Основы программирования в R

*Лабораторная работа 2.*

Распределения в R

В базовой версии R имеются функции для работы с целым рядом распространенных законов распределения вероятностей. В зависимости от назначения, имена этих функций начинаются с одной из следующих четырех букв:

* d (от "*density*", плотность): функции плотности вероятности ("функция распределения масс" для дискретных величин);
* p (от "*probability*", *вероятность*): кумулятивные функции распределения вероятностей;
* q (от "*quantile*", квантиль): функции для нахождения квантилей того или иного распределения;
* r (от "*random*", случайный): функции для генерации случайных чисел в соответствии с параметрами того или иного закона распределения вероятностей.

В частности, в базовой версии R реализованы следующие законы распределения вероятностей:

* [Распределение хи-квадрат](http://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D1%85%D0%B8-%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82) (chisq)
* [Распределение Фишера](http://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%A4%D0%B8%D1%88%D0%B5%D1%80%D0%B0) (f)
* [Нормальное распределение](http://ru.wikipedia.org/wiki/%D0%9D%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5) (norm)
* [Распределение Стьюдента](http://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%A1%D1%82%D1%8C%D1%8E%D0%B4%D0%B5%D0%BD%D1%82%D0%B0) (t)

Например, если мы ходим найти значение плотности стандартного нормального распределения в точке X= -1, то необходимо написать

dnorm(-1)

Получим

[**dnorm**](http://inside-r.org/r-doc/stats/dnorm)(-1)

[1] 0.2419707

Значение функции стандартного нормального распределения в точке X=-1 можно будет найти , как

[**pnorm**](http://inside-r.org/r-doc/stats/pnorm)(-1)

Получим

[**pnorm**](http://inside-r.org/r-doc/stats/pnorm)(-1)

[1] 0.1586553

Для вычисления [квантилей](http://ru.wikipedia.org/wiki/%D0%9A%D0%B2%D0%B0%D0%BD%D1%82%D0%B8%D0%BB%D1%8C) нормального распределения служит функция qnorm().

Например, значение квантиля уровня 0,75 стандартного нормального распределения можно будет найти qnorm(0.75).

[**qnorm**](http://inside-r.org/r-doc/stats/qnorm)(0.75)

[1] 0.6744898

Например, следующим образом вычислить нижний и верхний квартили стандартного нормального распределения:  
[**qnorm**](http://inside-r.org/r-doc/stats/qnorm)(p = [**c**](http://inside-r.org/r-doc/base/c)(0.25, 0.75))

[1] -0.6744898 0.6744898

Для функции dnorm() обязательным параметром является только x, для pnorm() — q, для qnorm() — p и для rnorm() — n. В этом случае используется стандартное нормальное распределение

Например, команда **dnorm(x, mean =** 5, **sd = 10)** **б**удет выдавать квантили нормального распределения со средним 5 и стандартным отклонением 10.

Задание 1:

Найти квантили:

* 1. Стандартного нормального распределения уровня 0,3, 0,5, 0,8, 0368, 0,9, 0,95, 0,975, 0,98.
  2. Стьюдента с 10 степенями свободы уровня 0,9, 0,95, 0,975
  3. Хи-квадрат с 15 степенями свободы уровня 0,05, 0,1, 0,9, 0,95
  4. Фишера с 3 и 5 степенями свободы уровня 0,9 и 0,95.
  5. Проверьте найденные значения по стат. таблицам.

Доверительные интервалы в R

Для того, чтобы строить доверительные интервалы, нам понадобится библиотека DescTools (но она не единственная в R библиотека, которая позволяет это делать). Установим ее:

install.packages("DescTools")

То, что R что-то пишет красным в процессе установки — это нормально. Это не ошибки, это просто задокументированные попытки R обратиться к сайту и загрузить необходимые компоненты библиотеки.

Если установить библиотеку не получается, посмотрите, что пишет R. Если что-то про отсутствие прав доступа, то следует запустить RStudio от имени администратора (кликнуть правой клавишей на значок RStudio и выбрать *Запуск от имени администратора*, на Mac OS сложнее, напишите мне). Если что-то похожее на Couldn't resolve host name, проверьте подключение к интернету.

Обратимся к библиотеке (без этого R не будет знать, откуда ему брать функции для доверительных интервалов, и будет выдавать ошибки в стиле «не знаю такой функции»):

**library**(DescTools)

Теперь у нас есть всё для работы, и давайте рассмотрим такую задачу.

***Доверительный интервал для доли***

Исследователь случайным образом опросил 100 человек и выяснил, что среди них 30 человек выступают в поддержку смертной казни. Давайте построим 90%-ный доверительный интервал для доли людей, которые поддерживают смертную казнь. Воспользуемся функцией BinomCI(). *CI* — сокращение от *confidence interval* (доверительный интервал), *Binom* — указание на биномиальную случайную величину, которая описывает число успехов — число интересующих нас людей.

*# n = 100, p = 0.3*

ci90 <- BinomCI(30, 100, conf.level = 0.90)

**Пояснения к коду:** на первом месте указано число успехов (число интересующих нас объектов), на втором - объем выборки. В conf.level мы указываем уровень доверия. Результат мы не выводим на экран, а сохраняем в переменную ci90, так что, прогоняя строчку кода, на экране мы ничего не увидим.

Посмотрим на результат:

ci90

## est lwr.ci upr.ci

## [1,] 0.3 0.230705 0.3798321

Что нам выдал R? Вектор (список) из трех чисел. Первое число - это выборочная доля (30/100). Второе — это нижняя граница доверительного интервала (lwr.ci — от *lower*), а третье — верхняя граница (upr.ci - от *upper*).

Проинтерпретируем полученный доверительный интервал. С 90%-ной уверенностью можно утверждать, что истинная доля людей, которые поддерживают смертную казнь, лежит в интервале от 0.23 до 0.38. Если независимо проводить аналогичные исследования на выборках одинакового размера много раз, в 90% случаев доверительный интервал накроет истинное значение доли людей, поддерживающих смертную казнь.

Если мы будем проводить аналогичное исследование на выборках одного и того же размера много раз, независимо друг от друга, в 90% случаев истинное значение доли сторонников смертной казни будет лежать в пределах от 0.23 до 0.38 (в предположении о том, что стандартная ошибка не изменяется от выборки к выборке).

Аналогичным образом построим 95%-ный и 99%-ный доверительные интервалы (интерпретировать не будем, интерпретация аналогична).

ci95 <- BinomCI(30, 100, conf.level = 0.95)

ci95

## est lwr.ci upr.ci

## [1,] 0.3 0.2189489 0.3958485

ci99 <- BinomCI(30, 100, conf.level = 0.99)

ci99

## est lwr.ci upr.ci

## [1,] 0.3 0.1974607 0.4274276

Видно, что выборочная доля у нас везде одинакова, но доверительные интервалы разные. Давайте сравним их длины. Чтобы определить длину доверительного интервала, нужно из верхней границы вычесть нижнюю. В нашем случае это соответствует тому, что из третьего числа в результате нужно вычесть второе.

*# длина 90%-ного интервала*

l90 <- ci90[3] - ci90[2]

l90

## [1] 0.1491272

*# длина 95%-ного интервала*

l95 <- ci95[3] - ci95[2]

l95

## [1] 0.1768997

*# длина 99%-ного интервала*

l99 <- ci99[3] - ci99[2]

l99

## [1] 0.229967

Как мы и обсуждали ранее, с увеличением уровня доверия длина доверительного интервала увеличивается, а точность (сконцентированность) оценки уменьшается: самый длинный интервал соответствует 99%-ному уровню доверия.

Пример.

Давайте теперь построим доверительный интервал для доли, но на реальных данных. Загрузим по ссылке уже знакомую нам базу с данными опроса перед референдумом в Чили в 1988 году: Chile.csv

df <- read.csv(file.choose())

Если вы загружайте файл по ссылке, скачивать его на компьютер необязательно (главное, иметь подключенный интернет).

Удалим из базы df строки с пропущенными значениями (NA) и сохраним изменения:

df <- na.omit(df)

Посмотрим на первые несколько строк в таблице:

head(df)

## X region population sex age education income statusquo vote

## 1 1 N 175000 M 65 P 35000 1.00820 Y

## 2 2 N 175000 M 29 PS 7500 -1.29617 N

## 3 3 N 175000 F 38 P 15000 1.23072 Y

## 4 4 N 175000 F 49 P 35000 -1.03163 N

## 5 5 N 175000 F 23 S 35000 -1.10496 N

## 6 6 N 175000 F 28 P 7500 -1.04685 N

Нас будет интересовать столбец **vote** — то, как респондент планирует голосовать на референдуме. Давайте посмотрим на уникальные значения в этом столбце:

table(df$vote)

##

## A N U Y

## 177 867 551 836

* Y — за статус-кво, за Пиночета у власти (*Yes*);
* N — против статус-кво, против Пиночета у власти (*No*);
* A — планирует не участовать (*Abstained*);
* U — еще не решил(а) (*Undecided*).

Построим доверительный интервал для доли людей, которые собирались голосовать за Пиночета. Исключим из рассмотрения респондентов, которые еще не определились/не собираются голосовать, оставим только тех, кто за и против. Сохраним их количество:

yes = 836

no = 867

Теперь построим 95%-ный доверительный интервал.

BinomCI(yes, yes + no, conf.level = 0.95)

## est lwr.ci upr.ci

## [1,] 0.4908984 0.4672024 0.5146353

Выборочная доля (p^) людей «за» равна 0.49. Проинтерпретируем полученный доверительный интервал.

*С уверенностью 95% можно утверждать, что истинная доля людей, которые планировали голосовать за Пиночета у власти, лежит в интервале от 0.47 до 0.52. Если проводить аналогичные исследования на выборках одинакового размера, в 95% случаев истинная доля людей, которые планировали голосовать за Пиночета, будет лежать в интервале от 0.47 до 0.52 при условии постоянства стандартной ошибки.*

Теперь посмотрим на доверительный интервал для среднего. Построим 95%-ный доверительный интервал для среднего возраста людей, которые планировали голосовать за Пиночета.

Сначала выберем из базы соответствующие строки:

voted <- df[df$vote == "Y", ]

**Пояснения к коду:** нас интересуют люди, которые голосовали “за”, следовательно, из базы df нужно взять строки, где df$vote равен "Y". Для проверки условий в программировании используется двойной знак =, чтобы отличать операцию сравнения от присваивания значений. Поэтому мы выбираем строки, соответствующие этим людям, и все столбцы (после запятой в квадратных скобках пустота — значит, берем все).

Теперь построим 95%-ный доверительный интервал для среднего – функция MeanCI():

MeanCI(voted$age)

## mean lwr.ci upr.ci

## 40.20335 39.17485 41.23185

**Обратите внимание:** мы не выставили уровень доверия conf.level = 0.95, но R не выдал нам ошибку. Наш код сработал, так как по умолчанию R в любом случае строит 95%-ный доверительный интервал.

Интерпретация полученного доверительного интервала:

*С 95%-ной уверенностью мы можем утверждать, что истинное значение среднего возраста людей, собиравшихся голосовать за Пиночета, лежит в интервале от 39 до 41 года.*

**Бонус:** так как данные по этому плебисциту в Чили можно считать историческими, у нас есть возможность проверить, накрыл ли посчитанный нами доверительный интервал для доли сторонников Пиночета значение, которое получилось по итогам плебисцита. Давайте допустим, что доля поддержавших оставление Пиночета у власти на референдуме и есть истинное значение доли в генеральной совокупности (конечно, это не совсем так — ведь не все население Чили принимало участие в голосовании).

Наш 90%-ный доверительный интервал был [0.47, 0.51], а 95%-ный был [0.47, 0.52]. По результатам референдума известно, что примерно 44% жителей проголосовали за оставление Пиночета у власти еще на 8 лет. Мы чуть-чуть промахнулись: наши доверительные интервалы попали в те 10% и 5% доверительных интервалов, которые не накрывают истинную долю в генеральной совокупности.

***Доверительный интервал для среднего***

Если мы хотим построить доверительный интервал для среднего, нам понадобится функция MeanCI(). Внутри нее достаточно указать показатель, среднее значение которого мы будем использовать для построения доверительного интервала.

Построим 95%-ный доверительный интервал для среднего значения дохода (income)

mci95 <- MeanCI(df$incom)

mci95

Выводы?

Давайте построим 99%-ный интервал для среднего значения дохода (income)

mci99 <- MeanCI(df$incom, conf.level = 0.99)

mci99

Выводы?

Также можно воспользоваться просто формулами и квантилями .

Задание 2:

Имеются следующие данные о баллах за экзамен: 10, 20, 30, 25, 20, 40, 30, 40, 20, 25.

1. Найти выборочные характеристики
2. Построить 95% ди для среднего балла, если данные имеют нормальное распределение с дисперсией 16.
3. Построить 95% ди для среднего балла, если данные имеют нормальное распределение.
4. Построить 95% ди для дисперсии балла, если данные имеют нормальное распределение.
5. Построить 95% ди для доли студентов с баллом больше, чем 25.

Задание 3:

Для выполнения задания Вам необходимо загрузить [базу данных](https://raw.githubusercontent.com/agconti/kaggle-titanic/master/data/train.csv) из файла Данные Титаник.xls, содержащую информацию по пассажирам “Титаника”.

Переменные, содержащиеся в файле:

* Pclass — класс пассажира (1 — высший, 2 — средний, 3 — низший);
* Sex — пол;
* Age — возраст;
* SibSp — количество братьев, сестер, сводных братьев, сводных сестер, супругов на борту титаника;
* Parch — количество родителей, детей (в том числе приемных) на борту титаника;
* Fare — плата за проезд;
* Embarked — порт посадки (C — Шербур; Q — Квинстаун; S — Саутгемптон).
* Survived – выжил пассажир или нет (1, если выжил)

В поле Age приводится количество полных лет. Для детей меньше 1 года — дробное. Если возраст не известен точно, то указано примерное значение в формате xx.5.

1. Постройте 95% доверительный интервал для среднего возраста пассажиров.
2. Постройте 95% доверительный интервал для дисперсии возраста пассажиров..
3. Постройте 95% доверительный интервал для среднего возраста пассажиров, если ср.квадратичное отклонение составляет 8.
4. Постройте 95% доверительный интервал для доли людей, которые выжили.