Козлов Даниил Николаевич
daniilkozlov@gmail.com

«Методы статистического анализа, картографирования и имитационнрого моделирования»

~~Компьютерные технологии и статистические методы в экологии и природопользовании~~

(геоэкологи, 98=13+13+82 часов)

1 курс магистратуры, весенний семестр 2016 г.

Преподаватель:

- Даниил Николаевич Козлов: daniilkozlov@landscape.edu.ru
- кафедра физической географии и ландшафтоведения

Информационная поддержка:

- http://landscape.edu.ru
- лекционные и практические материалы, задания, статьи, ссылки

Занятия:

- четверг 4 пара, 14:55-16:30, ауд. 2017

Задания:

- реферат статьи 2014-16 года из каталога ELSEVIER
- тематические задания

Проверка знаний:

- практические (70%), вопросы экзамена (20%), активная работа (10%)

ДОКЛАД ПО СТАТЬЕ

ФГМ

Астапенко Юлия	Spatial distribution of nitrate health risk associated with groundwater use as drinking water in Merida, Mexico (2015)

РАСПИСАНИЕ 2015

11.02	ЦЕЛИ, ЗАДАЧИ И СОДЕРЖАНИЕ КУРСА Экспертные и формальные модели.	ДЗ
18.02 25.02 03.03	**Проблемы статистического анализа данных в экологии и природопользовании**	ДЗ
10.03 17.03	Проблемы цифрового картографического моделирования: геостатистика и индикационное картографирование	ДЗ
24.03 31.03	Проблемы моделирования процессов самоорганизации в экологии и природопользовании	ДЗ
07.04 14.04	Семинар по проблемам (доклады по статьям)	
21.04	Резерв	

R – свободная программная среда вычислений

- язык программирования для статистической обработки данных и работы с графикой,
- свободная программная среда вычислений
- открытый исходный код под лицензией GNU GPL
- Росс Айхэк (англ. Ross Ihaka) и Роберт Джентлмен (англ. Robert Gentleman), статистический факультет Оклендского университета, Новая Зеландия
- R Foundation, Австрия
- http://www.r-project.org/

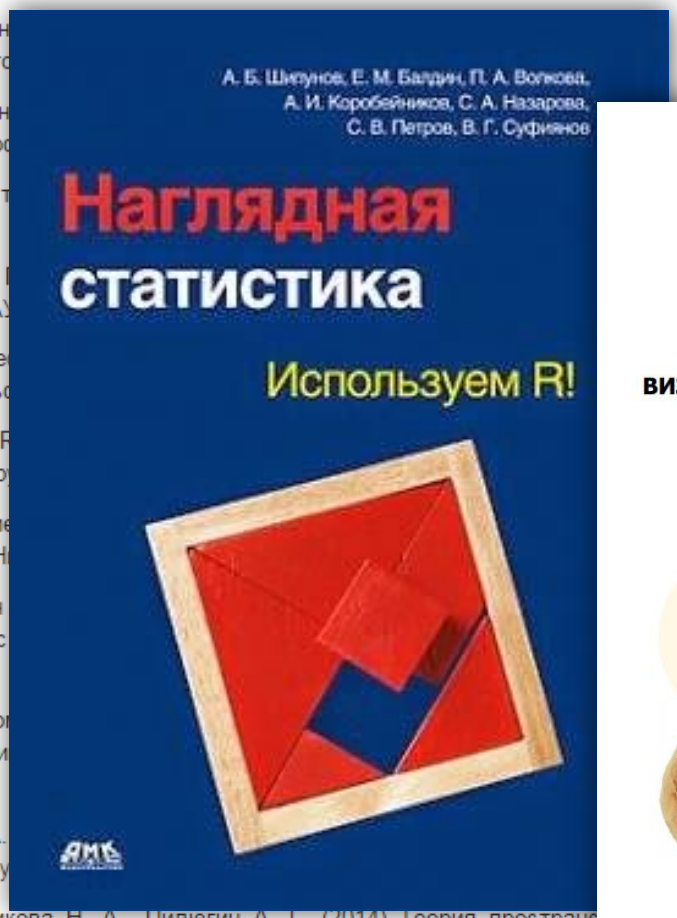
R – свободная программная среда вычислений

Ещё ▾ Следующий блог» Создать блог Войти

R: Анализ и визуализация данных

Книги и учебные пособия

- Мастицкий С. Э., Шитиков В. К. (2014) Статистический анализ и визуализация данных с помощью R. - Электронная книга, 400 с. PDF | Скрипты и данные
- Савельев А. А., Мухарамова С. С., Пилюгин … методическое пособие. Казань: Казанский го…
- Савельев А. А., Мухарамова С. С., Пилюгин … обработки данных. Учебно-методическое пос…
- Волкова П. А., Шипунов А. Б. (2008) Статист… работах. М.: Экопресс, 60 с. PDF
- Буховец А. Г., Москалев П. В., Богатова В. … системе R. Учебное пособие. Воронеж: ВГАУ…
- Зарядов И. С. (2010) Введение в статистиче… запись информации, графика. М.: Издательс…
- Зарядов И. С. (2010) Статистический пакет R… Издательство Российского университета др…
- Зорин А. В., Федоткин М. А. (2010) Введение… методическое пособие. Нижний Новгород: Н…
- Савельев А. А, Мухарамова С. С., Пилюгин … данных в экологии и природопользовании (с… университет, 120 с. PDF
- Шитиков В. К., Розенберг Г. С. (2012) Рандо… статистического анализа данных по биологии… PDF
- Шипунов А. Б., Балдин Е. М., Волкова П. А., … В. Г. (2012) Наглядная статистика. Использу…
- Савельев А. А., Мухарамова С. С., Чижикова Н. А., Пилюгин А. Г. (2014) Теория пространс… точечных процессов в задачах экологии и природопользования (с применением пакета R). - Каза… во Казан. ун-та, 146 с. PDF | Скрипты



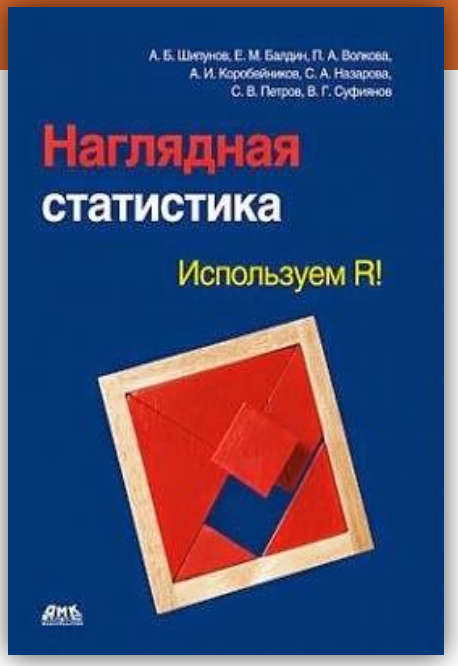
Поиск по блогу:

Поиск

распространенные методы выявления *необычных наблюдений*.

Диагностика линейных

ЛИТЕРАТУРА



А. Б. Шипунов, Е. М. Балдин, П. А. Волкова, А. И. Коробейников,С. А. Назарова, С. В. Петров, В. Г. Суфиянов

Наглядная статистика. Используем R!

http://ashipunov.info/shipunov/school/books/rbook.pdf

ПРОЧИТАТЬ РАЗДЕЛЫ

УСТАНОВИТЬ R и R-STUDIO

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function | Project: (None)

SoilProfiles.R* | pedons | Rplots.R | R_intro.R* | vars | pedproph

Source on Save | Run | Source

```
23  # write.table(data, "data/zhuki_new.txt", quote=FALSE) сохранить таблиц
24  # summary(data) сводная информация по таблице
25  # summary(data$VES) или отдельной колонке
26  dev.new()
27
28  data$var2.dx <- log(data$var2+1000) #создание новой колонки
29  # гистограмма
30  hist(data$var1, breaks=50, main="", col = "red")
31  hist(data$var2, breaks=50, main="", col = "blue")
32  hist(data$var1, breaks=50, main="", col = "lightblue", border = "darkgr
33  hist(data$var2.dx, breaks=50, main="", col = "blue", add=TRUE)
34  plot(density(data$var1, adjust=2), main="", col = "red")
35  plot(density(data$var1, adjust=2), main="", col = "blue", lwd=3, add=TR
36  rug(data$var1)
37
38  head(data)
39
40
41  # проверка на нормальность
42  qqnorm(data$var1)
43
```

32:1 (Top Level) R Script

Console

```
> hist(data$var2, breaks=50, main="", col = "blue", add=TRUE)
> help(hist)
> hist(data$var1, breaks=50, main="", col = "lightblue", border = "pink")
> h
Error: object 'h' not found
> h
Error: object 'h' not found
> hist(data$var1, breaks=50, main="", col = "red")
> hist(data$var1, breaks=50, main="", col = "lightblue", border = "pink")
> hist(data$var2, breaks=50, main="", col = "blue", add=TRUE)
> hist(data$var1, breaks=50, main="", col = "red")
> #hist(data$var1, breaks=50, main="", col = "lightblue", border = "pink")
```

Environment | History

Import Dataset | Clear | List

Global Environment

Data

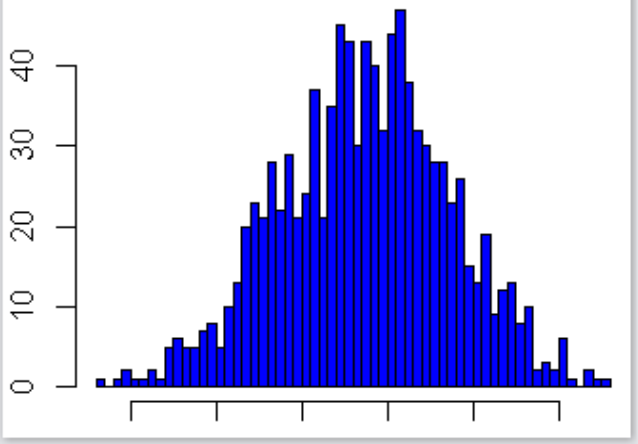
data	1000 obs. of 3 variable...
data.3	100 obs. of 3 variables
datagen	1000 obs. of 2 variable...
pedproph	150 obs. of 1 variables
vars	1000 obs. of 2 variable...

Values

HCLline	num [1:36] 70 90 115 55 8...
pedons	Formal class SoilProfileC...

Files | Plots | Packages | Help | Viewer

Zoom | Export | Clear All



ПРЕДЫДУЩЕЕ ДОМАШНЕЕ ЗАДАНИЕ

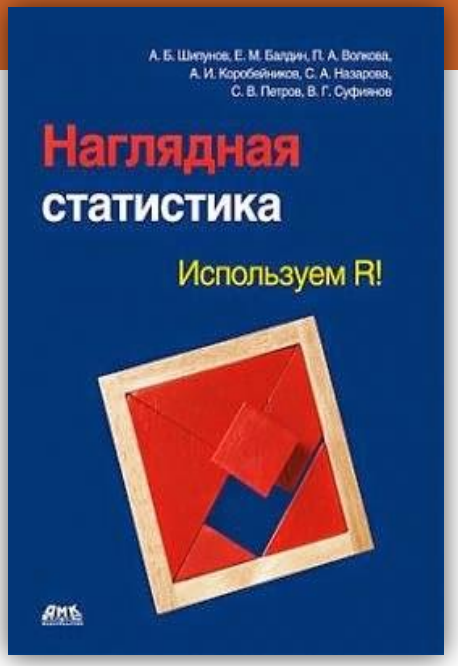
на основе темы магистерской диссертации

1. СФОРМУЛИРОВАТЬ ПРОБЛЕМУ И ЦЕЛЬ ИССЛЕДОВАНИЯ
2. СПРОЕКТИРОВАТЬ ТАБЛИЦУ ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА
3. ОБОСНОВАТЬ ШКАЛУ ИЗМЕРЕНИЙ КАЖДОГО СВОЙСТВА
4. ОБОСНОВАТЬ МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА (сравнение выборок, линейная связь между переменными)
5. ЗАПОЛНИТЬ ТАБЛИЦУ в Excel
6. СОХРАНИТЬ в формате .csv (разделитель “;”)
7. ИМПОРТИРОВАТЬ ТАБЛИЦУ в R

```
11  # загрузка файла данных
12  data <- read.csv2("D:/Dropbox/R/Monte.csv", header = TRUE, sep = ";",
13                    quote = "\"", dec = ".", fill = TRUE,
14                    comment.char = "", stringsAsFactors = FALSE)
15
16  View(data)  # показать таблицу данных
17  str(data) # показать структуру таблицы
18  head(data) # показать первые 6 строк таблицы
19  summary(data) # расчитать мин, среднее, медиану, макс.
```

8. ОФОРМИТЬ ПРЕЗЕНТАЦИЮ и выслать ее преподавателю

ЛИТЕРАТУРА к заданию #1



А. Б. Шипунов, Е. М. Балдин, П. А. Волкова, А. И. Коробейников,С. А. Назарова, С. В. Петров, В. Г. Суфиянов

Наглядная статистика. Используем R!

http://ashipunov.info/shipunov/school/books/rbook.pdf

ПРОЧИТАТЬ РАЗДЕЛЫ

Глава 2. Как обрабатывать данные

- 2.4 Применение, преимущества и недостатки R, стр. 25-26

Глава 4. Великое в малом: одномерные данные

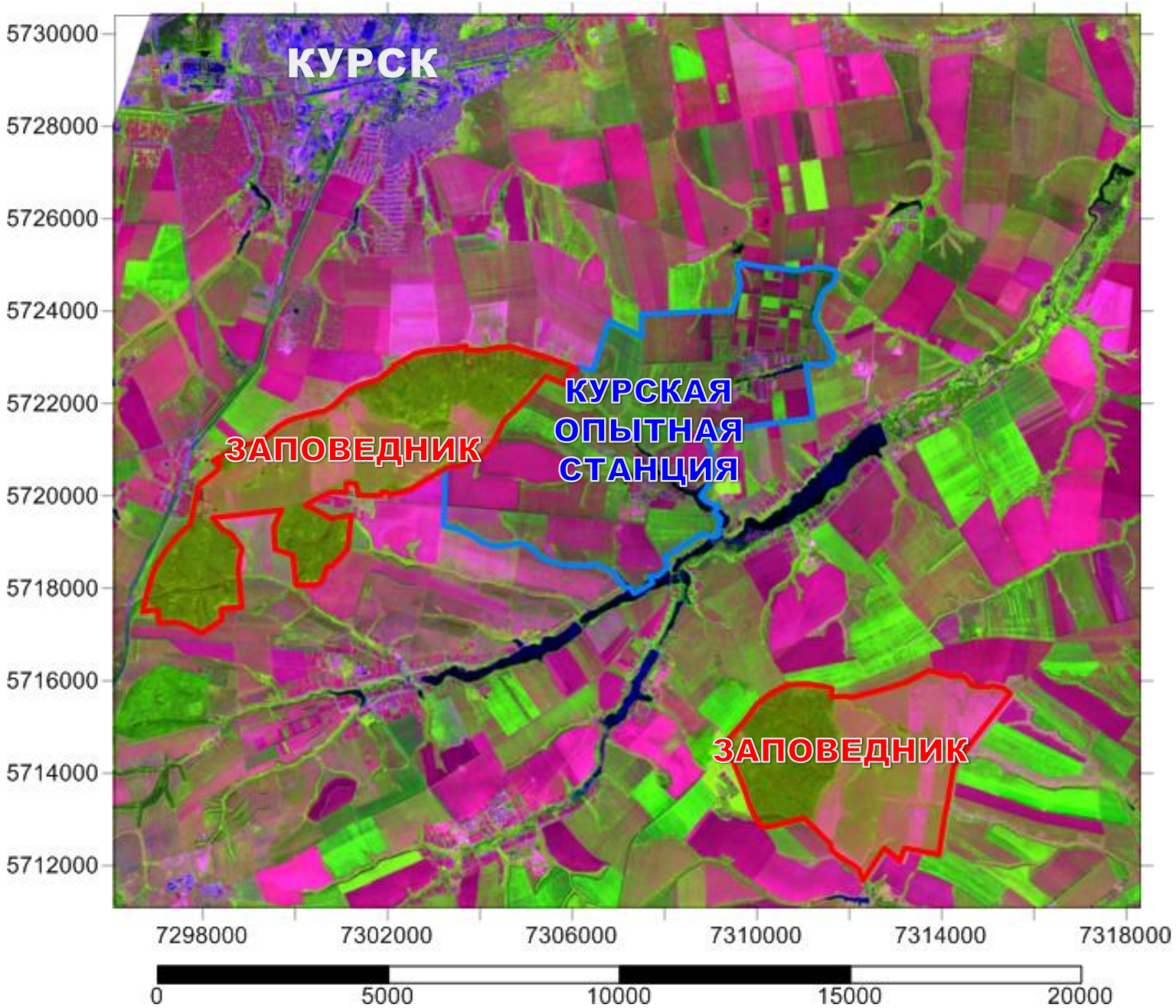
- 4.1. Как оценивать общую тенденцию, стр. 72-82
- 4.3. Одномерные статистические тесты, стр. 83-87

Глава 5. Анализ связей: двумерные данные

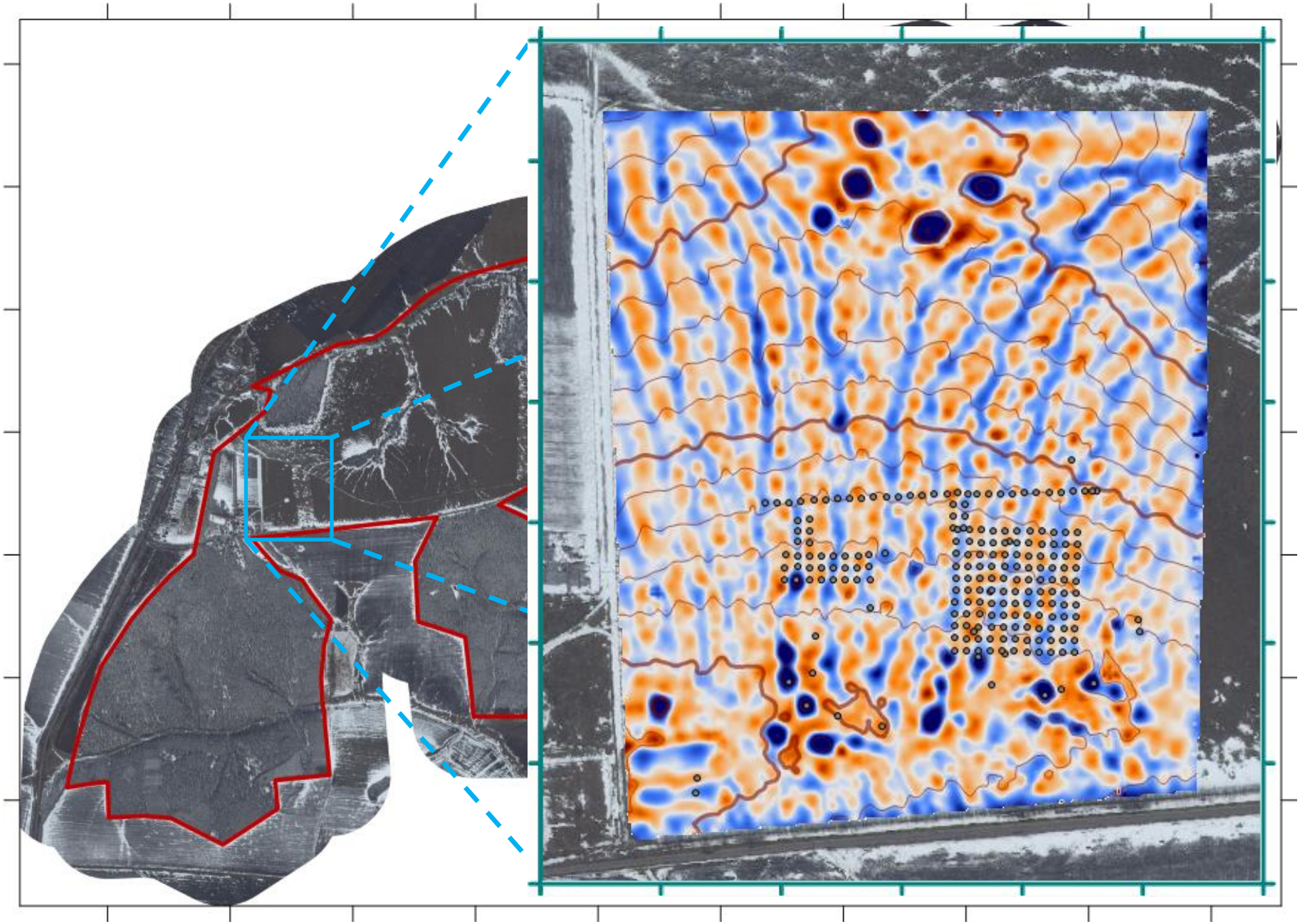
- 5.1. Что такое статистический тест, стр. 94-102
- 5.4. Есть ли взаимосвязь, или Анализ корреляций, стр. 109-114
- 5.5. Какая связь, или Регрессионный анализ, стр. 114-117

Приложение А. Пример работы в R, стр. 196-206

ТЕРРИТОРИЯ ИССЛЕДОВАНИЯ



КЛЮЧЕВОЙ УЧАСТОК «МЕТЕОСТАНЦИЯ»



ЗАДАНИЕ

1. СФОРМУЛИРОВАТЬ ПРОБЛЕМУ И ЦЕЛЬ ИССЛЕДОВАНИЯ
   ИССЛЕДОВАТЬ пространственную изменчивость почв северной лесостепи Среднерусской возвышенности (Центрально-Черноземный заповедник)
2. СПРОЕКТИРОВАТЬ ТАБЛИЦУ ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА

   строки – почвенные разрезы, колонки – свойства почвы (мощности горизонтов, таксон, запасы гумуса, кг/м2)
3. ОБОСНОВАТЬ ШКАЛУ ИЗМЕРЕНИЙ КАЖДОГО СВОЙСТВА

мощности горизонтов, запасы – интервальная, таксон - номинальная

4. ЗАПОЛНИТЬ ТАБЛИЦУ в Excel

id	A	AB	CaCO3	SID	HUM050	HUM100
m14-001	70	90	75	Чт	31.0	58.6
m14-002	75	90	120	Чтв	28.6	54.0
m14-003	85	95	135	Чтв	26.0	51.0
m14-004	70	85	80	Чл	38.0	62.0
…	…	…	…	…	…	…

ЗАДАНИЕ

5. ОБОСНОВАТЬ МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА (сравнение выборок, линейная связь между переменными)

- проверить H0 равенства интервальных значений для номинальных категорий (запасов органического вещества для четырех подтипов черноземов)
- проверить H0 отсутствия связи двух показателей (запасы органического вещества в слоях 0-50 и 0-100 см)

id	A	AB	CaCO3	SID	HUM050	HUM100
m14-001	70	90	75	Чт	31.0	58.6
m14-002	75	90	120	Чтв	28.6	54.0
m14-003	85	95	135	Чтв	26.0	51.0
m14-004	70	85	80	Чл	38.0	62.0
…	…	…	…	…	…	…

УСТАНОВИТЬ R и R-STUDIO

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function | Project: (None)

SoilProfiles.R* | pedons | Rplots.R | R_intro.R* | vars | pedproph

Source on Save | Run | Source

```
23  # write.table(data, "data/zhuki_new.txt", quote=FALSE) сохранить таблиц
24  # summary(data) сводная информация по таблице
25  # summary(data$VES) или отдельной колонке
26  dev.new()
27
28  data$var2.dx <- log(data$var2+1000) #создание новой колонки
29  # гистограмма
30  hist(data$var1, breaks=50, main="", col = "red")
31  hist(data$var2, breaks=50, main="", col = "blue")
32  hist(data$var1, breaks=50, main="", col = "lightblue", border = "darkgr
33  hist(data$var2.dx, breaks=50, main="", col = "blue", add=TRUE)
34  plot(density(data$var1, adjust=2), main="", col = "red")
35  plot(density(data$var1, adjust=2), main="", col = "blue", lwd=3, add=TR
36  rug(data$var1)
37
38  head(data)
39
40
41  # проверка на нормальность
42  qqnorm(data$var1)
43
```

32:1 (Top Level) R Script

Console

```
> hist(data$var2, breaks=50, main="", col = "blue", add=TRUE)
> help(hist)
> hist(data$var1, breaks=50, main="", col = "lightblue", border = "pink")
> h
Error: object 'h' not found
> h
Error: object 'h' not found
> hist(data$var1, breaks=50, main="", col = "red")
> hist(data$var1, breaks=50, main="", col = "lightblue", border = "pink")
> hist(data$var2, breaks=50, main="", col = "blue", add=TRUE)
> hist(data$var1, breaks=50, main="", col = "red")
> #hist(data$var1, breaks=50, main="", col = "lightblue", border = "pink")
```

Environment | History

Import Dataset | Clear | List

Global Environment

Data

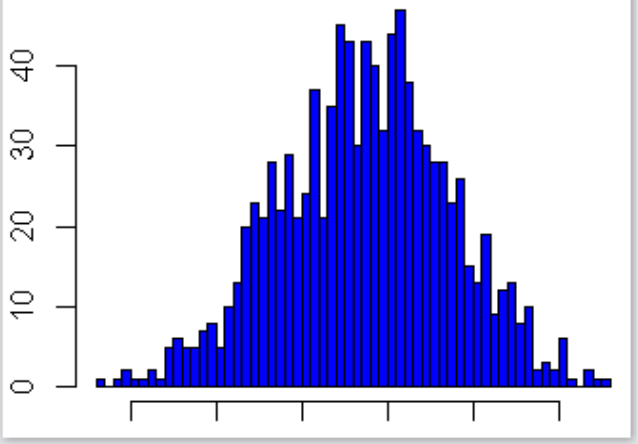
data	1000 obs. of 3 variable...
data.3	100 obs. of 3 variables
datagen	1000 obs. of 2 variable...
pedproph	150 obs. of 1 variables
vars	1000 obs. of 2 variable...

Values

HCLline	num [1:36] 70 90 115 55 8...
pedons	Formal class SoilProfileC...

Files | Plots | Packages | Help | Viewer

Zoom | Export | Clear All

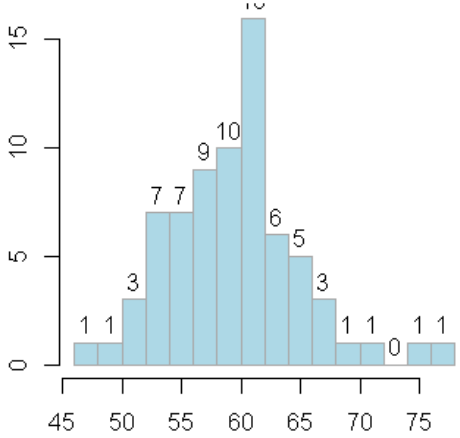


ПРИМЕР ОФОРМЛЕНИЯ ЗАДАНИЯ

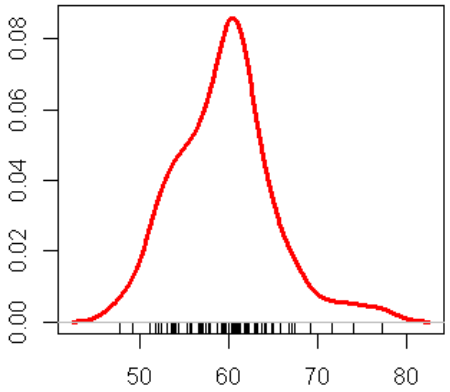
H0: запасы органического вещества в метровом слое черноземов равны

Чтк – типичный карбонатный, Чт – типичный, Чв – выщелоченный, Чл – лугово-черноземная почва

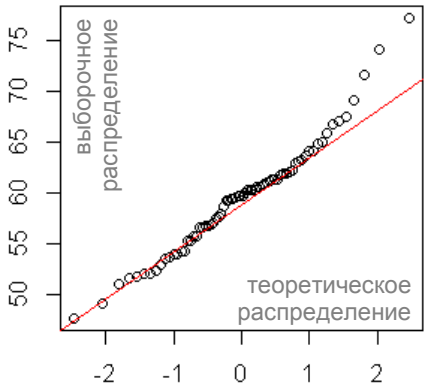
Порядок: 1. анализ распределений; 2. параметрический или непараметрический тест на различие средних



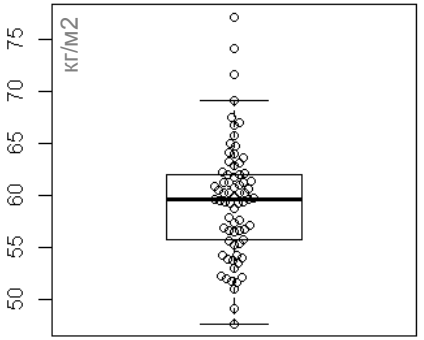
гистограмма распределения



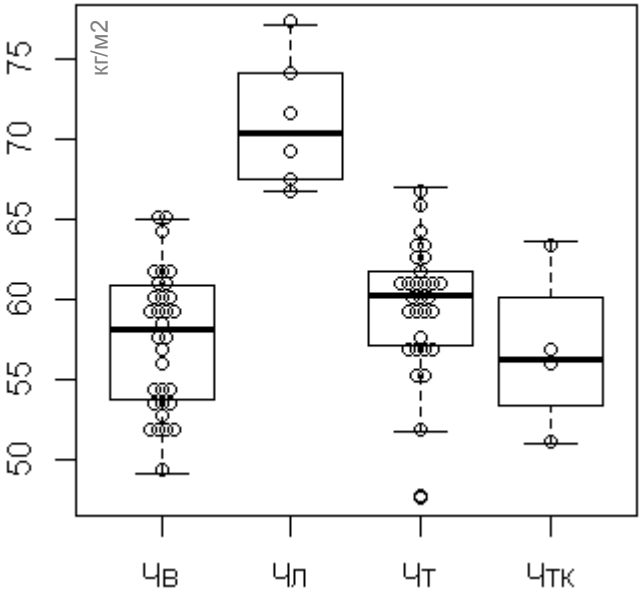
плотность распределения



нормальность распределения



ящик с усами и улий



почва	N	сред.	тесты, p-level		
			норм. распр.*	Чт**	Чв**
Чтк	4	56.8	недостаточно данных		
Чт	30	59.7	0.009	–	
Чв	32	57.6	0.209	0.047	–
Чл	6	71.1	0.668	0.001	0.0001

* тест на нормальность, ** - тест на достоверность различий средних

ВЫВОД: с уровнем значимости 0.95 запасы гумуса в метровом слое различаются Чл – 71.1 кг/м2; Чт – 59.7 кг/м2; Чв – 57.6 кг/м2; для статистически обоснованного суждения по Чтк недостаточно исходных данных

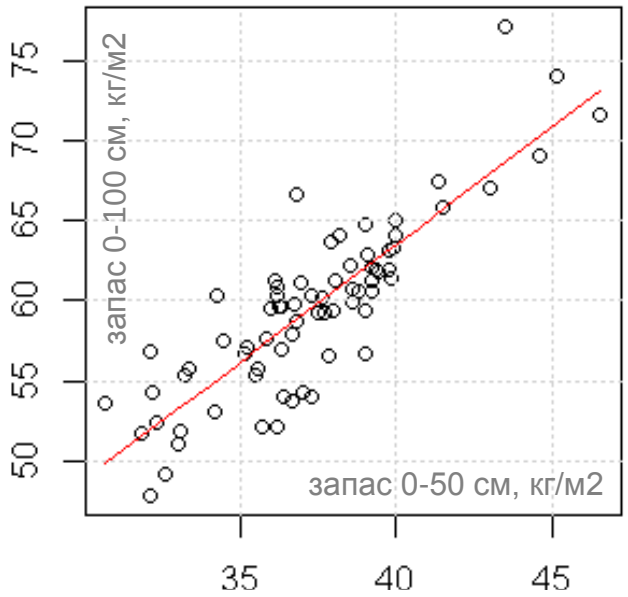
ПРИМЕР ОФОРМЛЕНИЯ ЗАДАНИЯ

H0: запасы органического вещества с глубиной меняются независимо

зависимая переменная (Y) — запасы гумуса в слое 0-100 см (h001m)

независимая переменная (X) — запасы гумуса в слое 0-50 см (h0050)

модель Y = b0 + b1*X




```
lm(formula = data$h001m ~ data$h0050)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7598 -1.5048 -0.0329  1.8413  8.4789

Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    4.4578     4.0250   1.108     0.272
data$h0050     1.4763     0.1075  13.733   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.873 on 70 degrees of freedom
Multiple R-squared:  0.7293
F-statistic: 188.6 on 1 and 70 DF,  p-value: < 2.2e-16
```

1. линейная модель вида *[запас 0-100 см] = 0 + 1.5 * [запас 0-50 см]* объясняет 72% (R2=0.72) изменчивости запасов органического вещества в метровом слое при среднеквадратической ошибке прогноза 2.9 кг/м2 (Residual standard error: 2.873);
2. значимость модели подтверждает высокое значение F-статистики, равное 189, и общий уровень значимости: p-value: 2.2e-16, что много меньше 0:001;
3. увеличение запасов в слое 0-50 на 10 кг соответствует увеличению запасов в слое 0-100 см примерно на 15 кг. Значение константы линейной модели (b0) недостоверно отличается от 0;
4. Наибольшее положительное отклонение истинного значения от модельного составляет 8.5 кг, наибольшее отрицательное -5.8 кг;
5. Почти половина остатков находится в пределах от первой квартили (1Q = -1.5 кг/м2) до третьей (3Q = 1.8 кг/м2)

ПРАКТИЧЕСКАЯ РАБОТА №1

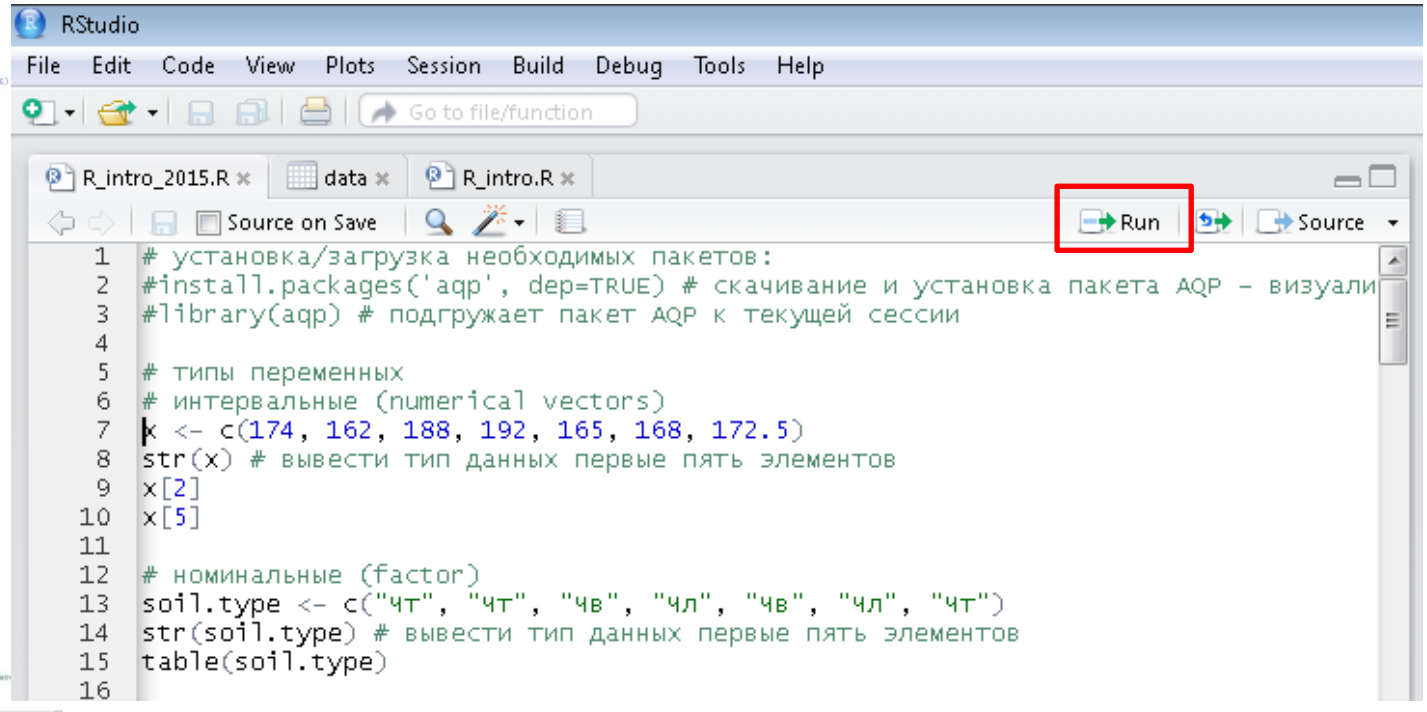
http://www.landscape.edu.ru/edu_help5_KiS.shtml#w1

Задание № 1. «Статистический анализ в среде R»

Цель: исследовать пространственную изменчивость почв северной лесостепи Среднерусской возвышенности (Центрально-Черноземный заповедник).

- инструкция к выполнению задания
- zip-архив с комплектом материалов

1. скачать и распаковать архив с материалами задания
2. открыть в R-Studio код **R_intro_2015.R**
3. выполнить задание: курсор в строку команды + Run



НУЛЕВАЯ ГИПОТЕЗА

РЕЗУЛЬТАТА
— КОНЕЧНОЙ ЦЕЛИ ИССЛЕДОВАНИЯ —
НЕ СУЩЕСТВУЕТ

НУЛЕВАЯ ГИПОТЕЗА (H_0) – основное проверяемое предположение об отсутствии различий, отсутствие влияние фактора, отсутствие эффекта, равенство нулю значений характеристик модели и т.п.

ОШИБКА ПЕРВОГО РОДА —

отказаться от H_0, в то время как она верна;

ОШИБКА ВТОРОГО РОДА —

принять H_0, когда она на самом деле не верна

Уровень значимости (*P-level*) – вероятность ошибки первого рода при принятии решения (вероятность ошибочного отклонения нулевой гипотезы), $p = 5\%, p = 1\%, p = 0.1\%$

МАЛО = НЕ ДОСТОВЕРНО vs. МНОГО = ТРУДОЕМКО

НУЛЕВАЯ ГИПОТЕЗА

		ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ («ИСТИНА»)	
		ВЕРНА НУЛЕВАЯ ГИПОТЕЗА (H0)	ВЕРНА АЛЬТЕРНАТИВНАЯ ГИПОТЕЗА (H1)
ВЫБОРКА («МОДЕЛЬ»)	ПРИНИМАЕМ НУЛЕВУЮ ГИПОТЕЗУ Ho	**ПРАВИЛЬНО !** H0 верно принята	H0 неверно принята **ОШИБКА ВТОРОГО РОДА**
	ПРИНИМАЕМ АЛЬТЕРНАТИВНУЮ ГИПОТЕЗУ H1	H0 неверно отвергнута **ОШИБКА ПЕРВОГО РОДА**	**ПРАВИЛЬНО !** H0 верно отвергнута

Уровень значимости (***p-level, p-value***) – вероятность ошибки первого рода при принятии решения (вероятность ошибочного отклонения нулевой гипотезы),
p = 0.05, p = 0.01, p = 0.001

H0: средние двух выборок (n=1000) не различаются (принадлежат одной генеральной совокупности)