

Методы проверки
статистических гипотез.

Корреляции и методы
сравнения. Дисперсионный
и регрессионный анализы.

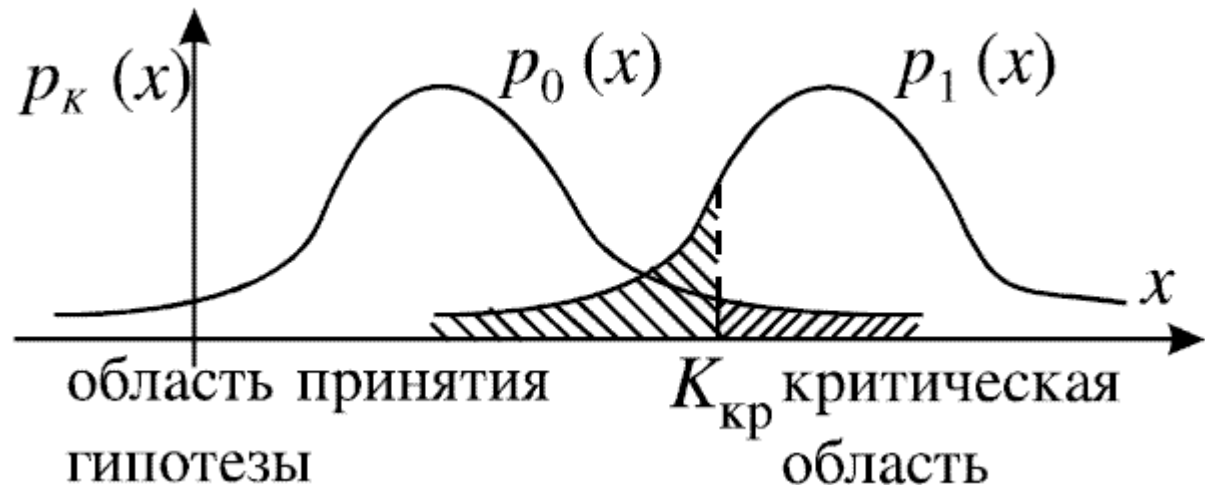
Долецкий А.Н.

ЗАДАЧИ СТАТ.АНАЛИЗА

- 1) выяснение свойств рассматриваемой генеральной совокупности по результатам измерения свойств части этой совокупности
- 2) прогнозирование характеристик исследуемого объекта

ЭТАПЫ ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

- 1) формирование случайной выборки из генеральной совокупности,
- 2) описание свойств выборки
- 3) обобщение результатов стат.обработки на генеральную совокупность (статистический вывод)



ЭТАПЫ РАБОТЫ С ВЫБОРКОЙ

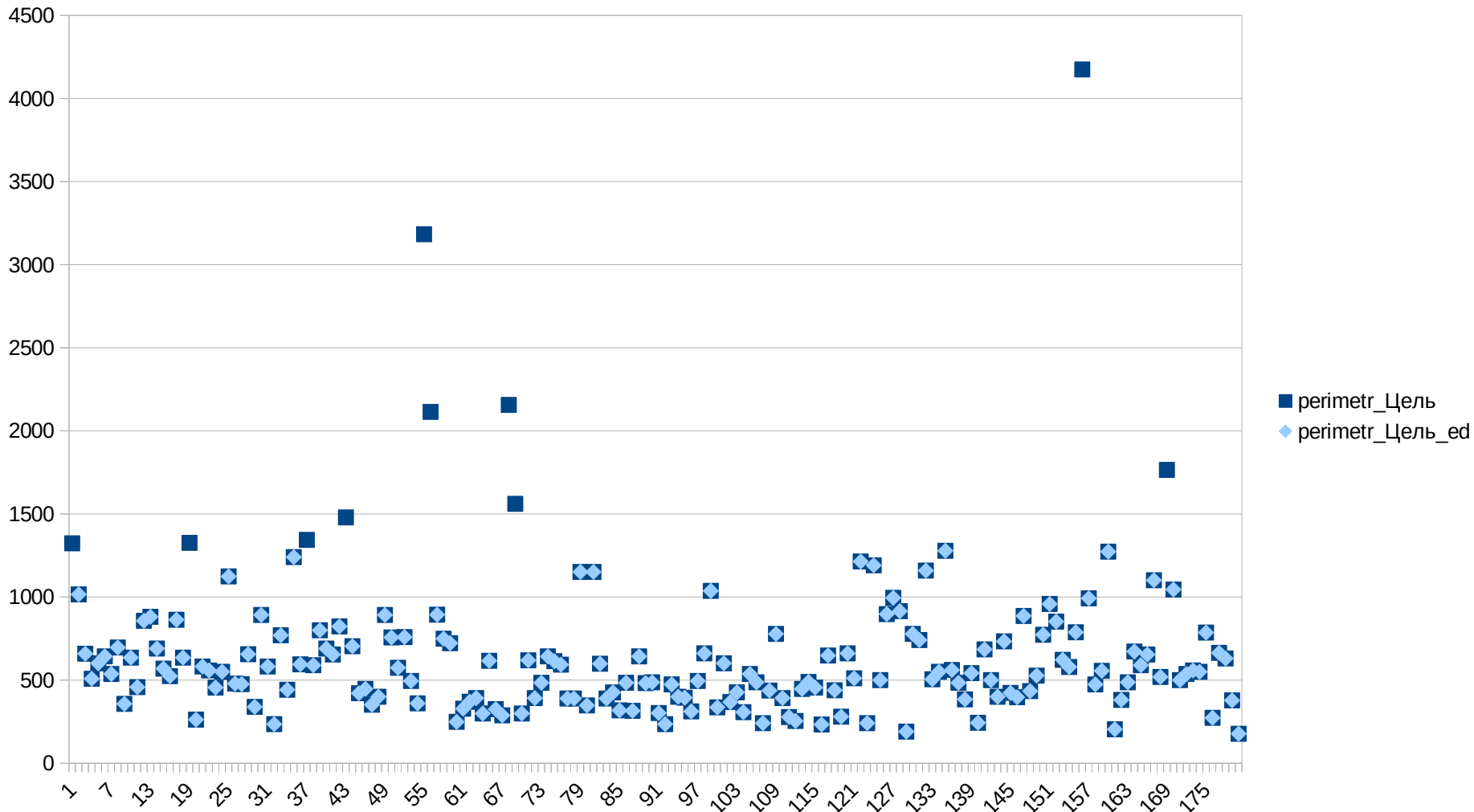


ИСТОЧНИКИ ПОГРЕШНОСТЕЙ ИЗМЕРЕНИЯ

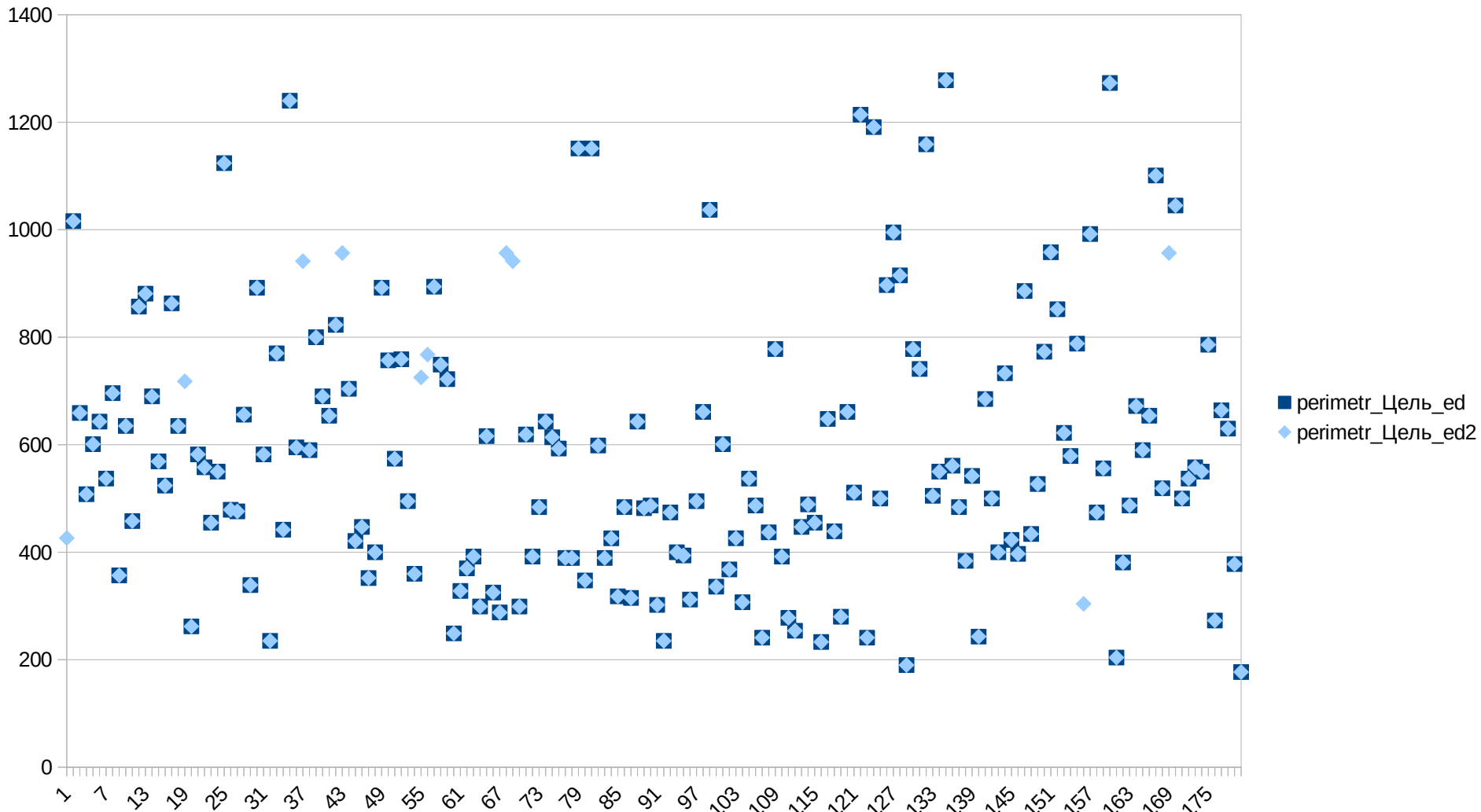
- 1) погрешности, вызванные объектом исследования (разложение материала, старение особей, разная плотность материала, изменение состояния животного, стрессы при вживлении датчиков и т.п.);
- 2) погрешности оператора (обучение, усталость, болезнь, реакция на внешние раздражители и т.п.);
- 3) инструментальные погрешности, связанные с погрешностями измерительных приборов;
- 4) методические погрешности, связанные с ошибочными представлениями о закономерностях (обоснованность допусков, полнота учета факторов и ограничений на них, эффекты взаимодействия объекта с измерительным комплексом и т.д.);
- 5) погрешности влияния внешней среды (температура, осадки, солнечная радиация, радиоактивность и т.п.).

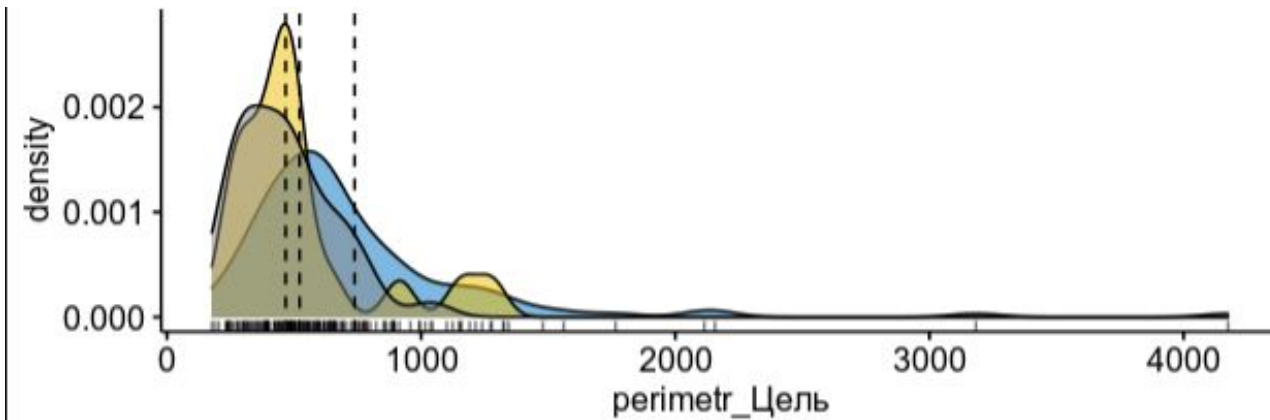
Качество исследования определяется тем, насколько удастся устранить или компенсировать воздействие вышеуказанных источников погрешностей на результат измерения

УДАЛЕНИЕ ВЫБРОСОВ



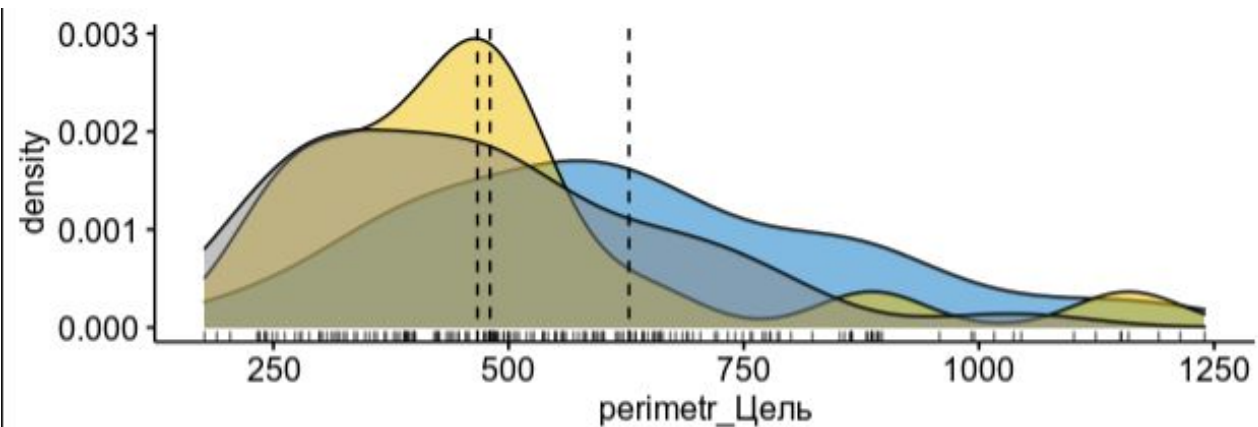
ЗАМЕНЯЕМ ПРОПУСКИ СРЕДНИМИ (МЕТОДОМ БОНФЕРРОНИ)





спортсмен	length	mean	sd
отсутствие регулярной физической активности	124	738.3597	519.8436
регулярная физическая активность	20	521.6000	283.3163
спортсмены	36	466.9444	188.9878

График плотности с таблицей значений



спортсмен	length	mean	sd
отсутствие регулярной физической активности	123	628.0119	236.4812
регулярная физическая активность	19	480.2728	223.4402
спортсмены	36	466.9444	188.9878

График плотности с таблицей значений

F5 `=ХИ2.ТЕСТ(C5:C10;D5:D10)`

	А	В	С	Д	Е	Г
1						
2						
3			Частоты			
4		Очки	Наблюдаемая О	Ожидаемая Е		
5		1	8	10		
6		2	12	10		
7		3	13	10		
8		4	7	10		
9		5	12	10		
10		6	8	10		
11		Итого	60	60		
12						

Тест χ^2
0,63857

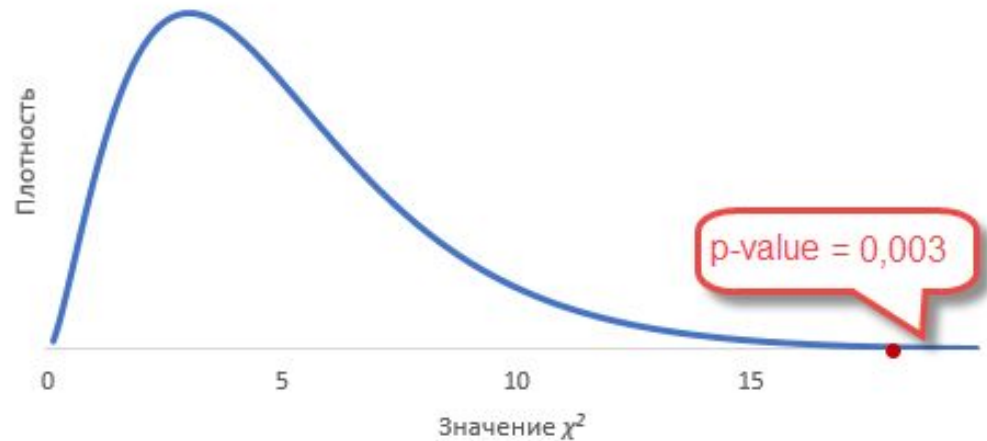
p-value

Критерий Пирсона хи-квадрат - критерий статистической значимости для категориальных переменных.

1) сопоставление эмпирического распределения признака с теоретическим - равномерным, нормальным или каким-то иным;

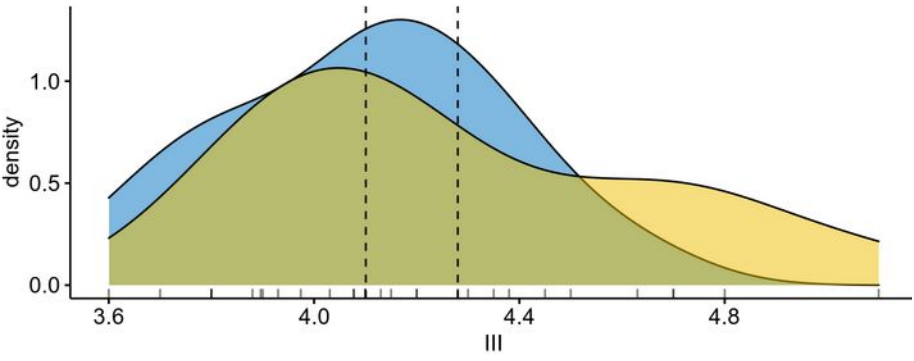
2) сопоставление двух, трех или более эмпирических распределений одного и того же признака (в скрипте до 10).

Распределение χ^2 (хи-квадрат)
d.f. = 5



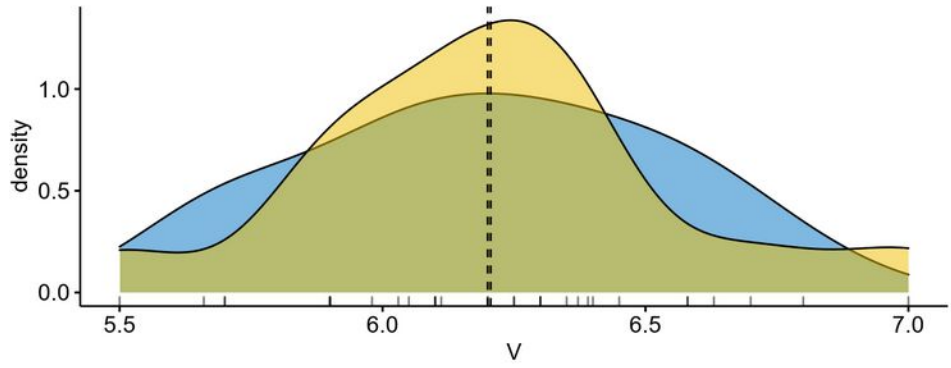
"Нормальность распределения количественных переменных (р-уровни теста shapiro.test)"

Возраст I III V I.V I.III
 "1.84e-12" "8.47e-10" "0.0351" "0.279" "9.56e-14" "2.91e-07"



группа	length	mean	sd
ЗПР	18	4.100815	0.2755547
СД	15	4.280043	0.3896974

График плотности с таблицей значений



группа	length	mean	sd
ЗПР	18	6.200035	0.3388074
СД	15	6.205553	0.3562510

График плотности с таблицей значений

Планирование исследования

Выбор статистического метода

Определение объема выборки

Расчет статистических величин:

Расчет относительных величин

Анализ динамического ряда

Прямой метод стандартизации

Показатели вариационного ряда

Расчёт демографических показателей

Сравнение совокупностей по качественным признакам:

Относительный риск

Отношение шансов

Анализ четырехпольной таблицы (критерий хи-квадрат)

Критерий хи-квадрат для произвольных таблиц

Критерий Мак-Немара

Выбор метода статистического анализа

Хотите, чтобы программа сама подобрала нужный метод анализа, исходя из Ваших данных? А потом еще и выполнила расчет и сама сделала корректное и полное описание полученных результатов в формате главы диссертации или соответствующего раздела статьи? Тогда обратите внимание на наш новый онлайн-сервис StatTech, который посчитает и опишет результаты за Вас!

Условия выбора

Число сравниваемых совокупностей:

одна



Цель статистической обработки:

определение нормальности распределения



Подходящие методы:

Критерий Шапиро-Уилка

Критерий Колмогорова-Смирнова

АЛГОРИТМ ДЕЙСТВИЙ

Сравнение двух независимых групп

Различается ли снижение систолического давления в зависимости от того, получал пациент препарат А или В?

Проверьте каждую из групп на нормальность распределения (*критерий Пирсона χ^2 , графический метод, тест Шапиро, критерий Колмогорова-Смирнова*).

Если распределение в каждой группе является нормальным — **t-критерий для независимых выборок**.

Если распределение отлично от нормального - воспользуйтесь **непараметрическими аналогами** - U критерием Манна-Уитни, критерием Колмогорова-Смирнова или критерием Вальда Вольфовица.

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

Подходят для малых выборок.

nonparametric methods = “distribution-free” tests.

- ✓ Свойства распределения неизвестны, **параметры** распределения (среднее, дисперсия и т. п.) **НЕ ИСПОЛЬЗУЮТСЯ**.
- ✓ Основной подход – **ранжирование** (*ranking*) наблюдений (выстраиваем их по порядку от самого маленького значения к наибольшему, или наоборот).
- ✓ Подразумевается, что **сравниваемые распределения имеют одинаковую форму** (и, иногда, дисперсию).



НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ

В некоторых книжках пишут, что параметрические методы не годятся, если данные **РАНГОВЫЕ**: неизвестно, насколько одно значение отличается от другого.

НО:

It is sometimes counseled that only nonparametric testing may be employed when dealing with ordinal-scale data, but such advice is based upon what Gaito (1980) calls “an old misconception”; this issue is also discussed by Anderson (1961), Gaito (1960), Savage (1957), and Stevens (1968). Interval-scale or ratio-scale measurements are not intrinsically required for the application of parametric testing procedures. Thus parametric techniques may be considered for ordinal-scale data *if* the assumptions of such methods are met—typically, random sampling from normally distributed populations with homogeneity of variances. But ordinal data often come

Zar, 2010; Quinn, Keough, 2002

Главное - проверить данные на соответствие нормальности и гомогенности.

СРАВНЕНИЕ 2-Х НЕЗАВИСИМЫХ ГРУПП

(аналоги *t*-test for independent samples)

В 5А и 5Б провели контрольную по математике и подсчитали баллы за задания.

Зависимая переменная – оценка (в баллах).

Независимая – класс.

Тест Манна-Уитни

(*Mann-Whitney U*-test)

H_0 : оценки в 5А И 5Б одинаковы.

H_1 : оценки в 5А И 5Б не одинаковы.

Параметры распределений **не упоминаются** в гипотезах.

Тест Манна-Уитни можно использовать и для ранговых, и для непрерывных переменных;

Можно использовать для **малых выборок** (от **5** измерений!).



Непараметрические критерии

- ✓ Если выборки удовлетворяют требованиям для параметрических тестов, мощность **теста Манна-Уитни** = **95% от мощности** t-теста.
- ✓ М-У тест - один из самых **мощных** среди непараметрических тестов.
- ✓ Альтернативная процедура – **ранговая трансформация** и последующий t-тест над рангами (Zar, 2010, Quinn, Keough, 2002).

Еще тесты для 2-х независимых групп:

Тест **Колмогорова-Смирнова** (*Kolmogorov-Smirnov two-sample test* – не тот, что для проверки на соответствие нормальному распределению): более чувствителен к различиям распределений **по форме**.

Тест **Вальда-Вольфовица** (*Wald-Wolfowitz Runs Test*) – оцениваются последовательности элементов (*runs*) из разных групп. Как и К-С тест, чувствителен к различиям распределений по форме.

Непараметрические критерии

U- критерий

$$r_s = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}$$

$$Z = \frac{U_{obs} - \mu_U}{\sigma_U}$$

n_1 и n_2 – размер выборок,

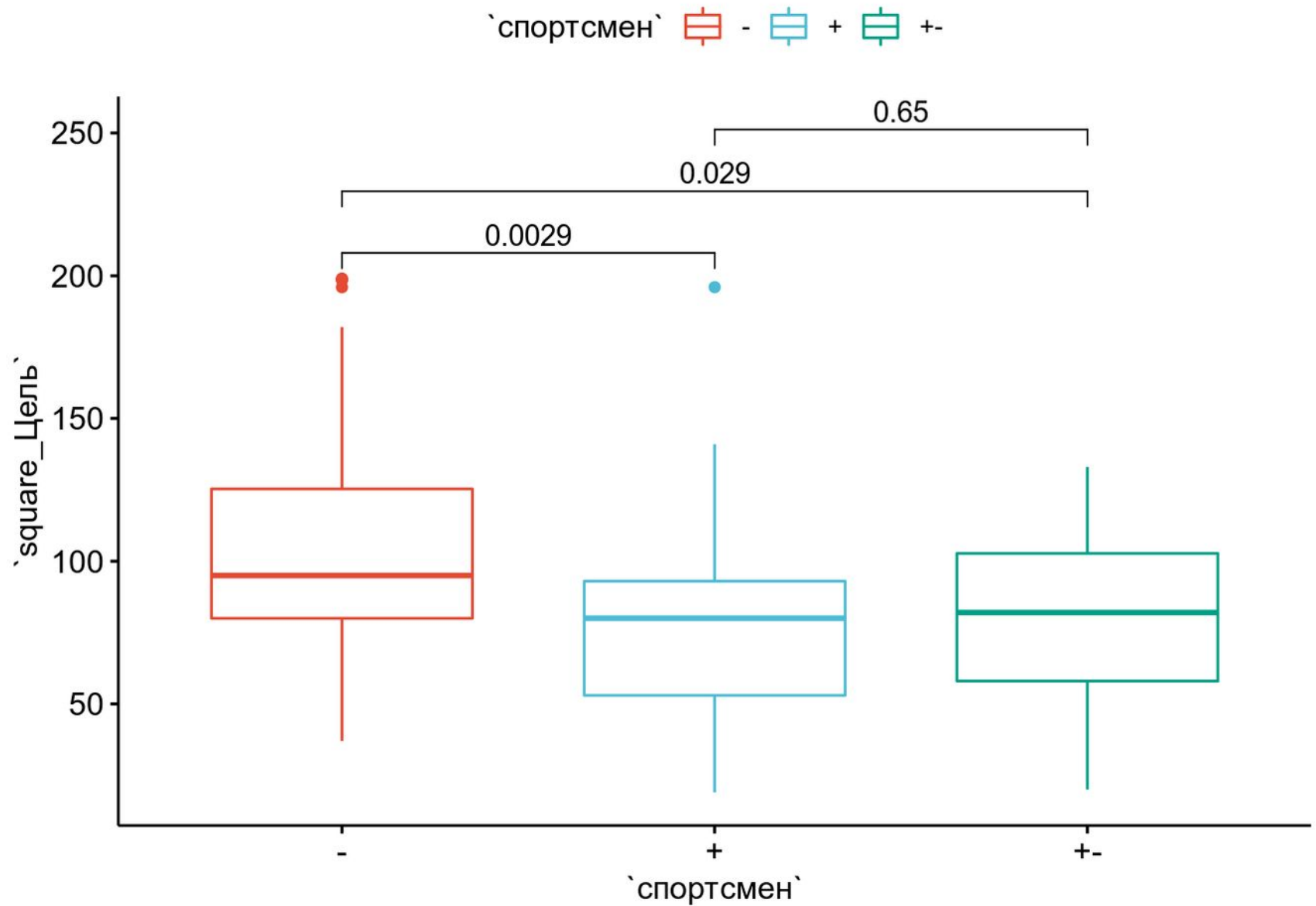
R_1 и R_2 – суммы рангов в выборках.

Статистикой критерия U_{obs} будет **меньшее** из этих двух значений.

H_0 отвергают в случае, если **U меньше** критического значения U_{cv} . (т.е., это исключение среди критериев).



Непараметрические критерии

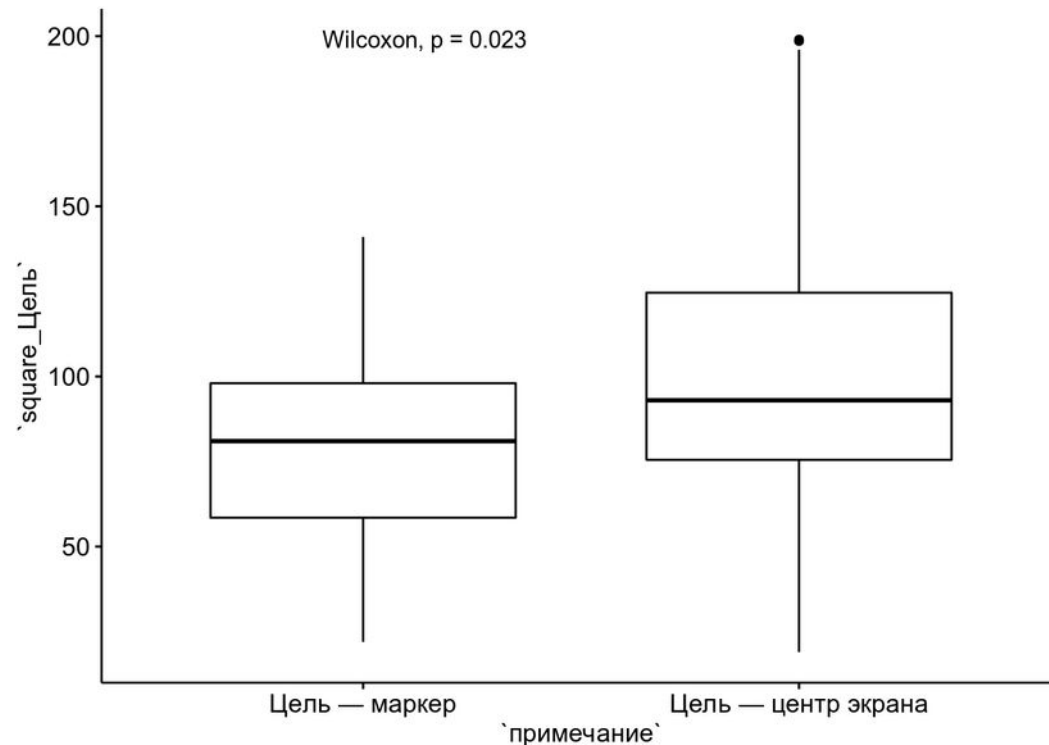


АЛГОРИТМ ДЕЙСТВИЙ

Вам нужно исследовать **одну группу пациентов** до и после лечения.

Если распределение в группе до и после лечения является нормальным - **t-критерий для зависимых выборок**.

Если распределение отлично от нормального - непараметрические аналоги - **критерий Вилкоксона, критерий знаков**.



АЛГОРИТМ ДЕЙСТВИЙ

Исследовать одну группу после **нескольких** видов измерений можно Дисперсионным анализом с повторными измерениями.

```
lm(formula = square_ОГ + square_3Г + square_Цель ~ вид.занятий,
```

```
## Residuals:                                ##                               Df Sum Sq Mean Sq F value Pr(>F)
##   Min   1Q Median   3Q   Max                                ## variable                14    442    31.57    19.28 <2e-16 ***
## -229.21 -104.00 -19.33  73.79 295.65                                ## отведение                20   1573    78.63    48.01 <2e-16 ***
##                                         ## name                      14   3367   240.50   146.83 <2e-16 ***
## Coefficients:                                ## Residuals             13955  22857    1.64
## ---                                           ## ---
##           Estimate Std. Error t value Pr(>|t|)                                ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
## (Intercept)           388.21    13.78  28.179 <2e-16 ***                                0.1 ' ' 1
## вид.занятийакробатика           -71.88    44.76  -1.606  0.111
## вид.занятийбаскетбол           -106.01    58.77  -1.804  0.074 .
## вид.занятийволейбол            -84.21    58.77  -1.433  0.155
## вид.занятийгандбол             -29.21    91.38  -0.320  0.750
## вид.занятийгреко-римская борьба -116.21    91.38  -1.272  0.206
## вид.занятийлегкая атлетика      -24.00    75.04  -0.320  0.750
## вид.занятийспортивная гимнастика -53.21    58.77  -0.905  0.367
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 127.8 on 109 degrees of freedom
## Multiple R-squared:  0.07312, Adjusted R-squared:  0.0136
## F-statistic: 1.228 on 7 and 109 DF, p-value: 0.2934
```

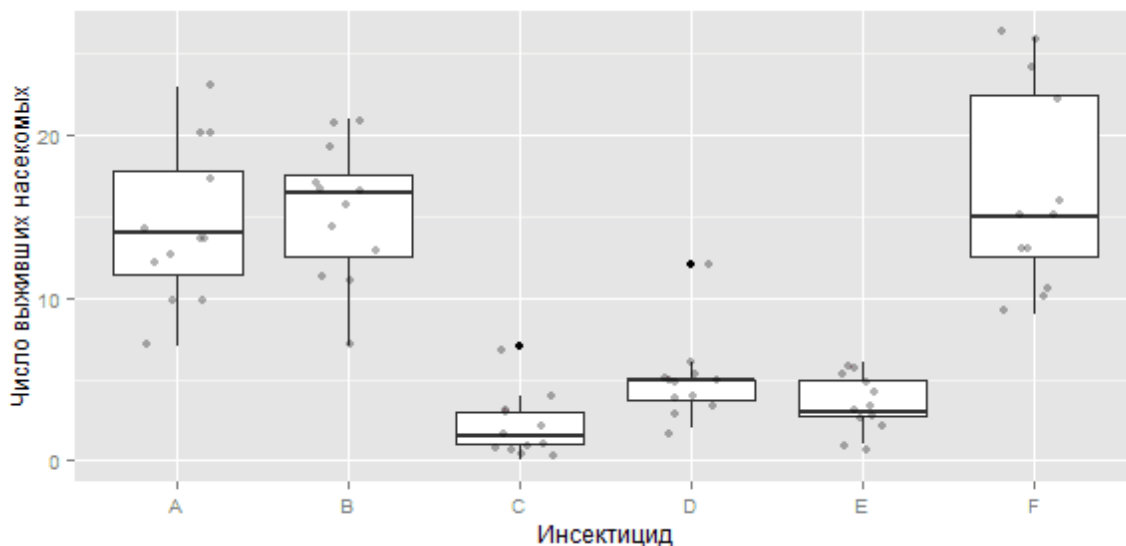
АЛГОРИТМ ДЕЙСТВИЙ

Условия применимости дисперсионного анализа

Во всех анализируемых группах значения зависимой переменной должны быть распределены нормально.

Дисперсия значений зависимой переменной должна быть одинакова во всех группах.

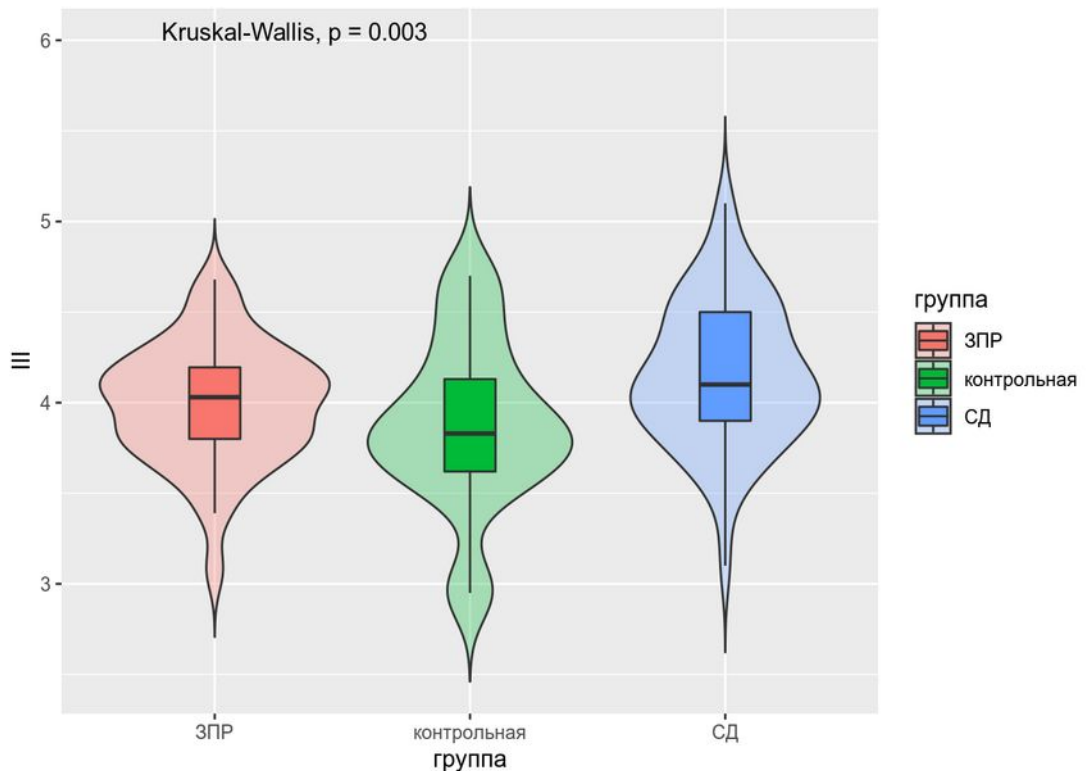
Во всех группах значения зависимой переменной должны быть независимы.



АЛГОРИТМ ДЕЙСТВИЙ

Условия применимости дисперсионного анализа

Условия не удовлетворены, воспользуйтесь **непараметрической альтернативой** - Q критерием Кохрена, ранговым дисперсионный анализом по Краскелу-Уоллису .



ANOVA Краскела-Уоллиса и медианный тест

Являются непараметрическими альтернативами однофакторного дисперсионного анализа.

Если переменных больше двух, то применяется дисперсионный анализ. Английское сокращение - ANOVA (analysis of variation).

Критерий Краскела-Уоллиса основан на рангах и предполагает, что рассматриваемая переменная непрерывна и измерена как минимум в порядковой шкале. Критерий проверяет гипотезу: имеют ли сравниваемые выборки одно и то же распределение или же распределения с одной и той же медианой.

Интерпретация схожа с интерпретацией параметрической однофакторной ANOVA за исключением того, что этот критерий основан на рангах, а не на средних значениях.

Алгоритм действий

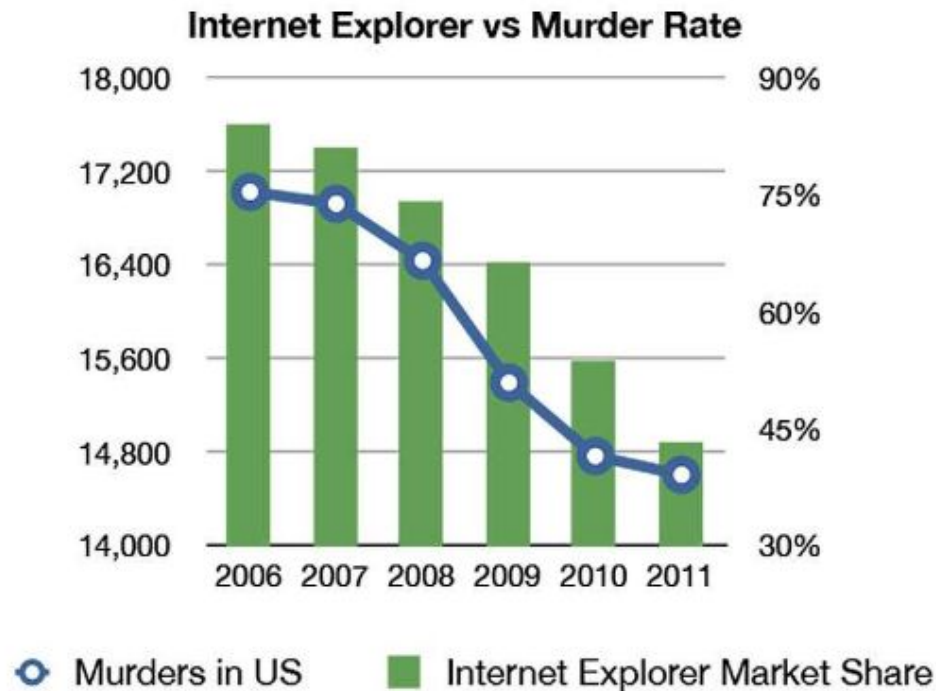
Вам необходимо сравнить группы, выраженные в категориальной шкале, т.е. значения исследуемой переменной может принимать только одно из заданных значений, например: пол (м, ж), степень тяжести (I, II III), инсульт (был, не был), наличие фибрилляций (да, нет) и т.п.

Анализ частот в группах (число пациентов, получающих препарат А и имеющих рецидив, число пациентов, получающих препарат А и не имеющих рецидива; аналогично для пациентов, не получающих препарат А - с рецидивом и без рецидива). Для проведения сравнения частот воспользуйтесь **хи-квадратом Пирсона**.

Если необходимо проанализировать связанные измерения (т.е. "до" и "после" лечения/вмешательства) для дихотомических переменных (это частный случай категориальной переменной, когда категорий всего две - 1, 0 или да, нет), используют критерий Мак-Немара.

КОРРЕЛЯЦИЯ

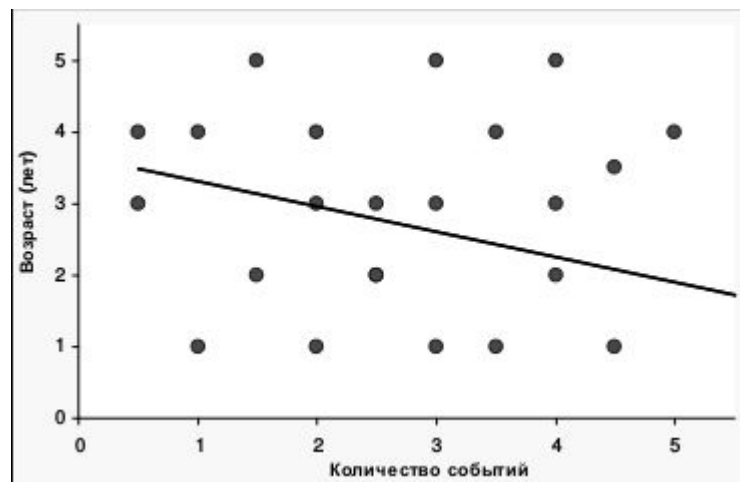
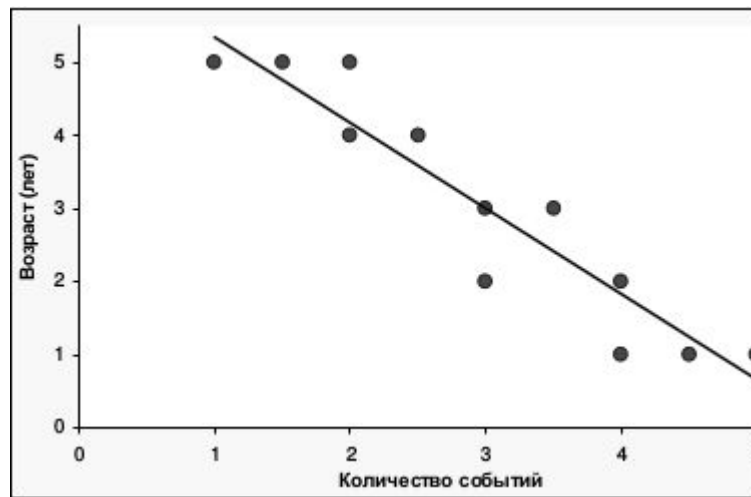
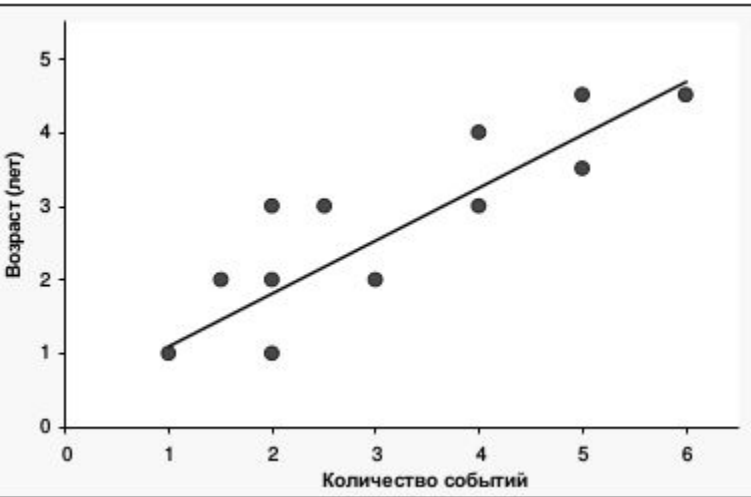
Корреляция — статистическая взаимосвязь между случайными величинами; не является достаточным условием причинно-следственной:



Другие примеры: <http://www.tylervigen.com/>

Пять видов связи между переменными

1. Прямая причинно-следственная связь
2. Обратная причинно-следственная связь
3. Связь вызвана третьей (скрытой) переменной
4. Взаимосвязь вызвана несколькими скрытыми переменными
5. Связи нет, наблюдаемая зависимость случайна



Корреляция Пирсона

Мера линейной связи между двумя переменными позволяет определить, насколько пропорциональна изменчивость двух переменных.

Если переменные пропорциональны друг другу, то графически связь можно представить в виде прямой линии с положительным (прямая пропорция) или отрицательным (обратная пропорция) наклоном.

Коэффициенты корреляции Спирмена и т-Кендалла (ранговые корреляции)

Если обе переменные, между которыми изучается связь, представлены в порядковой шкале, или одна из них - в порядковой, а другая - в метрической, то применяются ранговые коэффициенты корреляции: Спирмена или т-Кенделла. И тот, и другой коэффициент требует для своего применения предварительного ранжирования обеих переменных.

Алгоритм действий

Вам необходимо обнаружить связь между двумя непрерывными признаками - воспользуйтесь **корреляциями Пирсона** (в случае нормальности распределения переменных) или непараметрическими корреляциями, например, **R Спирмена**.

Если необходимо не просто проанализировать связь между непрерывными признаками, а использовать полученные данные для прогнозирования - воспользуйтесь **множественной регрессией**.

Выберите метод

Пример 1: изучалось воздействие марихуаны на скорость реакции. В качестве испытуемых были выбраны по 12 человек из каждой категории:

- никогда не пробовали марихуану;
- иногда употребляют марихуану;
- регулярно употребляют марихуану.

Испытуемые были разделены на две равные группы; половине из них дали выкурить две сигареты с марихуаной, вторая половина выкурила две обычные сигареты с запахом и вкусом марихуаны. Сразу после этого все испытуемые прошли тест на скорость реакции.

Требуется оценить влияние марихуаны на скорость реакции, учитывая фактор предыдущего опыта употребления.

ОЦЕНИТЕ РЕЗУЛЬТАТ

H_0^1 : средняя скорость реакции одинакова при употреблении и марихуаны, и сигарет.

H_0^2 : средняя скорость реакции не зависит от предыдущего опыта употребления марихуаны.

H_0^{12} : отсутствует межфакторное взаимодействие между употребляемым веществом и предыдущим опытом употребления марихуаны.

Source	SS	df	MS	F	Prob>F
Group	103041	1	103041	17.58	0.0002
Past use	23634.5	2	11817.2	2.02	0.1508
Interaction	23642.2	2	11821.1	2.02	0.1507
Error	175796.3	30	5859.9		
Total	326114	35			

Дополнительная литература

Онлайн курс "Методология научных исследования и котики", в которых объяснены методы сбора и анализа данных

<https://ru.coursera.org/learn/metodologiya-nauchnyh-issledovaniy-kotiki#syllabus>

Онлайн курс на Stepik "Основы статистики" – хороший онлайн курс по базовым понятиям и тестам

<https://stepik.org/course/76/syllabus>

"Discovering statistics using IBM SPSS", Andy Field – смешной и полезный учебник по статистике, у автора есть такой же веселый канал на YouTube с примерами на зомби и щеночках

<https://www.youtube.com/channel/UCakigkjm3vBzEHpFzECDXQQ>

Очень наглядный и интерактивный учебник "Seeing Theory" о том, как работают вероятности и регрессионный анализ

<https://seeing-theory.brown.edu/basic-probability/index.html>

Красивый учебник "Data Viz Project" о видах графиков и примеров из реальных источников

<https://datavizproject.com/>

