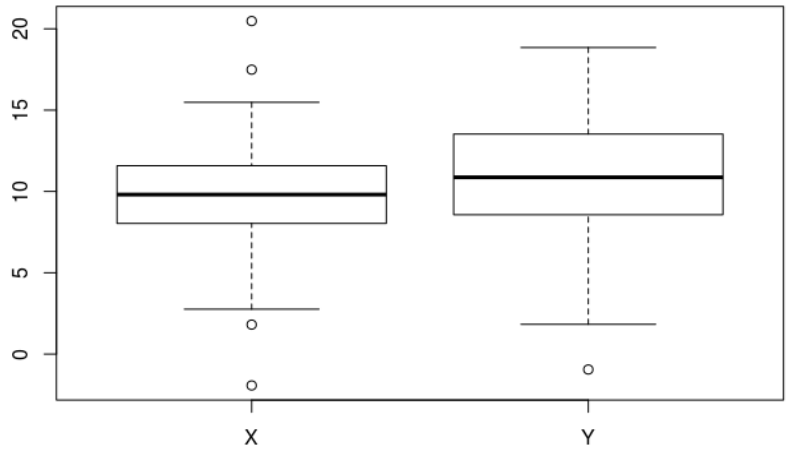
2 требуется пакет ggplot2

Диаграмма «Ящик с усами»

```
boxplot(x, y)
```

Важно указать название для каждой отдельной диаграммы, здесь это выборки X и Y:

```
boxplot(x,y, names=c('X', 'Y') )
```



Создание диаграммы boxplot с помощью пакета ggplot:

```
ggplot( stack(df), aes(x = ind, y = values) )  + geom_boxplot()
+ xlab('') + ylab('')
```

Функция stack(df) разбивает таблицу данных df на отдельные столбцы. При необходимости, можно указать только некоторые из них заменив df на df[c('X', 'Y')]. В квадратных скобках указывается вектор (массив) с именами интересующих столбцов.

6. **Проверка генеральной совокупности на «нормальность»**. Используем тест Шапиро-Уилка, нулевая гипотеза в котором формулируется так: случайная величина X распределена нормально, альтернативная гипотеза имеет прямо противоположный смысл. Применим тест к возможным значениям с.в. X. Здесь и далее договоримся использовать уровень значимости $\alpha=0.05$.

```
shapiro.test(x)
```

```
   Shapiro-Wilk normality test

data:  x
W = 0.99036, p-value = 0.6935
```

P-значение (p-value) больше заданного уровня значимости (0.6935 > 0.05) значит, нет оснований отклонить используемую в данном тесте нулевую гипотезу:

генеральная совокупность, из которой была сделана данная выборка распределена по нормальному закону.

7. Согласно тесту Шапиро-Уилка генеральная совокупность подчинятся нормальному закону распределения.
   Параметрами нормального распределения являются m — среднее, sd — стандартное отклонение. Эти значения уже вычислены в п. 5. Тогда **кривая плотности (теоретическое распределение)** , подходящая под данные выборки выглядит следующим образом:

$$f(x)=\frac{1}{2.87531\sqrt{2\pi}}e^{-\frac{(x-12.3393)^2}{2\cdot 2.87531^2}} \quad (2)$$

8. **Доверительные интервалы для математического ожидания**.
   Вычисляется с использованием функций получения аргумента по заданному значению вероятности. Получим доверительный интервал с достоверностью 0.95.

Вычислим полуширину интервала

```
delta = qnorm(0.95)*sdx/sqrt(N)
```

```
0.4729464
```

Левая граница

```
ml = mx - delta
```

Правая граница интервала

```
mr = mx + delta
```

9. Коэффициент корреляции.

```
  r = cor(x,y)
```

```
0.1895535
```

Проверим гипотезу о значимости коэффициента корреляции. В используемом критерии нулевая гипотеза формулируется так: коэффициент корреляции не значим, то есть в генеральной совокупности коэффициент корреляции равен нулю. Для проверки гипотезы используем данные выборок:

```
cor.test(x,y)
```

```
Pearson's product-moment correlation

data:  x and y
t = 1.9111, df = 98, p-value = 0.05891
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.007130066  0.372117067
sample estimates:
      cor
0.1895535
```

P-value (0.05891) получилось близким к уровню значимости. При возможности, в таких случаях следует увеличивать объём выборки чтобы иметь больше оснований для отклонения нулевой гипотезы. Однако, не располагая возможностью увеличить выборки, будем принимать решение исходя из полученных результатов: p-value > α, значит оснований для отклонения нулевой гипотезы нет. Значению коэффициента корреляции нельзя доверять, ибо оно получено именно таким случайно.

Построим уравнение линейной регрессии. Будем искать зависимость с.в. Y от с.в. X вида $y = kx + b$:

```
lm(y ~ x)
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     8.7892       0.1719
```

Уравнение линейной регрессии:

$$y = 0.1719\,x + 8.7892$$

10. Диаграмма рассеивания и линия регрессии

Приведём два способа построения диаграммы рассеивания и линии регрессии. Первый способ заключается в использовании стандартной библиотеки языка R:

Построим диаграмму рассеивания:

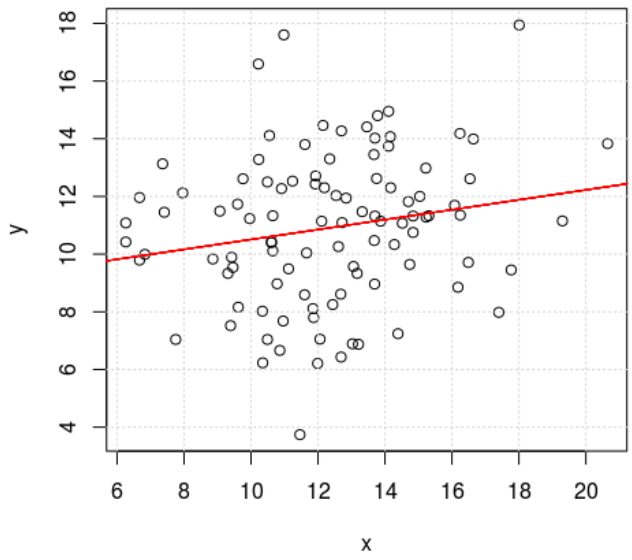
```
plot(x,y)
```

Добавим координатную сетку:

grid()

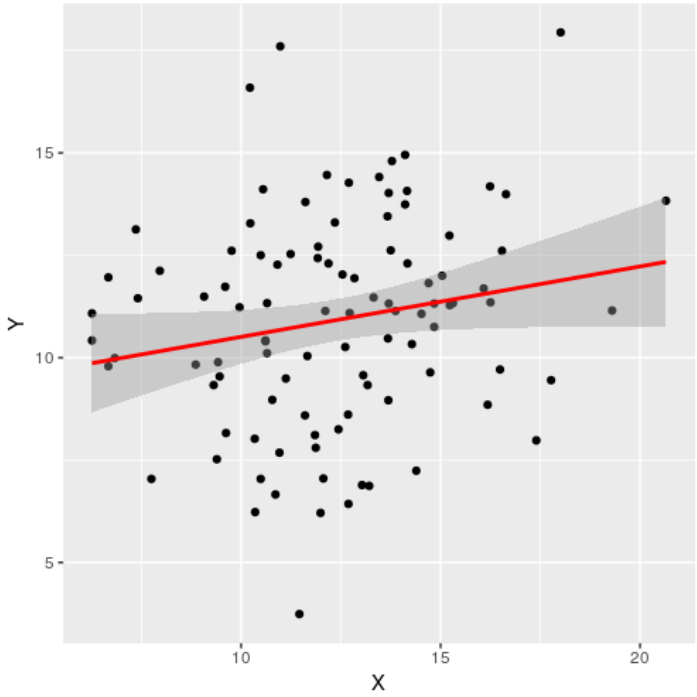
Изобразим линию регрессии (красным)

```
abline(lm(y ~ x), col='red')
```



Второй способ заключается в использовании библиотеки ggplot2, которая стоит более эстетичные графики и в дополнении позволяет изобразить семейство допустимых линий регрессии (на основании 95% доверительно интервала для коэффициента корреляции). Эта область обозначена серым, по обе стороны от линии регрессии.

```
ggplot(mydata, aes(x=mydata$X, y=mydata$Y))+geom_point()+
stat_smooth(method = "lm", col = "red") +
xlab('X')+ylab('Y')
```



Литература и ссылки

1. Гмурман В.Е. Теория вероятностей и математическая статистика
2. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга http://www.ievbras.ru/ecostat/Kiril/R/Mastitsky%20and%20Shitikov%202014.pdf
3. Ю. Е. Воскобойников, Е.И. Тимошенко Математическая статистика (с примерами в Excel) учебное пособие. http://window.edu.ru/resource/305/63305/files/stat_excel.pdf
4. R Studio, сайт разработчика https://www.rstudio.com/products/rstudio/download/
5. Computing in R: Frequency Tables http://courses.wccnet.edu/~palay/math160r/r_groups.htm
6. Установка R и R Studio https://www.youtube.com/watch?v=njtZ0yV8Nwo

Используемые данные

	X	Y									
0	15.04	12.0	31	16.25	11.35	62	9.31	9.33	93	12.44	8.25
1	16.24	14.18	32	10.34	8.02	63	14.52	11.07	94	12.68	8.61
2	7.96	12.12	33	14.74	9.64	64	10.65	11.33	95	20.65	13.83
3	17.4	7.98	34	9.07	11.49	65	9.76	12.61	96	10.55	14.11
4	6.83	9.99	35	8.86	9.83	66	9.39	7.52	97	11.87	7.8
5	9.42	9.89	36	6.26	10.42	67	10.61	10.4	98	12.11	11.14
6	14.11	14.95	37	13.87	11.14	68	13.67	13.45	99	12.19	12.3
7	19.3	11.15	38	6.26	11.08	69	10.91	12.27			
8	15.31	11.33	39	10.78	8.97	70	11.24	12.53			
9	11.12	9.49	40	13.21	6.87	71	12.61	10.26			
10	11.61	13.8	41	12.7	14.27	72	6.67	9.79			
11	14.84	11.32	42	10.86	6.66	73	6.67	11.96			
12	14.11	13.74	43	12.54	12.03	74	10.49	12.5			
13	11.66	10.04	44	13.06	9.57	75	10.98	17.6			
14	12.84	11.94	45	12.69	6.43	76	13.32	11.47			
15	12.06	7.05	46	14.39	7.24	77	9.62	8.16			
16	12.15	14.46	47	11.99	6.21	78	10.23	13.28			
17	18.01	17.94	48	16.08	11.69	79	13.69	8.96			
18	16.49	9.71	49	9.6	11.73	80	15.22	12.98			
19	13.03	6.89	50	14.7	11.82	81	10.49	7.04			
20	11.93	12.71	51	16.54	12.61	82	10.96	7.68			
21	16.64	13.99	52	7.41	11.45	83	11.46	3.74			
22	14.16	14.07	53	10.22	16.59	84	13.17	9.33			
23	11.6	8.59	54	7.36	13.13	85	13.68	10.47			
24	12.35	13.3	55	12.72	11.09	86	9.46	9.54			
25	13.46	14.41	56	13.78	14.8	87	15.23	11.28			
26	13.7	11.32	57	13.75	12.62	88	7.75	7.04			
27	10.65	10.11	58	11.85	8.11	89	9.96	11.23			
28	14.28	10.33	59	11.92	12.43	90	17.77	9.45			
29	16.18	8.85	60	10.61	10.42	91	10.35	6.23			
30	14.84	10.75	61	13.7	14.02	92	14.17	12.3			