КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет географии и экологии

Использование языка R для статистической обработки данных

Учебно-методическое пособие

КАЗАНЬ - 2007

Составители:
доктор биологических наук, доцент А.А.Савельев,
старший преподаватель С.С.Мухарамова,
старший преподаватель А.Г.Пилюгин

Учебно-методическое пособие предназначено для студентов естественных факультетов, изучающих курс «Теория вероятности и математическая статистика». Даются основные понятия языка R, разбираются примеры использования операторов, методы анализа и обработки предназначенной для выполнения практических заданий по курсам «ГЕОСТАТИСТИКА» и «Теория вероятности и математическая статистика». Печатается по решению учебно-методической комиссии факультета географии и экологии.

Введение

Цель этого пособия состоит в том, чтобы описать способы построения основных статистических моделей и использование стандартных статистических тестов для проведения статистичесого анализа с помощью системы R.

R – статистическая система анализа, созданная Россом Ихакой и Робертом Гентлеманом (1996, *J.Comput. Граф. Stat.*, 5: 299-314). R является и языком и программным обеспечением; его наиболее замечательный особенности:

- эффективная обработка данных и простые средства для сохранения результатов,
- набор операторов для обработки массивов, матриц, и других сложных конструкций,
- большая, последовательная, интегрированная коллекция инструментальных средств для проведения статистического анализа,
- многочисленные графические средства,
- простой и эффективный язык программирования, который включает много возможностей.

Язык **R** - рассматривают как диалект языка **S** созданный AT&T Бэлл Лаборатории. **S** доступен как программное обеспечение S-PLUS коммерческой системы MathSoft (см.http://www.splus.mathsoft.com для получения дополнительной информации). Есть существенные различия в концепции **R** и **S** (те, кто хочет знать больше об этом может читать статью, написанную Gentleman и Ihaka (1996) или R-FAQ (часто задаваемые вопросы) (http://cran.r-project.org/doc/FAQ/R-FAQ.html).

R доступен в нескольких формах: исходный текст программ, написанный на C (и некоторые подпрограммы в Fortran77) и в откомпилированном виде.

R – язык со многими функциями для выполнения статистического анализа и графического отображения результатов, которые визуализируется сразу же в собственном окне и могут быть сохранены в различных форматах (например, jpg, png, bmp, eps, или wmf под Windows, ps, bmp, pictex под Unix).

Результаты статистического анализа могут быть отображены на экране. Некоторые промежуточные результаты (*P*-values, коэффициент регрессии и т.п.) могут быть сохранены в файле и использоваться для последующего анализа.

R – язык, позволяющий пользователю использовать операторы циклов, чтобы последовательно анализировать несколько наборов данных. Также возможно объединить в отдельную программу различные статистические функции, для проведения более сложного анализа.

1. Статистические исследования в R

Широкий диапазон функций доступен в **base** пакете. Существует также большое количество других пакетов, которые увеличивают потенциальные возможности **R**. Они располагаются отдельно и должны быть загружены в память. Исчерпывающий список таких пакетов, вместе с их описаниями, можно найти в Интернете по адресу:
URL: http://cran.rproject.org/src/contrib/PACKAGES.html.

В пакете ***base*** пакете есть основные статистические модели:

lm линейные модели;

glm обобщенные линейные модели;

aov, anova дисперсионный анализ;

В пакете **stats** пакете есть дополнительные статистические модели, в первую очередь *glm* – обобщенная линейная модель, позволяющая, например, моделировать логистические или логарифмические зависимости. Пакеты **nlme**, **mgcv** позволяют строить нелинейные модели.

Например, пусть даны два вектора x и y с пятью наблюдениями каждый, и необходимо найти модель *линейной регрессии y на x*:

```
> x <-1:5
> y <-rnorm (5)
> lm (y~x)
Call:
lm (formula = y ~ x)
Coefficients(Коэффициенты:):
Intercept         x
0.2252    0.1809
```

Результат подгонки линейной модели **lm (y~x)** может быть скопирован в объект:

```
> mymodel<-lm (y~x)
```

Некоторые функции R позволяют пользователю отобразить полученной модели, среди которых **summary()** − выводит определенный набор статистических параметров (статистические тесты...), **residuals()** − отображает остатки регрессии, **predict()** −- прогнозные значения, и **coef()** − отображает вектор с оценками параметра.

```
> summary(mymodel)
 lm(formula = y ~ x)
Residuals:
1 2 3 4 5
1.0070 -1.0711 -0.2299 -0.3550 0.6490
```

Coefficients (Коэффициенты:):

Estimate Std. Error t value Pr(>|t|)(Оценка Станд. отклон t value P value (> |t |))

```
  (Intercept) 0.2252 1.0062 0.224 0.837
  x                          0.1809 0.3034 0.596 0.593
  Residual standard error(СКО): 0.9594 on 3 degrees of
freedom
```

Multiple R-Squared (Коэффициент детерминации R2)**: 0.1059, Adjusted R-squared** (Скорректированный R2)**: -0.1921**

```
  F-statistic: 0.3555 on 1 and 3 degrees of freedom(на 1
и 3 степени свободы), p-value: 0.593
```

```
> residuals (mymodel)
        1          2          3          4         5
1.0070047 -1.0710587 -0.2299374 -0.3549681 0.6489594
> predict(mymodel)
        1         2         3         4         5
0.4061329 0.5870257 0.7679186 0.9488115 1.1297044

> coef(mymodel)
(Intercept)           x
0.2252400   0.1808929
```

Эти значения можно использовать в последующих вычислениях, например:

```
> a <-coef (mymodel) [1]
> b <-coef (mymodel) [2]
> newdata <-c (11, 13, 18)
> a+ b*newdata
[1] 2.215062 2.576847 3.481312
```

Чтобы отобразить элементы результата анализа, можно использовать функцию **names ()**; фактически, эта функция может использоваться с любым объектом в **R**.

```
> names (mymodel)
[1] "coef" "residuals" "effects" "rank"
[5] "fitted.values" "assign" "qr" "df.residual"
[9] "xlevels" "call" "terms" "model"
> names(summary (mymodel))
[1] "call" "terms" "residuals" "coef"
[5] "sigma" "df" "r.squared" "adj.r.squared"
[9] "fstatistic" "cov.unscaled"
```

Сами элементы могут быть извлечены следующим способом:

```
> summary(mymodel) ["r.squared"]
$r.squared
[1] 0.09504547
```

Формулы – ключевые элементы в статистических анализах в **R.** Использование их одинаково для всех функций. Формула имеет форму *y ~ модель*, где *y* – проанализированный ответ и *модель* – набор условий, для которых некоторые параметры должны быть оценены.

Эти условия отделены арифметическими символами, но они имеют здесь особенное значение.

a+b – совокупный эффект a и b

a:b – интерактивный эффект между a и b

*a*b* – идентично a+b+a:b

poly(a,n) – полином от a степени n

^n – включает все взаимодействия до уровня n, то есть (a+b+c) ^2 идентичен a+b+c+a:b+a:c+b:c

b%in%a – эффекты b вложены в a (идентичный a+a:b)

a-b – удаляет эффект b, например: **(a+b+c) ^n-a:b** идентичен *a+b+c+a:c+b:c*, **y~x-1** выполняет регресс через начало координат (идентификатор. Для y~x+0, или 0+y~x)

Отсюда видно, что арифметические операторы в **R** имеют в формуле различные значения. Например, формула **y~x1+x2** определяет модель *y = b1x1 +b2x2 + a.* Для включения арифметических операций в формулу, можно использовать функцию **i()**: формула **y~I(x1+x2)** определяет модель *y = b (x1 + x2) + a*.

2. Статистические оценки

2.1. Выборочное среднее

Описание

Выборочной средней $x_{ср}$ называют среднее арифметическое значение признака выборочной совокупности. Если все значения x_1, x_2, …,x_n признака выборки объема n различны, то

$$\overline{x} = (x_1 + x_2 + ... + x_n)/n$$

Описание функции

mean(x, ...)

Параметры

x Вектор, матрица или data.frame.

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> mean(x)
7.12 (значение среднего)
```

2.2. Выборочная дисперсия и СКО

Описание

Характеризует рассеяние наблюдаемых значений количественного признака выборки вокруг своего среднего значения $x_{ср}$.

Выборочной дисперсией называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $x_{ср}$. Если в качестве оценки генеральной дисперсии принять выборочную дисперсию, то эта оценка будет приводить к систематическим ошибкам, давая заниженное значение генеральной дисперсии, поэтому используют исправленную дисперсию. Исправленная (несмещенная) выборочная дисперсия вычисляется по формуле

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Выборочное среднеквадратическое отклонение - СКО $S_x = \sqrt{S_x^2}$

Описание функции

var(x, y = NULL, na.rm = FALSE)

sd(x, na.rm = FALSE)

Параметры

x	Вектор, матрица или data.frame
y	NULL(по умолчанию) или вектор, матрица или data.frame такой же размерности, что и x
na.rm	Удалить данные, значения котрорых отсуствуют

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.9,6.5)
> var(x, y, na.rm = FALSE)
> sd(x, na.rm = FALSE)

[1] 2.9      (значение дисперсии)
[1] 2.459065 (значение СКО)
```

2.3. Медиана и мода

Описание

Медианой m_e называют варианту, которая делит вариационный ряд (упорядоченный по возрастанию) на две части, равные по числу вариант.

Если число вариант нечетно, т.е. $n = 2r+1$, то $m_e=x_{r+1}$;

при четном $n = 2r$, то me=(xr+xr+1)/2

Модой M_0 называют варианту, которая имеет наибольшую частоту. В R для этого можно использовать построение таблицы вариант

Описание функции

median(x, na.rm=FALSE)

Параметры

x	Вектор, матрица или data.frame
na.rm	Удалить отсутствующие данные?

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9,9.6,5.7,8.9,9.6)
> median(x, na.rm=FALSE)
8.9

> sort(unique(x))
3.6 7.8 9.6 5.7 8.9  уникальные значений вариант по
возрастанию

> x.t<-table(x)
> x.t
3.6 5.7 7.8 8.9 9.6  варианты (по возрастанию)
  1   2   1   2   3  сколько раз каждая встречалась

> order(x.t)
1 3 2 4 5            номера вариант (по частоте
встречаемости)

> sort(unique(x))[which.max(x.t)]
9.6                  мода
```

3. Проверка статистических гипотез

Приведем список некоторых основных пакетов, содержащих стандартные статистические тесты (многие критерии находятся в пакете stats, который загружается автоматически):

ctest - классические тесты (Фишера, "Стьюдента", Пирсона, Бартлетта, Колмогорова- Смирнова...)

eda - методы, используемые в “Разведочном анализе данных”

lqs - регрессия и оценка ковариации

modreg – современные методы построения регрессионных моделей: сглаживание и локальные регрессии

mva - многомерный анализ

nls – нелинейные модели регрессии

splines – сплайны

stepfun - эмпирические функции распределения

ts - исследования временных рядов

Для загрузки пакета используется фукция: **library()** с именем соответствующего пакета:

```
> library(eda)
```

3.1. Критерий X^2 Пирсона (Проверка гипотезы о нормальном распределении генеральной совокупности).

Описание

Критерий $\chi 2$ используется для анализа таблиц сопряженности признаков и сравнения законов распределения непрерывных случайных величин. Анализируются номинальные или приведенные к номинальной шкале данные, представленные в виде таблицы сопряженности признаков. Для непрерывных случайных величин используется принадлежность значений заданным интервалам, выбираемых таким образом, чтобы в каждом из них было не менее 5-7 значений (интервалы с меньшим числом значений объединяются). Простейшим выбором является равный шаг интервалов, равный $\lambda = \frac{x_{\max} - x_{\min}}{k}$, $k = 1 + 3.32 \cdot \lg(n)$ или $k = 5 \cdot \lg(n)$.

Вычисление критериальной статистики производится по формуле:

$$\chi^2 = \sum_i \frac{(n_i - n_i^{'})^2}{n_i^{'}}$$

где n_i- эмпирические частоты, $n_i^{'}$- теоретические частоты попадания элементов выборки в группы (заданные интервалы).

Число степеней свободы находят по формуле:

$f = k - 1 - r$

где k-число групп выборки, r-число параметров предполагаемого распределения, которые оценены по данным выборки.

Если предполагаемое распределение – нормальное, то по выборке оценивают два параметра (математическое ожидание и дисперсию), поэтому r=2 и $f = k - 3$.Одной из функций, осуществляющей проверку данного критерия в R является **`chisq.test()`**.

Описание функции

`chisq.test (x, y = NULL, p = rep(1/length (x), length (x)))`

Параметры

x	вектор или матрица.
y	вектор; игнорируемый, если x - матрица.
p	вектор теоретических вероятностей той же длины, что x.

Примечание

Если *x* − матрица с одной строкой или столбцом, или если *x* − вектор, и *y* не дан, *x* − одномерная таблица сопряженности признаков. В этом случае, проверенная гипотеза - равняются ли вероятности совокупности тем, что в *p*, или все равны, если *p* не дается. Если *x* - матрица с двумя строками (или столбцами), содержащими неотрицательные целые числа, то она рассматривается как таблица сопряженности признаков. Если *x* и *y* – два

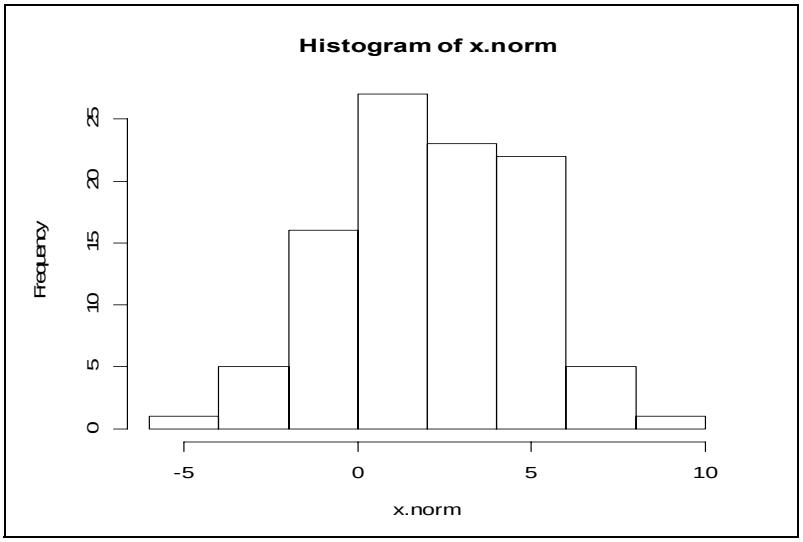
вектора, содержащих факторы (номинальные или ординальные значения), то по ним строится таблица сопряженности.

Критические значения (квантили) находятся с использованием функции **qchisq(p,df)** или по таблице χ2-распределения [2,стр.329].

Пример

Сгенерируем случайную выборку из нормального распределения, и проверим ее нормальность.

```
N<-100  # объем выборки
x.norm<-rnorm(N,mean=2,sd=2.5) # задаем среднее и СКО
```



Вычисляем квантили выборки с шагом 10% (по 10 элементов в интервале)

```
> x.norm.q <- quantile(x.norm,probs=seq(0,1,0.1))
> round(x.norm.q,2)
    0%   10%   20%   30%   40%   50%   60%   70%   80%
90%  100%
 -4.12 -1.51 -0.14  0.59  1.52  2.15  2.70  3.89  4.51
5.22  8.15
> summary(x.norm)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.4490  0.6675  2.0330  2.0670  3.1050  6.5830
```

Выбираем интервалы:

```
> k<-6    # число интервалов
> x.q <- c(-10, -1.0, 0.5, 2.0, 3.5, 5.0,12.0)
```

Вычисляем фактические частоты

```
> x.norm.hist<-hist(x.norm,breaks=x.q,plot=FALSE)
> x.norm.hist$counts
12 15 22 18 18 15
```

Вычисляем (по выборке) теоретические вероятности для каждого интервала

```
> x.q[1]<-(-Inf);x.q[k+1]<-(+Inf)#«раздвигаем» границы
до бесконечности
```

```
  > x.norm.p.theor<-
pnorm(x.q,mean=mean(x.norm),sd=sd(x.norm))
  > x.norm.p.theor<-(x.norm.p.theor[2:(k+1)]-
x.norm.p.theor[1:k])
  > round(x.norm.p.theor,2)
  0.12 0.15 0.21 0.22 0.16 0.14
```

Сравниваем фактические и теоретические частоты

```
  > chisq.test(x.norm.hist$counts,p=x.norm.p.theor)
          Chi-squared test for given probabilities
  data:  x.norm.hist$counts
  X-squared = 0.9691, df = 5, p-value = 0.965
```

Поскольку для проверки нулевой гипотезы H_0 о нормальности распределения генеральной совокупности в нашем случае используется правосторонний критерий, а уровень значимости (*p*-value) равен 0.965 (96.5%), то нужно допустить разрешить вероятность ошибки, равную 96.5%, чтобы считать выборку не принадлежащей нормальному распределению. Следовательно, гипотеза о нормальности принимается.

Приведем пример проверки случайности сопряжения признаков. Пусть первый признак является полом (закодированным символами «М» и «Ж»), а второй – наличием близорукости (закодированной числами 0, 1).

```
  > x<-
c("М","Ж","Ж","М","Ж","Ж","М","Ж","М","Ж","Ж","М","М")
  > y<-c( 0,  0,  1,  0,  0,  1,  0,  0,  1,  1,  0,  1,
0)
  > tbl<-table(x,y)
  > tbl
     y
  x  0 1
  Ж  4 3
  М  4 2
```

Оцениваем по выборке маргинальные (безусловные) вероятности для *x* и *y*:

```
  > p.x<-c(sum(tbl[1,])/sum(tbl),sum(tbl[2,])/sum(tbl))
  > p.y<-c(sum(tbl[,1])/sum(tbl),sum(tbl[,2])/sum(tbl))
```

или, что тоже самое (годится для таблиц любой размерности),

```
  > p.x<-apply(tbl,1,sum)/sum(tbl) # 1 - суммируем по
строкам
  > p.y<-apply(tbl,2,sum)/sum(tbl) # 2 - суммируем по
столбцам
```

Если сопряженности признаков нет (гипотеза H_0), то наблюдаемые относительные частоты должны совпадать с произведениями маргинальных вероятностей:

```
  >p.theor<-c(p.x[1]*p.y[1],p.x[2]*p.y[1],
p.x[1]*p.y[2],p.x[2]*p.y[2])
```

или, что тоже самое (годится для таблиц любой размерности),

```
> p.theor<- p.x %*% t(p.y)
> chisq.test(as.vector(t),p=p.theor)
        Chi-squared test for given probabilities

data:  as.vector(t)
X-squared = 0.1238, df = 3, p-value = 0.9888
```

Поскольку *p*-value близко к единице, то нулевая гипотеза принимается. Отметим, что поскольку в таблице сопряженности tbl есть ячейки с числами, меньшими 5 (в нашем случае – все ячейки такие), то аппроксимация хи-квадрат может быть неправильной, а полученный результат - неверным.

3.2. Критерий Фишера (Сравнение дисперсий двух нормальных генеральных совокупностей).

Описание

В качестве критериальной статистики для проверки гипотезы о равенстве генеральных дисперсий нормально распределенных генеральных совокупностей используют отношение выборочных дисперсий, т. е. случайную величину:

$$F = \frac{S_x^2}{S_y^2}$$

где S_x^2, S_y^2 – исправленные выборочные дисперсии для выборок объемом n_1 и n_2 соответственно**.** В качестве нулевой гипотезы H_0 формулируется гипотеза о равенстве генеральных дисперсий.

Величина F при условии справедливости нулевой гипотезы имеет распределение Фишера-Снедекора со степенями свободы $f_1 = n_1 - 1, f_2 = n_2 - 1$. Одной из функций, осуществляющей проверку данного критерия в R является **var.test()**

Описание функции

var.test(x,y,alternative = c("two.sided", "less","greater"), conf.level = 0.95, ...)

Параметры

x,y	вектор или объекты линейной модели (например, класса lm).
conf.level	доверительная вероятность
Alternative	альтернативная гипотеза. Может быть одна из **"two.sided"** (по умолчанию)-двустороняя критическая область, **"greater"** -правостороняя критическая область или **"less"-левостороняя** критическая область.

Пример

```
> x<-c(3.5, 3.6, 7.8, 9.6, 5.7, 8.9, 6.3)
> y<-c(1.0, 2.7, 8.9, 6.5, 8.9, 6.5,12.5,10.2, 1.2)
```

n_1=7, S_x^2=5.86; n_2=9, S_x^2=16.75

```
> var.test(x, y, alternative = c("two.sided"),
conf.level = 0.95)

        F test to compare two variances
data:  x and y
F = 0.3498, num df = 6, denom df = 8, p-value = 0.2174
alternative hypothesis: true ratio of variances is not
equal to 1
95 percent confidence interval:
 0.07519108 1.95855792
```

Результат теста:

F = 0.3498 (значение F статистики), число степеней свободы для x - 6, для y - 8

p-value = 0.2174, т.е. уровень ошибки, при котором можно отвергнуть гипотезу о равенстве дисперсий, равен 21.74%

95% доверительный интервал: (0.075, 1.959) - полученное нами значение *F*-статистики в него попадает, следовательно гипотеза о равенстве дисперсий принимается на 5% уровне значимости.

3.3. Критерий Стьюдента (Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых неизвестны и одинаковы).

Описание

Если предположить, что неизвестные генеральные дисперсии равны между собой, то для решения этой задачи можно применить критерий Стьюдента. Т.е. нужно, пользуясь критерием Фишера, предварительно проверить гипотезу о равенстве генеральных дисперсий. В случае независимых выборок в качестве критериальной статистики для проверки гипотезы принимают случайную величину:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)\cdot S_x^2 + (m-1)\cdot S_y^2}} \cdot \sqrt{\frac{n \cdot m \cdot (n+m-2)}{n+m}},$$

где $\bar{X}, \bar{Y}$ - выборочные средние, S_x^2, S_y^2 – выборочные дисперсии, n, m – объемы выборки и $f = n + m - 2$ - число степеней свободы для распределения критериальной статистики (если дисперсии не равны, то критерий остается применимым, но требует коррекции приведенной формулы и числа степеней свободы – необходимость такой коррекции указывается при вызове функции). Если выборки зависимые (парная выборка), то проверяется гипотеза о равенстве матожидания нулю для новой случайной величины $z_i = x_i - y_i$, также

имеющей нормальное распределение. В этом случае используется критериальная статистика Стьюдента

$$t = \frac{\overline{Z}}{\sqrt{S_z^2}} \cdot \sqrt{n}$$

Одной из функций, осуществляющей проверку данного критерия в R является **t.test()**

Описание функции

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

Параметры

x	Числовой вектор значений.
y	Числовой вектор значений (используется для парного теста, см. ниже).
paired	Признак парного теста: проверяется гипотеза для x-y, поэтому вектор y должен присутствовать и соответствовать по длине вектору x.
alternative	Символьная строка, определяющая альтернативную гипотезу, должна быть одна из **"two.sided"** (по умолчанию)-двустороняя критическая область, **"greater"** -правостороняя критическая область или **"less"-левостороняя** критическая область.
var.equal	Логическая переменная, указывающая на равенство дисперсий
conf.level	Доверительная вероятность.

Примечание

По умолчанию var.equal=FALSE (дисперсии предполагаются неравными), в этом случае для вычислений используется оценка Велча (Welch).

Пример (используем те же выборки, что и для сравнения дисперсий)

```
> x<-c(3.5, 3.6, 7.8, 9.6, 5.7, 8.9, 6.3)
> y<-c(1.0, 2.7, 8.9, 6.5, 8.9, 6.5,12.5,10.2, 1.2)
> t.test(x,y,alternative=c("two.sided"),var.equal=TRUE,conf.level=0.95)

        Two Sample t-test

data:  x and y
t = -0.0018, df = 14, p-value = 0.9986
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 -3.760075  3.753726
sample estimates:
mean of x mean of y
 6.485714  6.488889
```

Значения

t = -0.0018 (значение критериальной статистики), число степеней свободы равно 14.

p-value = 0.9986, т.е. чтобы отвергнуть гипотезу, нужно допустить 99.86% ошибки.

95% доверительный интервал (-3.760075, 3.753726). Поскольку наше значение в него попадает, то нулевая гипотеза принимается на 5% уровне значимости.

Если равенство дисперсий не проверялось, или гипотеза о равенстве не принимается, то вызов критерия выглядит так:

```
> 
t.test(x,y,alternative=c("two.sided"),var.equal=FALSE,
conf.level=0.95)

        Welch Two Sample t-test

data:  x and y
t = -0.0019, df = 13.242, p-value = 0.9985
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -3.545004  3.538655
sample estimates:
mean of x mean of y
 6.485714  6.488889
```

Число степеней свободы теперь 13.242 вместо 14, и границы доверительного интервала несколько изменились.

Приведем пример для проверки нулевой гипотезы о равенстве матожиданий для парной выборки (выборки должны быть одинаковой длина):

```
> x<-c(3.5, 3.6, 7.8, 9.6, 5.7, 8.9, 6.3, 8.3, 4.5)
> y<-c(1.0, 2.7, 8.9, 6.5, 8.9, 6.5,12.5,10.2, 1.2)
> t.test(x,y,alternative=c("two.sided"),var.equal=TRUE,
paired=TRUE)

        Paired t-test

data:  x and y
t = -0.0202, df = 8, p-value = 0.9843
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
```

```
 -2.554391  2.509946
sample estimates:
mean of the differences
            -0.02222222
```

Значения:

t = -0.0202 (значение критериальной статистики), число степеней свободы равно 8.

p-value = 0.9943, т.е. чтобы отвергнуть гипотезу, нужно допустить 99.86% ошибки.

95% доверительный интервал (-2.554391, 2.509946). Поскольку наше значение в него попадает, то нулевая гипотеза принимается на 5% уровне значимости.

3.4. Критерии Бартлетта и Кохрана (Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам).

Описание

Критерий Барлетта используется для проверки гипотезу об однородности (равенстве) нескольких дисперсий, полученных по выборкам разного объема. Для этого рассчитывают среднюю арифметическую исправленных дисперсий, взвешенную по числам степеней свободы:

$$\overline{S}^2 = \frac{1}{f}\sum_{i=1}^{k} f_i \cdot S_i^2 ,$$

где $f_i = n_i - 1$ число степеней свободы для *i*-й выборки объема n_i, S_i^2 - выборочная дисперсия дисперсия *i*-й выборки, $f = \sum_i f_i$ - общее число степеней свободы, и k - число выборок.

В качестве критериальной статистики для проверки гипотезы об однородности дисперсий используют критерий Бартлетта:

$$B = V \big/ C ,$$

имеющая распределение χ^2, где

$$V = 2.303 \cdot [f \cdot \lg \overline{S}^2 - \sum_{i=1}^{l} f_i \cdot \lg S_i^2] ,$$

$$C = 1 + \frac{1}{3(k-1)} \cdot \left[\left(\sum_{i=1}^{k} \frac{1}{f_i} \right) - \frac{1}{f} \right]$$

Одной из функций, осуществляющей проверку данного критерия в R является **`bartlett.test()`**

Описание функции

`bartlett.test (x, g...)`

Параметры

x числовой вектор значений, или список числовых значений векторов , или объекты линейной модели (класса "lm").

g вектор или фактор, дающий группу для соответствующих элементов x. Игнорируемый, если x - список.

Примечание

Если *x* - список, его элементы будут взяты как выборки и *g* игнорируется, и можно просто использовать **bartlett.test(x)**. Если выборки еще не содержатся в списке, используют **bartlett.test(list (x...))**.

Критические значения (правосторонний критерий) находятся по таблице распределения χ^2 с $k-1$ степенями свободы [2,стр.329] или используют функцию вычисления квантилей распределения Хи-квадрат **qchisq(p,df)**.

Пример

```
> x1<-c(3.5, 3.6, 7.8, 9.6, 5.7, 8.9, 6.3)
> x2<-c(1.0, 2.7, 8.9, 6.5, 8.9, 6.5,12.5,10.2, 1.2)
> x3<-c(3.6,7.8,9.6,5.7,8.9)
> x4<-c(2.7,8.9,6.5,8.9)
```

Дисперсии выборок равны соотвественно 5.86, 16.75, 6.05 и 8.57, нулевая гипотеза H_0 –дисперсии всех генеральных совокупностей равны между собой, уровень значимости – 5%.

```
> bartlett.test(list(x1,x2,x3,x4))
        Bartlett test of homogeneity of variances

data:  list(x1, x2, x3, x4)
Bartlett's K-squared = 2.2368, df = 3, p-value = 0.5247
```

Значения

Bartlett's K-squared = 2.2368 (значение критериальной статистики теста Бартлетта), число степеней свободы 3,

p-value = 0.5247, т.е. отвергнуть гипотезу H_0 можно только при допустимой ошибке в 52.47%. Следовательно, гипотеза об однородности дисперсий принимается на 5% уровне значимости.

Если объем выборок (примерно) одинаковый, то может использоваться тест экстремальных значений Кохрана (Cochran) из пакета **outliers**, реализуемый функцией **cochran.test().**

Описание функции

cochran.test(object,data)

Параметры

object числовой вектор, содержащий значения дисперсий для каждой выборки S_i^2

data числовой вектор, содержащий объем каждой выборки

В в качестве критериальной статистики используется

$$C = \frac{\max_i \left\{ S_i^2 \right\}}{\sum_i S_i^2}$$

а для вычисления критических значений – функция вычисления квантилей распределения Кохрана qcochran(p, n, k) из того же пакета, где *p*-доверительная вероятность, *n* – объем одной выборки (если объемы различаются, то берется среднее значение), *k* – число выборок.

Пример

Используем в примере те же выборки, что и в предыдущем случае, объем выборок 7, 9, 5 и 4 элементов соответственно. Нулевая гипотеза H_0 – дисперсии всех генеральных совокупностей равны между собой, уровень значимости – 5%.

```
> cochran.test(object=
c(var(x1),var(x2),var(x3),var(x4)), data=c(7,9,5,4))
        Cochran test for outlying variance

data:  c(var(x1), var(x2), var(x3), var(x4))
C = 0.4499, df = 6.25, k = 4.00, p-value = 0.3083
alternative hypothesis: Group 2 has outlying variance
```

Значения

Cochran C = 0.4499 (значение критериальной статистики теста Кохрана), число степеней свободы (средний объем выборки) 6.25, число групп 4, *p*-value 0.3083. Альтернативная гипотеза – дисперсия второй выборки значительно больше остальных (является «выбросом»). Поскольку *p*-value = 0.3083, то отвергнуть гипотезу H_0 можно только при допустимой ошибке в 30.83%. Следовательно, гипотеза об однородности дисперсий принимается на 5% уровне значимости.

4. Дисперсионный анализ

Описание

Данный метод основан на разложении общей дисперсии численного признака на составляющие ее компоненты (отсюда и название метода ANalysis Of VAriance или ANOVA), сравнивая которые с друг другом посредством F-критерия Фишера можно определить, какую долю (по отношению к совокупности случайных причин) общей вариации признака обуславливает действие на него известных величин (факторов).

Метод основан на сравнении межгрупповой и внутригрупповой изменчивости признака. Каждую группу образуют значения признака при фиксированных значениях (уровнях) известных факторов, поэтому единственным источником дисперсии (изменчивости) внутри каждой группы является суммарное воздействие совокупности случайных причин. Общая

модель дисперсионного анализа (на примере двух факторов) выглядит следующим образом:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

где μ - среднее значение признака, α_i - влияние первого фактора на *i*-м уровне (при *i*-м значении), β_j - влияние второго фактора на *j*-м уровне (при *j*-м значении), $(\alpha\beta)_{ij}$ - влияние взаимодействия факторов на указанных уровнях (если факторы не независимы), и ε_{ijk} - суммарное влияние на признак случайных факторов, имеющее нормальное распределение с нулевым матожиданием и дисперсией $\sigma^2_{ош}$. Предполагается, что ε_{ijk} не зависит от уровней факторов, поэтому общая дисперсия признака (точнее, общая сумма квадратов $SS_{общ} = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}...)^2$, где точки в индексе среднего показывают, по каким из них проводилось осреднение, может быть разложена на компоненты (частные суммы), соотвествущие вкладу в общую дисперсию каждой составляющей.

В простейшем случае, если имеется всего один фактор, такое разложение представляется в виде таблицы дисперсионного анализа:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Источник дисперсии	SS сумма квадратов	Степеней свободы	Средний квадрат	F статистика
Фактор (межгрупповая)	$SS_{факт} = \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}..)^2$	$k-1$	$S^2_{факт} = \frac{SS_{факт}}{k-1}$	$F = \frac{S^2_{факт}}{S^2_{ош}}$
Случайная составляющая	$SS_{ош} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i.)^2$	$\sum_{i=1}^{k} (n_i - 1)$	$S^2_{ош} = \frac{SS_{ош}}{\sum_{i=1}^{k} (n_i - 1)}$	
Общая	$SS_{общ} = \sum_i \sum_j (y_{ij} - \bar{y}..)^2$	$N-1$	$S^2_{общ} = \frac{SS_{общ}}{N-1}$	

где k - число групп, n_i - число наблюдений в *i*-ой группе, $N = \sum_i n_i$ - общее число наблюдений.

Для проведения однофакторного дисперсионного анализа в R используется линейная модель, в которой единственной независимой переменной выступает этот фактор.

Описание функции

anova(object)

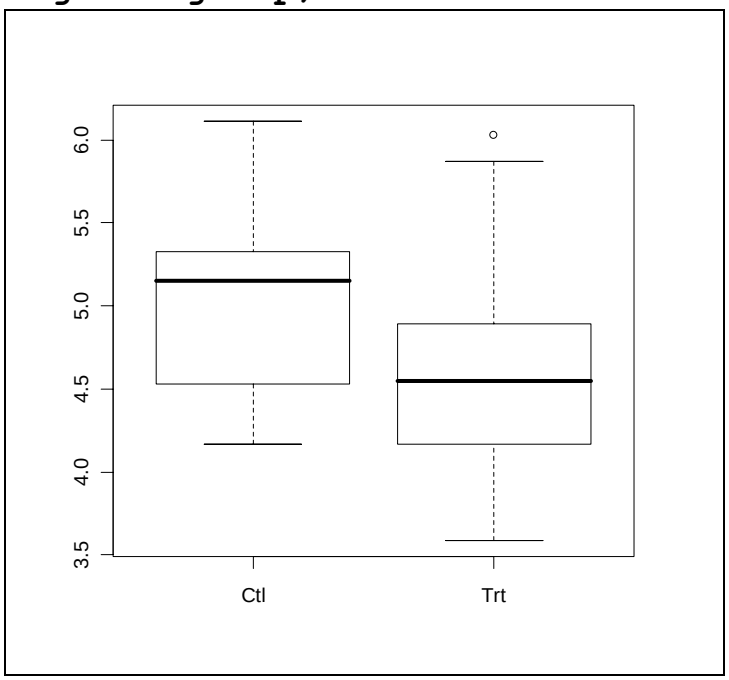
Параметры

`object` Объект класса lm, glm .

В примере ниже формируется набор данных, включающий 20 значений признака (вектор weight, по 10 значений для каждого из двух уровней фактора и вектор значений фактора group), строится модель зависимости признака от фактора, и выполняется дисперсионный анализ. В случае однофакторной модели таблица дисперсионного анализа совпадает с результатом дисперсионного анализа модели – сравнением остаточной и модельной дисперсий (последняя является суммарным вкладом всех факторов).

Пример

```
> ctl <-
c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
> trt <-
c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
> group <- gl(2,10,20, labels=c("Ctl","Trt"))
> weight <- c(ctl, trt)
> boxplot(weight ~ group)
```




```
> lm.D <- lm(weight ~ group)
> summary(lm.D)
Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4938  0.0685  0.2462  1.3690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8465     0.1557  31.124   <2e-16 ***
group1       -0.1855     0.1557  -1.191    0.249

Residual standard error: 0.6964 on 18 degrees of
freedom
Multiple R-Squared: 0.07308,    Adjusted R-squared:
0.02158
```

```
  F-statistic: 1.419 on 1 and 18 DF,  p-value: 0.249

  > anova(lm.D)
  Analysis of Variance Table

  Response: weight
            Df Sum Sq Mean Sq F value Pr(>F)
  group      1 0.6882  0.6882  1.4191  0.249  факторная
дисперсия
  Residuals 18 8.7293  0.4850
внутригрупповая дисперсия
```

Значения

Как информация по модели **summary()**, так и вызов **anova()** выдают для однофакторной модели те же результаты: значение F-статистики 1.419 при 1 и 18 степенях свободы, нулевая гипотеза H0 гласит, что фактор group не влияет на признак weight, уровень значимости (*p*-value) – 0.249, что означает, что гипотеза может быть отвергнута только если допустить 24.9% ошибки. Таким образом, гипотеза об остуствии вличния фактора принимается на 5% уровне значимости.

5. Корреляционный анализ

5.1. Коэффициент корреляции и проверка гипотезы о его значимости.

Описание

Выборочный коэффициент корреляции Пирсона определяется равенством:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{n \cdot S_x \cdot S_y}$$

x_i, y_i - варианты зависимой (парной) выборки $\langle x, y \rangle$, т.е. выборочные значения признаков X и Y, выбираемых парами, n-объем выборки, S_x и S_y - выборочные средние квадратические отклонения, $\overline{x}$ и $\overline{y}$ - выборочные средние.

Известно, что если величины Y и X независимы, то коэффициент корреляции генеральных совокупностей ρ_{xy}=0; если ρ_{xy}=$\pm$1, то Y и X связаны линейной функциональной зависимостью, т.е. коэффициент корреляции ρ_{xy} (а значит и выборочный коэфициент корреляции r_{xy}, являющийся оценкой для ρ_{xy}) измеряют силу линейной связи между Y и X. В системе R для вычисления выборочного коэфициента корреляции используется функция cor().

Если основная гипотеза H_0 говорит о независимости Y и X (т.е. коэфициент корреляции равен нулю), то в качестве критериальной статистики используется статистика

$$t_p = \sqrt{n-2} \cdot \frac{r_{xy}}{\sqrt{1-r_{xy}^2}},$$

имеющая распределение Стьюдента с $n-2$ степенями свободы. Если объем выборки большой ($n \sim 100$), то можно считать, что статистика имеет нормальное распределение.

Проверка гипотезы о значимости показателя ранговой корреляции осуществляется функцией **cor.test()**

Описание функции

cor(x, y, method = c("pearson", "kendall", "spearman"))

Параметры

X Вектор, матрица или data.frame

Y Второй вектор (или NULL, если первый аргумент – матрица или фрейм данных)

method Вычисляемый коэфициент корреляции (по умолчанию – pearson)

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.8,6.4)
> cor(x, y)
0.4668
```

5.2. Показатель ранговой корреляции

В качестве критериев оценки назависимости могут применяться и другие коэффициенты корреляции, например показатель ранговой корреляции Спирмена, позволяющий оценить нелинейную, но монотонную зависимость: в этом случае вычисляется кореляция не самих значений, а их рангов (порядковых номеров при упорядочении). Другим ранговым критерием является τ-критерий Кендалла.

Проверка по нескольким критериям может быть использована для приблизительной оценки оценки вида зависимости: если ранговая корреляция большая (статистически значимая), а линейная – маленькая (статистически не значимая), то зависимость нелинейная; если обе корреляции большие, то зависимость линейная; если обе корреляции маленькие, что либо зависимости нет, либо она немонотонная.

Если основная гипотеза гласит, что коэфициент корреляции равен не нулю, а некоторому отличному от нуля числу, то в качестве критериальной статистики используется *z*-преобразование Фишера:

$$z(r_{xy}) = \frac{1}{2}\ln\frac{1+r_{xy}}{1-r_{xy}}$$

Эта величина распределена примерно нормально для всех значений коэффициента корреляции генеральных совокупностей, ее матожидание равно $z(\rho_{xy})$, а дисперсия $1/(n-3)$, где n - объем выборки. Поэтому границы доверительного интервала для $z(\rho_{xy})$ находят с использованием квантилей нормального распределения; получить границы для ρ_{xy} можно обратным преобразованием.

Описание функции

cor.test(x, y,alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), conf.level = 0.95, ...)

Параметры

x, y	Числовые вектора x и y одинаковой длины .
alternative	Выбирает альтернативную гипотезу одну из "two.sided" (по умолчанию)-двустороняя критическая область, "greater" -правостороняя критическая область или "less"-левостороняя критическая область.
method	Выбирает какой коэфициент корреляции используется в тесте. Один из"pearson", "kendall", или "spearman".
conf.level	Доверительная вероятность

Примечание

Для проверки нулевой гипотезы H0 о равенстве показателя корреляции нулю необходимо в *alternative* выбрать *"two.sided".*

Критическое значение находят по таблице критических точек распределения Стьюдента с числом степей свободы $f = n-2$ (в R используется функция вычисления квантилей распределения Стьюдента **qt(p,df)**).

Пример

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.8,6.4)
> cor.test(x,y ,alternative = c("two.sided"),
method = c("pearson"))

        Pearson's product-moment correlation

t = 0.9142, df = 3, p-value = 0.428
95 percent confidence interval: -0.7063858  0.9555364
```

```
  sample estimates: cor = 0.4667999

  > cor.test(x,y,alternative= c("two.sided"),
method=c("spearman"))

          Spearman's rank correlation rho

  S = 16, p-value = 0.7833
  sample estimates: rho = 0.2
```

Значение

Для обычной линейной корреляции (Пирсона) мы получили выборочное значений 0.4668, значение *t*- статистики 0.9142 при 3 степенях свободы, и *p*-value равное 0.428. Это означает, что отвергнуть нулевую гипотезу можно только при допущении ошибки в 42.8%. 95% доверительный интервал равен (-0.7063858, 0.9555364) и поскольку он содержит ноль, то нулевая гипотеза принимается на 5% уровне значимости.

Для ранговой корреляции Спирмена выборочное значений коэффициента корреляции еще меньше (0.2), а *p*-value еще больше (0.7833). Поэтому и по ранговому критерию мы отвергаем наличие связи между X и Y.

6. Линейная регрессия

Описание

Линейная зависимость между переменными описывается уравнением общего вида $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n + \varepsilon$ где y - зависимая переменная, $\alpha_0, \alpha_1, ..., \alpha_n$ - неизвестные константы, $x_1, x_2, ..., x_n$ - известные (независимые) переменные, и ε - нормально распределенная случайная величина с нулевым матожиданием и дисперсией σ^2_{err}. Задачей построения линейной среднеквадратической модели регрессионной зависимости переменной y от независимых переменных является получение оценки параметров $\alpha_0, \alpha_1, ..., \alpha_n$ и оценка адекватности построенной модели вида

$\hat{y} = a_0 + a_1 x_1 + ... + a_n x_n$

где $a_0, a_1, ..., a_n$ - оценки параметров $\alpha_0, \alpha_1, ..., \alpha_n$.

Рассмотрим простейший случай одной независимой переменной:

$\hat{y} = a + bx$ В этом уравнении модели линейной регрессии a - свободный член, а параметр b определяет наклон линии регрессии по отношению к осями координат. Параметры a и b определяются методом наименьших квадратов, который приводит к формуле:

$$b = r_{xy} \frac{S_y}{S_x},\ a = \overline{y} - b\overline{x},$$

где

$\overline{y}, \overline{x}$ - выборочные средние арифметические;

S_x, S_y - выборочные средние квадратичные отклонения;

r_{xy} - выборочный коэффициент корреляции.

Для построения линейной модели регрессии используется функция **lm(formula=f)**, которая в простейшем случае содержит только формулу от переменных (векторов, содержащих элементы парной выборки); запись *y~x* означает, что строится модель зависимости *y* от *x*.

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.8,6.4)
> p.lm<-lm(formula=x~y)
> summary(p.lm)

Residuals:
      1       2       3       4       5
-1.7151 -0.3409  2.5529 -2.3954  1.8985

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0845     3.5050   1.165    0.328
y             0.4558     0.4985   0.914    0.428

Residual standard error: 2.511 on 3 degrees of freedom
Multiple R-Squared: 0.2179,     Adjusted R-squared: -
0.0428
F-statistic: 0.8358 on 1 and 3 DF,  p-value: 0.428
```

Команда **summary()** выдает полную информацию о построенной модели:

значения остатков (residuals - разность модельных и истинных значений переменной *y*). Если объем выборки большой, то печатается оценка распределения остатков (квартили).

коэфициенты модели и оценку их значимости по критерию Стьюдента (в нашем случае все коэфициенты не значимы, поскольку все вероятности (0.328 и 0.428) больше 0.05 - т.е. нельзя считать, что существует линейная зависимость между x и *y*).

Оценку значимости зависимости по критерию Фишера и квадрат коэфициента корреляции (R-squared), который показывает долю дисперсии *y*, объясненной с использованием модели (исправленное значение для R2 равно 0, статистика Фишера *F*=0.8358, уровень значимости критерия Фишера 42.8%, т.е. зависимость отсуствует).

Для визуализации построенной модели можно использовать вспомогательные функции:

Описание функций

abline(a, b, untf = FALSE, ...)

abline(h=, untf = FALSE, ...)

abline(v=, untf = FALSE, ...)

Параметры

a,b Параметры в линейном уравнении

untf Если TRUE, то рисует линию в преобразованных координатах

h,v Y и X значения для горизонтальной и вертикальной линии соответственно

plot(x, y, xlim=range(x),ylim=range(y),type="p", main, xlab, ylab, ...)

Параметры

X,Y Координаты точек x и y.

xlim, ylim Значения для осей x и y.

Type Тип графика(“ p” для точек)

Main **Название графика**

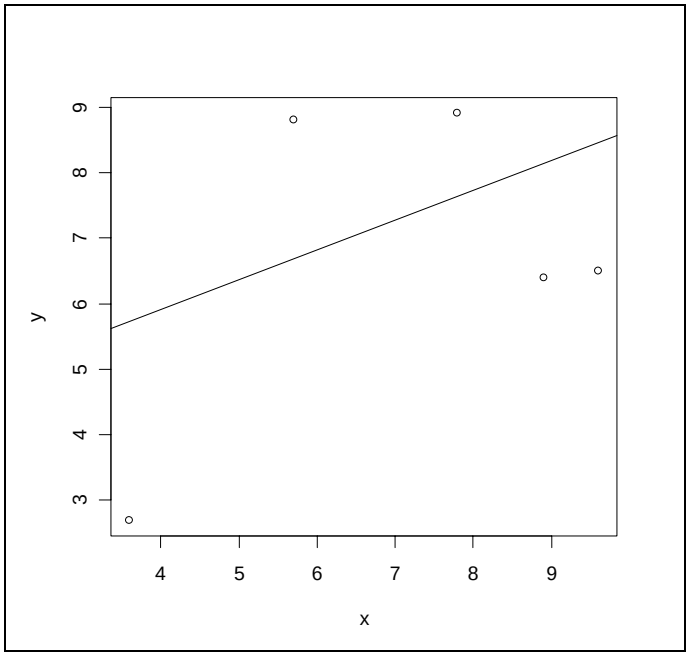
Xlab,ylab Название осей.

Функция **abline()** строит прямую по найденным a и b.

Функция **plot()** строит экспериментальные точки.

Пример

```
plot(x,y)
abline(lm(x~y))
```



Список литературы.

1. **Гмурман В.Е.** Теория вероятностей и математическая статистика/ В.Е.Гмурман.М.:Высшая школа, 2000.-479с.
2. **Лакин Г.Ф.** Биометрия/ Г.Ф. Лакин. М: Высшая школа, 1990.-352с.
3. Теория вероятностей и математическая статистика/ Под редакцией В.А. Колемаева. М: Высшая школа, 1991.-400с.
4. **Гайдышев И.** Анализ и обработка данных: специальный справочник - СПб: Питер, 2001.-752с.
5. **Бейли Н**. Статистические методы в биологии/Н.Бейли.М.:Мир,1963.-272с.
6. **Гланц С.** Медико-биологическая статистика/ С. Гланц. М: Практика, 1999.-449с.
7. **А.А.Савельев, С.С.Мухарамова, А.Г.Пилюгин, Е.А.Алексеева** Основные понятия языка R / А.А.Савельев, С.С.Мухарамова, А.Г.Пилюгин, Е.А.Алексеева К ффф 2007.-28с