Занимательная статистика

Манга

ОДЭКА

マンガでわかる
統計学

高橋　信／著
トレンド・プロ／マンガ制作

Ohmsha

Манга

Занимательная СТАТИСТИКА

Син Такахаси

Перевод с японского

Захаровой Е. А., Коги Муцуми

ОДЭКА

Москва
Издательский дом «Додэка-XXI»
2010

УДК 311
ББК 60.6
Т15

Такахаси, Син.
Т15 Занимательная статистика. Манга / Син Такахаси ; пер. с яп. Захаровой Е. А., Коги Муцуми. — М. : Додэка-XXI, 2010. — 224 с. : ил. — (Серия «Образовательная Манга»). — Доп. тит. л. яп. — ISBN 978-5-94120-244-7.
I. Захарова, Е. А., пер.

Если тебя интересует статистика, или тебе просто нужно как-то обработать данные, то «Занимательная статистика» поможет тебе преодолеть чувство, что «ты плохо знаешь математику». Этот иллюстрированный путеводитель легко и непринуждённо проведёт тебя по пути познания статистики. А полученные знания ты сможешь закрепить с помощью упражнений, без которых, как известно, не обходится ни одна книга по математике.

Последуй за всегда невозмутимым Ямамото и ты увидишь, как он научит Руи:

- рассчитать среднее значение, медиану и стандартное отклонение результатов в боулинге;
- построить гистограмму цен на китайскую лапшу рамэн;
- определить вероятность получения проходного балла на экзаменах по математике;
- вычислить коэффициент Крамера, чтобы узнать, как предпочитают признаваться в любви юноши и девушки;
- узнать, как нормируются результаты тестов, когда учителя оценивают успеваемость.

Эти и другие примеры из реальной жизни позволят тебе с лёгкостью усвоить то, что многие находят трудным для понимания.

Если ты хочешь разобраться в статистике, но от обычных учебников статистики у тебя пухнет голова и клонит в сон, или если тебе просто нужно освежить забытые знания, пусть Ямамото-сан и Руи будут твоими гидами.

Книга будет полезна учащимся старших классов средних школ и колледжей, студентам вузов, а также всем, кто интересуется статистикой и хочет, чтобы обучение было лёгким и увлекательным.

УДК 311
ББК 60.6

ISBN 978-5-94120-244-7 (рус.)
ISBN 978-4-27406-570-5 (яп.)

Содержание

Предисловие

Данная книга — наглядное учебное пособие по статистике, которое, в первую очередь, предназначается тем, кому приходится заниматься анализом различных данных, а также тем, кто пока такой анализ не проводит, но хотел бы знать, что же такое статистика.

Автору также будет весьма приятно, если книга окажется интересной и для тех, кто уже изучал эту дисциплину.

Статистика — одна из областей математики, тесно связанная с жизнью и работой.

Если овладеть всеми премудростями этой науки, то можно, например:

- предусмотреть, сколько коробок жареной лапши будет продано в студенческом киоске, который планируется открыть на университетском празднике;
- оценить вероятность успешной сдачи квалификационного экзамена;
- сравнить вероятность выздоровления, если принимать лекарство X и не принимать это лекарство.

Книга содержит 7 глав. За некоторым исключением, главы книги построены следующим образом:

- манга (комикс);
- объяснение, дополняющее мангу;
- упражнения и ответы;
- выводы.

Книга написана так, что читатель может усвоить материал, прочитав только мангу. А вот чтобы получить более глубокие знания, придётся прочитать и всё остальное.

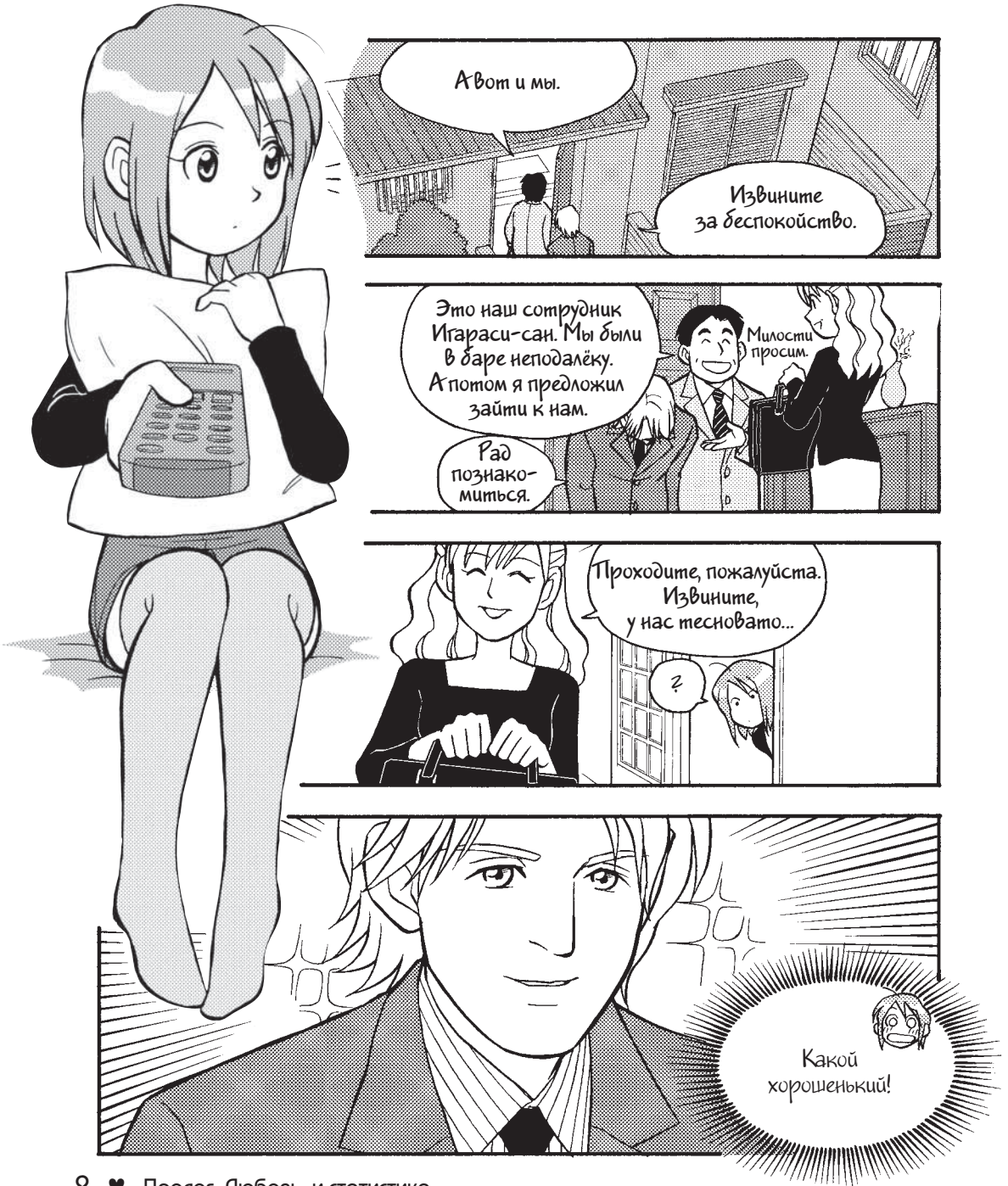
Предел мечтаний автора — читатель, который, перевернув последнюю страницу книги, скажет: «Статистика — это так интересно! Но это ещё и полезно! Да это просто здорово!».

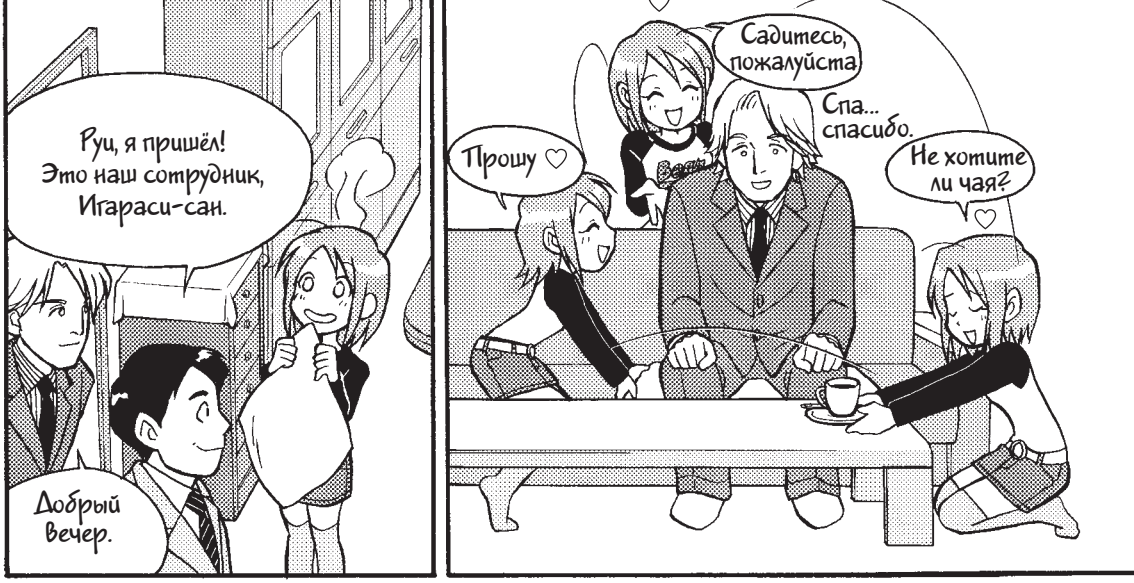
Я бесконечно благодарен всем сотрудникам редакции издательства Ohmsha за предоставленную мне возможность написать эту книгу, а также всем сотрудникам компании Trend-Pro. Я глубоко признателен г-ну Ре Акино, автору сценария, и г-ну Ироха Иноуэ, воплотившему этот сценарий в виде рисунков, за те титанические усилия, которые им пришлось приложить, чтобы на основе моей рукописи создать потрясающий комикс. Я также благодарен г-ну Фумитакэ Сакаи (социологический факультет университета Риккё), советами которого я руководствовался во время работы над этой книгой.

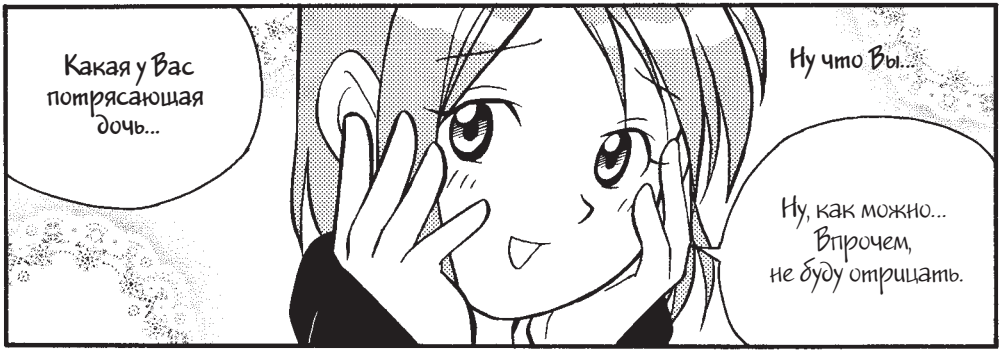
Син Такахаси
Июль, 2004 год

Пролог

Любовь и статистика


А вот и мы.
Извините за беспокойство.
Это наш сотрудник Игараси-сан. Мы были в баре неподалёку. А потом я предложил зайти к нам.
Милости просим.
Рад познакомиться.
Проходите, пожалуйста. Извините, у нас тесновато...
?
Какой хорошенький!


Руи, я пришёл! Это наш сотрудник, Игараси-сан.
Добрый вечер.
Садитесь, пожалуйста
Спа... спасибо.
Прошу ♡
Не хотите ли чая? ♡


Какая у Вас потрясающая дочь...
Ну что Вы...
Ну, как можно... Впрочем, не буду отрицать.


Игараси-сан, а чем Вы занимаетесь?
Работаю в одной компании с твоим отцом.
Проще говоря, занимаюсь маркетингом.
Маркетингом?


Другими словами, провожу маркетинговые исследования и использую для этого статистические методы.
Возможно, вы еще не проходили в школе, что такое маркетинг?
Не помню... Кажется, нет.


Честная девочка. Ну, а что такое статистика, знаешь?
Э-э-э...


Вижу, что нет. В общем, статистика — наука, изучающая большие совокупности однородных объектов на основании их выборочного исследования.


Что-то я слишком загнул.
НЕПОСТИЖИМО...
Эй, Руи! Что с тобой?
А, кстати, ...

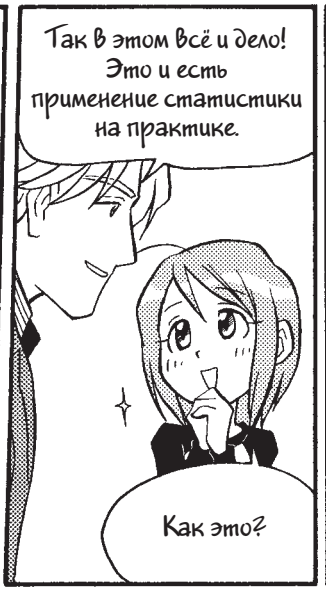

ВЕДОМОСТИ
... как раз в сегодняшней газете есть информация о рейтинге кабинета министров.

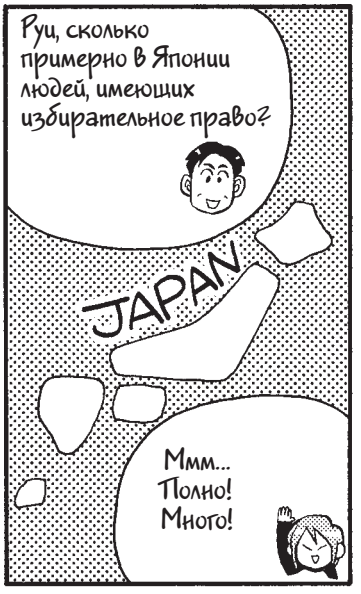

По исследованиям, проведённым газетой "Ведомости", рейтинг кабинета министров среди избирателей 39%.
И что это значит?


Но сотрудники газеты моего мнения не спрашивали.
А как насчет Вас, Такацу-сан?

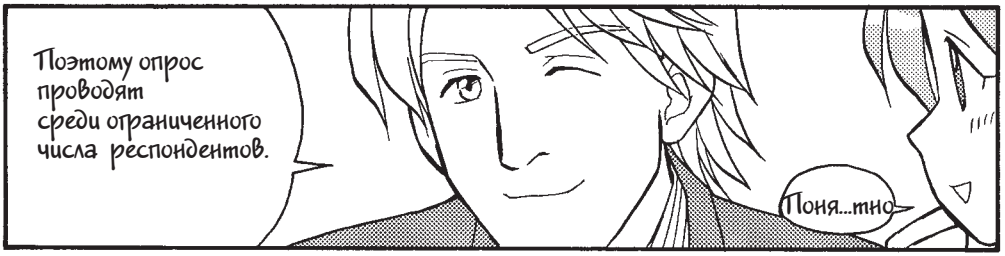

Меня...
Нет, меня тоже не спрашивали.


Хм-м... Мнением двоих не поинтересовались, а рейтинг опубликовали. Как такое может быть?
Между тем, избирательного права вас никто не лишал. Более чем странно...


Так в этом всё и дело! Это и есть применение статистики на практике.
Как это?

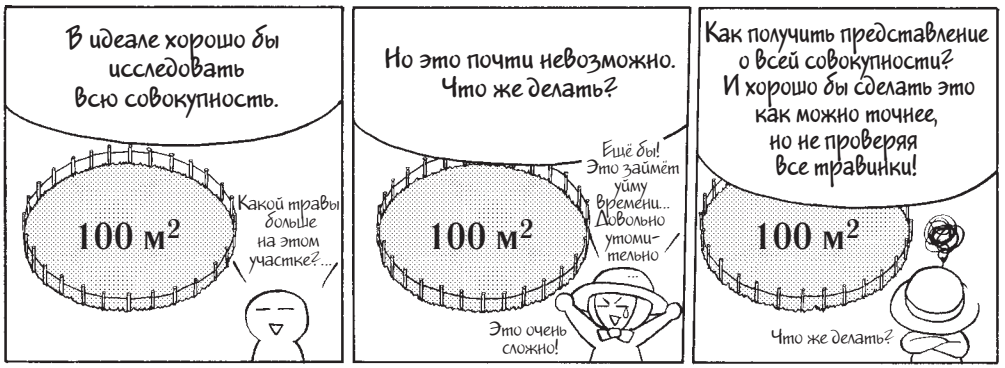

Руи, сколько примерно в Японии людей, имеющих избирательное право?
JAPAN
Ммм... Полно! Много!


Верно, действительно много. Если определять рейтинг на основе опроса всего населения, цифры будут более точными.
Да.
ЗА
ПРОТИВ
Один, два, три, четыре
Один, два, три, четыре
ЗА
ПРОТИВ
Однако нереально опросить всех.
Согласна.
Невозможно!
И даже бессмысленно!
Согласен, невозможно
ЗА
ПРОТИВ

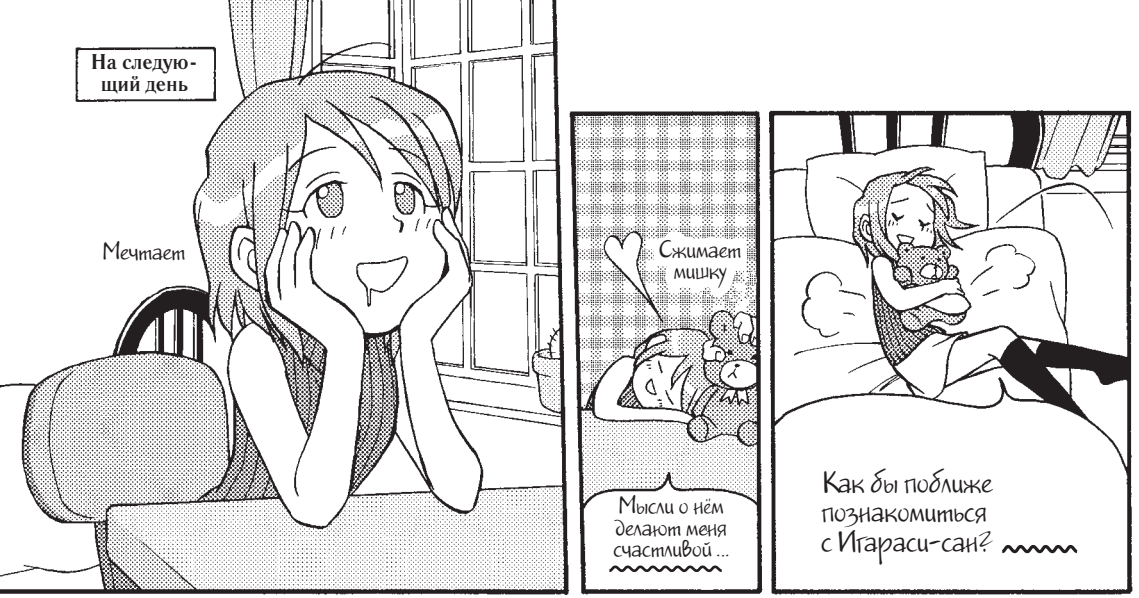

Поэтому опрос проводят среди ограниченного числа респондентов.
Поня...тно

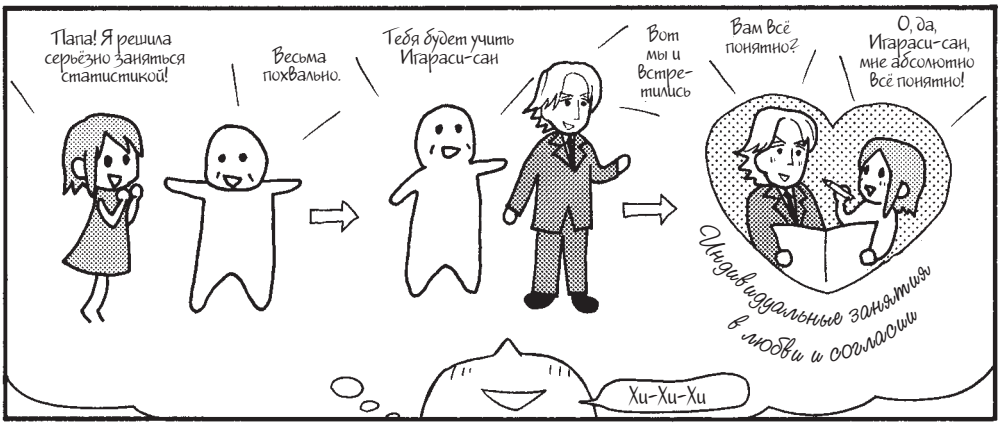

В статистике исследуемые объекты или явления называются генеральной совокупностью, а часть объектов такой совокупности, отобранных для её изучения, называют выборочной совокупностью (или выборкой).
Генеральная совокупность...
Выборочная совокупность...
Папа, ты что, надо мной издеваешься, говоря такие сложные вещи?!
Плачет
Я не хотел тебя расстраивать

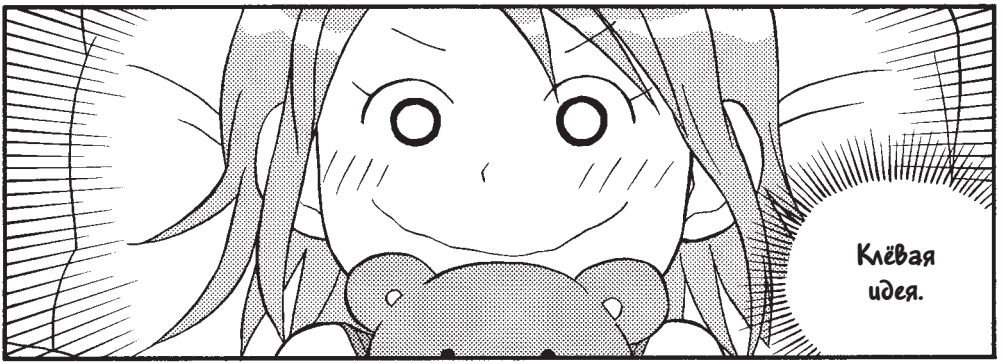

Подожди, всё не так сложно. Смотри, если вернуться к разговору о рейтинге кабинета министров, генеральная совокупность — это все люди, обладающие избирательным правом.
Было опрошено 2000 человек, и, следовательно, они представляют выборочную совокупность.
Все, имеющие избирательное право.
Выборка
2000 человек
Генеральная совокупность
Выборочная совокупность
Вот теперь понятно!

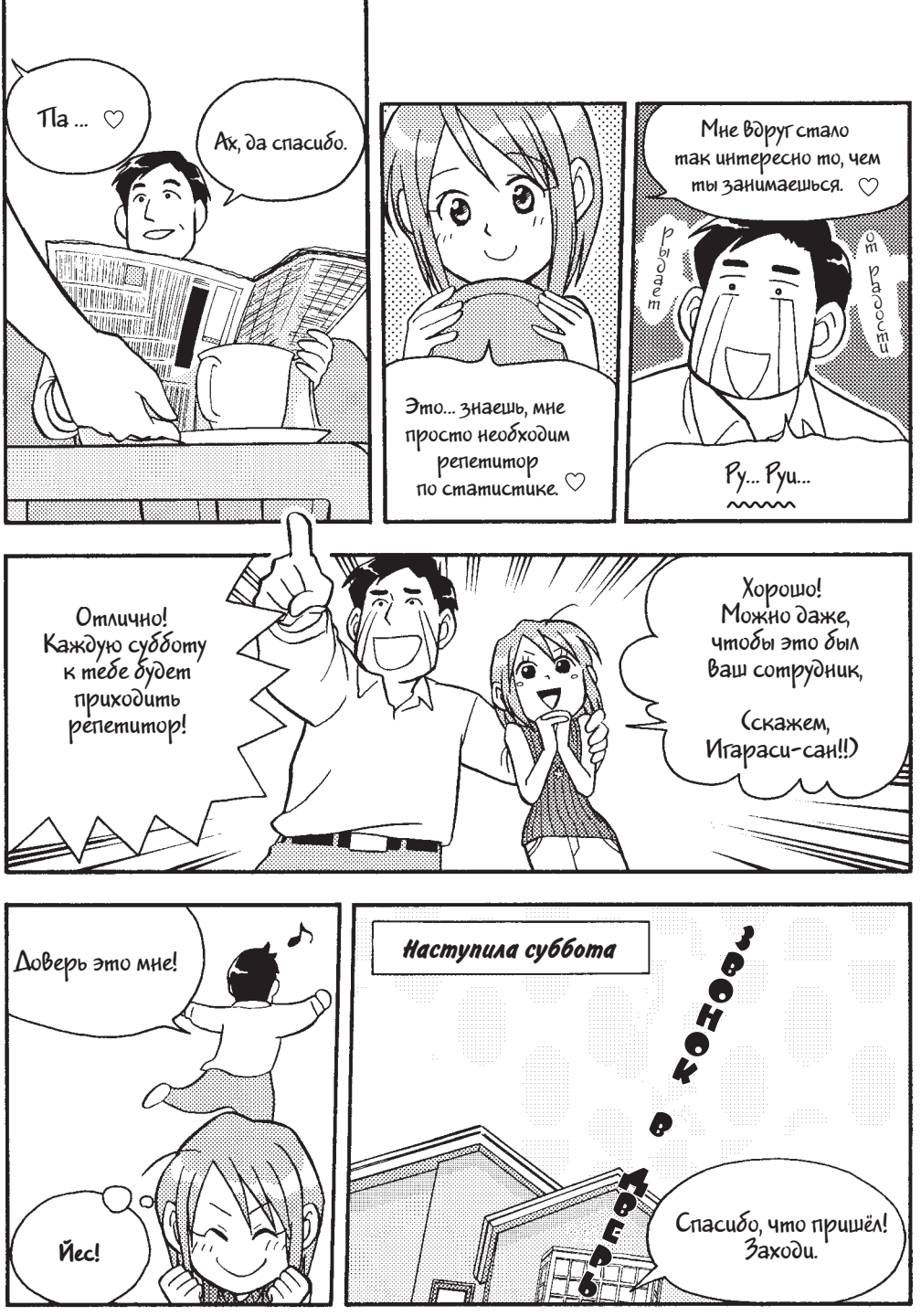

В идеале хорошо бы исследовать всю совокупность.
100 м²
Какой травы больше на этом участке?...
Но это почти невозможно. Что же делать?
100 м²
Ещё бы! Это займёт уйму времени... Довольно утомительно
Это очень сложно!
Как получить представление о всей совокупности? И хорошо бы сделать это как можно точнее, но не проверяя все травинки!
100 м²
Что же делать?


1 м²
Ограничимся 1 м² и воспользуемся помощью статистики!
Кажется, я начинаю понимать. И что дальше?
Нет-нет. Это в следующий раз.
УЛЫБАЕТСЯ
Какой же он симпатяга...

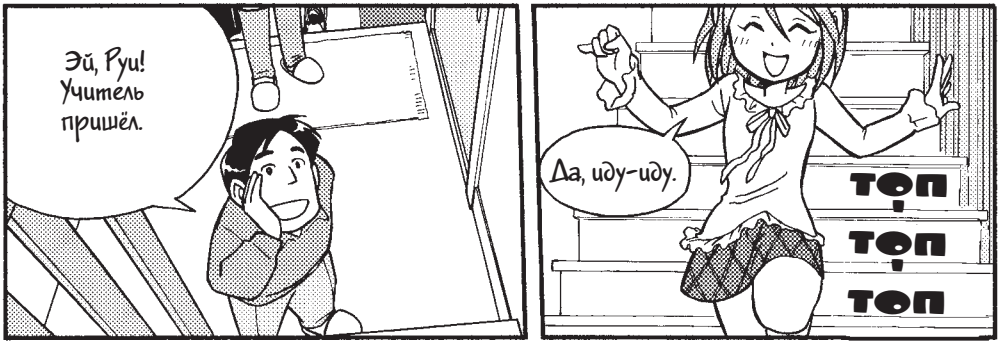

На следующий день
Мечтает
Сжимает мишку
Мысли о нём делают меня счастливой...
Как бы поближе познакомиться с Игараси-сан?

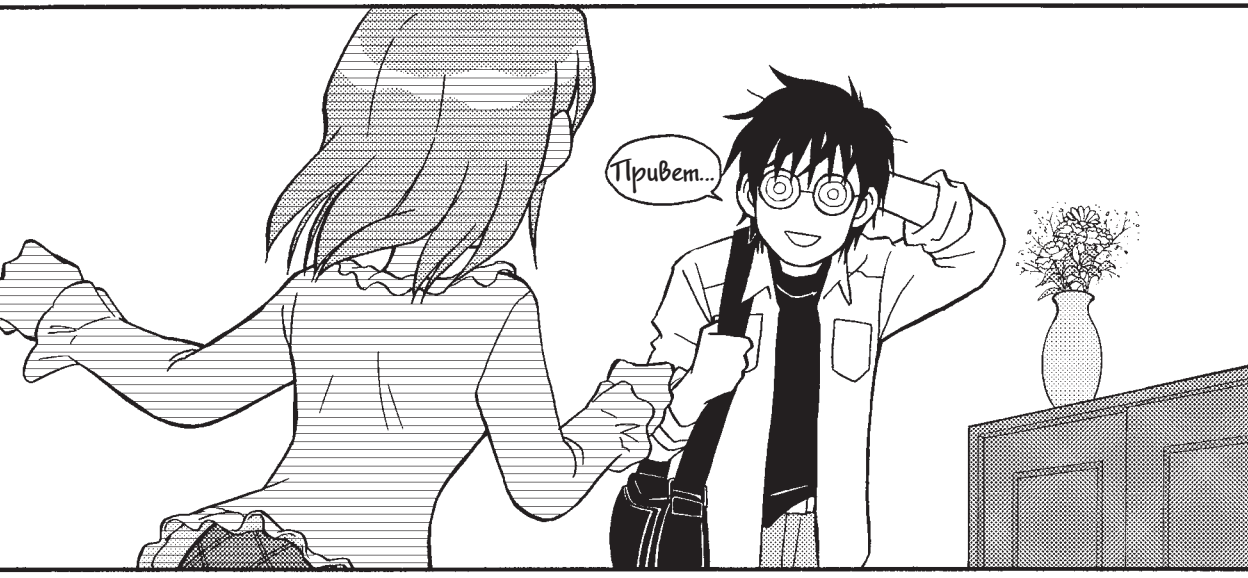

Папа! Я решила серьёзно заняться статистикой!
Весьма похвально.
Тебя будет учить Игараси-сан
Вот мы и встретились
Вам всё понятно?
О, да, Игараси-сан, мне абсолютно всё понятно!
Индивидуальные занятия в любви и согласии
Хи-Хи-Хи


Клёвая идея.


Па... ♡
Ах, да спасибо.
Это... знаешь, мне просто необходим репетитор по статистике. ♡
Мне вдруг стало так интересно то, чем ты занимаешься. ♡
рыдает
от радости
Ру... Руи...
Отлично! Каждую субботу к тебе будет приходить репетитор!
Хорошо! Можно даже, чтобы это был ваш сотрудник,
(скажем, Игараси-сан!!)
Доверь это мне!
Йес!
Наступила суббота
звонок в дверь
Спасибо, что пришёл! Заходи.

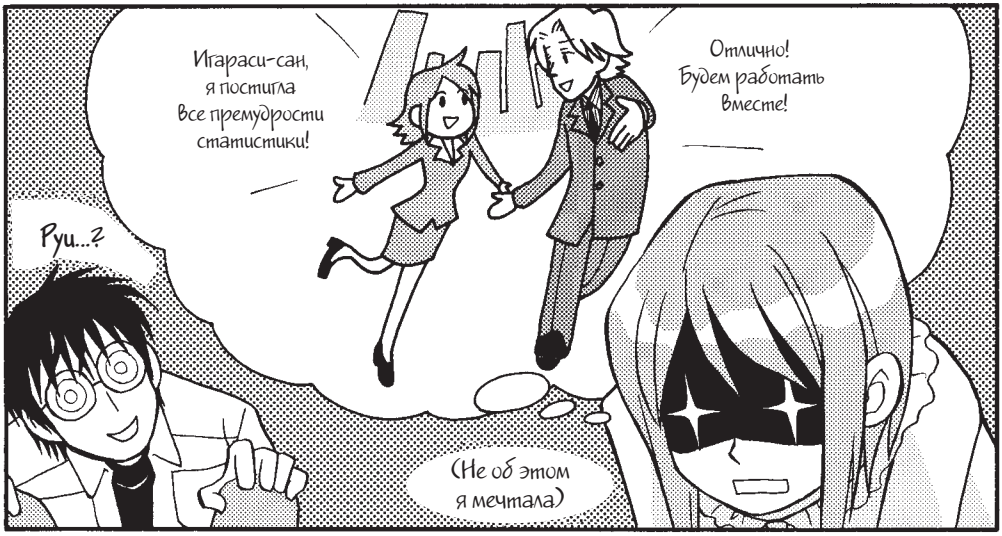

Пришёл!


Эй, Руи! Учитель пришёл.
Да, иду-иду.
топ
топ
топ


Привет...


Кто этот парень?!!
О, нет!
Руи, это наш сотрудник, Мамору Ямамото.
Какой-то растрёпанный!
Очень приятно!
Папа, а Игараси-сан?
Причём тут Игараси? Мамору и живёт ближе, и учить будет не хуже.
Ну, я пошёл, счастливо позаниматься.
Хе-хе-хе!


Ну что, Руи, начнём?
Это дурной сон
У-У-У...

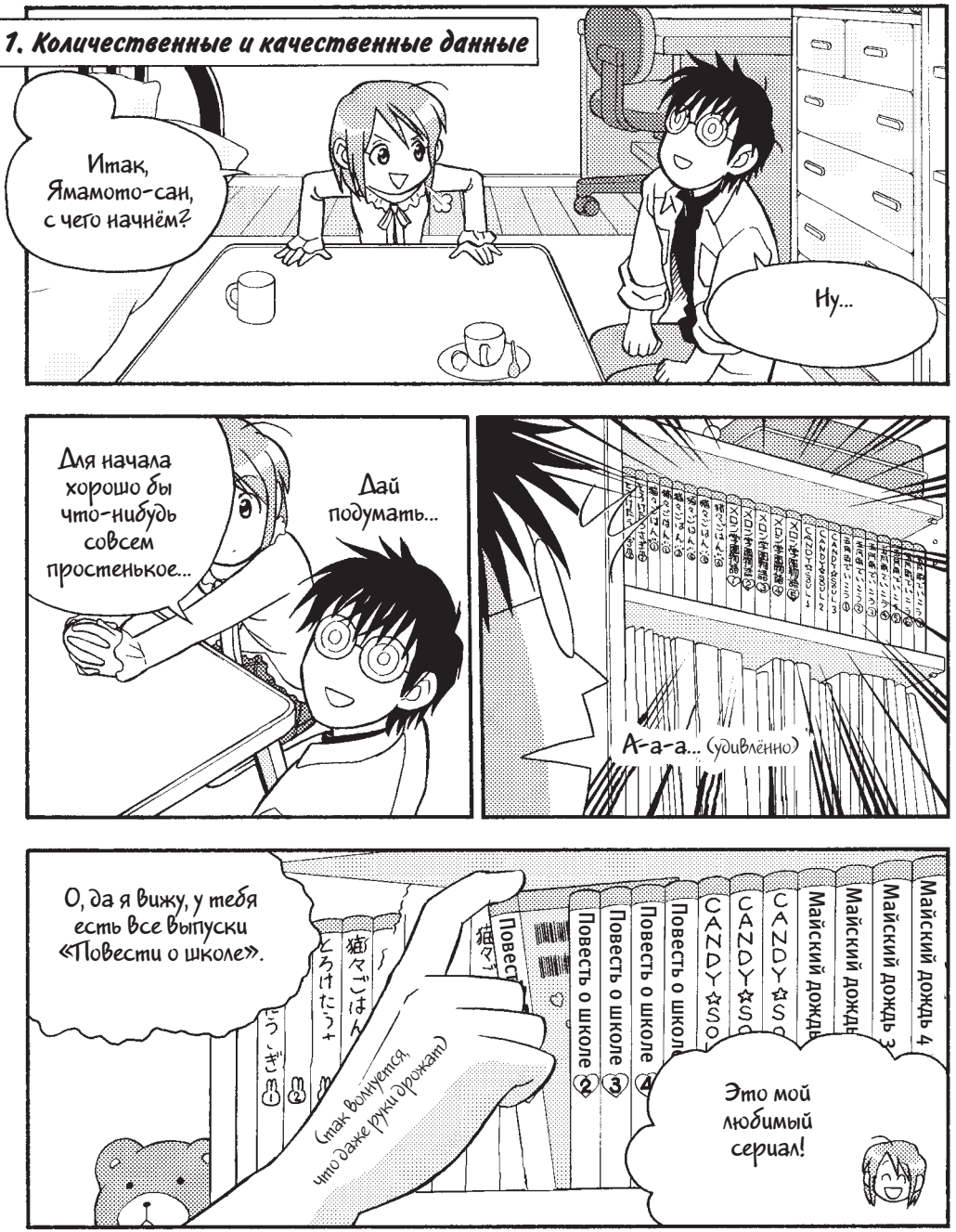

Игараси-сан, я постигла все премудрости статистики!
Руи...?
Отлично! Будем работать вместе!
(Не об этом я мечтала)

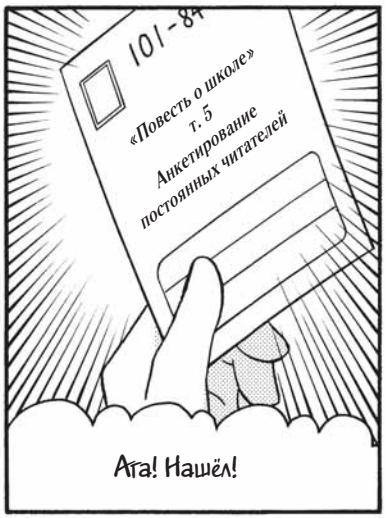

Идея!
?
Он поможет мне выучить статистику, и я стану ближе к Игараси-сан !!!
Никогда не сдавайся!
Начнём прямо сейчас!
О ... Ok.
Так начались их занятия...

Глава 1

Разберёмся с типами данных

1. Количественные и качественные данные





Повесть о школе, т. 5
Анкета постоянных читателей

Вопрос 1. Ваше мнение о 5-м томе «Повести о школе»?

1. Очень интересно
2. Довольно интересно
3. Так себе
4. Скучновато
5. Совершенно неинтересно

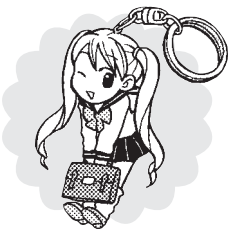
Вопрос 2. Ваш пол?

1. ж 2. м

Вопрос 3. Ваш возраст? ___ лет

Вопрос 4. Сколько выпусков журнала вы приобретаете в месяц? ___ шт

Среди участников анкетирования будут разыграны 30 брелоков «Рина»!

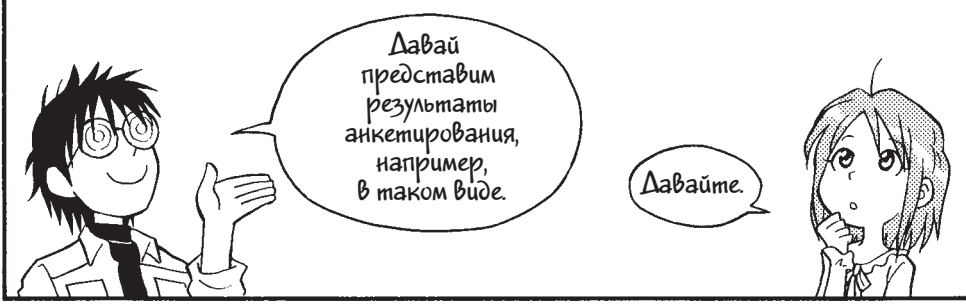


Спасибо за Ваши ответы. Ваше ценное для нас мнение мы учтём в последующих изданиях, а также при разработке будущих проектов.

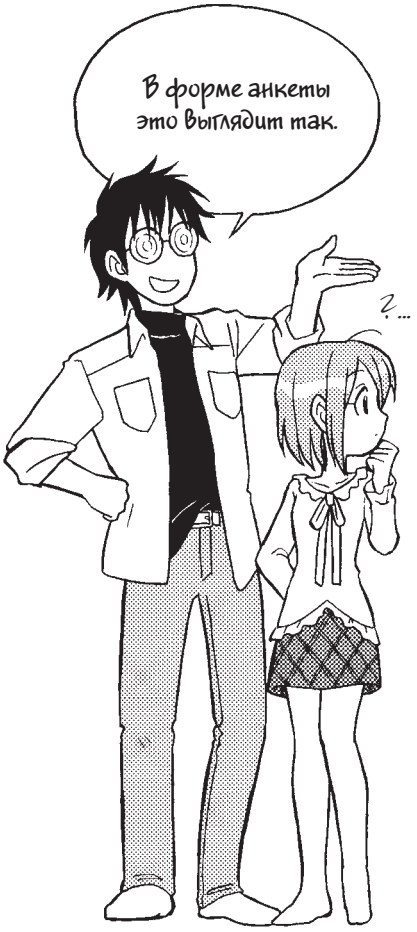

Кому нужна эта анкета?
Если её отправить ...
Вспышка
101-8460
... можно получить, брелок «Рина» если, конечно, повезёт.
Какой ты глупый! А я тебе поверила.
Я это я сделала.
Я... постараюсь...
Но ты мне нравишься...
Как же хочется получить брелок... Да, Руи?
М-ммм... Да мне как-то...
Вернёмся к статистике ...
Так сразу?..
Для начала ответь на вопросы анкеты.
Постараюсь.
Готово!
(что-то пишет)
Вот и хорошо.

Анкета постоянных читателей

Респондент	Ваше мнение о «Повести о школе»	Пол	Возраст, лет	Кол-во приобретаемых в месяц выпусков, шт.
Фуи	очень интересно	ж	17	2
А	довольно интересно	ж	17	1
Б	так себе	м	18	5
В	скучновато	м	22	7
Г	довольно интересно	ж	25	4
Д	совершенно неинтересно	м	20	3
Е	очень интересно	ж	16	1
Ж	довольно интересно	ж	17	2
З	так себе	м	18	0
И	так себе	ж	21	3
⋮	⋮	⋮	⋮	⋮




Меня всё-таки беспокоит, что думают читатели об этой книге!
Прямо одержимый какой-то...
Ну, хорошо, оставим это. Тема сегодняшнего занятия — типы данных.
(садится)
Что это значит?
Данные можно поделить на две группы:
которые нельзя измерить ...
... и которые можно измерить
КАЧЕСТВЕННЫЕ
КОЛИЧЕСТВЕННЫЕ



«Повесть о школе», т. 5
Анкета постоянных читателей

Вопрос 1. Ваше мнение о 5-м томе «Повести о школе»?

1. Очень интересно
2. Довольно интересно
3. Так ...
4. Скучновато
5. Совершенно неинтересно

Качественные данные

Вопрос 2. Ваш пол?

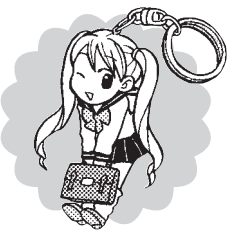
1 (ж) 2. м

Вопрос 3. Ваш возраст? 17 лет

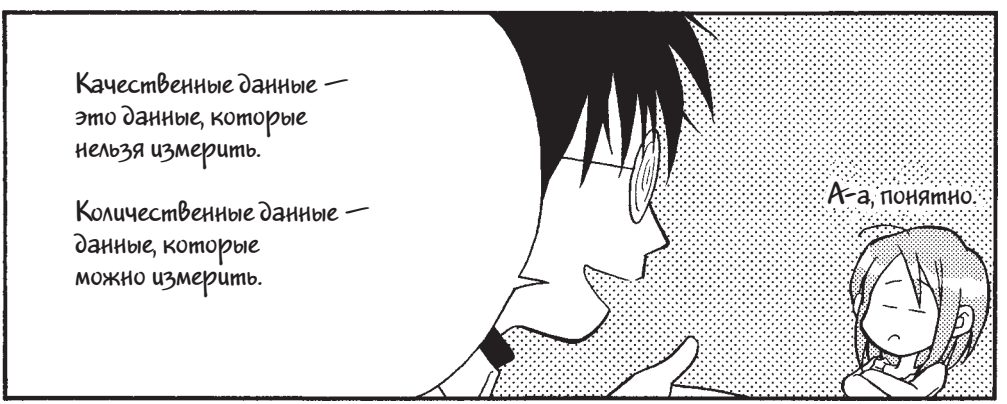
Количественные данные

Вопрос 4. Сколько выпусков журнала ... приобретаете в месяц? 2 шт

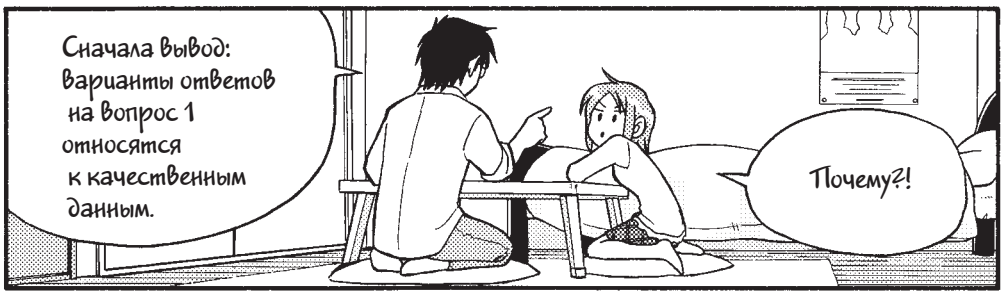
Среди участников анкетирования будут разыграны 30 брелоков «Рина»!

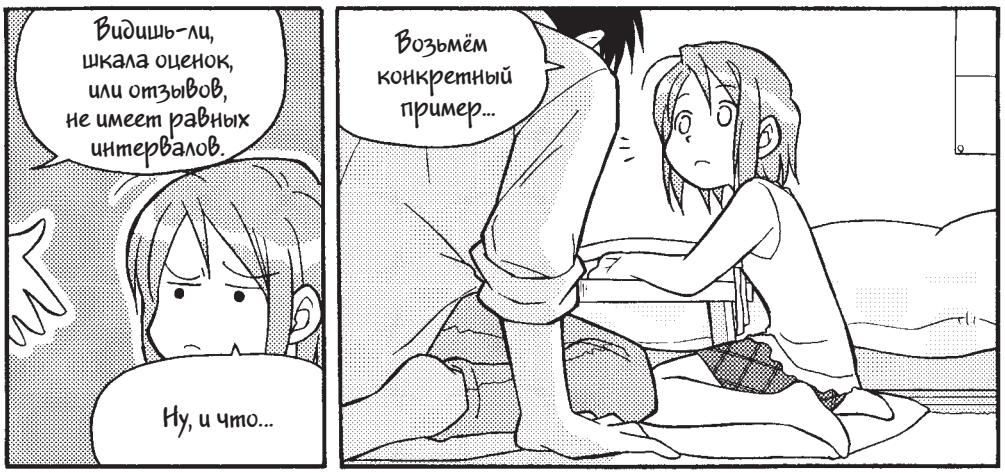


Спасибо за Ваши ответы. Ваше ценное для нас мнение мы учтём в последующих изданиях, а также при разработке будущих проектов.

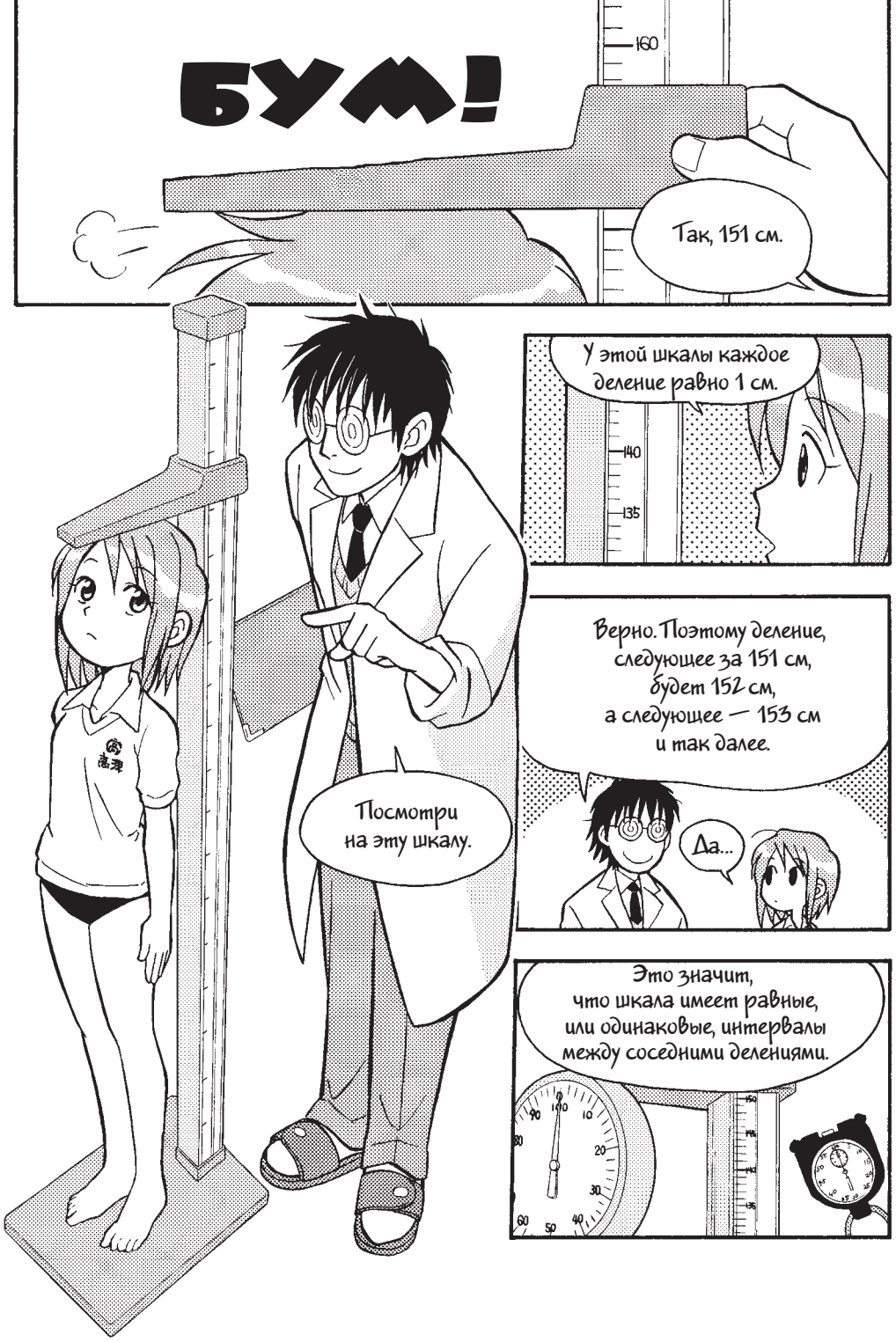


2. Примеры качественных данных






Руи ...
Что??
Сколько ты весишь?
Что?!
Да как Вы смеете задавать девушке такие вопросы? Это просто возмутительно!
Извини.
А какой у тебя рост, можно спросить?
151 см!
Вот с этого и надо было начинать!
Слава богу, ты уже не сердишься.
Для измерения роста используют так называемый ростомер, так?
Да!

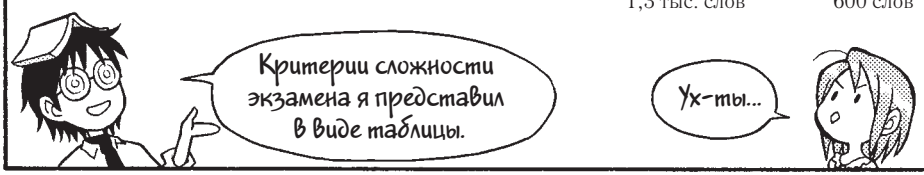

БУМ!
160
Так, 151 см.
У этой шкалы каждое деление равно 1 см.
140
135
Верно. Поэтому деление, следующее за 151 см, будет 152 см, а следующее — 153 см и так далее.
Да...
Посмотри на эту шкалу.
Это значит, что шкала имеет равные, или одинаковые, интервалы между соседними делениями.

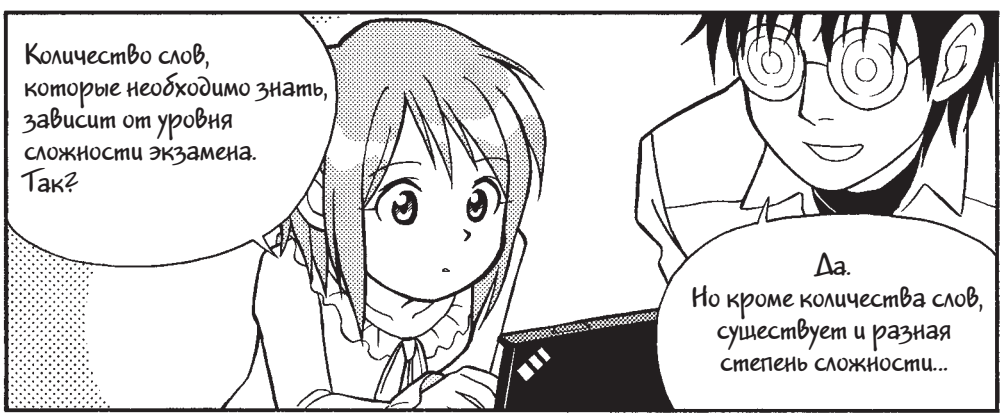

Следовательно, рост можно измерить и он, таким образом, относится к количественным данным.
Ура, поняла!
Теперь перейдём к следующему примеру.
Что теперь?
Открываем...
Так...
Google
Ой, у тебя обои из «Повести о школе»
Руи, ты же сдавала квалификационный экзамен по английскому языку, который проводит общество для тестирования знания английского языка?
Об этом мне сказал твой папа.
Постойте, Вы это о чём?
Короче, ты сдавала экзамен Eiken?
Да, у меня второй уровень.

Критерии сложности экзамена Eiken

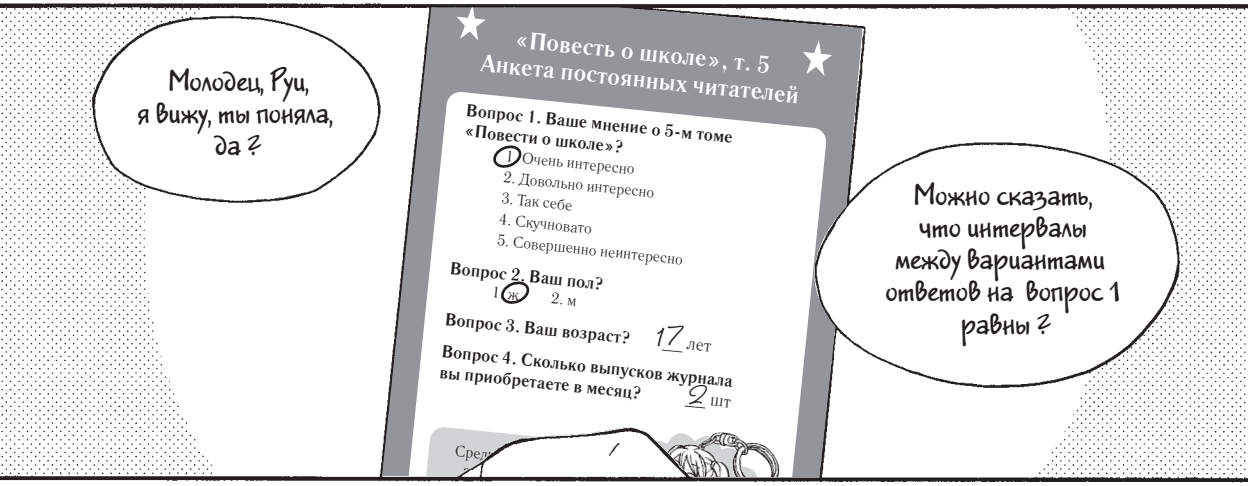
http://www.eiken.or.jp/

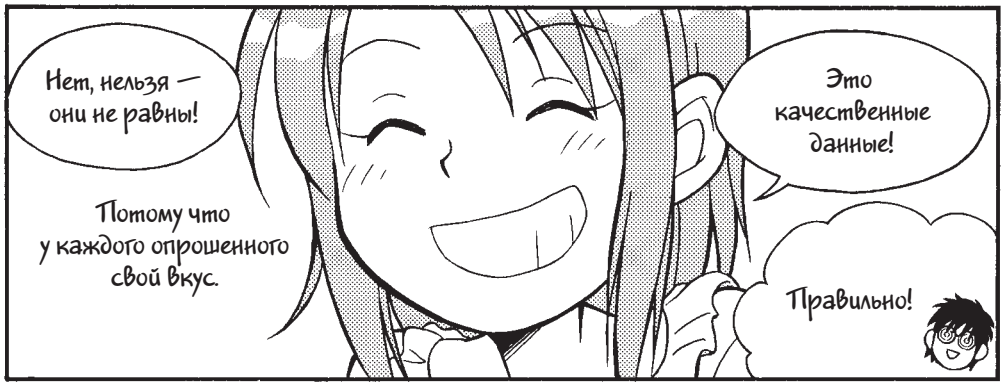
1-й уровень	2-й уровень	3-й уровень	4-й уровень	5-й уровень
Продвинутый, соответствует программе вуза; словарный запас примерно 10 —15 тыс. слов	Соответствует программе средней школы; словарный запас примерно 5,1 тыс. слов	Соответствует программе 3-го года обучения средней школы; словарный запас примерно 2,1 тыс. слов	Средний, соответствует программе 2-го года обучения средней школы; словарный запас примерно 1,3 тыс. слов	Начальный, соответствует программе 1-го года обучения средней школы; словарный запас примерно 600 слов

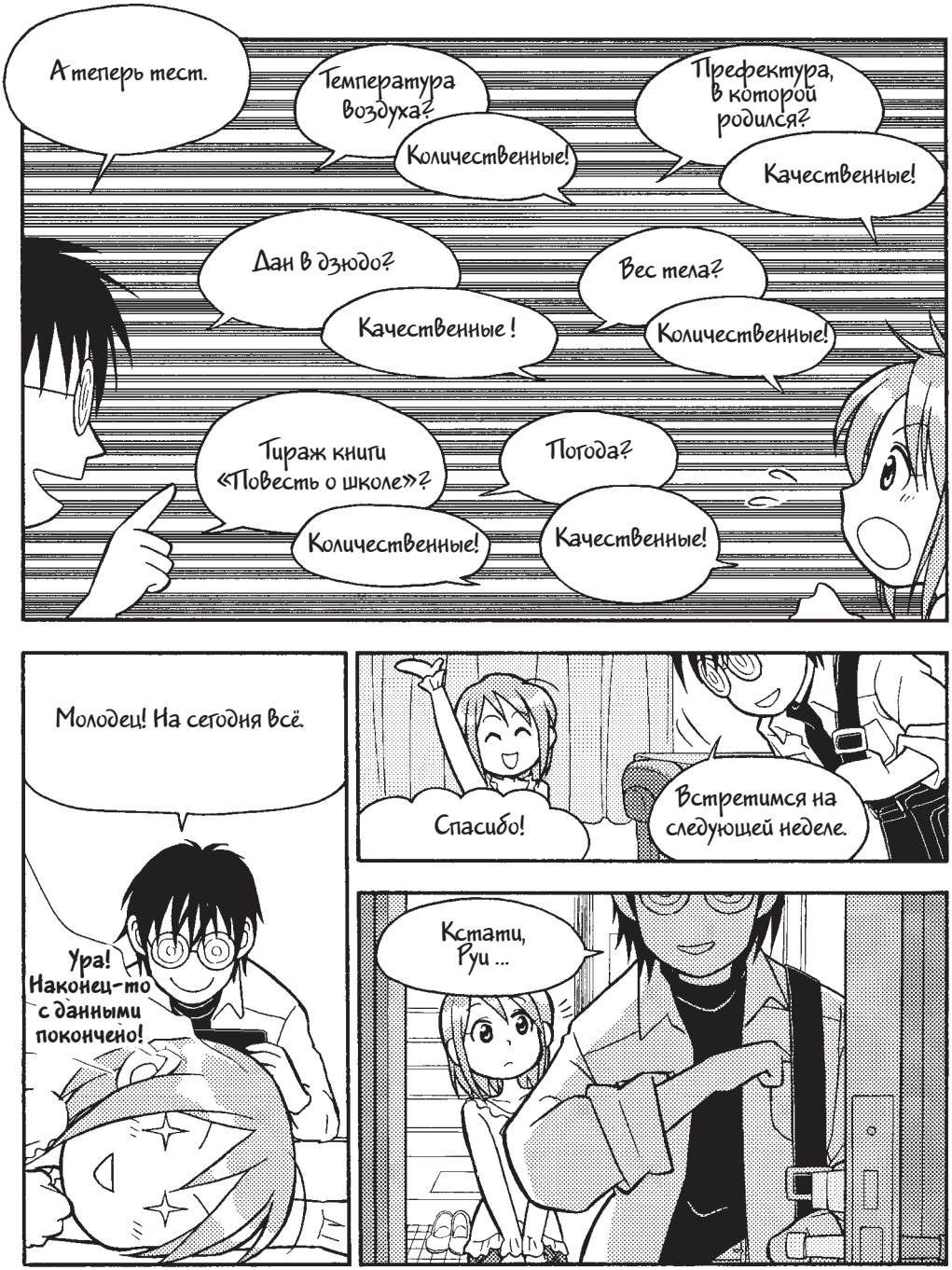


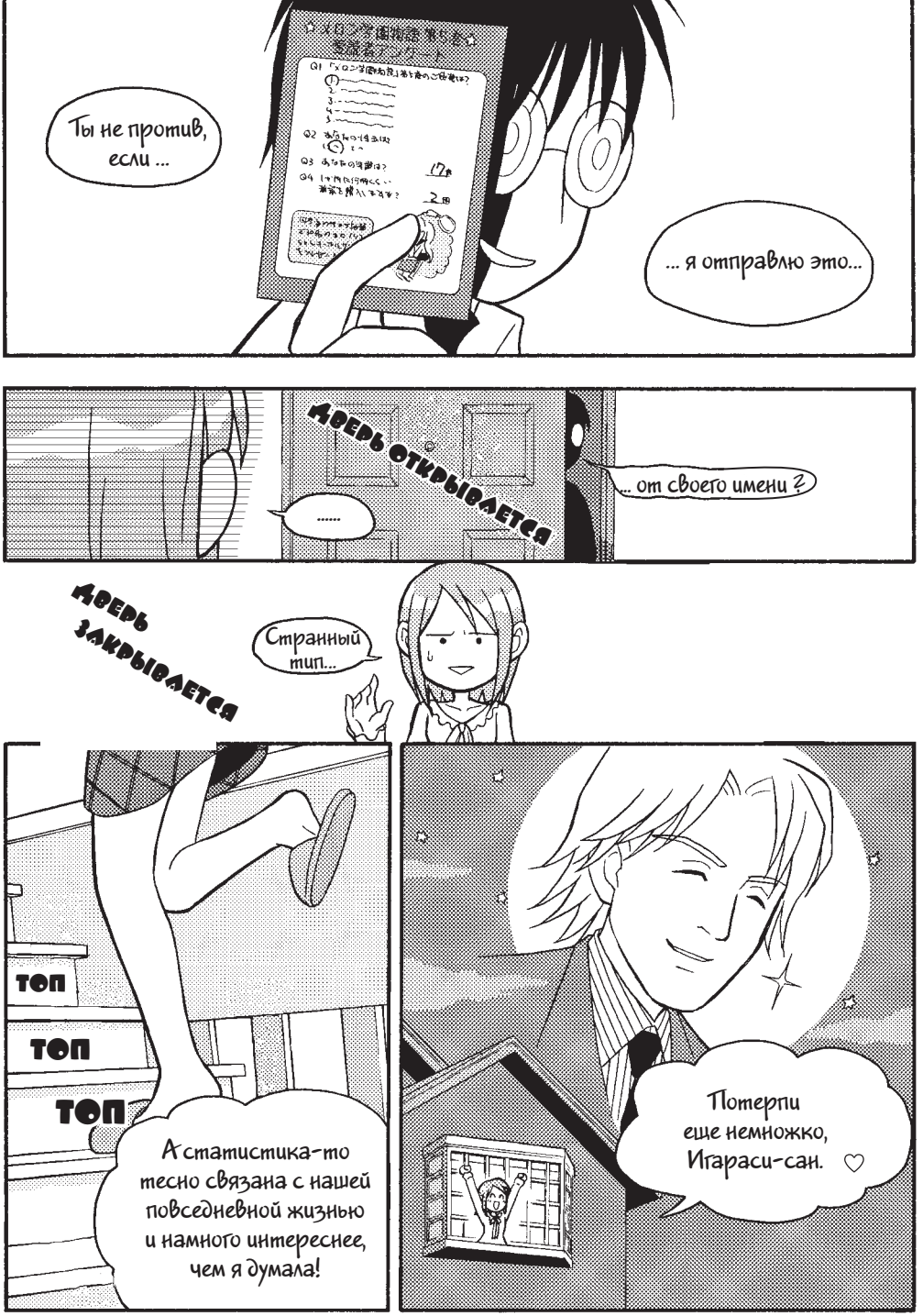



Вот поэтому интервалы между уровнями не могут быть одинаковыми.
Значит, ...
... уровни экзамена нельзя измерить и они являются качественными данными!
Ясно!


Молодец, Руи, я вижу, ты поняла, да?
«Повесть о школе», т. 5
Анкета постоянных читателей
Вопрос 1. Ваше мнение о 5-м томе «Повести о школе»?
1. Очень интересно
2. Довольно интересно
3. Так себе
4. Скучновато
5. Совершенно неинтересно
Вопрос 2. Ваш пол?
1. Ж 2. М
Вопрос 3. Ваш возраст? 17 лет
Вопрос 4. Сколько выпусков журнала вы приобретаете в месяц? 2 шт
Можно сказать, что интервалы между вариантами ответов на вопрос 1 равны?


Нет, нельзя — они не равны!
Потому что у каждого опрошенного свой вкус.
Это качественные данные!
Правильно!


А теперь тест.
Температура воздуха?
Количественные!
Префектура, в которой родился?
Качественные!
Дан в дзюдо?
Качественные !
Вес тела?
Количественные!
Тираж книги «Повесть о школе»?
Количественные!
Погода?
Качественные!
Молодец! На сегодня всё.
Ура! Наконец-то с данными покончено!
Спасибо!
Встретимся на следующей неделе.
Кстати, Руи ...


Ты не против, если ...
... я отправлю это...
Дверь открывается
.....
... от своего имени?
Дверь закрывается
Странный тип...
ТОП
ТОП
ТОП
А статистика-то тесно связана с нашей повседневной жизнью и намного интереснее, чем я думала!
Потерпи еще немножко, Игараси-сан. ♡

3. Использование многовариантных ответов на практике

Как было показано, вопрос 1 анкеты постоянных читателей относится к качественным данным. Однако на практике, например при опросе потребителей, эти данные часто рассматриваются как количественные. Другими словам, возможны случаи, когда это может выглядеть так:

	балл
очень интересно	5
довольно интересно	4
так себе	3
скучновато	2
совершенно неинтересно	1

или так:

	балл
очень интересно	2
довольно интересно	1
так себе	0
скучновато	−1
совершенно неинтересно	−2

Существуют мир теории и мир практики, точнее, теоретический мир и реальный. Поэтому одни и те же данные могут рассматриваться и как количественные, и как качественные: всё зависит от того, где они используются — в теории или на практике.

Упражнение

Посмотрите на таблицу:

Респондент	Группа крови	Оценка вкусовых качеств спортивного коктейля X	Комфортная комнатная температура при работающем кондиционере, °C	Лучший результат бега на 100 м, с
А	B (III)	невкусно	25	14,1
Б	A (II)	вкусно	24	12,2
В	AB (IV)	вкусно	25	17,0
Г	O (I)	так себе	27	15,6
Д	A (II)	невкусно	24	18,4
...	...	...	...	...

Определите, к каким категорям данных относятся графы: «Группа крови», «Оценка вкусовых качеств спортивного коктейля X», «Комфортная комнатная температура при работающем кондиционере» и «Лучший результат бега на 100 м».

Ответ

«Группа крови» и «Оценка вкусовых качеств спортивного коктейля X» относятся к качественным данным.
«Комфортная комнатная температура при работающем кондиционере» и «Лучший результат бега на 100 м» относятся к количественным данным.

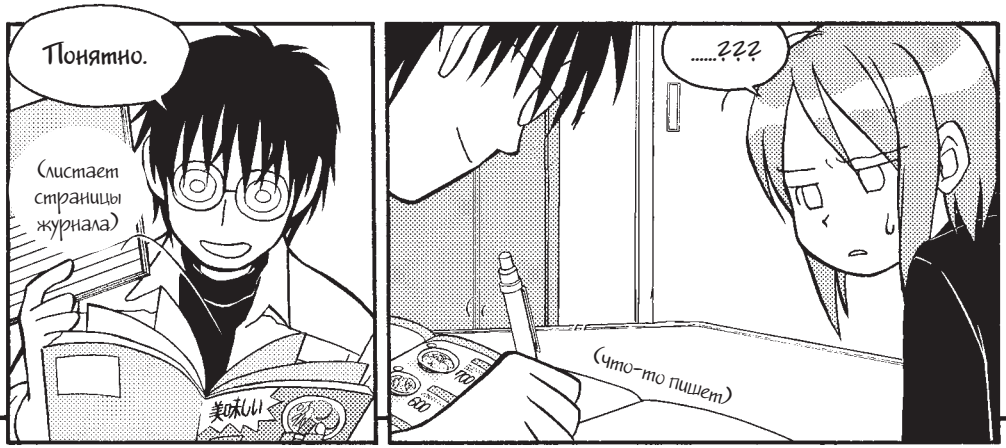
Выводы

Данные делятся на количественные и качественные. Такие категории данных, как, например «Очень интересно», ..., «Совершенно неинтересно» с точки зрения теории относятся к качественным данным. Однако на практике эти же данные могут рассматриваться и как количественные.

Глава 2

Знакомимся с количественными данными


1. Ряды распределения и гистограммы
ВАЖНО
* Вкусный рамэн**. 50 лучших ресторанов
美味しい
ラーメン
best 50*
Дверь открывается
Привет, Руи.
** Рамэн — китайская лапша, блюдо ресторанов быстрого питания.
Очень! Посмотрела этот журнал и думаю, в какой ресторан лучше пойти.
A-a-a... (удивляется)
Всё так аппетитно, правда?
渋谷
500
890
880
700
890
Здравствуйте, Ямамото-сан.
Да, никак ты любишь рамэн?



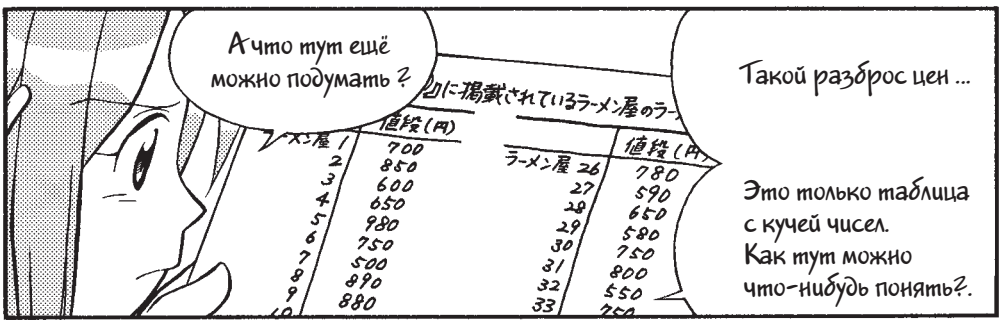
Для начала я свёл в таблицу цены на рамэн.

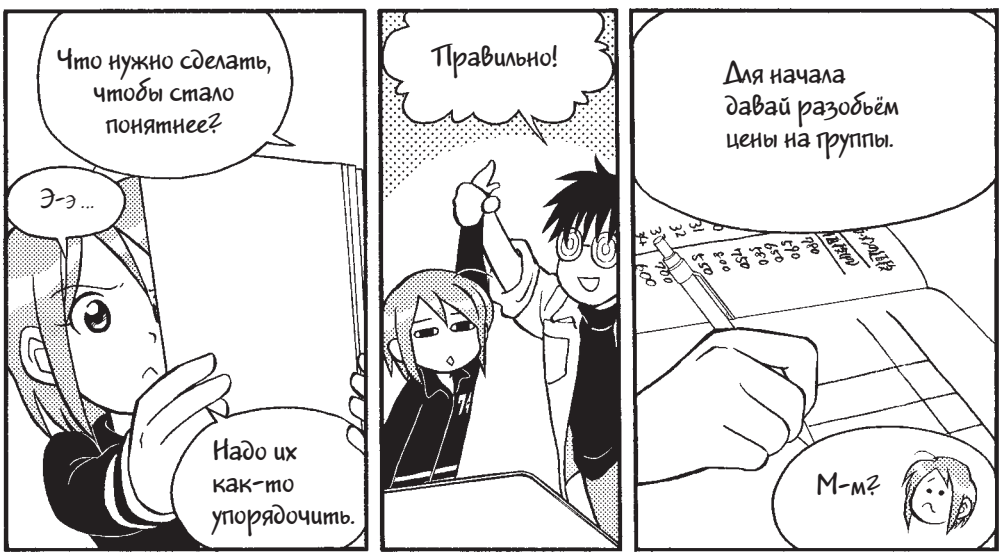
Цены на рамэн в 50 лучших ресторанах
(от. журнал «Вкусный рамэн. 50 лучших ресторанов»)

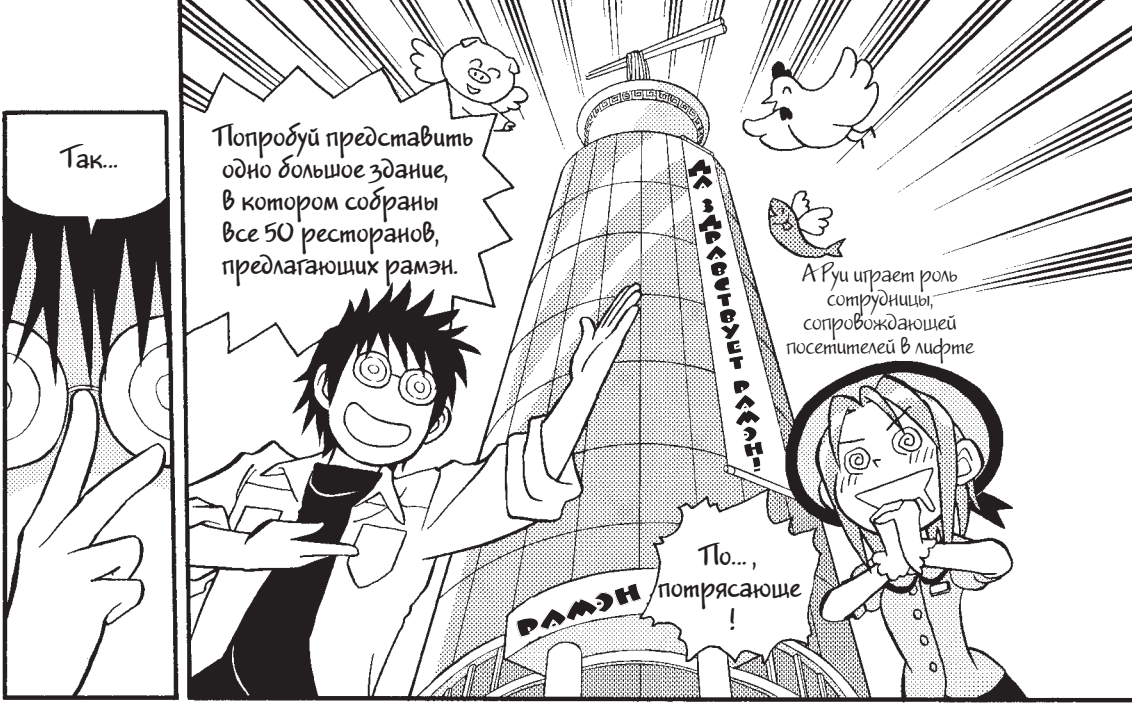
Ресторан	Цена, йены	Ресторан	Цена, йены
1	700	26	780
2	850	27	590
3	600	28	650
4	650	29	580
5	980	30	750
6	750	31	800
7	500	32	550
8	890	33	750
9	880	34	700
10	700	35	600
11	890	36	800
12	720	37	800
13	680	38	880
14	650	39	790
15	790	40	790
16	670	41	780
17	680	42	600
18	900	43	670
19	880	44	680
20	720	45	650
21	850	46	890
22	700	47	930
23	780	48	650
24	850	49	777
25	750	50	700

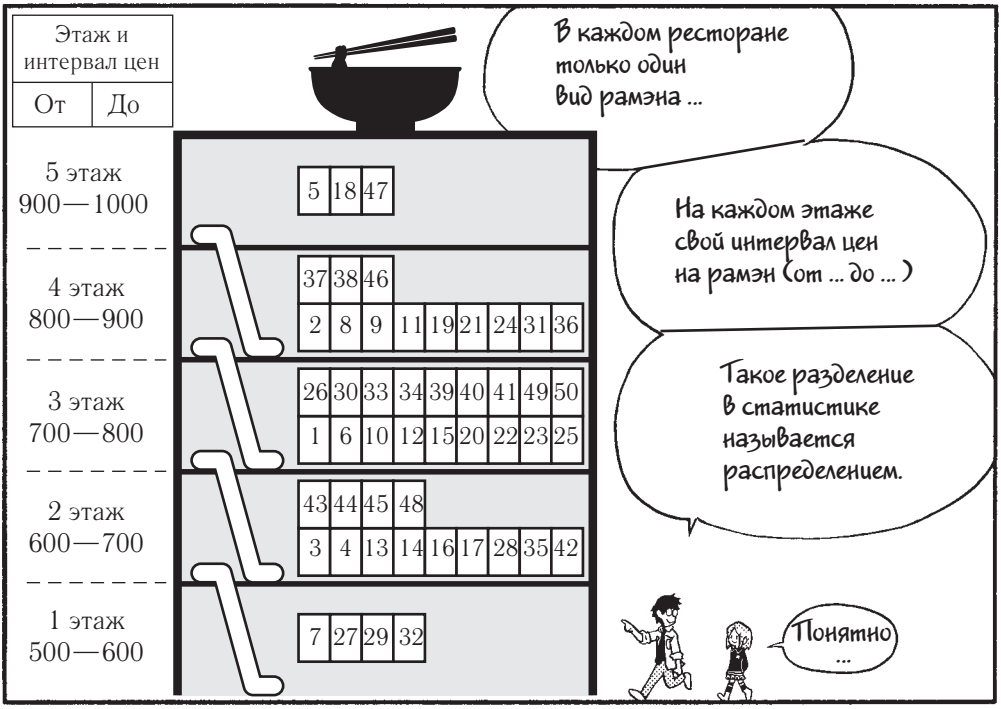
От обсуждения ресторанов плавно перешли к занятию... Всё-таки странный тип...


О чем ты думаешь, когда смотришь на эту таблицу?
М-ммм...
Хочется поесть рамэн
Что?!
И это всё? Подумай ещё.
Ещё?...


А что тут ещё можно подумать?
Такой разброс цен ...
Это только таблица с кучей чисел. Как тут можно что-нибудь понять?.


Что нужно сделать, чтобы стало понятнее?
Э-э...
Надо их как-то упорядочить.
Правильно!
Для начала давай разобьём цены на группы.
М-м?


Так...
Попробуй представить одно большое здание, в котором собраны все 50 ресторанов, предлагающих рамэн.
Я обожаю рамэн!
А Руи играет роль сотрудницы, сопровождающей посетителей в лифте
РАМЭН
По..., потрясающе!


Этаж и интервал цен
От
До
5 этаж 900—1000
5 18 47
4 этаж 800—900
37 38 46
2 8 9 11 19 21 24 31 36
3 этаж 700—800
26 30 33 34 39 40 41 49 50
1 6 10 12 15 20 22 23 25
2 этаж 600—700
43 44 45 48
3 4 13 14 16 17 28 35 42
1 этаж 500—600
7 27 29 32
В каждом ресторане только один вид рамэна ...
На каждом этаже свой интервал цен на рамэн (от ... до ...)
Такое разделение в статистике называется распределением.
Понятно ...


На каждом этаже есть вывеска, на которой указана средняя цена на рамэн на этом этаже.
2 этаж 650
ЛАПША
Добро пожаловать!
Рамэн 670
На втором этаже цены варьируются от 600 до 700 йен. Значит, средняя цена равна 650 йен!
Путеводитель по этажам
Номер этажа и интервал цен
Рестораны, расположенные на каждом этаже
Цифра на вывеске (середина интервала)
5 этаж 900...1000
950
4 этаж 800...900
850
3 этаж 700...800
750
2 этаж 600...700
650
1 этаж 500...600
550
Это называется серединой интервала
(смеется)
(Ямамото-сан в статусе сотрудницы, сопровождающей посетителей по этажам)
Рестораны распределены по этажам в соответствии с ценами на рамэн. На каждом этаже может быть разное количество ресторанов.
3F 700-800Y
750
2F 600-700Y
650
1F 500-600Y
Действительно.
На 1-м этаже — 4, на 2-ом этаже — 13, и т.д.
Число ресторанов, расположенных на каждом этаже, называют частотой.
74
На 3-м этаже больше всего ресторанов — 18!
Теперь вычислим относительную частоту ресторанов на 3-м этаже!


Относительная частота похожа на процентное отношение, которое должно быть тебе знакомо.
Относительную частоту ???
Это доля от всей совокупности, если всю совокупность принять за 1.

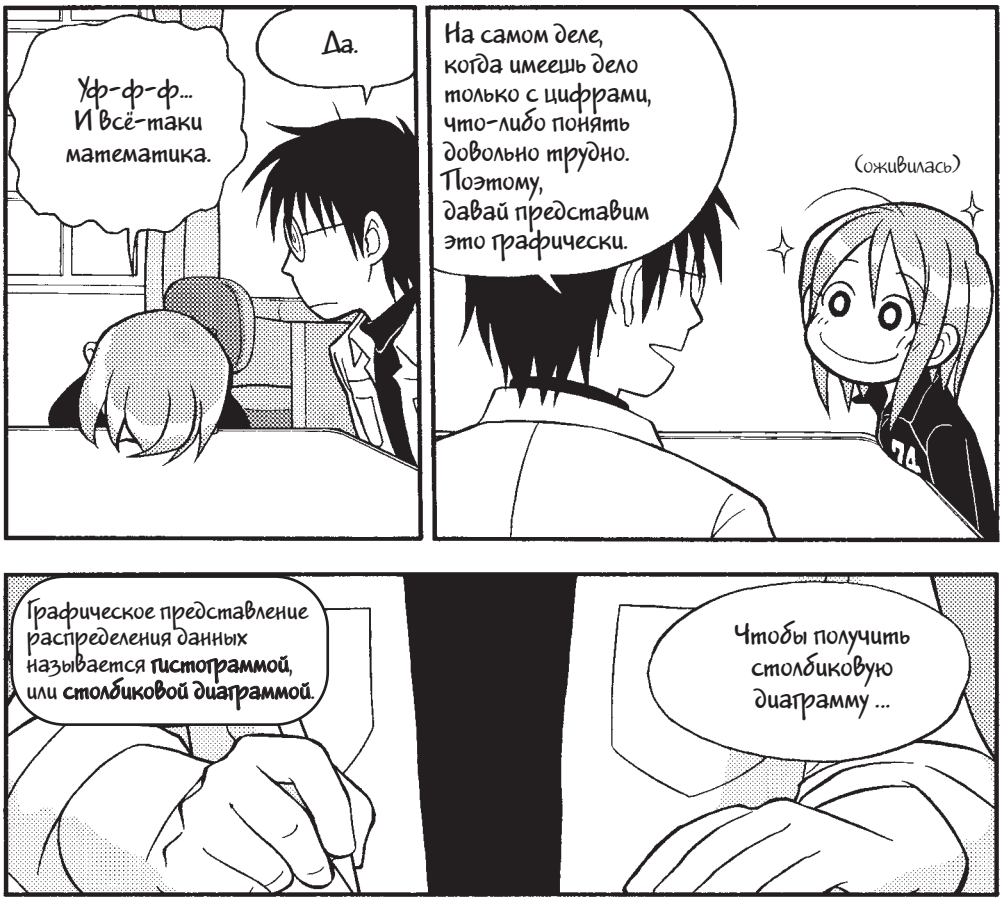

Есть вот такая формула:
Относительная частота = Часть совокупности / Вся совокупность


Так... Число ресторанов, расположенных на 3-м этаже, равно 18. Общее число ресторанов = 50. Следовательно...
(считает в уме)
18/50 = 36/100 = 0,36
Правильно?


Да!
Относительная частота ресторанов, предлагающих рамэн по цене от 700 до 800 йен, т.е. по средней цене 750 йен, равна 0,36.
Если данное значение умножить на 100, поскольку это доля, то получим 36%.
О, нет!
Это же математика ...

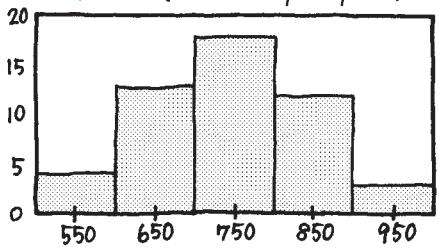
Распределение по цене 50 лучших ресторанов, предлагающих вкусный рамэн

Интервал цен	Середина интервала	Кол-во ресторанов, частота	Количество ресторанов, относительная частота
500 ~ 600	550	4	0.08
600 ~ 700	650	13	0.26
700 ~ 800	750	18	0.36
800 ~ 900	850	12	0.24
900 ~ 1000	950	3	0.06
Итого:		50	1.00

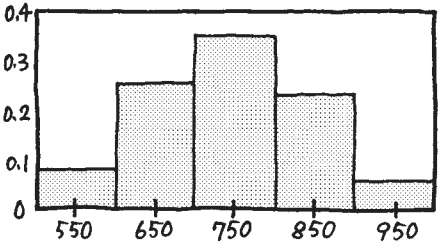


Гистограммы построены на основе таблицы распределения 50 лучших ресторанов, предлагающих рамэн

Гистограмма 1.
Частота (количество ресторанов)



Гистограмма 2.
Относительная частота



... по горизонтали, (на оси x), откладывают переменные.

В нашем случае это будут цены на рамэн.

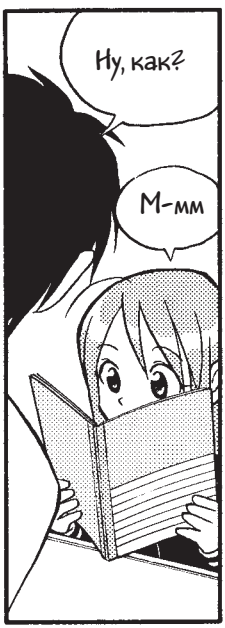
Ширина столбца равна величине интервала.

В середине столбца ставят значение, равное середине интервала.

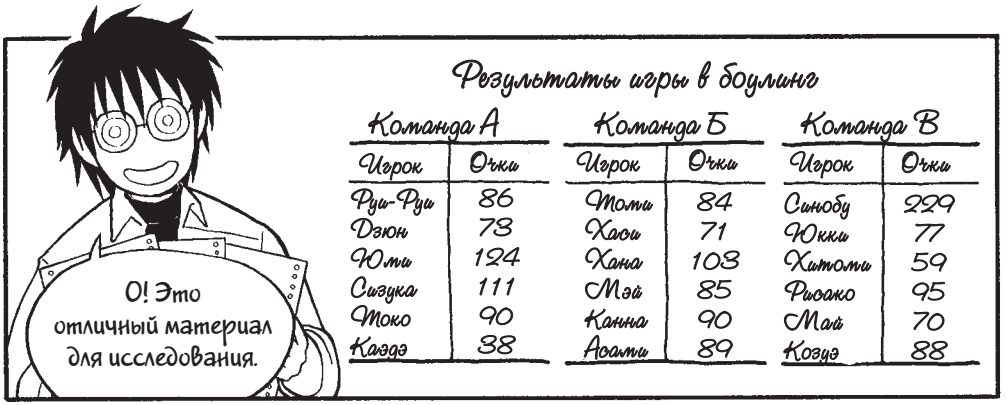
По вертикали (оси y) отложены:

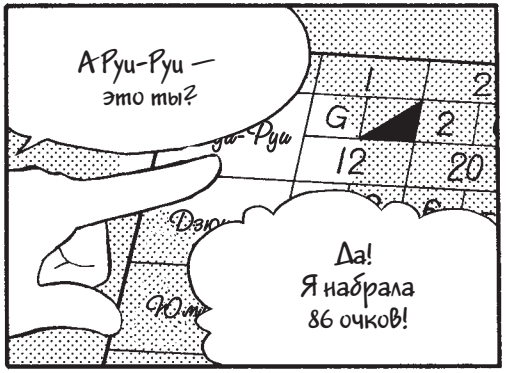
на верхнем рисунке — частота,

на нижнем рисунке — относительная частота




2. Средняя величина
Мы недавно с девочками из моего класса ходили в боулинг...
во время перерыва на чай
Удалось сбить хоть одну кеглю?
Что?! Да я ...
Я очень хорошо играю в боулинг!
74
Я пошутил.
Если все девочки класса, это довольно много, не так ли?
А
Б
В
Да, 18 девчонок. Поэтому мы разделились на три команды по 6 человек.
Смотри, вот таблица результатов игры.
(быстро достаёт таблицу)

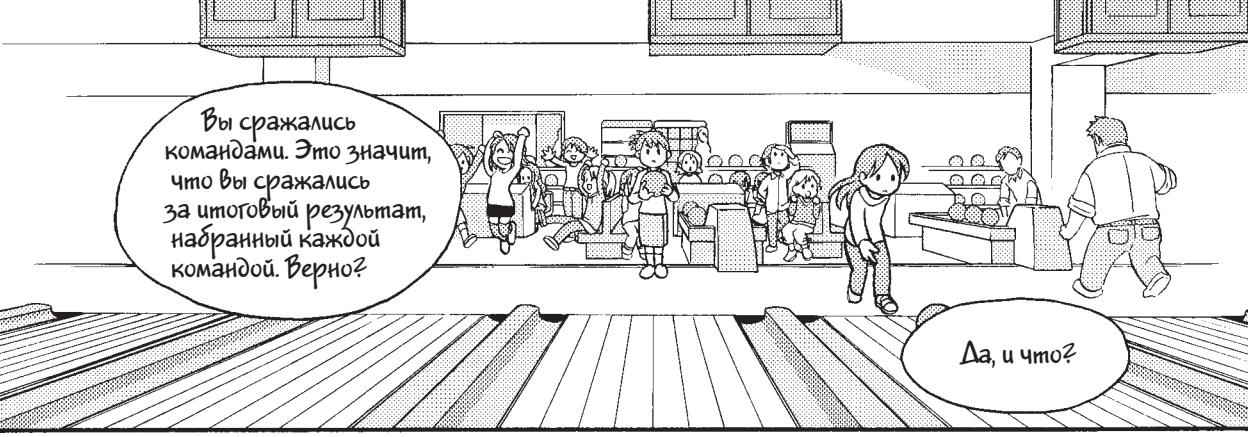

О! Это отличный материал для исследования.
Результаты игры в боулинг
Команда А
Игрок | Очки
Руи-Руи | 86
Дзюн | 73
Юми | 124
Сизука | 111
Токо | 90
Каэдэ | 38
Команда Б
Игрок | Очки
Томи | 84
Хаси | 71
Хана | 108
Мэй | 85
Канна | 90
Аоами | 89
Команда В
Игрок | Очки
Синобу | 229
Юкки | 77
Хитоми | 59
Рисако | 95
Маи | 70
Козуэ | 88


А Руи-Руи — это ты?
Да! Я набрала 86 очков!


Беглый просмотр результатов позволяет сделать вывод, что у тебя, Руи, был средний результат в команде, да?
И что с того?!


Средний означает результат, который в среднем набрал один человек в каждой команде. Понятно?
Понятно. Это результат, который находится посередине других результатов, набранных игроками команды, так?


Если мой результат окажется выше среднего, Вы угостите меня пирожным! ♡
Может быть ...
Давай попробуем вычислить среднюю величину.



Команда А

$$\frac{86+73+124+111+90+38}{6}=\frac{522}{6}=87$$

Команда Б

$$\frac{84+71+103+85+90+89}{6}=\frac{522}{6}=87$$

Команда В

$$\frac{229+77+59+95+70+88}{6}=\frac{618}{6}=103$$


И, пожалуйста, не называйте меня, Руи-Руи!
Извини...
Ус...успокойся.
Слушай, а что если вместо пирожного я тебя «угощу» кое-чем другим.
Например?
Например, средней величиной, о которой я тебе только что рассказывал. Кстати, она называется средней арифметической.
Кроме неё есть ещё средняя геометрическая и средняя гармоническая. Формулы оставим на десерт, а сами термины тебе придется запомнить. Хорошо?
Средняя геометрическая
$\sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$
Средняя гармоническая
$\frac{1}{\left(\frac{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}{n}\right)}$
Лучше бы Вы угостили меня пирожным.

Давай посмотрим на команду В.

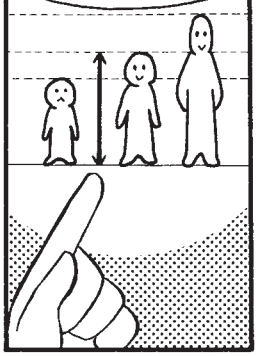
Как ты думаешь,

Результаты игры в боулинг

Команда А		Команда Б		Команда В	
Игрок	Очки	Игрок	Очки	Игрок	Очки
Фуи-Фуи	86	Тоти	84	Синобу	229
Дэюн	73	Хаси	71	Юкки	77
Юми	124	Хана	103	Хитоми	59
Сизука	111	Мэй	85	Рюако	95
Токо	90	Канна	90	Май	70
Каэдэ	38	Асами	89	Козуэ	88

правильно ли считать средним результатом команды В 103 очка?

Медиана — значение, которое приходится на середину ряда, если расположить данные в порядке возрастания (или убывания).



Для начала попробуем расположить в ряд очки, набранные игроками каждой команды.

Команда А

38	73	86	90	111	124

Команда Б

71	84	85	89	90	103

Команда В

59	70	77	88	95	229

Ряд с нечётным числом элементов

−1041,6	−39,0	**−5,7**	60,4	77,3

медиана

Ряд с чётным числом элементов

−0,4	35,2	**37,8**	**42,2**	46,1	910,3

Среднее значение этих двух элементов будет медианой

Если ряд состоит из нечётного числа элементов, медианой будет значение, находящееся точно посередине, ...

... а если ряд с чётным числом элементов, как в случае с боулингом, медианой будет среднее значение между 3-м и 4-м элементами.

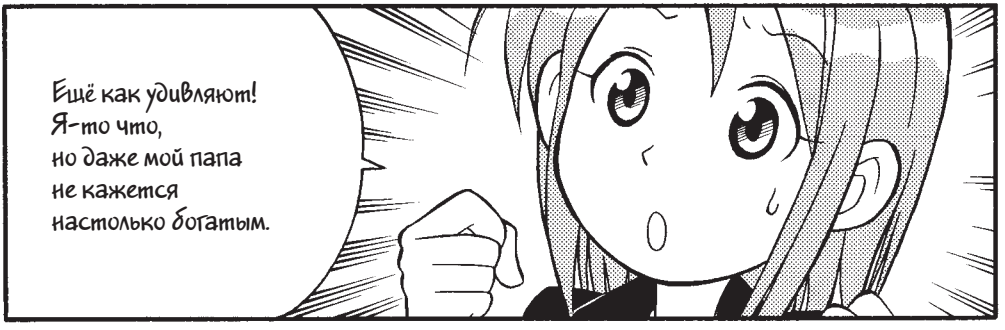
А теперь попробуем вычислить медиану для команды В.

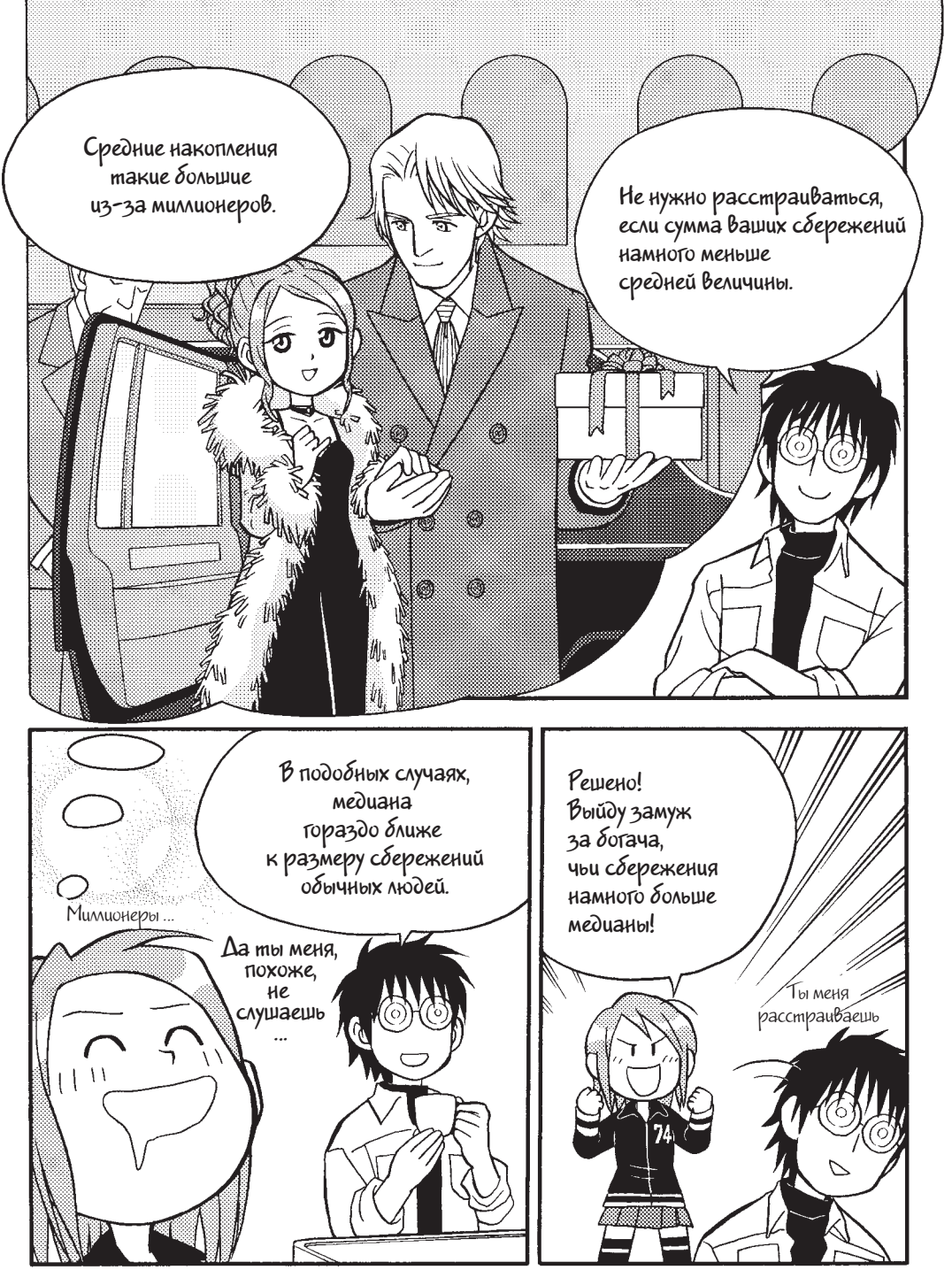
$\frac{77 + 88}{2} = 82{,}5$

Вот что получается.

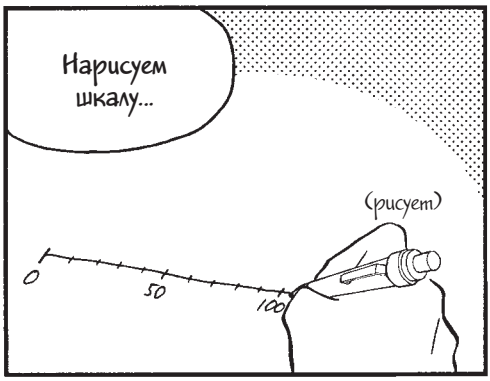
Правильно!

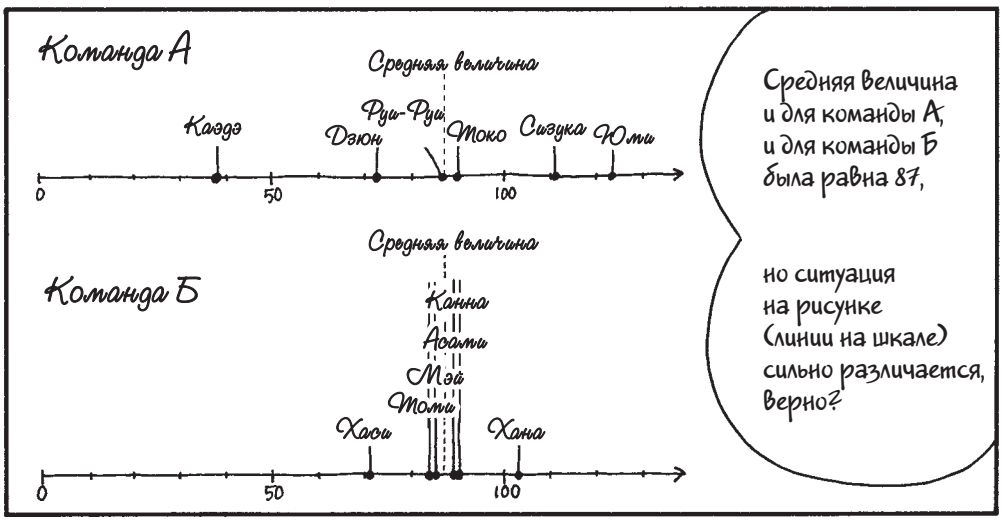
* Меньше 3500 рублей.




Средние накопления такие большие из-за миллионеров.
Не нужно расстраиваться, если сумма ваших сбережений намного меньше средней величины.
В подобных случаях, медиана гораздо ближе к размеру сбережений обычных людей.
Миллионеры ...
Да ты меня, похоже, не слушаешь ...
Решено! Выйду замуж за богача, чьи сбережения намного больше медианы!
Ты меня расстраиваешь
74

4. Стандартное отклонение






Да уж.
У команды А есть и низкие, и высокие результаты, а в команде Б все примерно одинаковы.
Чтобы описать разброс значений, используется стандартное отклонение (называемое также средним квадратическим отклонением).
Что это такое?
Если коротко, это показатель отклонения отдельных значений от их средней величины.
М-м-ммм... (непонятно)
Стандартное отклонение не можеть быть меньше 0. Чем больше стандартное отклонение, тем больше отдельные значения могут отличаться от средней величины.
0 (минимум)
Нет разброса значений — все значения равны
Разброс значений
Как ты думаешь, в какой команде стандартное отклонение больше?
М-м-м... Может, в А?

Правильно! А точная формула имеет вид

И опять математика...

$$\text{Стандартное отклонение} = \sqrt{\frac{(i\text{-е значение} - \text{среднее значение})^2}{\text{кол-во значений}}}$$

Да не переживай ты так! Всего-то и надо — подставить в эту формулу конкретные числа. Хочешь, попробуем вместе?

Хочу.

Сначала команда А

Команда А

$$\sqrt{\frac{(86-87)^2+(73-87)^2+(124-87)^2+(111-87)^2+(90-87)^2+(38-87)^2}{6}} =$$

$$= \sqrt{\frac{(-1)^2+(-14)^2+37^2+24^2+3^2+(-49)^2}{6}} =$$

$$= \sqrt{\frac{1+196+1369+576+9+2401}{6}} =$$

$$= \sqrt{\frac{4552}{6}} =$$

$$= \sqrt{758.6\cdots} =$$

$$\fallingdotseq 27.5$$

Так-то наверно и я смогу.

Тогда попробуй сама посчитать стандартное отклонение для команды Б.

Готово!

Команда Б

$$\sqrt{\frac{(84-87)^2+(71-87)^2+(103-87)^2+(85-87)^2+(90-87)^2+(89-87)^2}{6}}$$

$$=\sqrt{\frac{(-3)^2+(-16)^2+16^2+(-2)^2+3^2+2^2}{6}}=$$

$$=\sqrt{\frac{9+256+256+4+9+4}{6}}=$$

$$=\sqrt{\frac{538}{6}}=$$

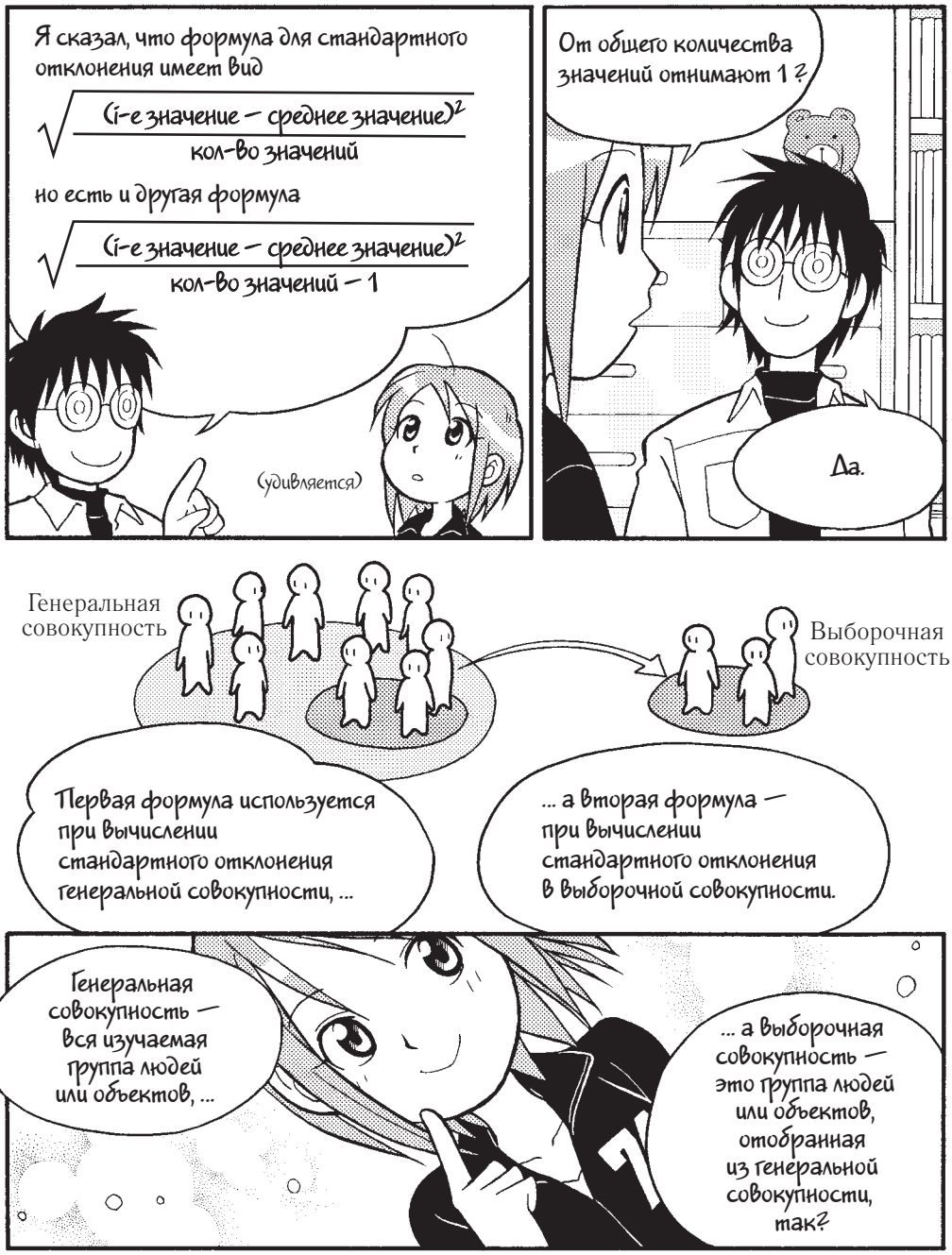
$$=\sqrt{89.6\cdots}=$$

$$\fallingdotseq 9.5$$

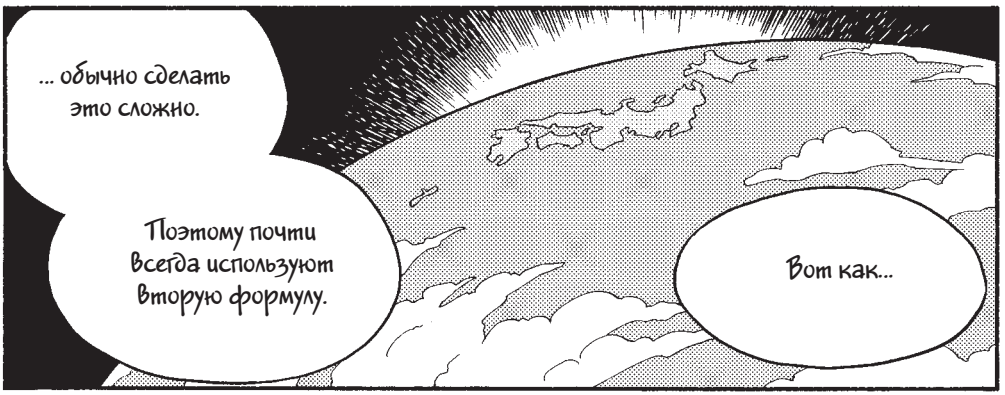
Правильно! Видишь, справилась же!

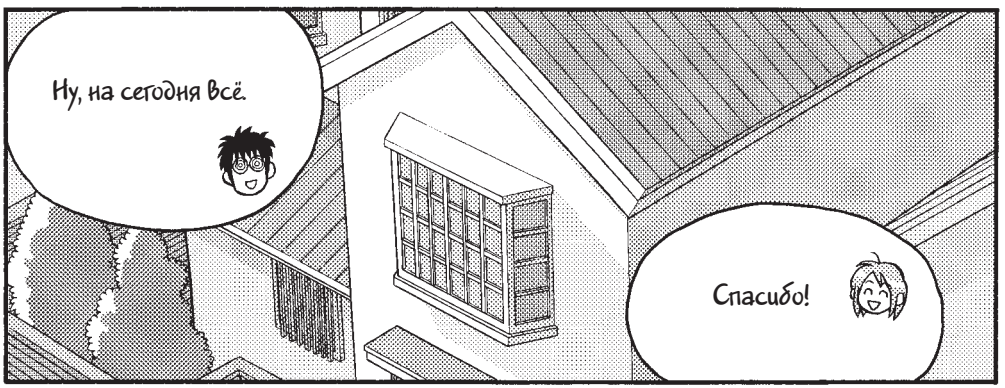
Стандартное отклонение:
Команда А — 27,5
Команда Б — 9,5

На самом деле у всех игроков команды **Б** похожие результаты. Стандартное отклонение здесь меньше, чем в команде **А**


Я сказал, что формула для стандартного отклонения имеет вид
√((i-е значение − среднее значение)²) / кол-во значений
но есть и другая формула
√((i-е значение − среднее значение)²) / кол-во значений − 1
(удивляется)
От общего количества значений отнимают 1?
Да.
Генеральная совокупность
Выборочная совокупность
Первая формула используется при вычислении стандартного отклонения генеральной совокупности, ...
... а вторая формула — при вычислении стандартного отклонения в выборочной совокупности.
Генеральная совокупность — вся изучаемая группа людей или объектов, ...
... а выборочная совокупность — это группа людей или объектов, отобранная из генеральной совокупности, так?


Да, верно. Хорошо, когда есть возможность получить данные обо всех объектах совокупности, как в случае с твоей командой. Но ...


... обычно сделать это сложно.
Поэтому почти всегда используют вторую формулу.
Вот как...


Ну, на сегодня всё.
Спасибо!

Возможно, не все читатели до конца разобрались с понятиями «ряды распределения» и «гистограммы».
Поэтому автор предлагает еще раз рассмотреть таблицу на **стр. 38**.

Таблица 2.1. Распределение 50 лучших ресторанов, предлагающих вкусный рамэн (по цене на рамэн)

Интервалы цен, йены	Середина интервала	Количество ресторанов, частота	Кол-во ресторанов, относительная частота
500—600	550	4	0,08
600—700	650	13	0,26
700—800	750	18	0,36
800—900	850	12	0,24
900—1000	950	3	0,06
Всего:		**50**	**1,00**

Как следует из таблицы, величина интервала равна 100. Это значение не является стандартом в математике. Просто так захотел Ямамото-сан. Решение о выборе интервала принимает тот, кто анализирует данные.

Не исключено, что среди читателей найдутся и такие, которым не дает покоя вопрос: «Ряды распределения, построенные на основе субъективных решений, неубедительны. Я не смогу показать их другим. Нет ли математического способа определения величины интервала?». Конечно, есть. Покажем, как можно вычислить величину интервала для **Табл. 2.1**.

Шаг 1

Количество интервалов определяется по формуле Стерджесса:

$$\text{Кол-во интервалов} = 1 + \frac{\log_{10} N}{\log_{10} 2},$$

где N — количество значений в совокупности.

$$1 + \frac{\log_{10} 50}{\log_{10} 2} = 1 + 5{,}6438\ldots = 6{,}6438\ldots \approx 7$$

Шаг 2

Величина интервала определяется по формуле:

$$\frac{\text{MAX} - \text{MIN}}{\text{Кол-во интервалов}},$$

где MAX — максимальное значение в совокупности,
MIN — минимальное значение в совокупности.

$$\frac{980 - 500}{7} = \frac{480}{7} = 68{,}5714\ldots \approx 69$$

Ниже приведена таблица распределения 50 лучших ресторанов по цене на рамэн с величиной интервала, рассчитанной по формуле, данной на **Шаге 2**.

Таблица 2.2. Распределение 50 лучших ресторанов, предлагающих вкусный рамэн (по цене на рамэн)

Интервал цен, йены	Середина интервала	Количество ресторанов, частота	Количество ресторанов, относительная частота
500—569	534,5	2	0,04
569—638	603,5	5	0,10
638—707	672,5	15	0,30
707—776	741,5	6	0,12
776—845	810,5	10	0,20
845—914	879,5	10	0,20
914—983	948,5	2	0,04
Итого		**50**	**1,00**

Ну и как? Не исключено, что некоторым данная таблица может показаться менее убедительной, чем **Табл. 2.1**.

При этом могут возникнуть такие вопросы, как: «Почему величина интервала равна именно 69 йенам?», «И что это за формула, ну, этого, как его, Стер... ? Да я вообще такого не знаю!» и «Почему интервалы распределены таким непонятным образом?!». Кроме того, среди читателей найдутся и такие, которые не рискнут самостоятельно определить величину интервала.

Случаи, когда распределение непонятно, даже если величина интервала определена математическим способом, встречаются довольно часто. И здесь уместно вспомнить то, о чем шла речь в начале этой главы: таблицы (ряды) распределения позволяют систематизировать данные наблюдения и интуитивно понять общую ситуацию. Следовательно, вполне достаточно выбрать такую величину интервалов, которая будет понятна тем, кто проводит статистический анализ.

6. Теория оценивания и описательная статистика

Объясняя Руи, что такое статистика, Игараси-сан определил её как науку, изучающую «большие совокупности однородных объектов на основании их выборочного исследования». Это не совсем так.

В статистике можно выделить два раздела: теорию оценивания и описательную статистику. В прологе речь шла о теории оценивания. Тогда что же такое описательная статистика? Это набор методов по упорядочиванию данных с целью наиболее простого и ясного восприятия этих данных. Можно считать, что описательная статистика рассматривает выборку как генеральную совокупность.

Возможно, такое определение кому-то покажется абстрактным и сложным для понимания. Поэтому приведём пример. Ямамото-сан вычислял среднее значение очков, набранное игроками команды Руи, и стандартное отклонение, чтобы представить положение в команде Руи в наглядном виде. Именно такая статистика и есть описательная.

Упражнение

Результаты забега на 100 метров приведены в следующей таблице:

Участник забега	Результат бега на 100 м, с
А	16,3
Б	22,4
В	18,5
Г	18,7
Д	20,1

Рассчитайте на основании этих результатов
- средний результат,
- медиану,
- отклонение.

Ответ

$$\text{Средний результат} = \frac{16{,}3 + 22{,}4 + 18{,}5 + 18{,}7 + 20{,}1}{5} = \frac{96}{5} = 19{,}2$$

Медиана: 16,3 18,5 **18,7** 20,1 22,4

Стандартное отклонение =

$$= \sqrt{\frac{(16{,}3 - 19{,}2)^2 + (22{,}4 - 19{,}2)^2 + (18{,}5 - 19{,}2)^2 + (18{,}7 - 19{,}2)^2 + (20{,}1 - 19{,}2)^2}{5}} =$$

$$= \sqrt{\frac{(-2{,}9)^2 + 3{,}2^2 + (-0{,}7)^2 + (-0{,}5)^2 + 0{,}9^2}{5}} =$$

$$= \sqrt{\frac{20{,}2}{5}} =$$

$$= \sqrt{4{,}04} =$$

$$\approx 2{,}01$$

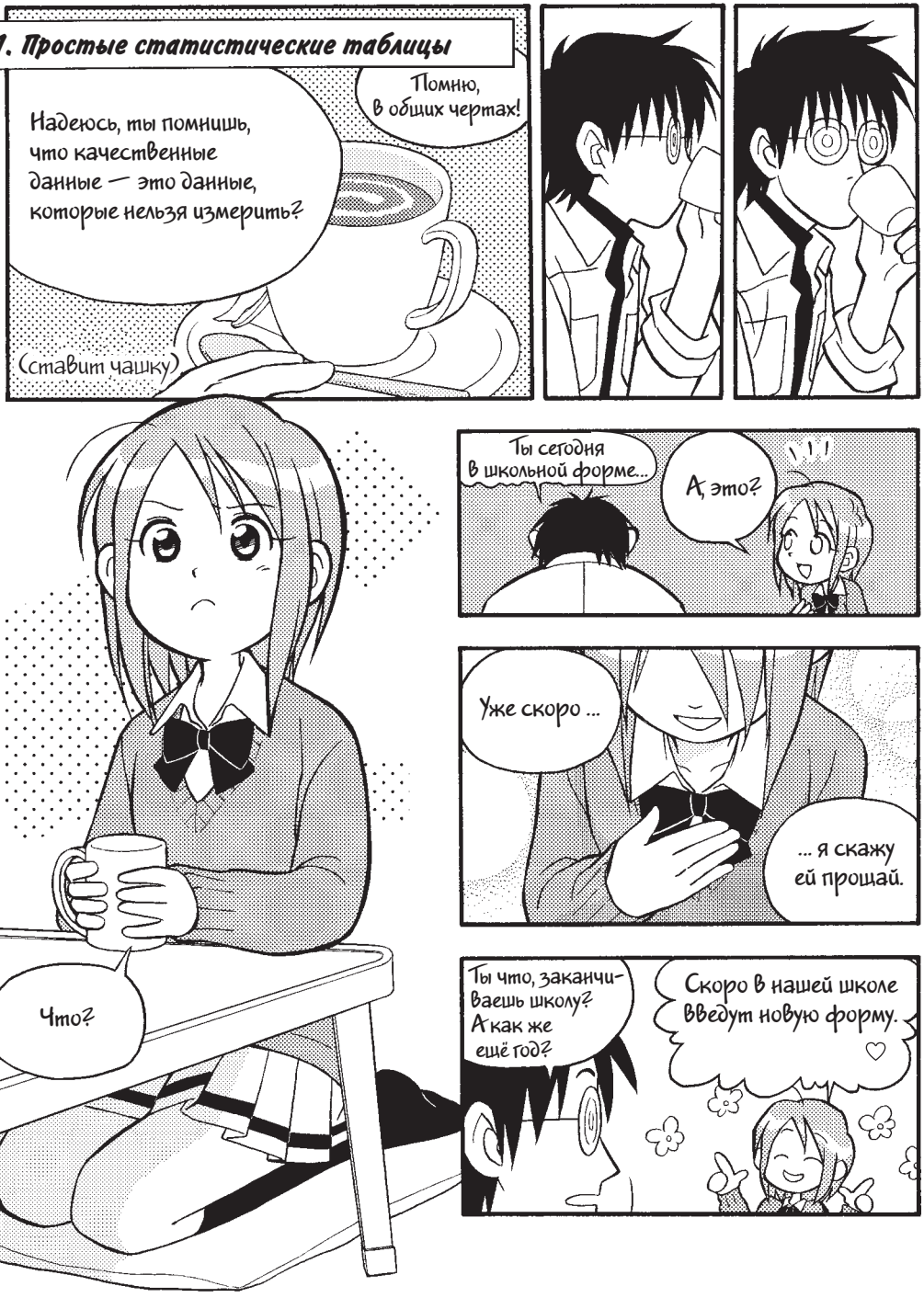
Выводы

- **Чтобы «интуитивно» понять общую ситуацию с данными, строят ряды распределения.**
- **Величина интервала в рядах распределения определяется по формуле Стерджесса.**
- **Чтобы «математически» понять ситуацию с данными, вычисляют среднюю величину, медиану и стандартное отклонение.**
- **Если ряд распределения содержит слишком большие или слишком малые значения, рассчитывают медиану, а не среднюю величину.**
- **Стандартное отклонение — показатель, отражающий степень разброса (рассеяния) значений.**

Глава 3

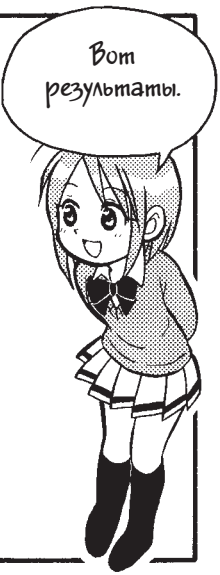
Знакомимся с качественными данными

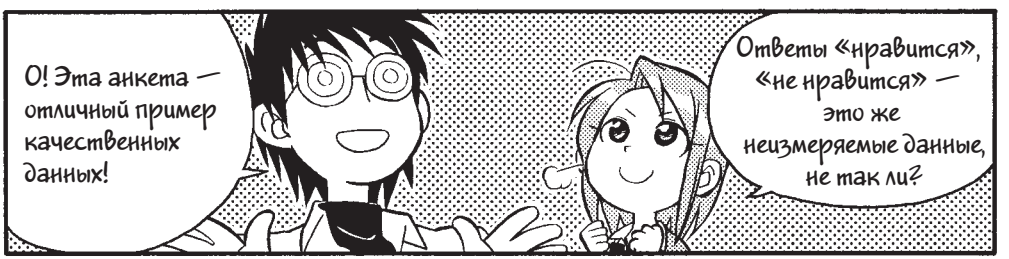
1. Простые статистические таблицы

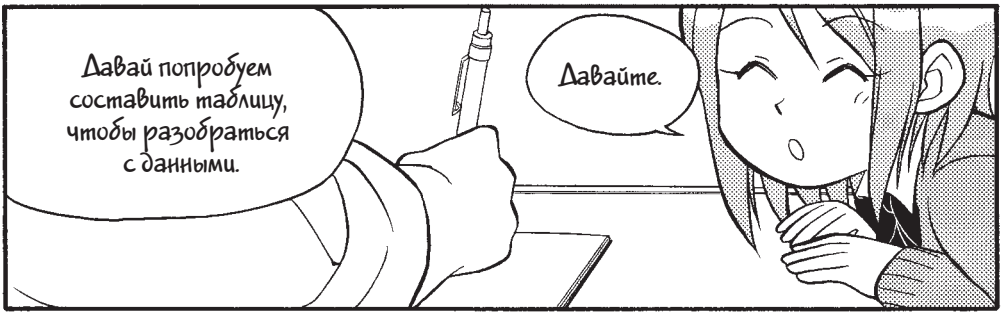


Анкета. Нравится ли вам новая форма?

	Новая форма...		Новая форма...		Новая форма...
1	нравится	16	так себе	31	так себе
2	так себе	17	нравится	32	так себе
3	нравится	18	нравится	33	нравится
4	так себе	19	нравится	34	не нравится
5	не нравится	20	нравится	35	нравится
6	нравится	21	нравится	36	нравится
7	нравится	22	нравится	37	нравится
8	нравится	23	не нравится	38	нравится
9	нравится	24	так себе	39	так себе
10	нравится	25	нравится	40	нравится
11	нравится	26	нравится		
12	нравится	27	не нравится		
13	так себе	28	нравится		
14	нравится	29	нравится		
15	нравится	30	нравится		



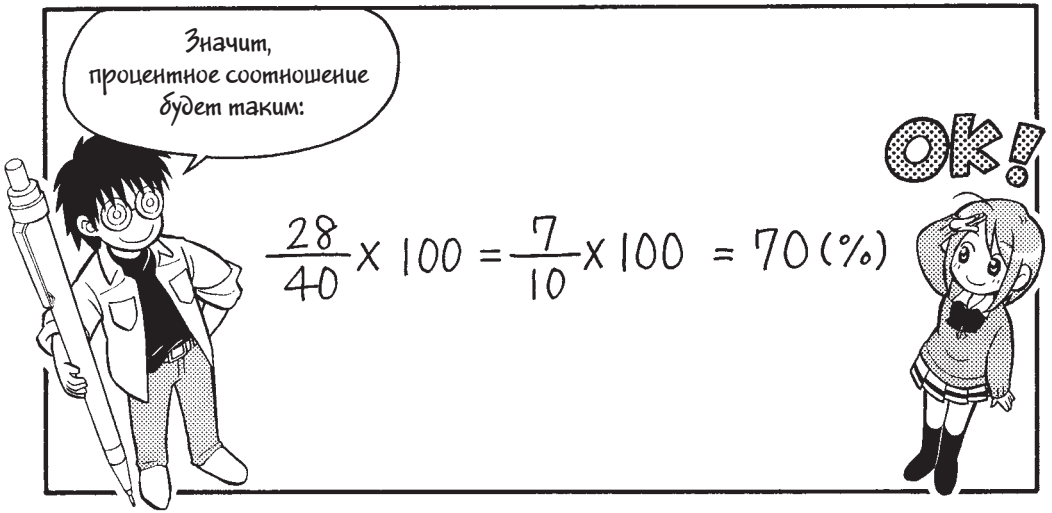



Давай попробуем составить таблицу, чтобы разобраться с данными.
Давайте.


Оценка новой школьной формы
Ответ | Кол-во | %
нравится | 28 | 70
так себе | 8 | 20
не нравится | 4 | 10
Итого: | 40 | 100
Это простая статистическая таблица.

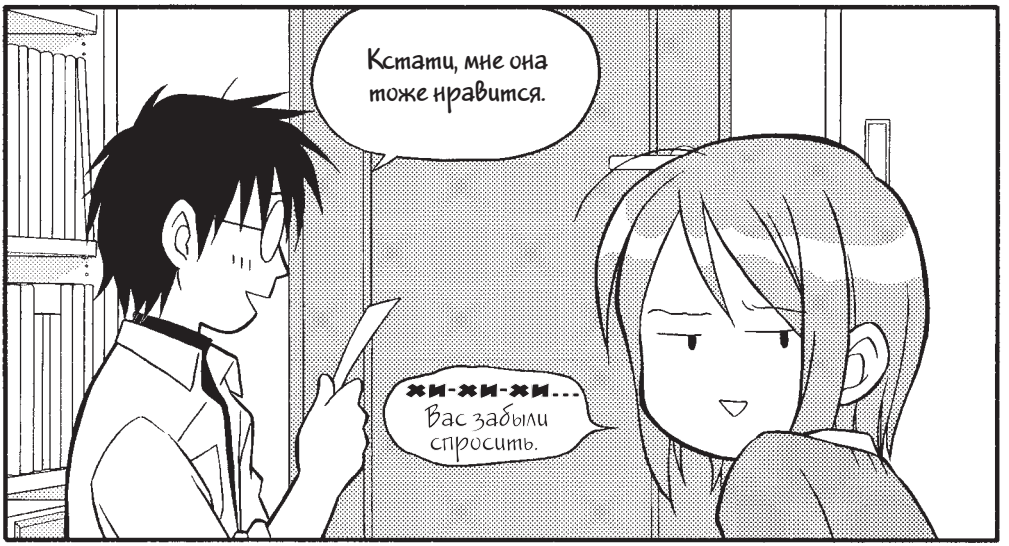

Между прочим, Руи, а ты как ответила?
Нравится! ♡


Теперь ещё раз. Какова частота ответов «нравится»?
У нас 28 ответов «нравится» следовательно, 28!


Значит, процентное соотношение будет таким:
$\frac{28}{40} \times 100 = \frac{7}{10} \times 100 = 70(\%)$
OK!


Новая форма...
нравится
так себе
не нравится
70
20
10
n
40
Чтобы было более понятно, нарисуем диаграмму.
Хорошо. Так действительно намного понятнее.


Из диаграммы следует, что большую часть занимают ответы «нравится», что, в свою очередь, свидетельствует о популярности новой школьной формы.
Я так и знала! Потому что форма классная!


Кстати, мне она тоже нравится.
Хи-хи-хи... Вас забыли спросить.

Упражнение

Одна газета провела анкетирование по вопросу, какая из двух политических партиий победит на следующих выборах.
Результаты анкетирования приведены в виде следующей таблицы:.

Респондент	Победит партия А или Б
1	победит Б
2	победит Б
3	победит Б
4	не знаю
5	победит А
6	победит Б
7	победит А
8	не знаю
9	победит Б
10	победит Б

Используя результаты анкетирования,
постройте простую статистическую таблицу.

Ответ

Простая статистическая таблица выглядит так:

Оценка партии	Частота	%
Победит А	2	20
Не знаю	2	20
Победит Б	6	60
Итого	10	100

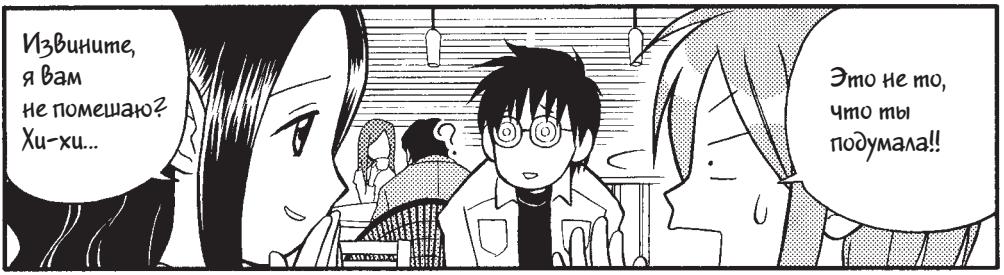
Выводы

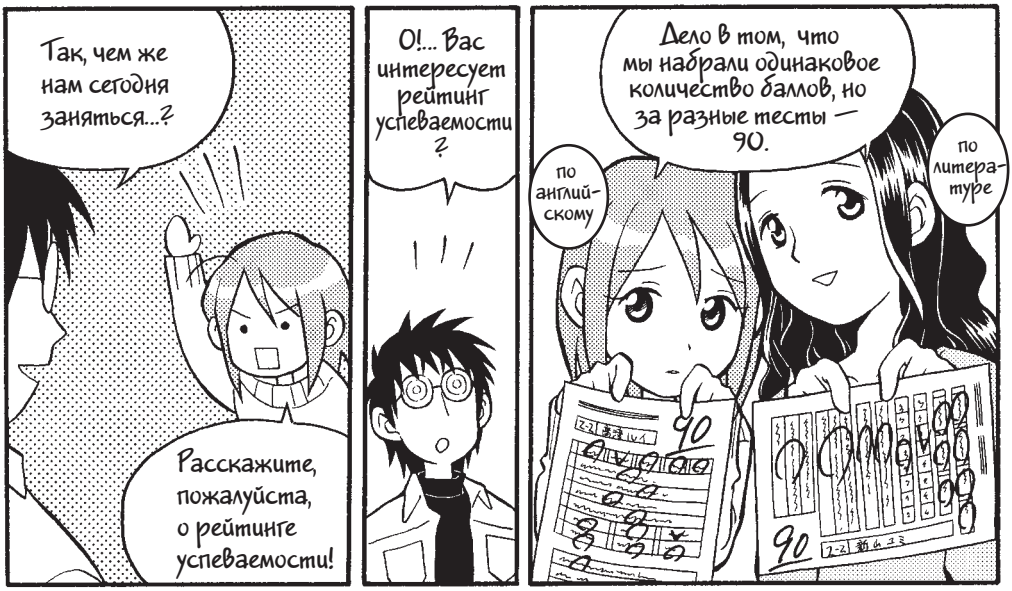
- **Чтобы понять общую ситуацию с данными, строят простые статистические таблицы.**

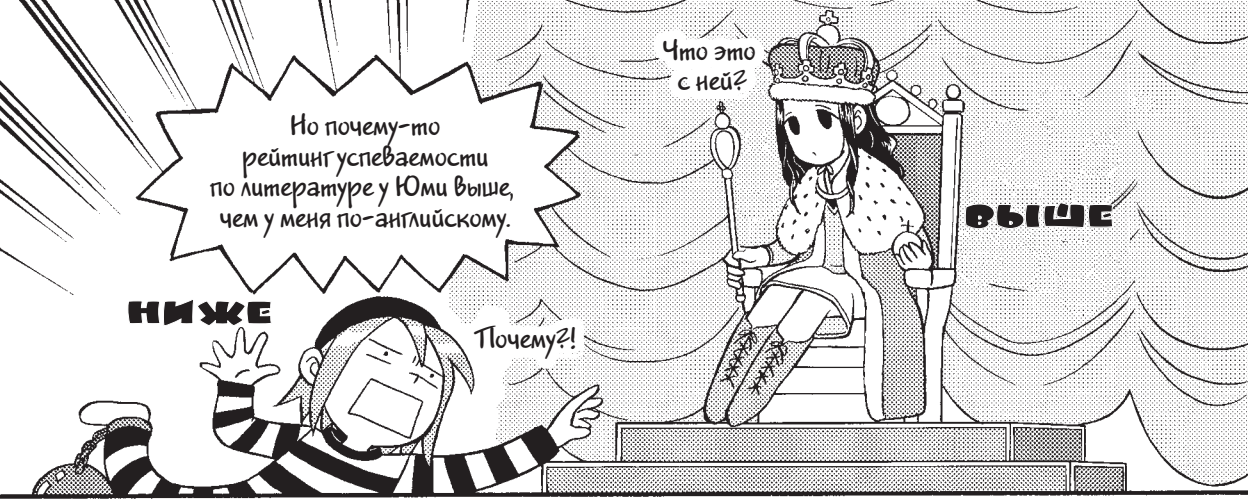
Глава 4

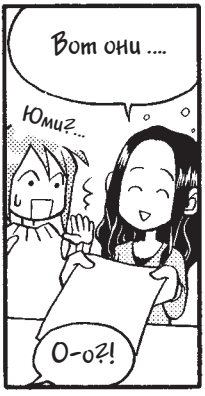
Нормированное отклонение и рейтинг успеваемости

1. Нормирование и нормированное отклонение





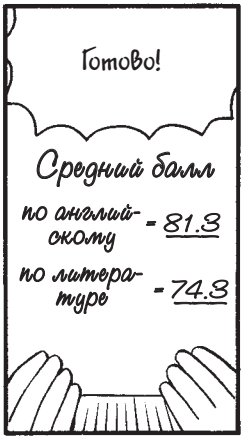


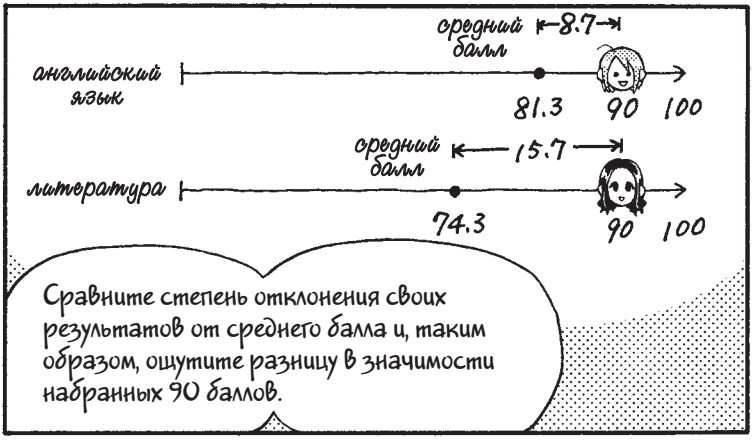


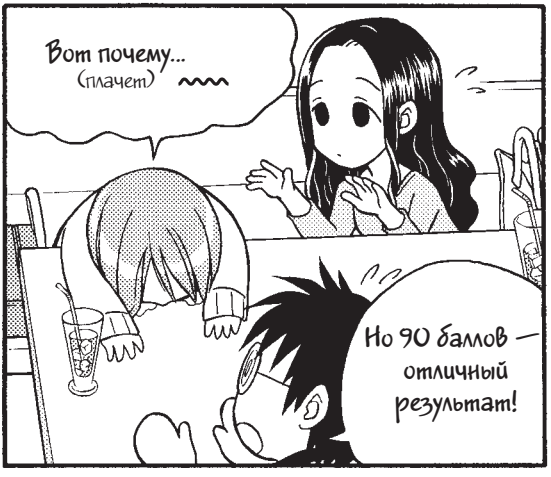
Результаты тестов (максимальное количество баллов — 100)

Ученик	Английский язык	Литература	Ученик	Английский язык	Литература
Фуми	90	71	З	67	85
Юми	81	90	И	87	93
А	73	79	К	78	89
Б	97	70	Л	85	78
В	85	67	М	96	74
Г	60	66	Н	77	65
Д	74	60	О	100	78
Е	64	83	П	92	53
Ж	72	57	Р	86	80

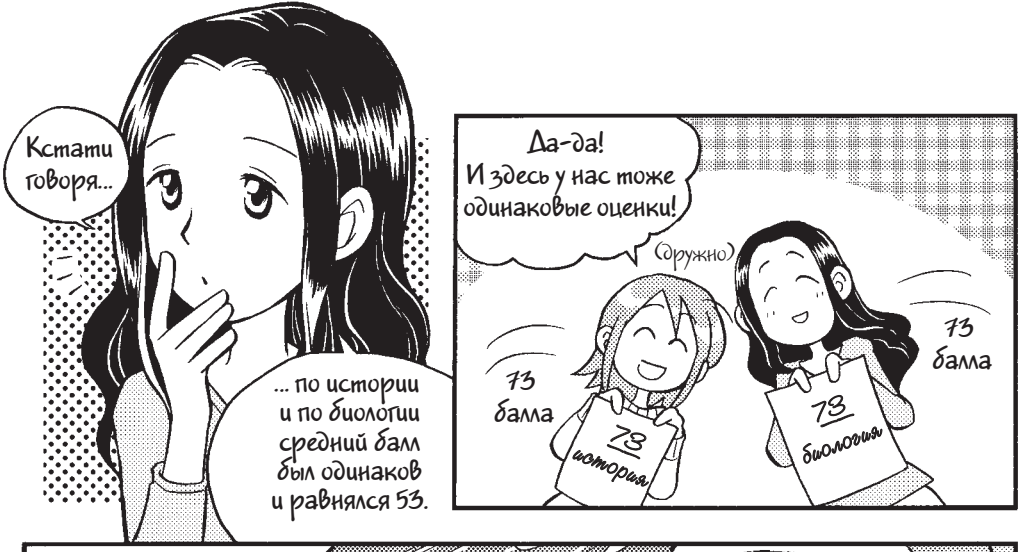

Попробуй вычислить средний балл по каждому предмету.
Хорошо.

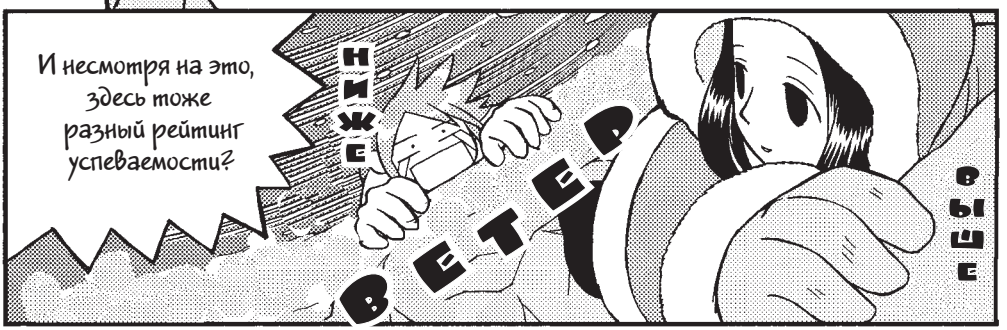

Готово!
Средний балл
по английскому - 81.3
по литературе - 74.3


средний балл
8.7
английский язык
81.3
90
100
средний балл
15.7
литература
74.3
90
100
Сравните степень отклонения своих результатов от среднего балла и, таким образом, ощутите разницу в значимости набранных 90 баллов.


Вот почему...
(плачет)
Но 90 баллов — отличный результат!


Вы обе заслужили награду в виде пирожных.
Ура!!!
Пирожные!
Пирожные!





Хотя отклонение от среднего результата одинаковое.

Угу, понятно.

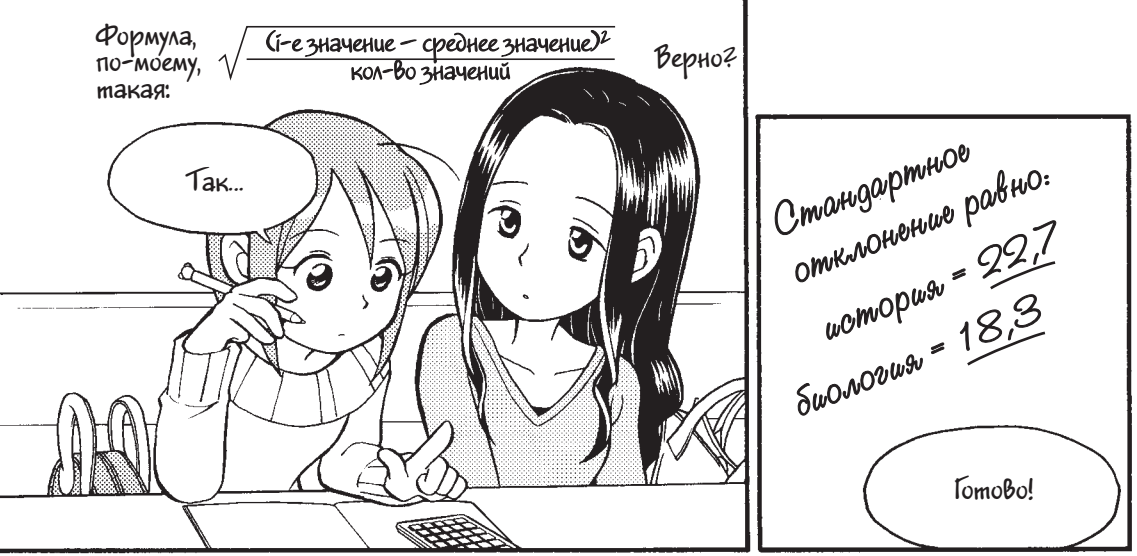
Ученик	История	Биология
Рю	73	59
Юта	61	73
А	14	47
Б	41	38
В	49	63
Г	87	56
Д	69	15
Е	65	53
Ж	36	80

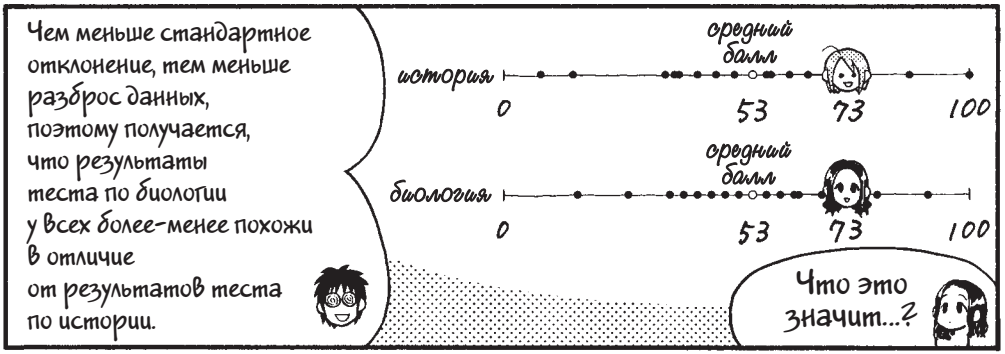
Ученик	История	Биология
З	7	50
И	53	41
К	100	62
Л	57	44
М	45	26
Н	56	91
О	34	35
П	37	53
Р	70	68
Средний балл:	53	53

А как насчёт стандартного отклонения по этим предметам?

А-а! Степень разброса значений!

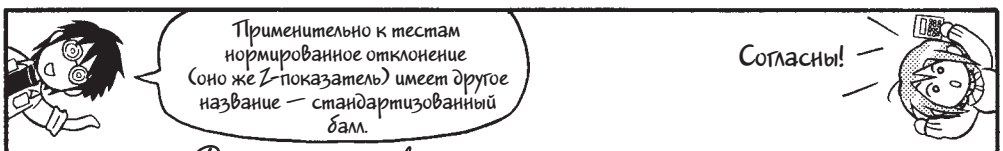
Ну, Рю, ты даёшь!


Формула, по-моему, такая:
$\sqrt{\frac{(i\text{-е значение} - \text{среднее значение})^2}{\text{кол-во значений}}}$
Верно?
Так...
Стандартное отклонение равно:
история = 22,7
биология = 18,3
Готово!


Чем меньше стандартное отклонение, тем меньше разброс данных, поэтому получается, что результаты теста по биологии у всех более-менее похожи в отличие от результатов теста по истории.
история
средний балл
0
53
73
100
биология
средний балл
0
53
73
100
Что это значит...?


Если бы я готовился к поступлению в институт, я бы поднажал на биологию, так как в тесте по биологии **значимость балла выше**, чем в тесте по истории.
Какие-то 1—2 балла могут существенно повлиять на конечный результат.
Как ему идёт форма старшеклассника...
Хи-Хи-Хи


Хотя количество баллов у нас и одинаковое – 73, в биологии значимость баллов выше.
Руи, не унывай
(расстроилась)
Я не думала, что сравнивать баллы так сложно.
Да, это так...
(недовольна)
Вот почему придумали нормирование (называемое также нормировка, стандартизация, нормализация)!
Нормирование?
Это преобразование значений, проводимое на основе стандартного отклонения и отклонения от среднего значения. Его ещё называют Z-преобразованием.
Такое преобразование делает более простым изучение значимости баллов!



Результаты тестов по истории и биологии

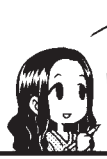
Ученик	История	Биология
Фуи	73	59
Юми	61	73
А	14	47
Б	41	38
В	49	63
Г	87	56
Д	69	15
Е	65	53
Ж	36	80
З	7	50
И	53	41
К	100	62
Л	57	44
М	45	26
Н	56	91
О	34	35
П	37	53
Р	70	68
Средний балл	53	53
Станд. отклонение	22.7	18.3

Нормированные отклонения

История	Биология
0.88	0.33
0.35	1.09
-1.71	-0.33
-0.53	-0.82
-0.18	0.55
1.49	0.16
0.70	-2.08
0.53	0
-0.75	1.48
-2.02	-0.16
0	-0.66
2.07	0.49
0.18	-0.49
-0.35	-1.48
0.13	2.08
-0.84	-0.98
-0.70	0
0.75	0.82
0	0
1	1

Нормированное отклонение Фуи по истории $= \frac{78-53}{22.7} = \frac{20}{22.7} = 0.88$

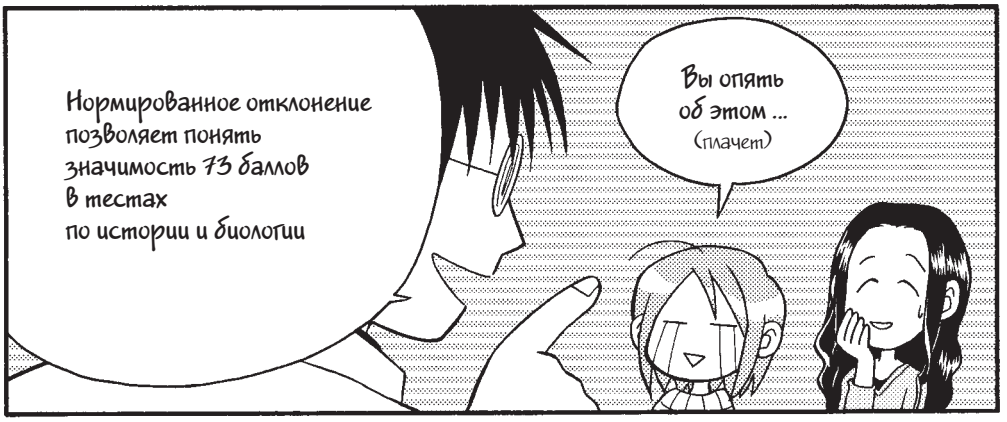
Нормированное отклонение Юми по биологии $= \frac{78-53}{18.3} = \frac{20}{18.3} = 1.09$



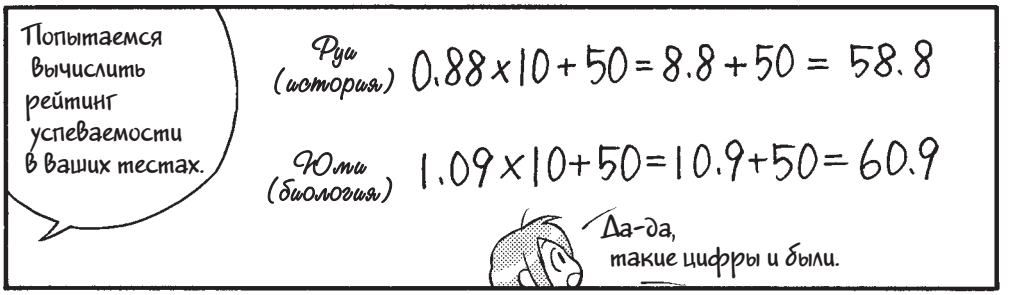
2. Свойства нормированного отклонения

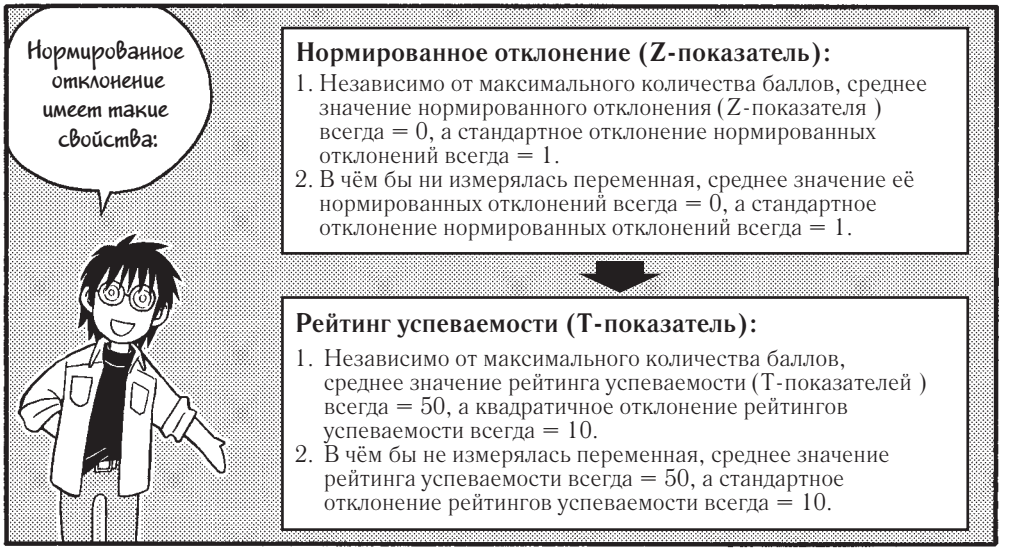
1. Независимо от максимального количества баллов, среднее значение нормированного отклонения (Z-показателя) всегда равно **0**, а стандартное отклонение нормированных отклонений всегда равно **1**.

2. В чём бы ни измерялась переменная, среднее значение её нормированных отклонений всегда равно **0**, а стандартное отклонение нормированных отклонений всегда равно **1**.



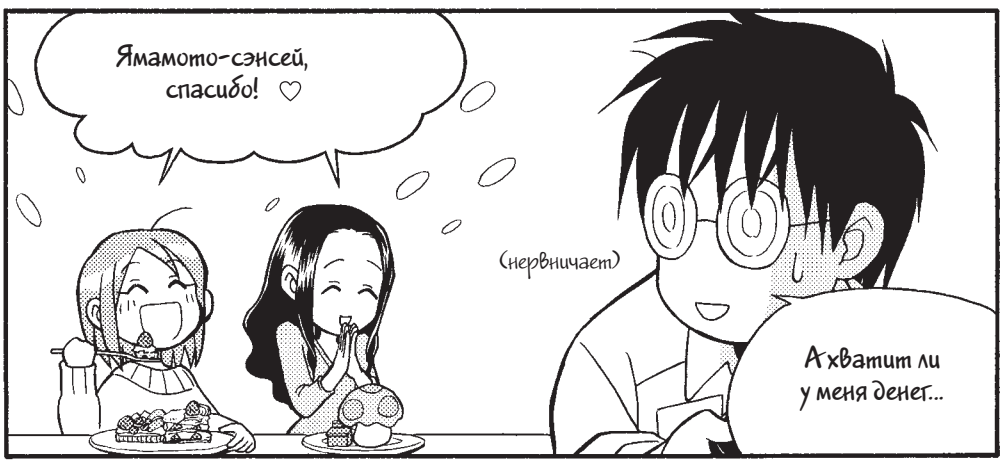
3. Рейтинг успеваемости






Вот как...
В тестах, а также в других аналогичных исследованиях, значимость каждого балла имеет важное значение, поэтому и используют рейтинг успеваемости.
Может быть, на этом сегодня закончим?
Ура!
Пирожные!


Я хочу это и это...
Фруктовая тарталетка
¥1.200-
Слоёный торт с клубникой
¥1,500-
(не ожидал)
А я хочу попробовать вот это...
«Сказочный грибок»
¥2.500-


Ямамото-сэнсей, спасибо! ♡
(нервничает)
А хватит ли у меня денег...

4. Что такое рейтинг успеваемости?

В общем случае это Т-показатель, который вычисляется по формуле:

$$\text{Т-показатель} = \text{Нормальное отклонение} \times 10 + 50 =$$

$$= \frac{\text{Значение} - \text{Среднее значение}}{\text{Стандартное отклонение}} \times 10 + 50.$$

В классе Руи 40 учеников, и из них 18 девочек. Рейтинги успеваемости были приведены только для двух девочек. Если объектом исследования были бы все ученики класса, среднее значение и стандартное отклонение были бы совсем другими, и, естественно, рейтинги успеваемости Руи и Юми тоже были бы другими. Если бы рассчитывались рейтинги успеваемости всех учеников класса, показатель Руи был бы выше.

Результаты теста всех учеников представлены в Табл. 4.1. Обязательно попробуйте вычислить рейтинги успеваемости. Забегая вперёд, скажу что у Руи по истории он равен 59,1, а у Юми по биологии он равен 56,7.

Теперь представим, что одинаковые тесты проводились в двух классах. Сначала вычислили отдельно средний балл и стандартное отклонение для учеников 1-го класса, и затем на их основе рассчитали рейтинг успеваемости. То же проделали и для 2-го класса. Показатель успеваемости ученика А из 1-го класса был равен 57, и рейтинг успеваемости ученика Б из 2-го класса тоже был равен 57. На первый взгляд кажется, что у них одинаковые знания. Однако, такие необходимые для расчёта рейтинга успеваемости данные, как средний балл и стандартное отклонение, в 1-м и 2-м классах были разными. Следовательно, нельзя сравнивать рейтинги успеваемости этих двух учеников.

И еще один пример. Ученик **А** в апреле сдавал пробный экзамен на подготовительных курсах. Его рейтинг успеваемости был равен 54. Все лето **А** продолжал заниматься на других курсах. А в сентябре, чтобы убедиться, что его труды не пропали даром, он решил сдать еще один пробный экзамен, но уже на других курсах. На этот раз рейтинг успеваемости был 62. На первый взгляд кажется, что знания **А** улучшились. Но он сдавал пробные экзамены на разных курсах, и, следовательно, организаторы проведения экзамена в апреле и в сентябре были разные. Это, в свою очередь, означает, что средний результат и стандартное отклонение, т.е. показатели, необходимыедля вычисления рейтинга успеваемости, в апреле и сентябре были разные, и, следовательно, нельзя сравнивать два значения рейтинга успеваемости **А**.

Ну как? Понятие показатель успеваемости довольно-таки сложное.

Таблица 4.1. Результаты теста по истории и биологии (всех учеников класса Руи)

Девочка	История	Биология
Руи	73	59
Юми	61	73
А	14	47
Б	41	38
В	49	63
Г	87	56
Д	69	15
Е	65	53
Ж	36	80
З	7	50
И	53	41
К	100	62
Л	57	44
М	45	26
Н	56	91
О	34	35
П	37	53
Р	70	68

Мальчик	История	Биология
а	54	2
б	93	7
в	91	98
г	37	85
д	44	100
е	16	29
ж	12	57
з	44	37
и	4	95
к	17	39
л	66	70
м	53	14
н	14	97
о	73	39
п	6	75
р	22	80
с	69	77
т	95	14
у	16	24
ф	37	91
х	14	36
ц	88	76

Средний балл по всему классу	**48,0**	**54,9**
Стандартное отклонение по всему классу	**27,5**	**26,9**

Упражнение

В упражнении на стр. 57 приведены результаты бега на 100 метров.

Участник	Результат бега на 100 м, с
А Б В Г Д	16,3 22,4 18,5 18,7 20,1
Среднее значение	19,2
Стандартное отклонение	2,01

Проверьте с помощью этой таблицы:
1. Равно ли 0 среднее значение нормированных отклонений.
2. Равно ли 1 стандартное отклонение нормированных отклонений.

Ответ

1. Среднее значение нормированных отклонений =

$$= \frac{\left(\frac{16{,}3-19{,}2}{2{,}01}\right)+\left(\frac{22{,}4-19{,}2}{2{,}01}\right)+\left(\frac{18{,}5-19{,}2}{2{,}01}\right)+\left(\frac{18{,}7-19{,}2}{2{,}01}\right)+\left(\frac{20{,}1-19{,}2}{2{,}01}\right)}{5} =$$

$$= \frac{\left\{\frac{(16{,}3-19{,}2)+(22{,}4-19{,}2)+(18{,}5-19{,}2)+(18{,}7-19{,}2)+(20{,}1-19{,}2)}{2{,}01}\right\}}{5} =$$

упорядочили числитель

$$= \frac{\left\{\frac{16{,}3+22{,}4+18{,}5+18{,}7+20{,}1-19{,}2-19{,}2-19{,}2-19{,}2-19{,}2}{2{,}01}\right\}}{5} =$$

отдельно индивидуальные значения, отдельно средние значения $(-19{,}2)$.

$$= \frac{\left\{\frac{96-19{,}2\times 5}{2{,}01}\right\}}{5} =$$

$$= \frac{\left\{\frac{96-96}{2{,}01}\right\}}{5} =$$

$$= \frac{0}{5} =$$

$$= 0$$

2. Стандартное отклонение нормированных отклонений =

$$= \sqrt{\frac{\left(\frac{16{,}3-19{,}2}{2{,}01}-0\right)^2+\left(\frac{22{,}4-19{,}2}{2{,}01}-0\right)^2+\left(\frac{18{,}5-19{,}2}{2{,}01}-0\right)^2+\left(\frac{18{,}7-19{,}2}{2{,}01}-0\right)^2+\left(\frac{20{,}1-19{,}2}{2{,}01}-0\right)^2}{5}} =$$

$$= \sqrt{\frac{\left(\frac{16{,}3-19{,}2}{2{,}01}\right)^2+\left(\frac{22{,}4-19{,}2}{2{,}01}\right)^2+\left(\frac{18{,}5-19{,}2}{2{,}01}\right)^2+\left(\frac{18{,}7-19{,}2}{2{,}01}\right)^2+\left(\frac{20{,}1-19{,}2}{2{,}01}\right)^2}{5}} =$$

$$= \sqrt{\frac{\left\{\frac{(16{,}3-19{,}2)^2+(22{,}4-19{,}2)^2+(18{,}5-19{,}2)^2+(18{,}7-19{,}2)^2+(20{,}1-19{,}2)^2}{2{,}01^2}\right\}}{5}} =$$

упорядочили числитель

$$= \sqrt{\frac{1}{2{,}01^2}\times\frac{(16{,}3-19{,}2)^2+(22{,}4-19{,}2)^2+(18{,}5-19{,}2)^2+(18{,}7-19{,}2)^2+(20{,}1-19{,}2)^2}{5}} =$$

$$= \frac{1}{2{,}01}\times\sqrt{\frac{(16{,}3-19{,}2)^2+(22{,}4-19{,}2)^2+(18{,}5-19{,}2)^2+(18{,}7-19{,}2)^2+(20{,}1-19{,}2)^2}{5}} =$$

$$= \frac{1}{\text{Стандартное отклонение}}\times \text{Стандартно отклонение} =$$

$$= 1$$

Внимательно посмотрите на таблицу на стр. 78.

Выводы

- **Нормирование (нормировка, стандартизация, нормализация), или Z-преобразование, — преобразование значений, проводимое на основе данных о степени разброса (рассеяния) и отклонения от среднего значения. Нормирование позволяет оценить значимость значений.**

- **Нормирование позволяет сравнивать различные переменные величины, например:**
  - **величины, имеющие разный размах (разность между максимальным и минимальным значениями);**
  - **величины, имеющие разные единицы измерения.**

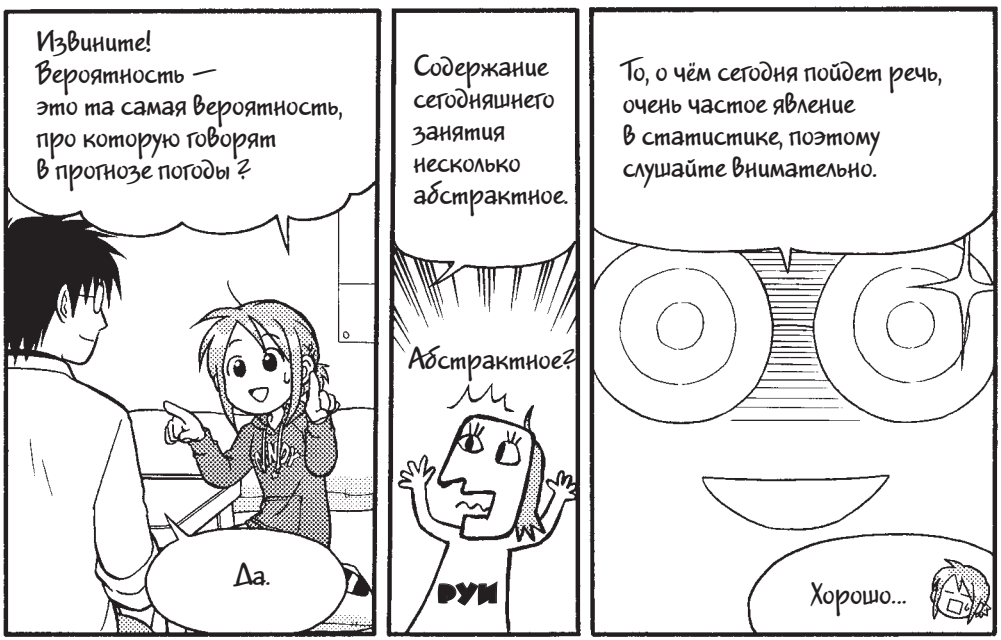
- **Нормированные данные — нормированные отклонения отдельных значений.**

- **Рейтинг успеваемости рассчитывается по формуле Т-показателя на основе нормированного отклонения.**

Глава 5

Вычислим вероятность

1. Функция распределения плотности вероятности



Представим, что все ученики 11-х классов Центрального округа ...

Результаты теста по английскому языку всех одиннадцатиклассников школ Центрального округа

Ученик	Балл
1	42
2	91
...	⋮
10421	50
Средний балл	53
Стандартное отклонение	10

... сдавали экзамен на подготовительных курсах.

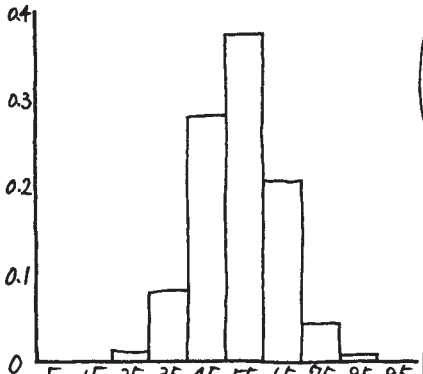
Вы сегодня очень хорошо подготовлены.

Ха-Ха-Ха. Мы только начали.

(перевернул страницу)

Если эту таблицу представить в виде гистограммы, то получим ...

Гистограмма
«Результаты теста по английскому языку». Величина интервала равна 10.



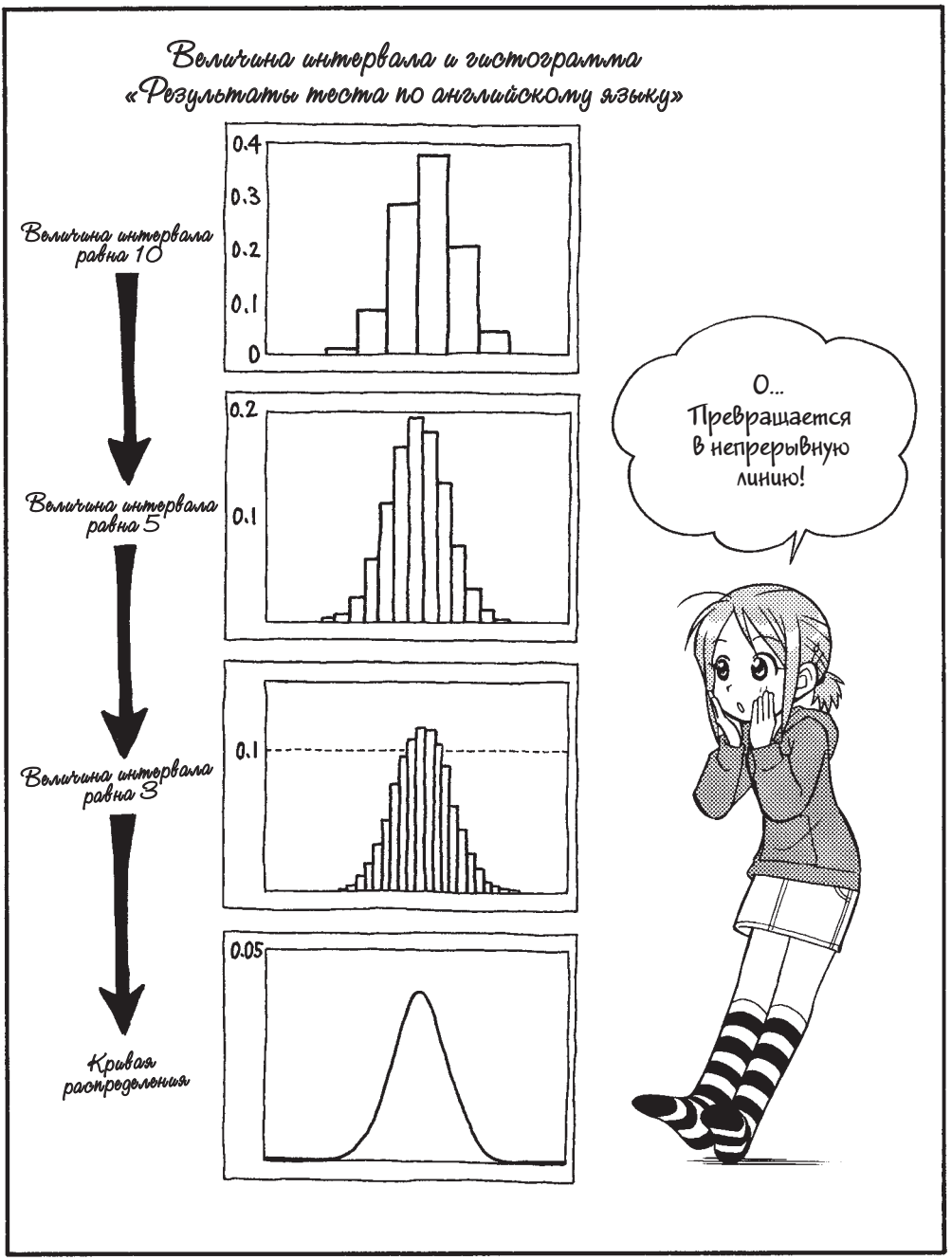
О-о-о... Действительно, если нарисовать гистограмму, можно что-то понять.

Потому что наглядно.

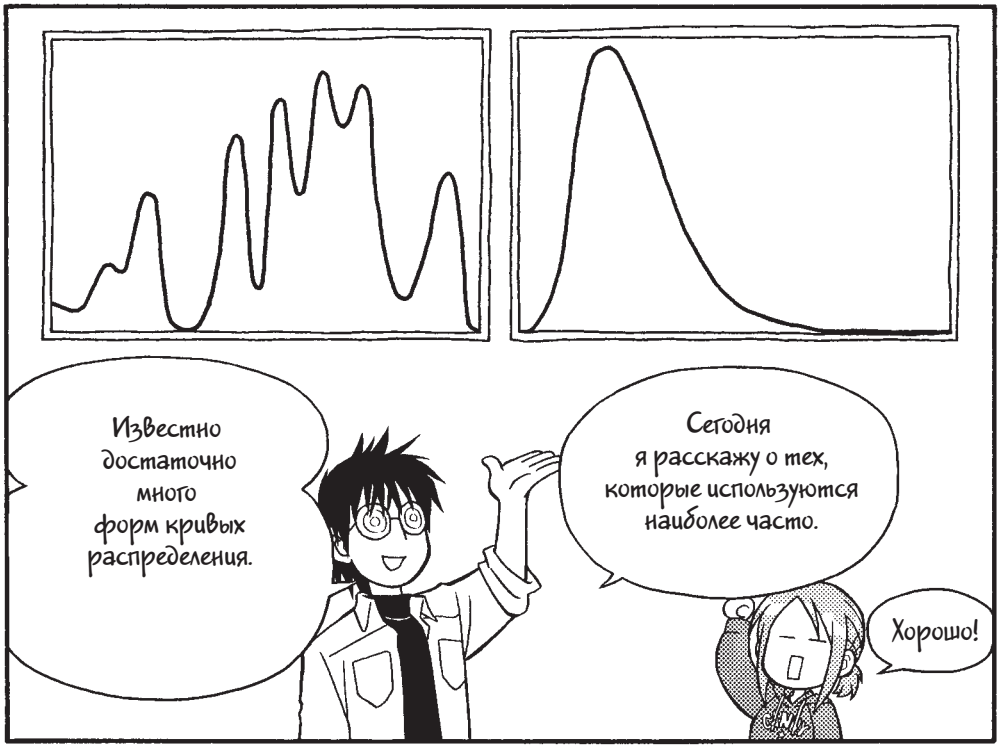
Что произойдет, если на этой гистограмме уменьшить величину интервала?

Что?

(Ямамото-сан уменьшился)


Величина интервала и гистограмма
«Результаты теста по английскому языку»
Величина интервала равна 10
Величина интервала равна 5
Величина интервала равна 3
Кривая распределения
0.4
0.3
0.2
0.1
0
0.2
0.1
0.1
0.05
О...
Превращается в непрерывную линию!


Если величина интервала на гистограмме стремится к нулю, то получается кривая распределения ...
... а функция, описываемая этой кривой, в статистике называется функцией распределения вероятностей.


Известно достаточно много форм кривых распределения.
Сегодня я расскажу о тех, которые используются наиболее часто.
Хорошо!

2. Нормальное распределение

$$f(x) = \frac{1}{\sqrt{2\pi} \times \text{Стандартное отклонение}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\text{Стандартное отклонение}}\right)^2}$$

где $\bar{x}$ — средняя величина x (средняя арифметическая ряда)

Вот.

Что это?!!

Ха-Ха-Ха

Это — часто встречающаяся в статистике функция распределения вероятности.

А что это за знак «e»?

Символ «e» — математическая константа, основание натурального логарифма; его иногда называют числом Эйлера или числом Непера. e = 2,7182.

ОСНОВАНИЕ

Можешь думать, что это как число π

Фу...

Ну, тогда ладно...

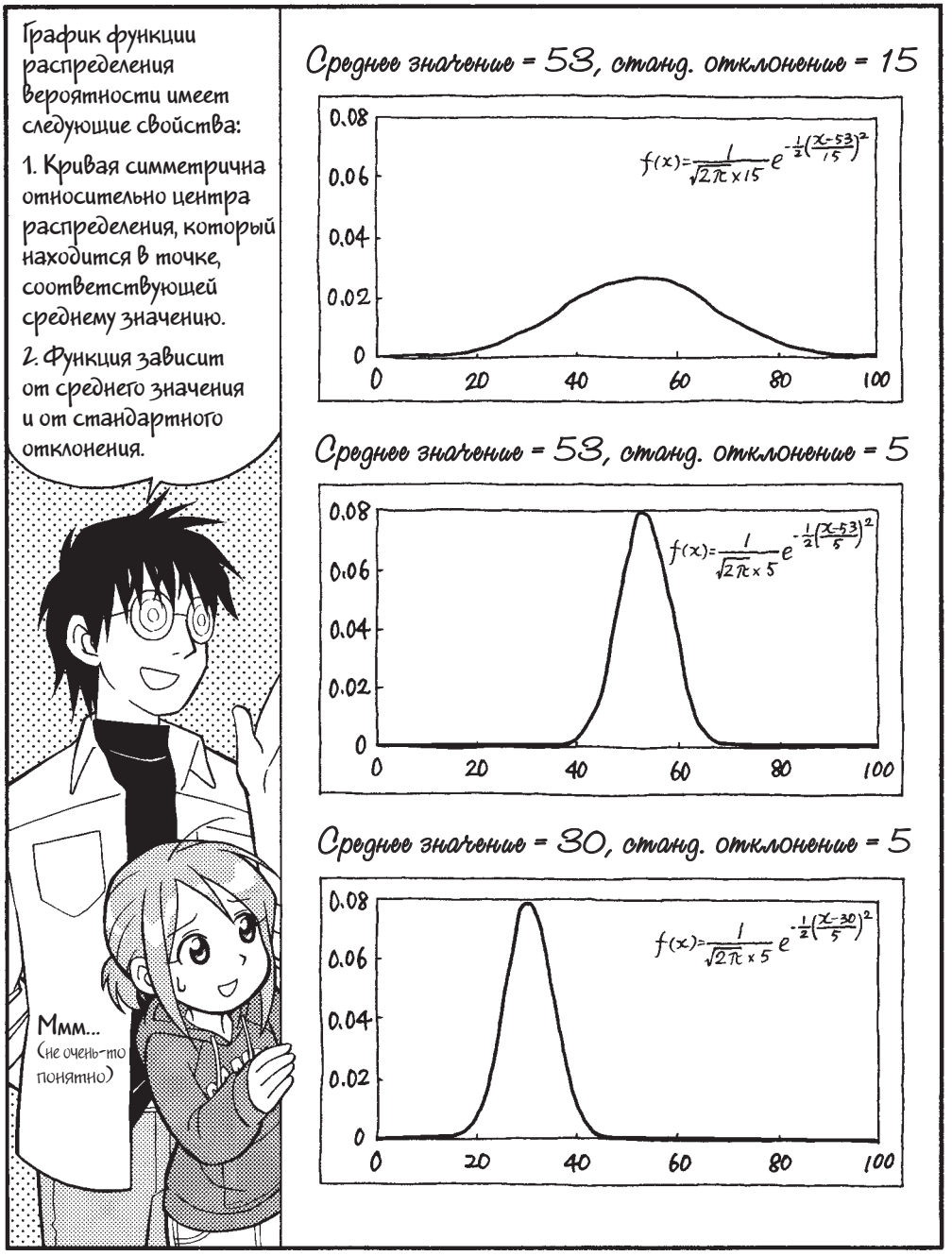
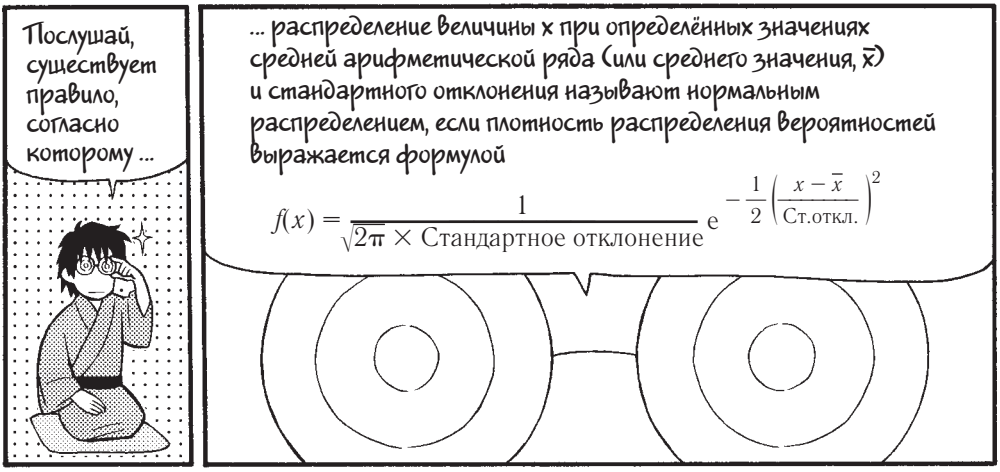
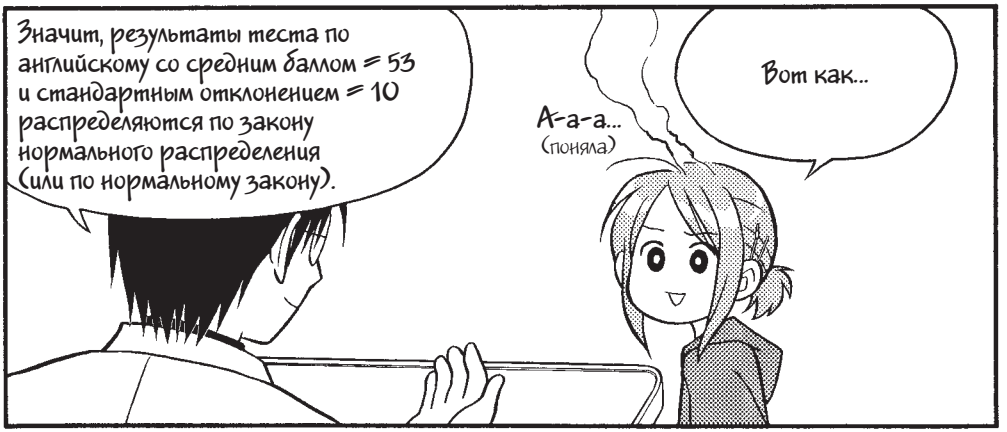


График функции распределения вероятности имеет следующие свойства:
1. Кривая симметрична относительно центра распределения, который находится в точке, соответствующей среднему значению.
2. Функция зависит от среднего значения и от стандартного отклонения.
Ммм... (не очень-то понятно)
Среднее значение = 53, станд. отклонение = 15
$f(x)=\frac{1}{\sqrt{2\pi}\times 15}e^{-\frac{1}{2}\left(\frac{x-53}{15}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
Среднее значение = 53, станд. отклонение = 5
$f(x)=\frac{1}{\sqrt{2\pi}\times 5}e^{-\frac{1}{2}\left(\frac{x-53}{5}\right)^2}$
Среднее значение = 30, станд. отклонение = 5
$f(x)=\frac{1}{\sqrt{2\pi}\times 5}e^{-\frac{1}{2}\left(\frac{x-30}{5}\right)^2}$


Послушай, существует правило, согласно которому ...
... распределение величины x при определённых значениях средней арифметической ряда (или среднего значения, $\bar{x}$) и стандартного отклонения называют нормальным распределением, если плотность распределения вероятностей выражается формулой
$f(x) = \frac{1}{\sqrt{2\pi} \times \text{Стандартное отклонение}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\text{Ст.откл.}}\right)^2}$


Чего ???!!!
... называют нормальным распределением... !?
Ничего не понимаю...
Правило довольно специфическое, поэтому просто запомни его.


Возьмём для примера предыдущий тест. Если кривая функции распределения вероятностей результатов теста по английскому имеет следующий вид...
... то имеет место нормальное распределение при среднем балле 53 и стандартном отклонении 10.
$f(x) = \frac{1}{\sqrt{2\pi} \times 10} e^{-\frac{1}{2}\left(\frac{x-53}{10}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100



3. Стандартное нормальное распределение

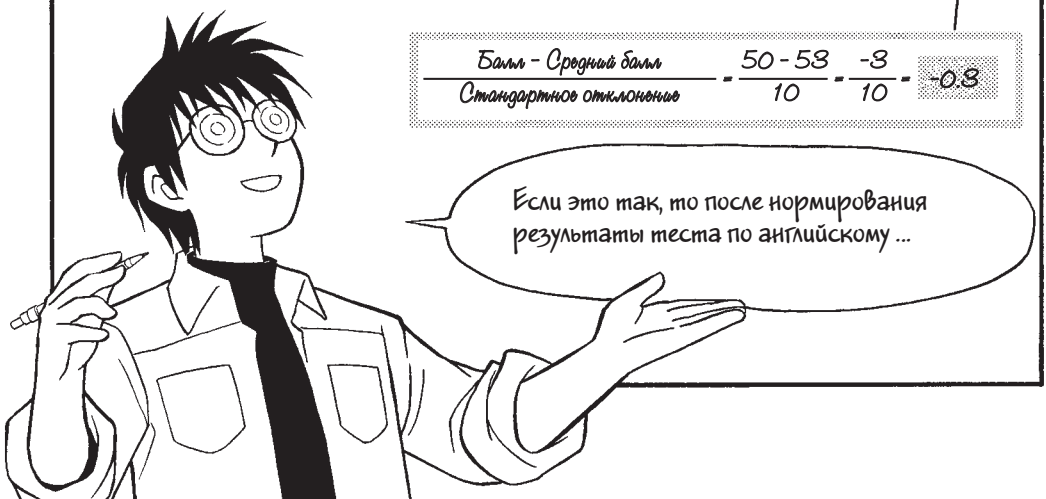
А если плотность распределения вероятностей выражается формулой

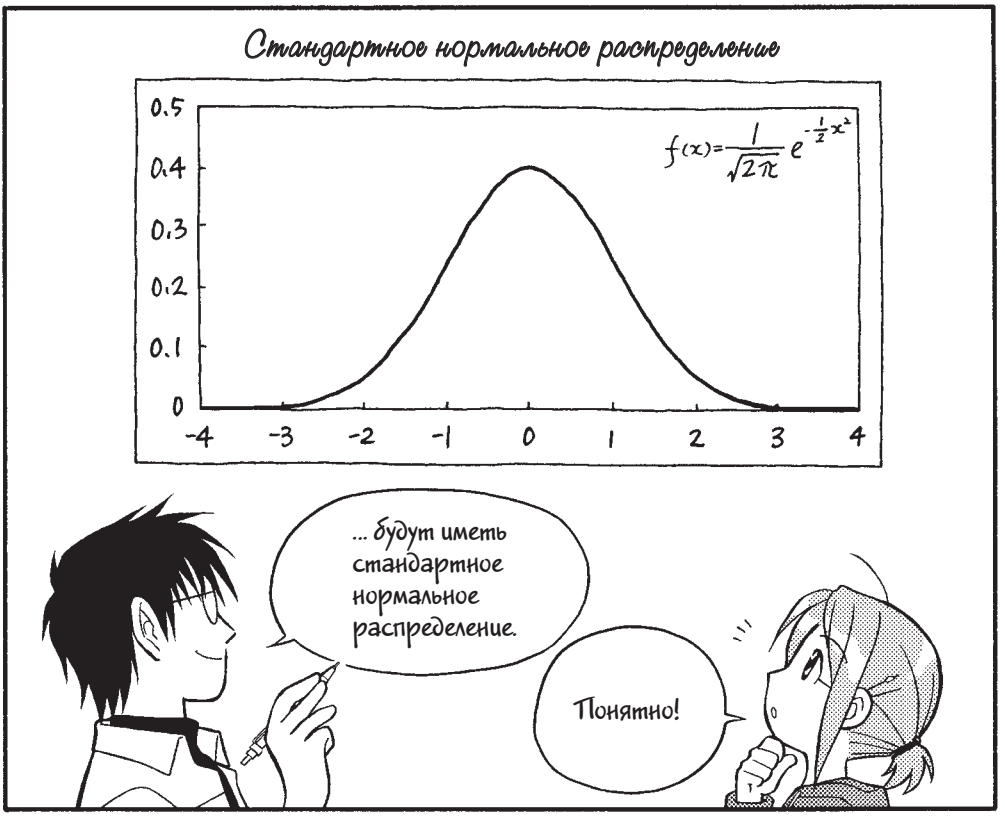
$$f(x) = \frac{1}{\sqrt{2\pi} \times \text{Ст.откл.}} \, e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\text{Ст.откл.}}\right)^2} = \frac{1}{\sqrt{2\pi} \times 1} \, e^{-\frac{1}{2}\left(\frac{x-0}{\text{Ст.откл.}}\right)^2} = \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}x^2}$$

то в статистике не говорят, что величина x имеет нормальное распределение при значении средней арифметической ряда = 0 и стандартном отклонении = 1. Принято говорить, что величина x имеет стандартное нормальное распределение.

Ученик	Балл	Нормированное отклонение
1	42	-1.1
2	91	3.8
...	⋮	⋮
10421	50	-0.3
среднее значение	53	0
стандартное отклонение	10	1

$$\frac{\text{Балл} - \text{Средный балл}}{\text{Стандартное отклонение}} = \frac{50-53}{10} = \frac{-3}{10} = -0.3$$




Стандартное нормальное распределение
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$
... будут иметь стандартное нормальное распределение.
Понятно!

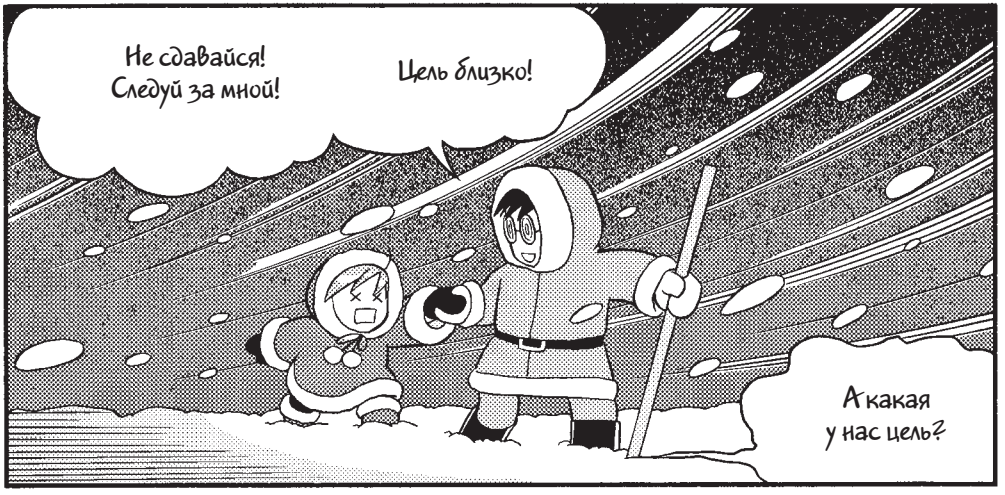
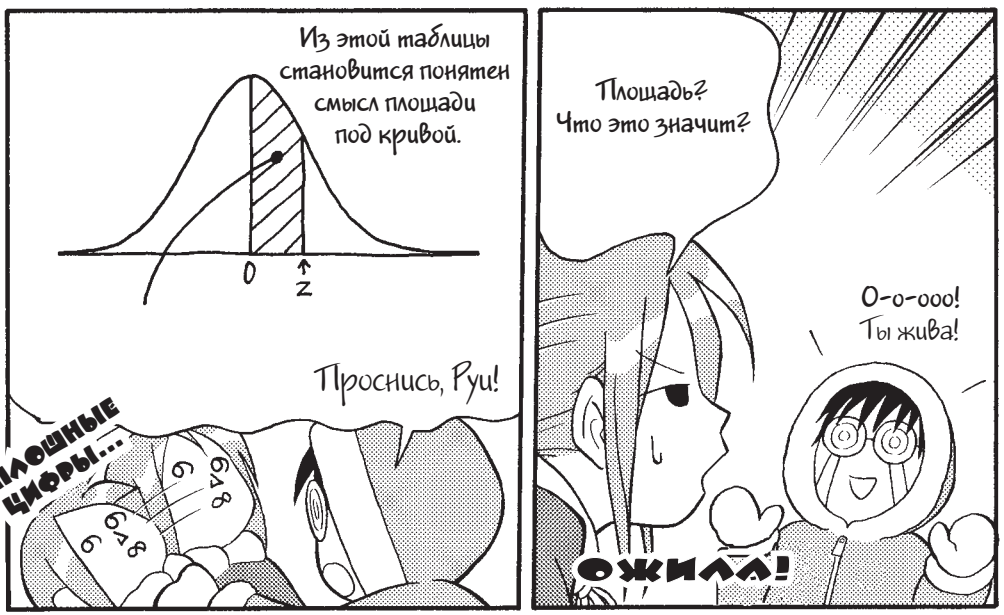
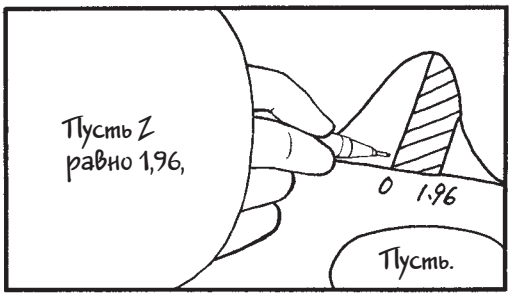
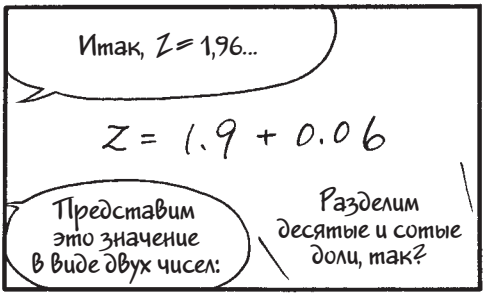

Не сдавайся! Следуй за мной!
Цель близко!
А какая у нас цель?

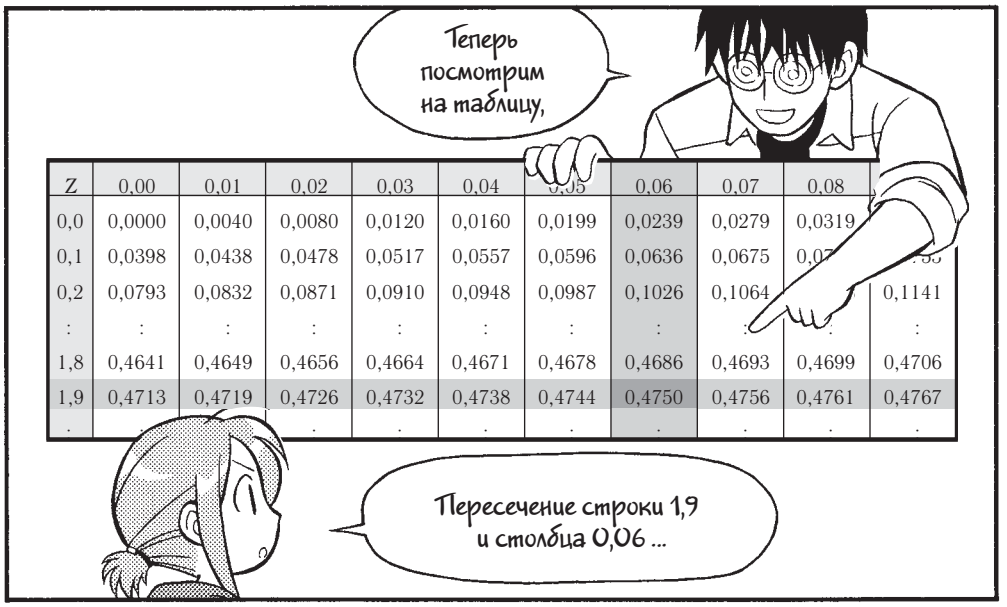
Таблица стандартного нормального распределения

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
:	:	:	:	:	:	:	:	:	:	:
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
:	:	:	:	:	:	:	:	:	:	:

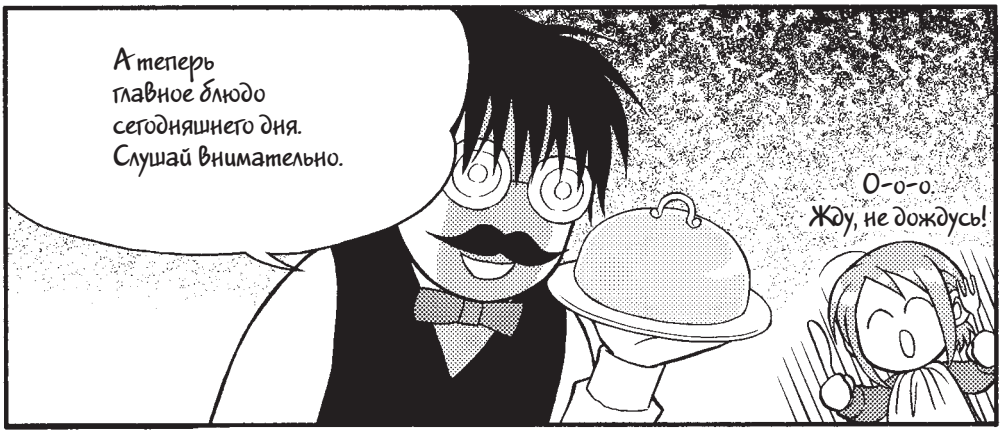








Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,07	
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064		0,1141
:	:	:	:	:	:	:	:	:	:	:
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
.	.	.	.	.	.	.	.	.	.	.


А теперь
главное блюдо
сегодняшнего дня.
Слушай внимательно.
О-о-о.
Жду, не дождусь!

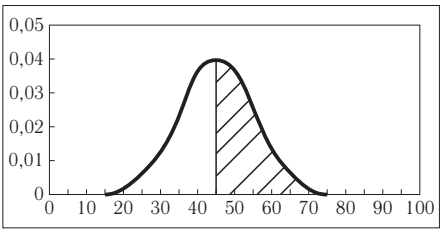

Площадь области,
ограниченной кривой стандартного
нормального распределения и осью x,
равна доле и вероятности.


Сейчас приведу два примера,
постарайся понять.

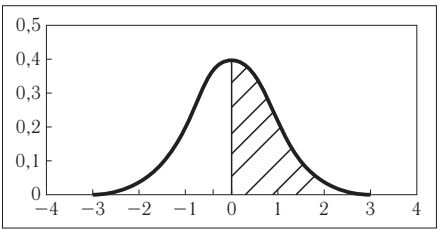
Пример 1

Все десятиклассники школ Восточного округа сдавали тест по математике. Когда им поставили оценки, стало ясно, что распределение результатов теста можно считать нормальным при среднем балле 45, и стандартном отклонении 10.
Теперь хорошенечко подумай.
Следующие пять выводов имеют один и тот же смысл:

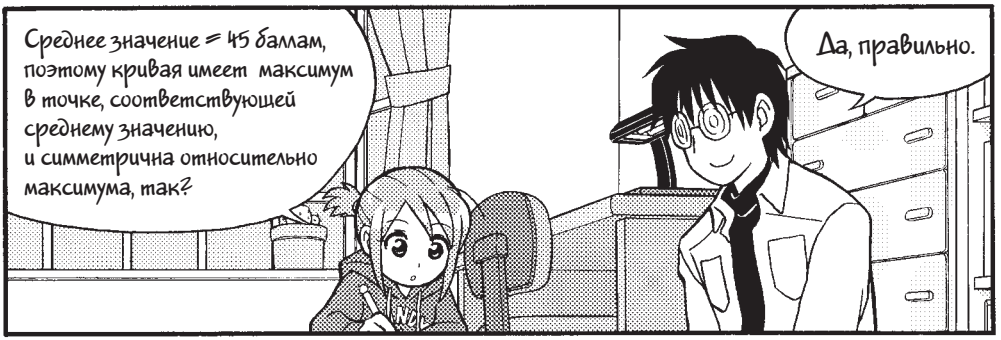
1. При нормальном распределении, когда среднее значение = 45 и стандартное отклонение = 10, площадь заштрихованной области = 0,5.

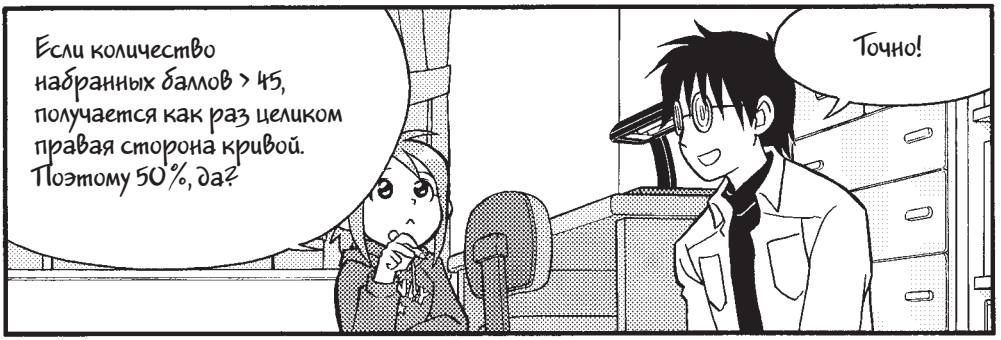


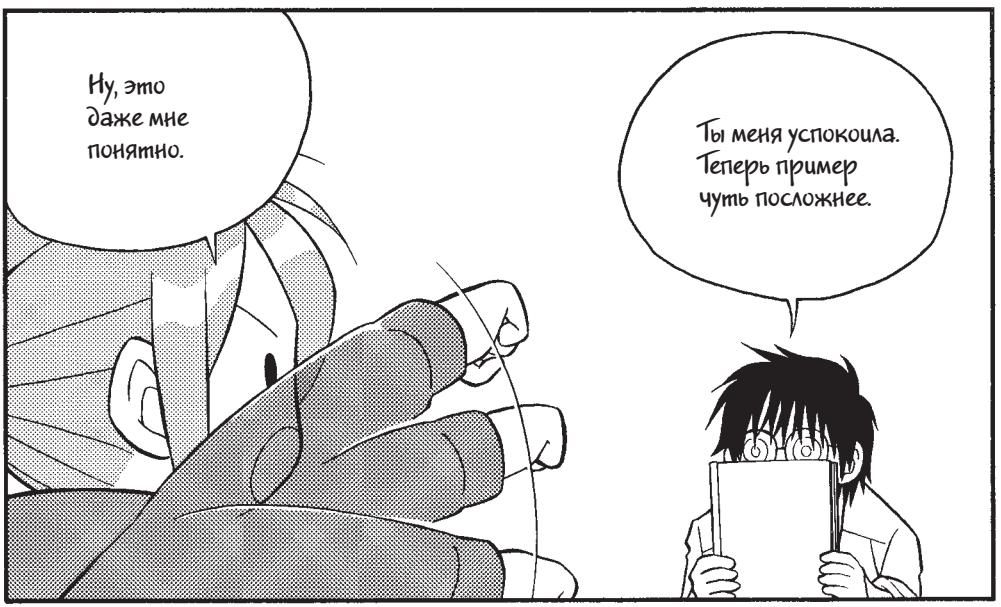
2. Доля учеников, чей результат был > 45 баллов, составляет 50% от общего числа участников теста.
3. Предположим, что из общего числа учеников произвольно выбрали одного. Вероятность того, что он набрал > 45 баллов, равна 50%.
4. Для стандартного нормального распределения, полученого после нормирования результатов теста, доля учеников с результатом > 0 составляет 50% от общего числа участников теста.



5. Предположим, что из общего числа учеников произвольно выбрали одного. При стандартном нормальном распределении, полученном после нормирования результатов теста по математике, вероятность того, что он набрал положительный балл, составляет 50%.


Среднее значение = 45 баллам, поэтому кривая имеет максимум в точке, соответствующей среднему значению, и симметрична относительно максимума, так?
Да, правильно.

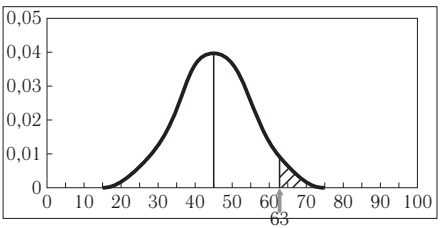

Если количество набранных баллов > 45, получается как раз целиком правая сторона кривой. Поэтому 50%, да?
Точно!


Ну, это даже мне понятно.
Ты меня успокоила. Теперь пример чуть посложнее.

Пример 2

Все десятиклассники школ Восточного округа сдавали тест по математике. Когда им поставили оценки, стало ясно, что распределение результатов теста можно считать нормальным при среднем балле = 45 и стандартном отклонении = 10. Теперь напряги мозги. Как и в предыдущем примере, все следующие пять выводов имеют один и тот же смысл. Но на этот раз ты для начала прочти вывод 4.

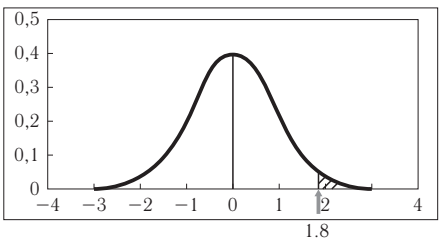
1. При нормальном распределении, когда среднее значение = 45 и стандартное отклонение = 10, площадь заштрихованной на графике области = 0,5 − 0,4641 = 0,0359.



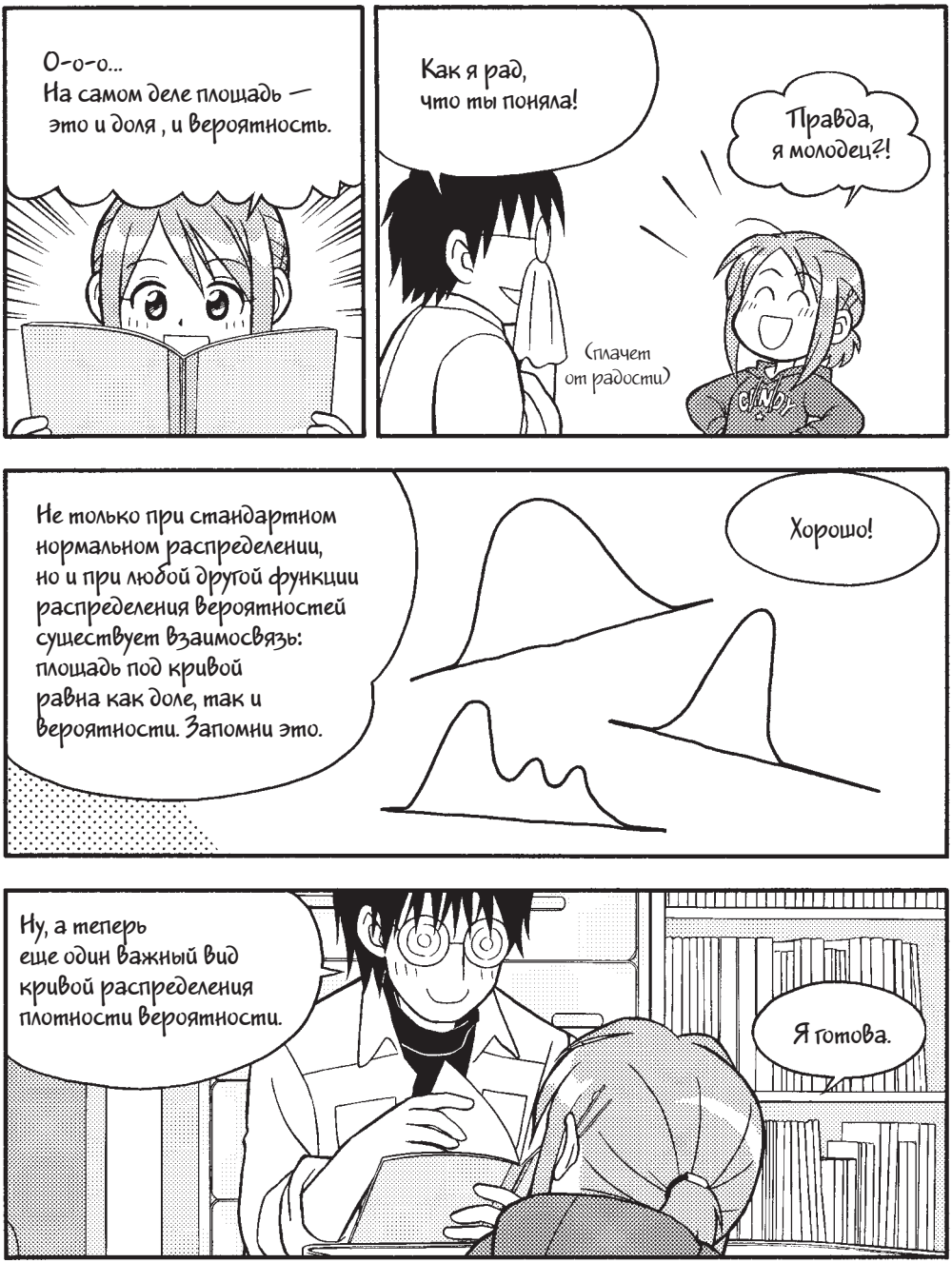
2. Доля учеников, чей результат > 63 баллов, = 0,5 − 0,4641= 0,0359 или 3,59% от общего числа сдававших экзамен.
3. Предположим, что из общего числа учеников был произвольно выбран один. Вероятность того, что он набрал > 63 баллов, = 0,5 − 0,4641 = 0,0359 или 3,59%.
4. При нормальном распределении доля учеников с нормированным отклонением

$$\text{балла} > 1.8 = \frac{18}{10} = \frac{63-45}{10} = \frac{\text{Значение} - \text{Среднее значение}}{\text{Стандартное отклонение}}$$

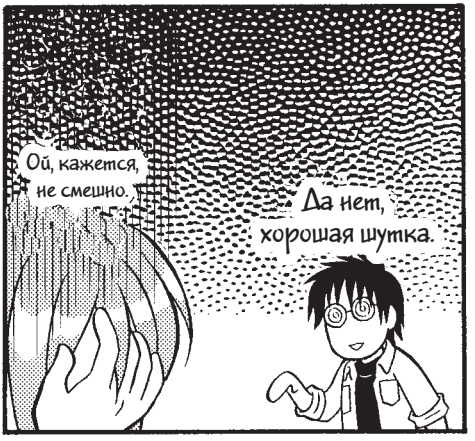
составляет 3,59% (0,5 − 0,4641 = 0,0359)
(см. Таблицу стандартного нормального распределения).



5. Предположим, что оценки учеников после нормирования распределены по стандартному нормальному закону. Вероятность того, что нормированное отклонение произвольно выбранного ученика > 1,8 равна 3,59%.


О-о-о...
На самом деле площадь — это и доля, и вероятность.
Как я рад, что ты поняла!
Правда, я молодец?!
(плачет от радости)
Не только при стандартном нормальном распределении, но и при любой другой функции распределения вероятностей существует взаимосвязь: площадь под кривой равна как доле, так и вероятности. Запомни это.
Хорошо!
Ну, а теперь еще один важный вид кривой распределения плотности вероятности.
Я готова.

4. Распределение хи-квадрат

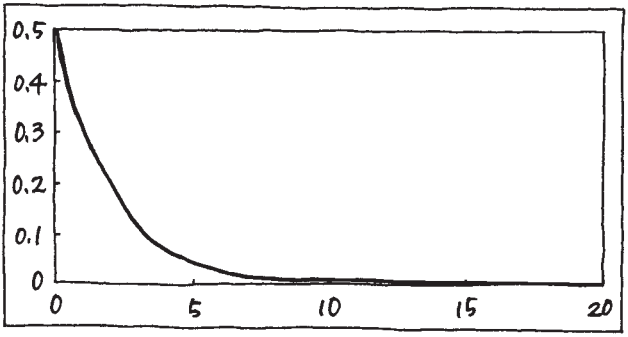


Если функция распределения вероятностей выражается формулой

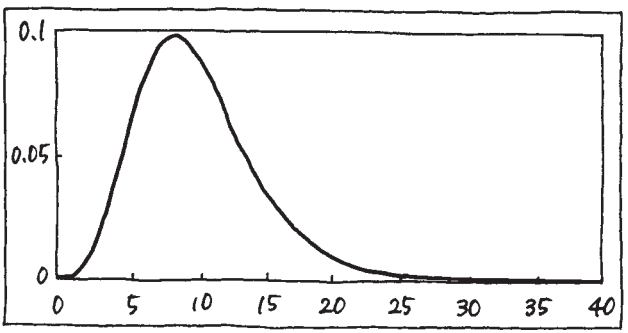
$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}} \times \int_0^{\infty} x^{\frac{n}{2}-1} e^{-x}\, dx} \times x^{\frac{n}{2}-1} \times e^{-\frac{x}{2}} & при\ x > 0, \\ 0 \quad при\ x \leqslant 0, \end{cases}$$

то в статистике говорят, что величина x имеет распределение хи-квадрат с числом степеней свободы n.

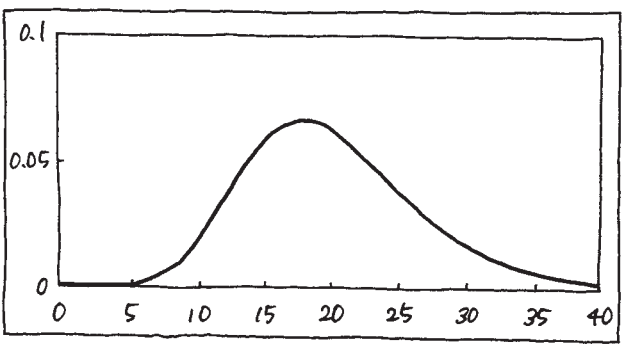
Число степеней свободы равно 2

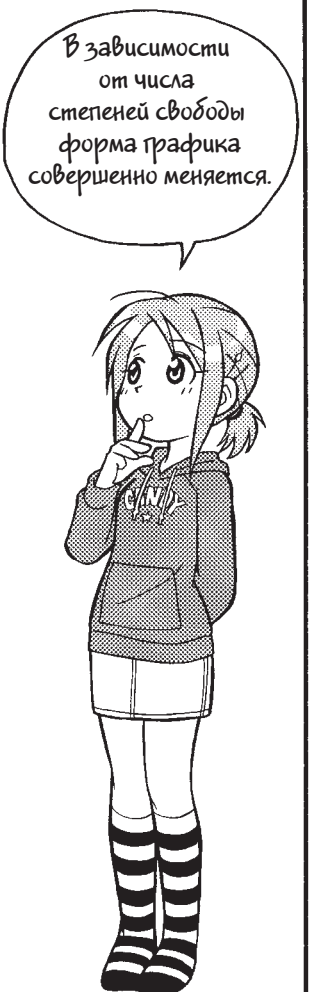


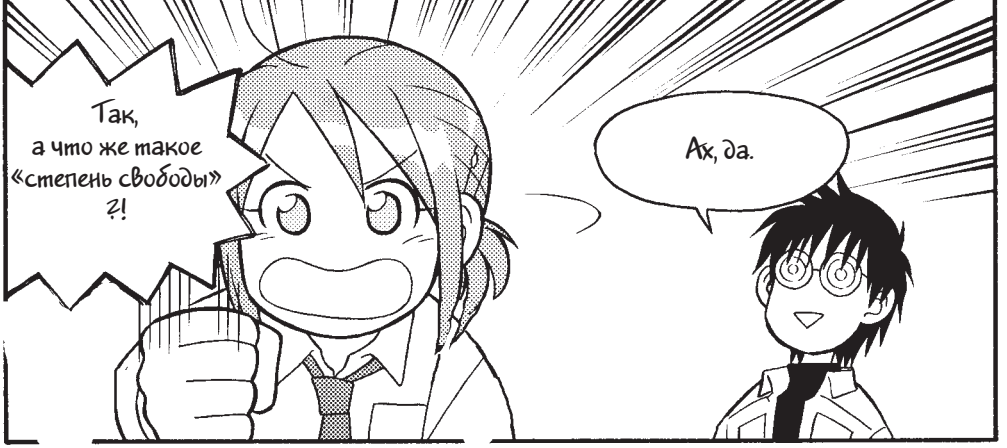
Число степеней свободы равно 10



Число степеней свободы равно 20



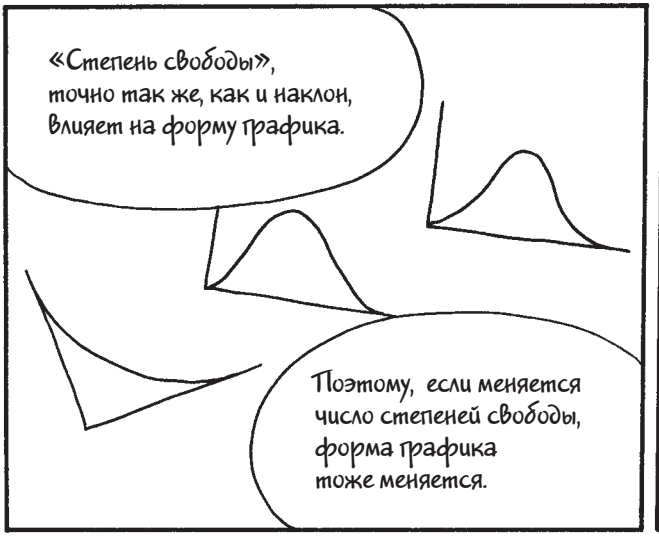



Так,
а что же такое
«степень свободы»
?!
Ах, да.

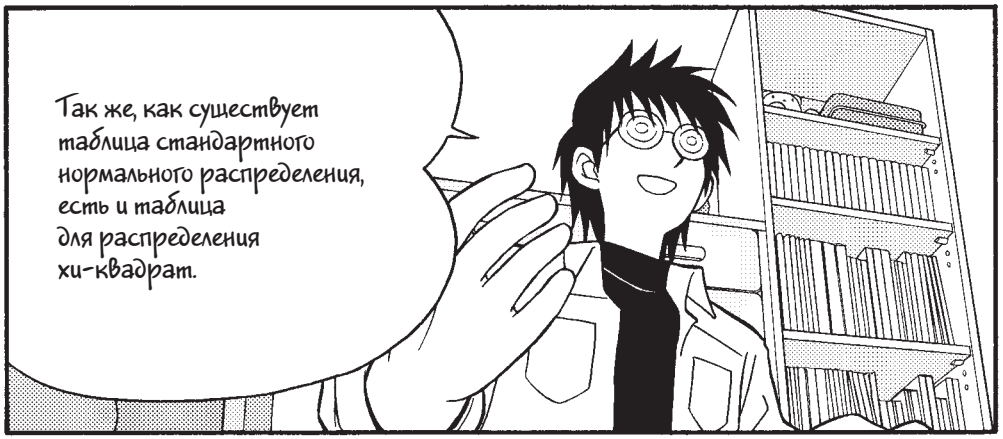

Что такое «a» в линейной функции f(x) = ax + b ?
А почему это ты меня об этом спрашиваешь?


b
f(x)=ax+b
Может, наклон?


Правильно!
Если меняется значение a,
меняется и наклон прямой на графике.
Угу.


«Степень свободы»,
точно так же, как и наклон,
влияет на форму графика.
Поэтому, если меняется
число степеней свободы,
форма графика
тоже меняется.


Так вот
что такое
«степень
свободы»!
Она также
зависит от
размера выборки.
Чем больше
выборка, тем
больше степеней
свободы.


Так же, как существует таблица стандартного нормального распределения, есть и таблица для распределения хи-квадрат.

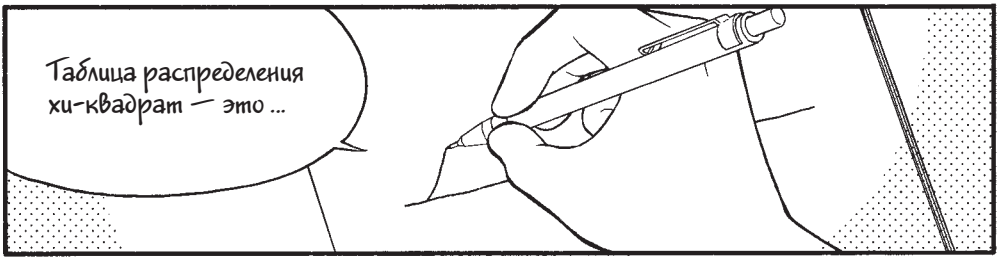
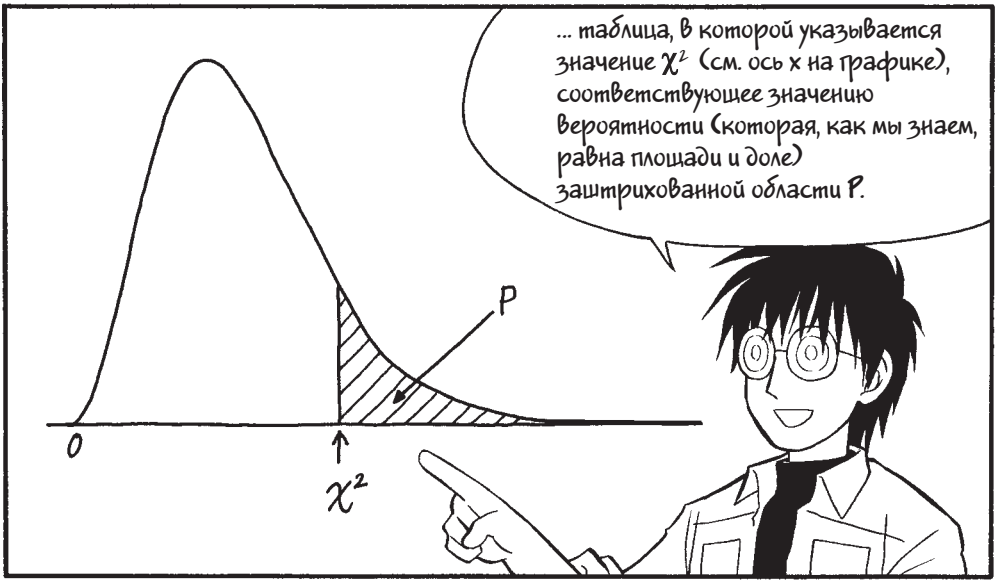


Таблица распределения хи-квадрат — это ...


... таблица, в которой указывается значение χ^2 (см. ось x на графике), соответствующее значению вероятности (которая, как мы знаем, равна площади и доле) заштрихованной области P.
P
0
χ^2

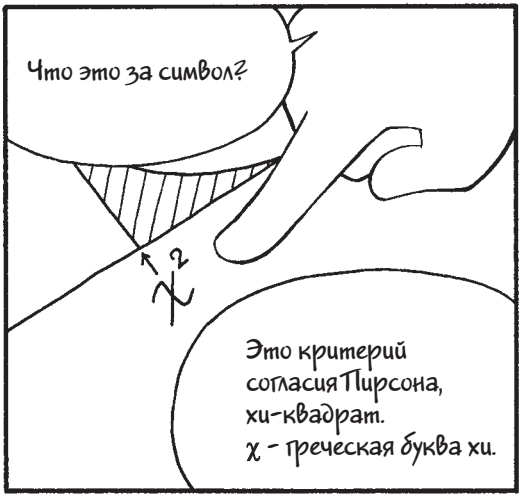
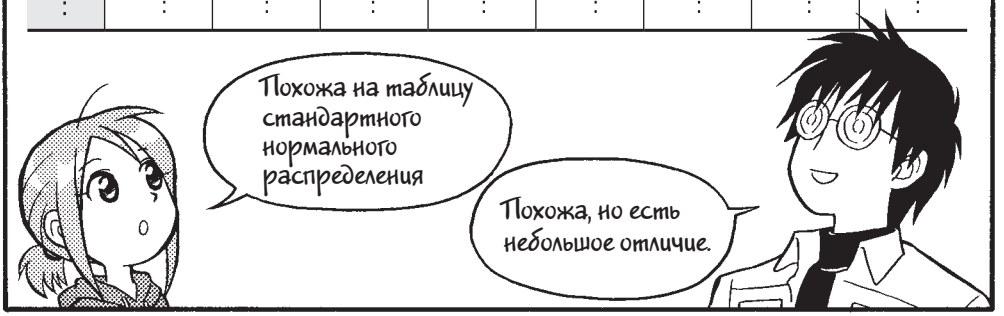


Таблица распределения хи-квадрат

Степень свободы \ P	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000039	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2104	10.5965
3	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449	12.8381
4	0.2070	0.2971	0.4844	0.7107	9.4877	11.1433	13.2767	14.8602
5	0.4118	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119	18.5475
7	0.9893	1.2390	1.6899	2.1673	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	15.5073	17.5345	20.0902	21.9549
9	1.7349	2.0879	2.7004	3.3251	16.9190	19.0228	21.6660	23.5893
10	2.1558	2.5582	3.2470	3.9403	18.3070	20.4832	23.2093	25.1881
:	:	:	:	:	:	:	:	:



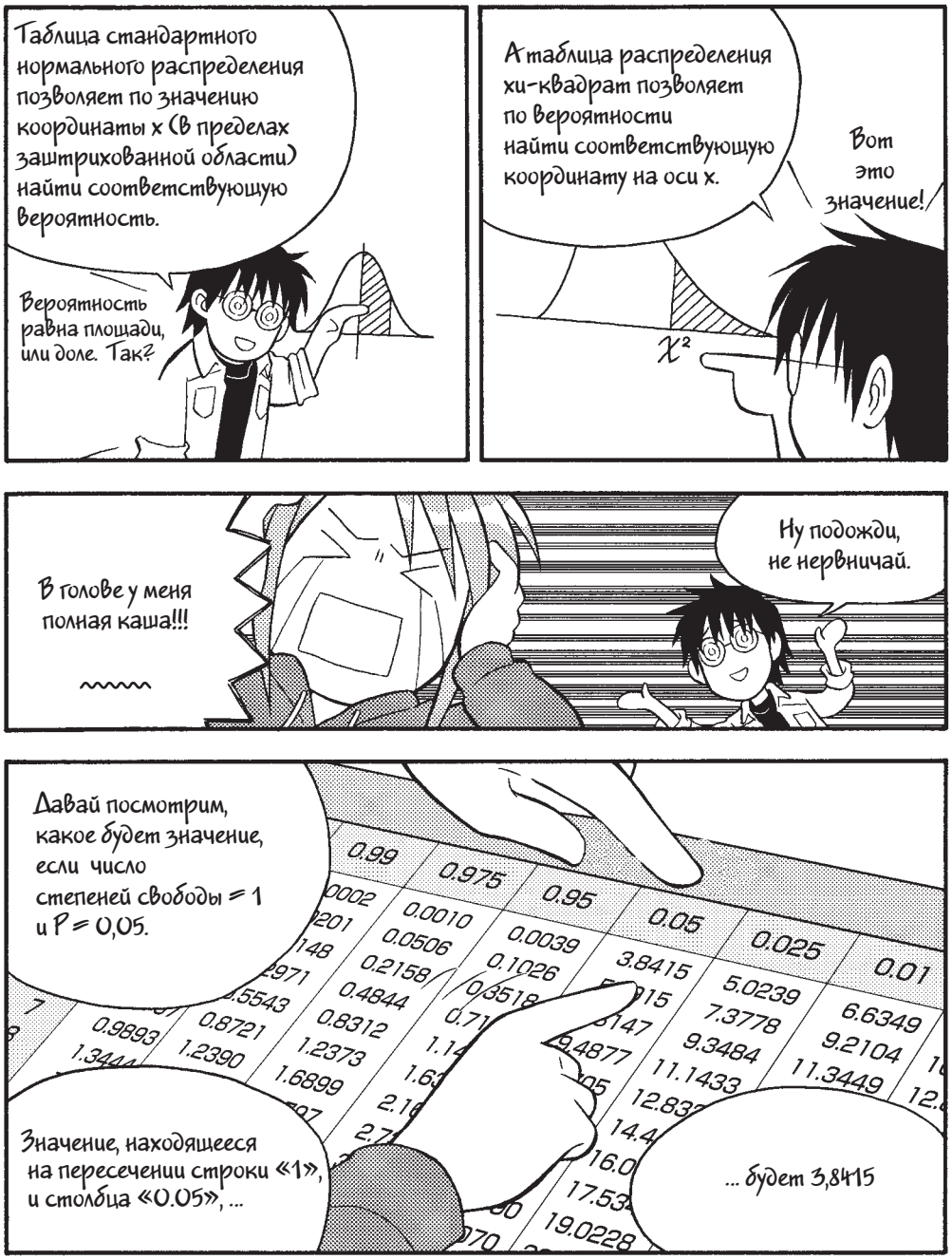
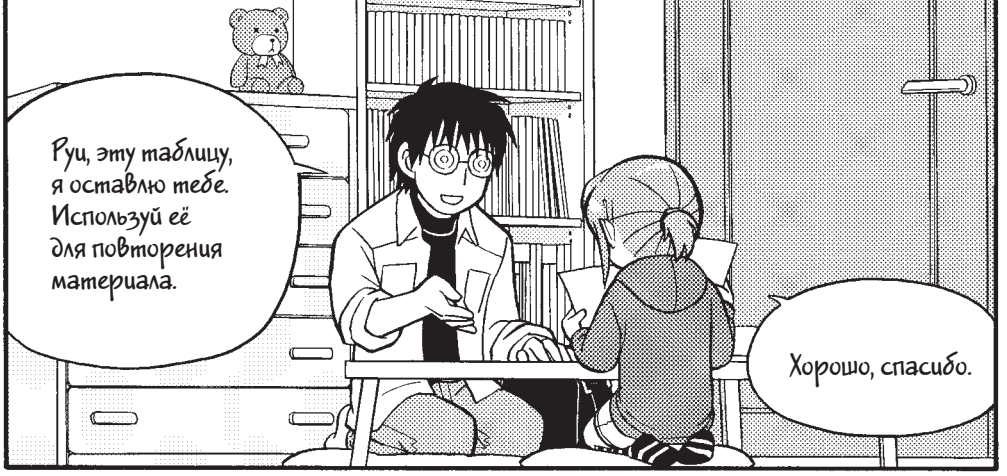
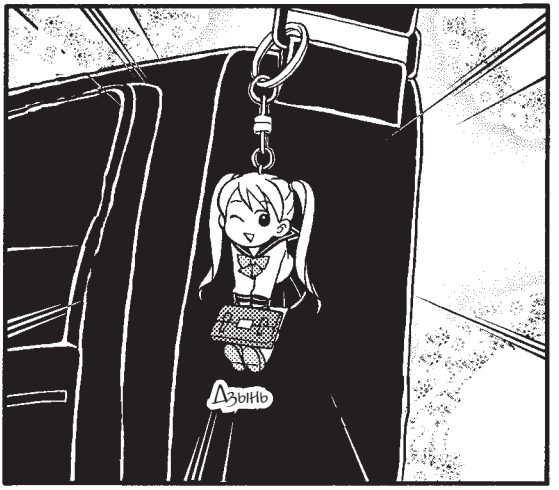
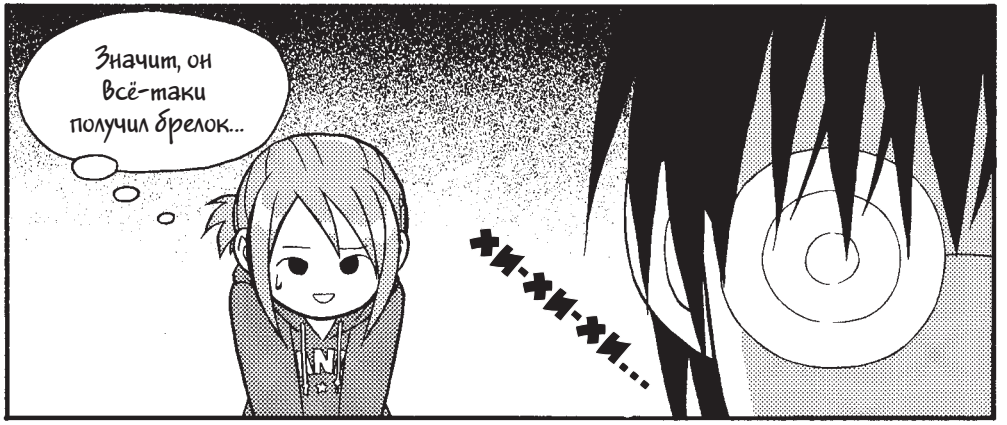


Таблица стандартного нормального распределения позволяет по значению координаты x (в пределах заштрихованной области) найти соответствующую вероятность.
Вероятность равна площади, или доле. Так?
А таблица распределения хи-квадрат позволяет по вероятности найти соответствующую координату на оси x.
Вот это значение!
χ^2
В голове у меня полная каша!!!
Ну подожди, не нервничай.
Давай посмотрим, какое будет значение, если число степеней свободы = 1 и P = 0,05.
0.99
0.975
0.95
0.05
0.025
0.01
0.0002
0.0010
0.0039
3.8415
5.0239
6.6349
0.0201
0.0506
0.1026
7.3778
9.2104
0.148
0.2158
0.3518
9.3484
11.3449
0.9893
1.3440
0.5543
0.8721
0.4844
1.2390
0.8312
1.2373
1.6899
11.1433
12.83
14.4
16.0
17.53
19.0228
Значение, находящееся на пересечении строки «1», и столбца «0.05», ...
... будет 3,8415


Руи, эту таблицу, я оставлю тебе. Используй её для повторения материала.
Хорошо, спасибо.


Ну, тогда на сегодня всё.
Э-эх... (встаёт)
Ты хорошо сегодня потрудилась.


Дзынь


Значит, он всё-таки получил брелок...
Хи-хи-хи...

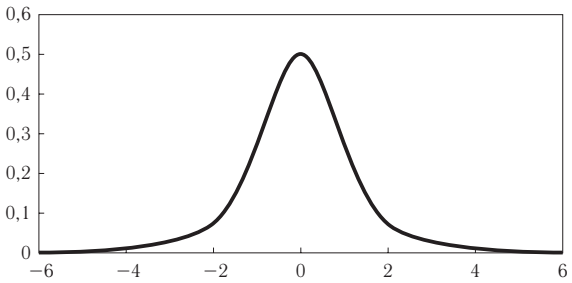
5. Распределение Стьюдента

В статистике часто используется и такая формула распределения вероятностей:

$$f(x) = \frac{\int_0^\infty x^{\frac{n+1}{2}-1} e^{-x} dx}{\sqrt{n\pi}\ \int_0^\infty x^{\frac{n}{2}-1} e^{-x} dx} \times \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

где n — число степеней свободы. Если плотность распределения вероятностей можно выразить с помощью этой формулы, в статистике это означает, что величина x имеет распределение Стьюдента с числом степеней свободы n.

Распределение Стьюдента с числом степеней свободы 5



6. Распределение Фишера, или F-распределение

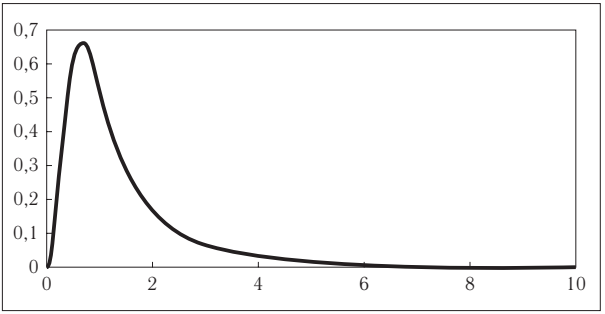
Не менее часто используется и такая функция распределения плотности вероятности:

$$f(x) = \begin{cases} \dfrac{\left(\int_0^\infty x^{\frac{n+m}{2}-1} e^{-x} dx\right) \times n^{\frac{n}{2}} \times m^{\frac{m}{2}}}{\left(\int_0^\infty x^{\frac{n}{2}-1} e^{-x} dx\right) \times \left(\int_0^\infty x^{\frac{m}{2}-1} e^{-x} dx\right)} \times \dfrac{x^{\frac{n}{2}-1}}{(nx+m)^{\frac{n+m}{2}}} & \text{при } x > 0, \\ 0 \quad \text{при } x \leqslant 0, \end{cases}$$

где n и m — число степеней свободы величины x.

Если формула плотности распределения вероятности имеет такой вид, в статистике это означает, что x имеет F-распределение с числом степеней свободы n и m.

Случай, когда степень свободы n = 10, а m = 5



7. Распределения и Excel

До начала 90-х вычисление вероятности и значения *x* было настолько сложным и трудоёмким, что его можно было выполнить только с помощью таблиц стандартного нормального распределения и распределения хи-квадрат. Однако, по мере развития компьютерных технологий, необходимость в этих таблицах отпала, поскольку величины, которые указаны в таблицах, можно получить с помощью программы Excel.

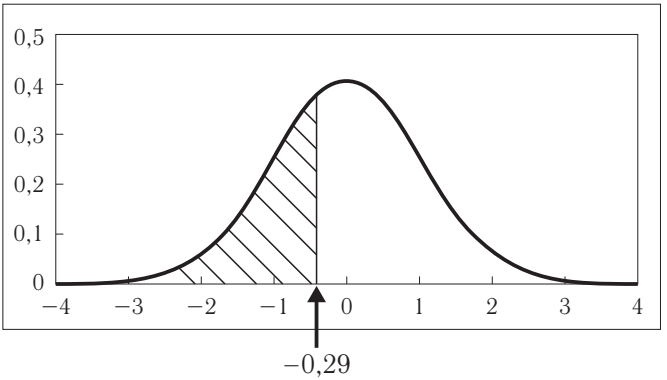
Таблица 5.1.

Распределение	Функция		Значение функции
	русский	английский	
Нормальное	НОРМРАСП	NORMDIST	Вероятность для заданного *x*
распределение*	НОРМОБР	NORMINV	Значение *x* для заданной вероятности
Стандартное нормальное	НОРМСТРАСП	NORMSDIST	Вероятность для заданного *x*
распределение	НОРМСТОБР	NORMSINV	Значение *x* для заданной вероятности
Распределение	ХИ2РАСП	CHIDIST	Вероятность для заданного *x*
хи-квадрат	ХИ2ОБР	CHIINV	Значение *x* для заданной вероятности
Распределение	СТЬЮДРАСП	TDIST	Вероятность для заданного *x*
Стьюдента	СТЬЮДРАСПОБР	TINV	Значение *x* для заданной вероятности
Распределение Фишера	FРАСП	FDIST	Вероятность для заданного *x*
	FРАСПОБР	FINV	Значение *x* для заданной вероятности

* Функция распределения вероятностей (речь идёт о нормальном распределении) зависит от таких параметров, как среднее значение и стандартное отклонение. Поэтому таблицу нормального распределения, даже при большом желании, невозможно составить. Но с помощью Excel можно сформировать таблицу, соответствующую таблице нормального распределения, что очень удобно.

Упражнение

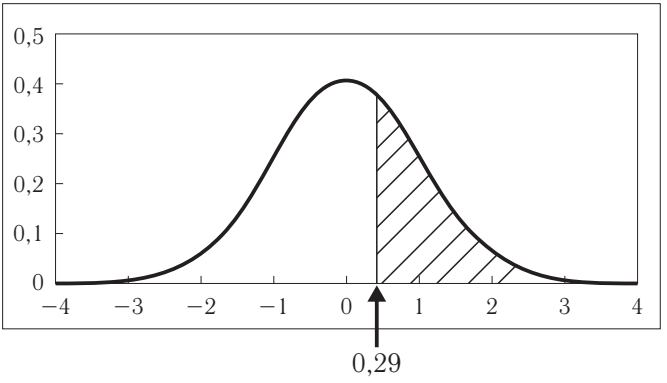
1. Используя таблицу стандартного нормального распределения на стр. 93, вычислите вероятность для заштрихованной на графике области.



2. Используя таблицу распределения хи-квадрат на стр. 103, вычислите значение χ^2, если число степеней свободы равно 2, а $P = 0{,}05$.

Ответ

1. Искомая вероятность равна вероятности заштрихованной области.



Вероятность при $z = 0{,}29 = 0{,}2 + 0{,}09$, как следует из таблицы стандартного нормального распределения, равна 0,1141. Следовательно, искомая вероятность будет равна: $0{,}5 - 0{,}1141 = 0{,}3859$.

2. Как следует из таблицы распределения хи-квадрат, значение величины χ^2, которое нужно найти, равно 5,9915.

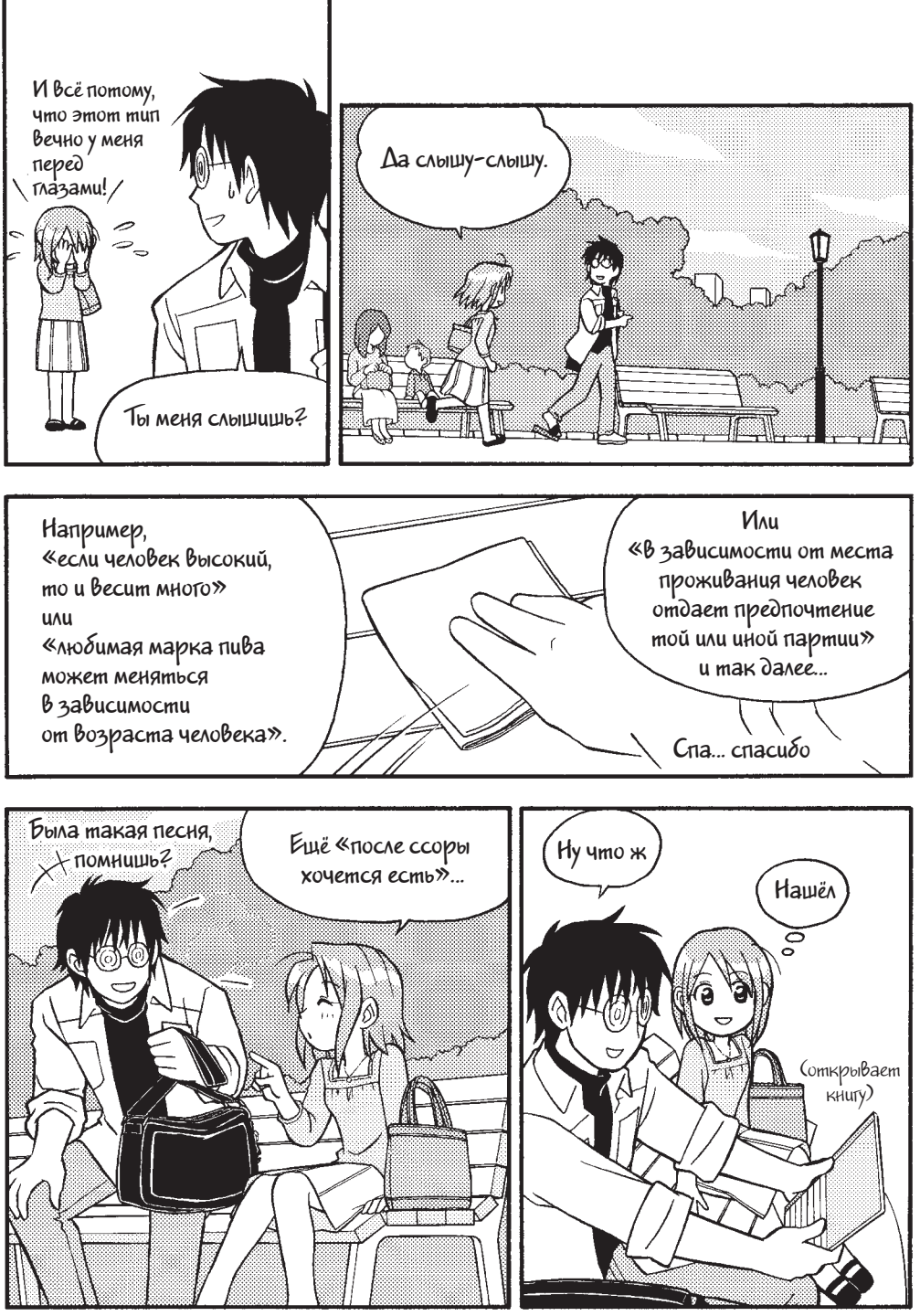
Выводы

- К широко используемым видам функции распределения вероятностей можно отнести:
  - нормальное распределение;
  - стандартное нормальное распределение;
  - распределение хи-квадрат;
  - распределение Стьюдента;
  - распределение Фишера.
- Площадь области, ограниченной осью x и кривой распределения вероятностей, равна 1.
- Площадь области, ограниченной осью x и кривой распределения вероятностей, тождественна и доле, и вероятности.
- С помощью таблиц распределений или программы Excel можно вычислить:
  - вероятность при определенном значении x;
  - значение x при определенном значении вероятности.

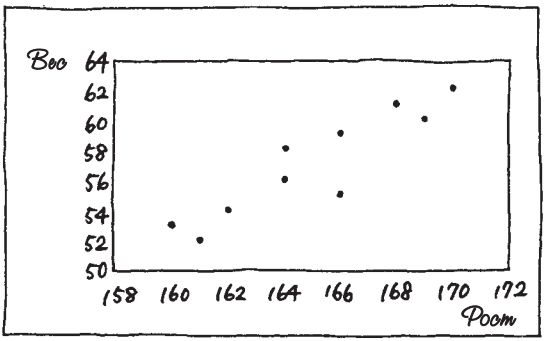
Глава 6

Что может связывать две переменные


(кто-то идёт, шаркая ногами)
На самом деле ...
Что это у Вас за вид? Ботинки разные!!
А-а-а, действительно.
... хорошо иногда провести урок на улице
Ах, если бы это был Игараси-сан, было бы так здорово...
Игараси-сан...
Игараси-сан... Э-э-э ?
Чёрт... Я уже стала забывать, как он выглядит...
Сегодня я тебе расскажу про связь между двумя переменными

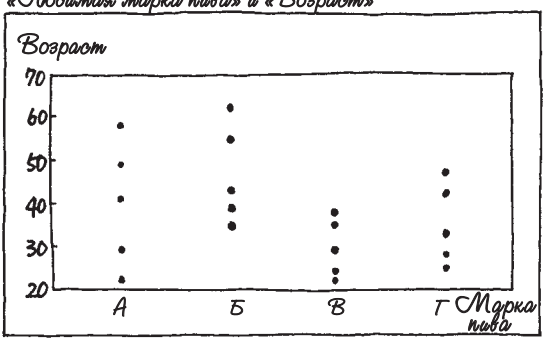

И всё потому, что этот тип вечно у меня перед глазами!
Ты меня слышишь?
Да слышу-слышу.
Например, «если человек высокий, то и весит много» или «любимая марка пива может меняться в зависимости от возраста человека».
Или «в зависимости от места проживания человек отдает предпочтение той или иной партии» и так далее...
Спа... спасибо
Была такая песня, помнишь?
Ещё «после ссоры хочется есть»...
Ну что ж
Нашёл
(открывает книгу)

Точечная диаграмма «Рост» и «Вес»



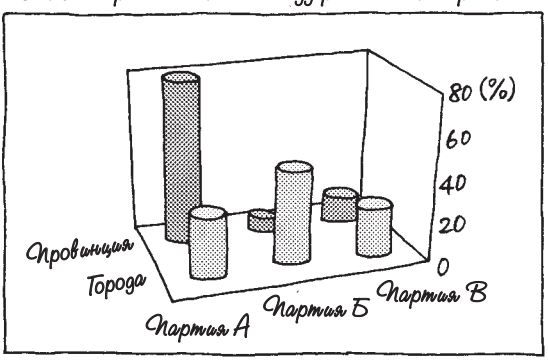
Количественные данные и количественные данные

Точечная диаграмма «Любимая марка пива» и «Возраст»



Количественные данные и качественные данные

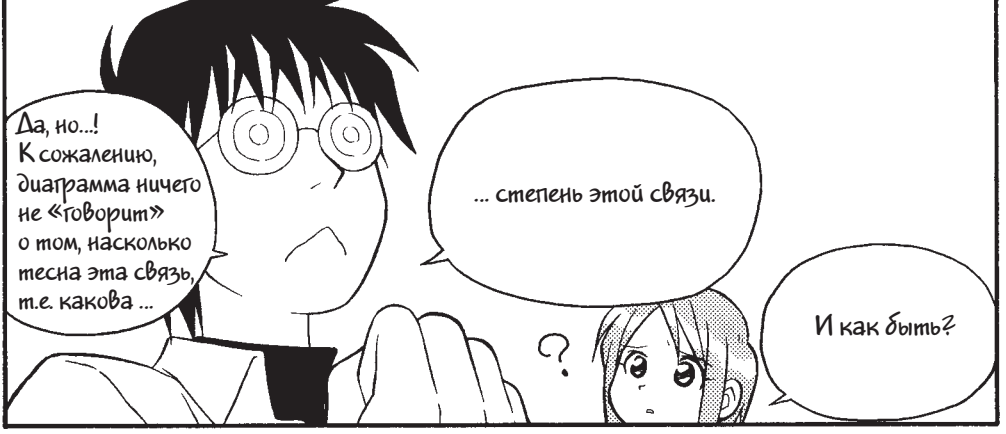
Столбиковая диаграмма «Место проживания» и «Поддерживаемая партия»

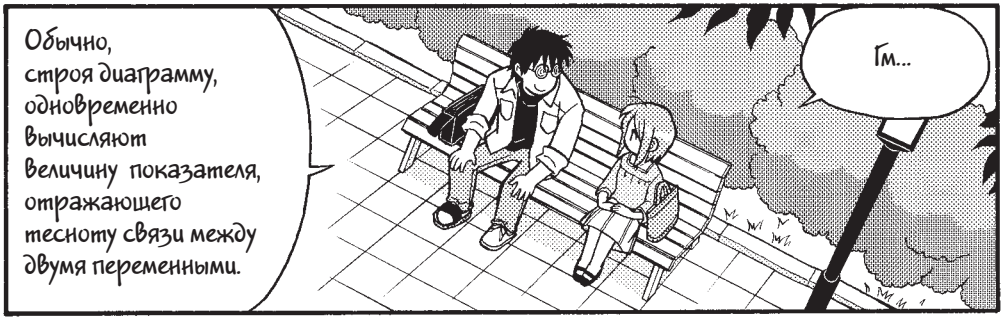


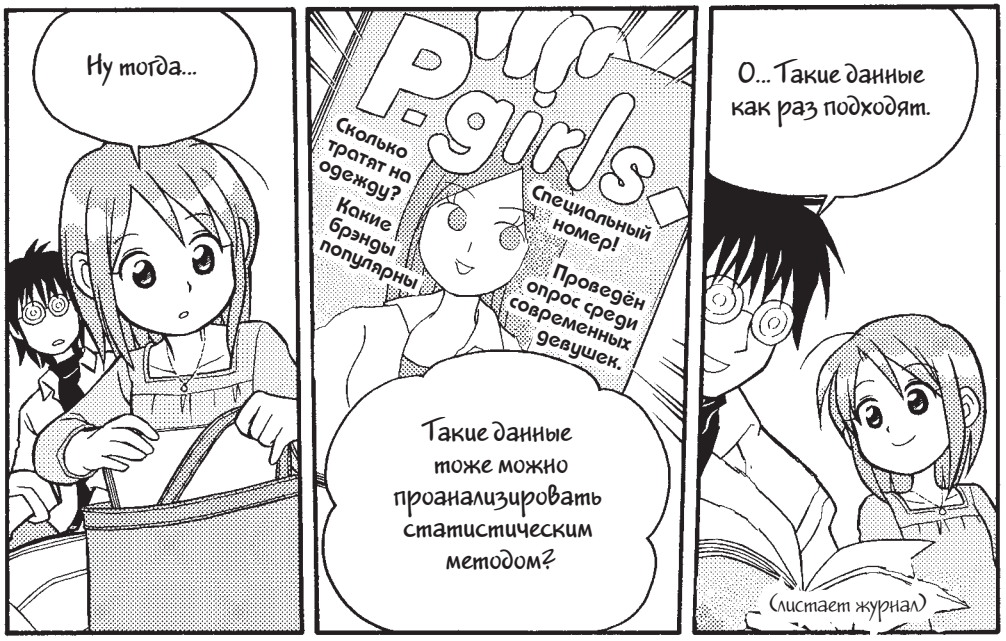
Качественные данные и качественные данные

Когда построишь диаграмму, сразу ясно, есть или нет взаимосвязь между двумя переменными.

Ага


Да, но...! К сожалению, диаграмма ничего не «говорит» о том, насколько тесна эта связь, т.е. какова ...
... степень этой связи.
И как быть?


Обычно, строя диаграмму, одновременно вычисляют величину показателя, отражающего тесноту связи между двумя переменными.
Гм...


Ну тогда...
P-girls
Сколько тратят на одежду?
Специальный номер!
Какие брэнды популярны
Проведён опрос среди современных девушек.
Такие данные тоже можно проанализировать статистическим методом?
О... Такие данные как раз подходят.
(листает журнал)

1. Коэффициент линейной корреляции

Ой, есть анкета: «расходы на косметику» и «расходы на одежду».

Количественные и количественные данные

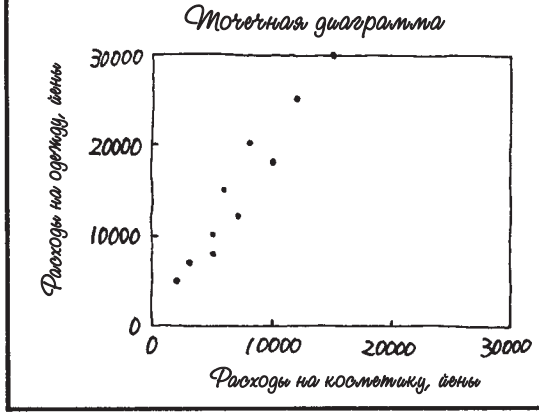
Расходы на косметику и Расходы на одежду (в месяц)

Ответы 10 женщин в возрасте от 20 до 29 лет

Респондент	Расходы на косметику, йены	Расходы на одежду, йены
А	3000	7000
Б	5000	8000
В	12000	25000
Г	2000	5000
Д	7000	12000
Е	15000	30000
Ж	5000	10000
З	6000	15000
И	8000	20000
К	10000	18000

Для начала построим диаграмму.

Ага.

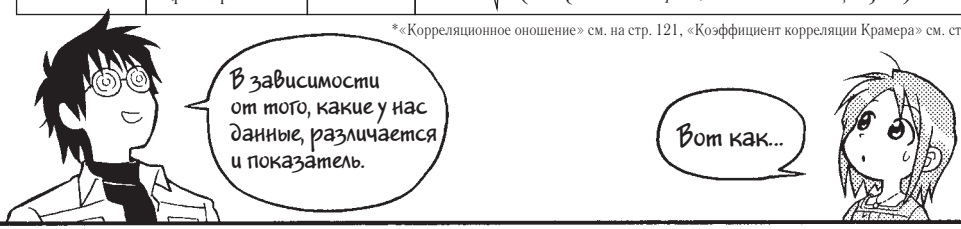


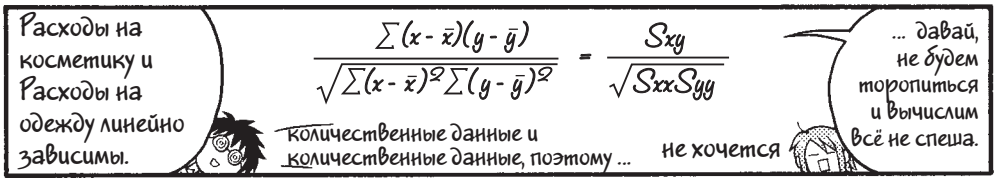
Понятно. Похоже, что тот, кто много тратит на косметику, много тратит и на одежду.

А теперь определим тесноту связи.

Тип данных	Показатель	Значение	Формула
Количеств. и количеств.	Коэффициент линейной корреляции	−1 ... 1	$\frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{Sxy}{\sqrt{Sxx\ Syy}}$
Количеств. и качеств.	Корреляционное отношение	0 ... 1	$\frac{\text{Межгрупповая дисперсия}}{\text{Внутригрупповая дисперсия} + \text{Межгрупповая дисперсия}}$
Качеств. и качеств.	Коэффициент корреляции Крамера	0 ... 1	$\sqrt{\frac{\chi_0^2}{n\left(\min\{\text{кол-во строк, кол-во столбцов}\}-1\right)}}$

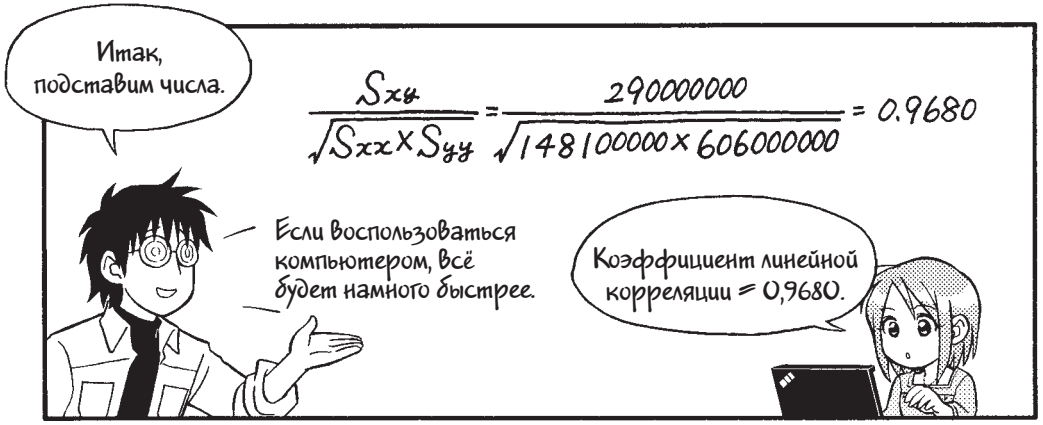
*«Корреляционное соношение» см. на стр. 121, «Коэффициент корреляции Крамера» см. стр. 127.





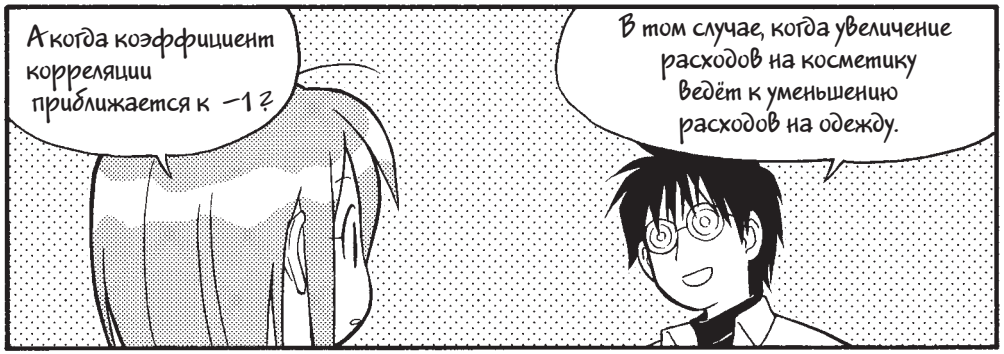
Порядок вычисления коэффициента линейной корреляции (для определения тесноты связи между «расходами в месяц на косметику» и «расходами в месяц на одежду»)

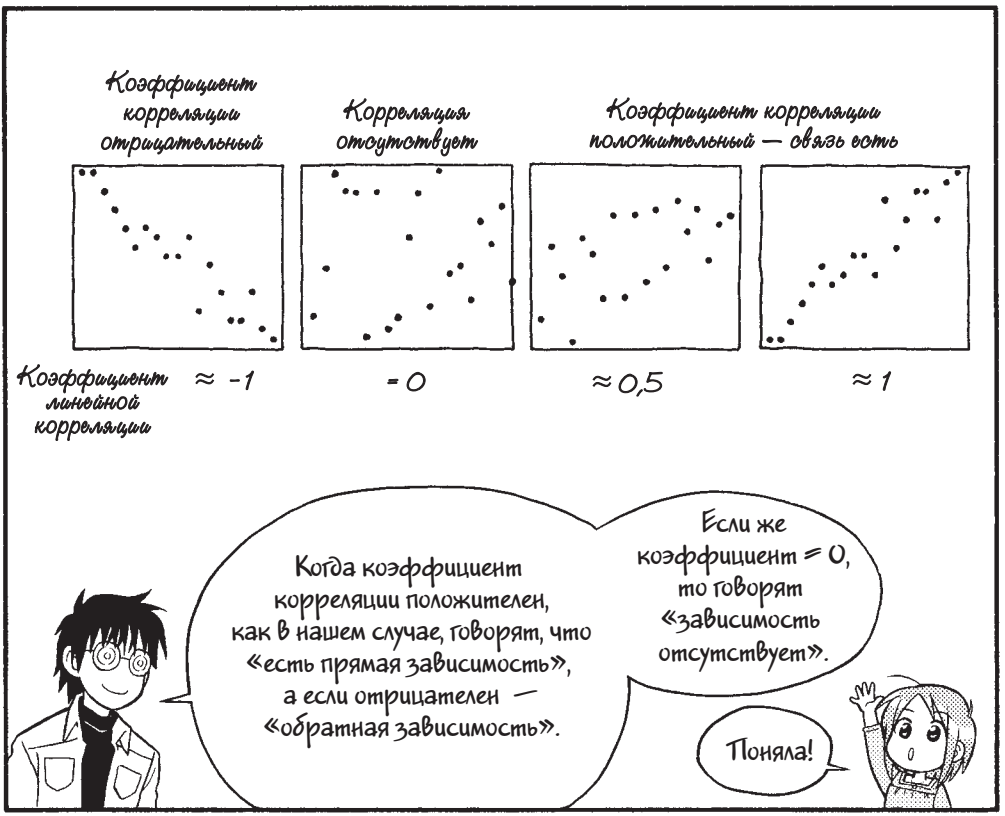
Респондент	Расходы на косметику, йены	Расходы на одежду, йены					
	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
А	3000	7000	−4300	−8000	18490000	64000000	34400000
Б	5000	8000	−2300	−7000	5290000	49000000	16100000
В	12000	25000	4700	10000	22090000	100000000	47000000
Г	2000	5000	−5300	−10000	28090000	100000000	53000000
Д	7000	12000	−300	−3000	90000	9000000	900000
Е	15000	30000	7700	15000	59290000	225000000	115500000
Ж	5000	10000	−2300	−5000	5290000	25000000	11500000
З	6000	15000	−1300	0	1690000	0	0
И	8000	20000	700	5000	490000	25000000	3500000
К	10000	18000	2700	3000	7290000	9000000	8100000
Сумма	73000	150000	0	0	148100000	606000000	290000000
Ср. знач.	7300	15000			↓	↓	↓
	↳ $\bar{x}$	↳ $\bar{y}$			Sxx	Syy	Sxy


Итак, подставим числа.
$\frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{290000000}{\sqrt{148100000 \times 606000000}} = 0.9680$
Если воспользоваться компьютером, всё будет намного быстрее.
Коэффициент линейной корреляции = 0,9680.


Когда между двумя переменными существует тесная связь, коэффициент корреляции приближается к ±1, а когда связь слабая, он приближается к 0.
-1 ← 0 → +1
Гм...

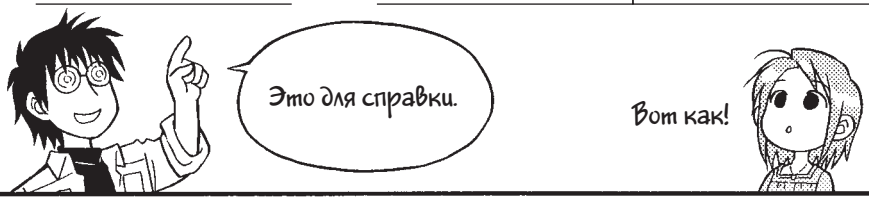

Этот результат довольно близок к 1, т.е. между расходами на косметику и расходами на одежду существует тесная связь, так?
P.girls
Примерно так.


А когда коэффициент корреляции приближается к −1?
В том случае, когда увеличение расходов на косметику ведёт к уменьшению расходов на одежду.


Коэффициент корреляции отрицательный
Корреляция отсутствует
Коэффициент корреляции положительный — связь есть
Коэффициент линейной корреляции
≈ -1
= 0
≈ 0,5
≈ 1
Когда коэффициент корреляции положителен, как в нашем случае, говорят, что «есть прямая зависимость», а если отрицателен — «обратная зависимость».
Если же коэффициент = 0, то говорят «зависимость отсутствует».
Поняла!
Кстати, кое-что ещё о коэффициенте линейной корреляции.
К сожалению, в статистике нет какого-то определённого значения коэффициента корреляции, свыше которого две переменные считаются тесно связанными.
Как-то уклончиво ...

Критерии величины коэффициента линейной корреляции

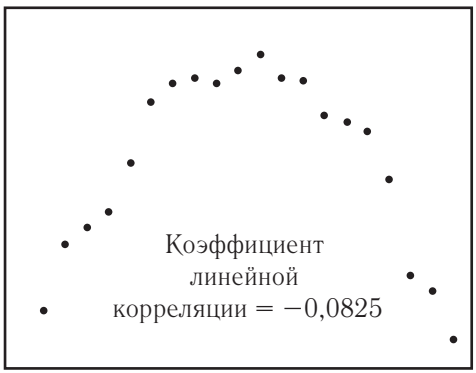
Абсолютная величина коэфф-та лин. корреляции		Вывод о степени взаимосвязи	Вывод о наличии взаимосвязи
1,0—0,9	⇒	Очень тесная	Есть
0,9—0,7	⇒	Достаточно тесная	
0,7—0,5	⇒	Слабая	
< 0,5	⇒	Очень слабая	Нет



Примечание

До этого я говорил, что коэффициент линейной корреляции — это показатель тесноты связи между количественными данными. Строго говоря, это не совсем так. Коэффициент корреляции еще показывает, является эта зависимость линейной или нет.

Пример нелинейной зависимости

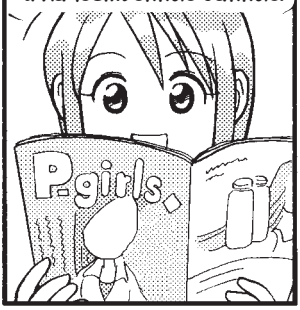


Например, эта диаграмма демонстрирует очевидную связь между двумя переменными.
Но поскольку зависимость нелинейная, коэффициент линейной корреляции практически равен нулю.

2. Коэффициент корреляции между данными разных типов

Теперь, перейдём к следующей теме! Здесь есть данные о «возрасте» и «любимом бренде одежды».

Это количественные и качественные данные.



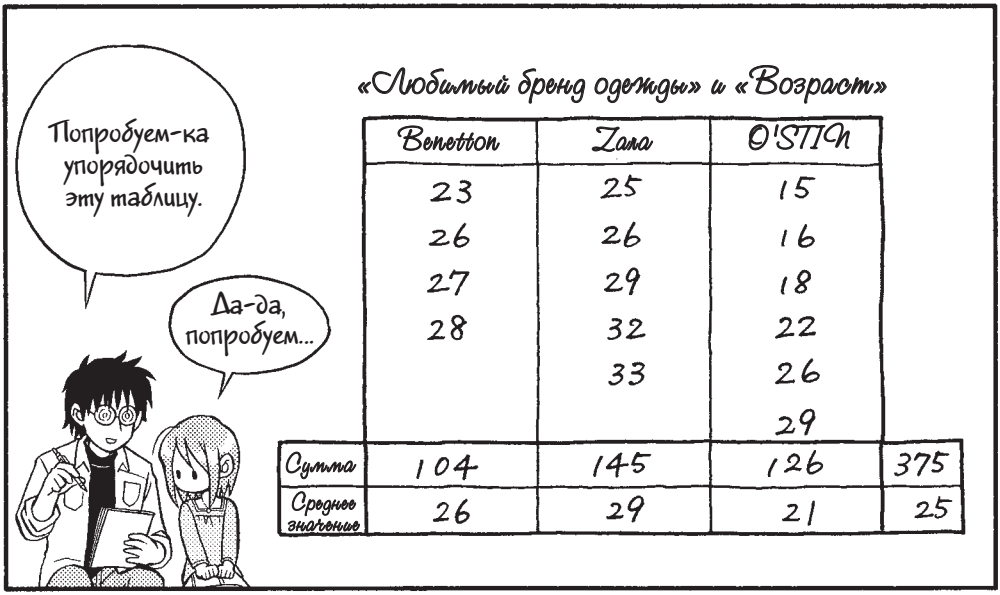
Возраст и Любимый бренд одежды

Респондент	Возраст	Бренд
А	27	Benetton
Б	33	Zara
В	16	O'STIN
Г	29	Бапари
Д	32	Zara
Е	23	Benetton
Ж	25	Zara
З	28	Benetton
И	22	O'STIN
К	18	O'STIN
Л	26	Zara
М	26	Benetton
Н	15	O'STIN
О	29	Zara
П	26	O'STIN

Для случая, когда одни данные количественные, а другие качественные, вычисляется коорреляционное отношение, которое может принимать значения от 0 до 1.

А в этом случае приближение значения к 1 тоже свидетельствует о наличии тесной связи между данными?

Да, это так.


Попробуем-ка упорядочить эту таблицу.
Да-да, попробуем...
«Любимый бренд одежды» и «Возраст»
Benetton
Zara
O'STIN
23
26
27
28
25
26
29
32
33
15
16
18
22
26
29
Сумма
104
145
126
375
Среднее значение
26
29
21
25

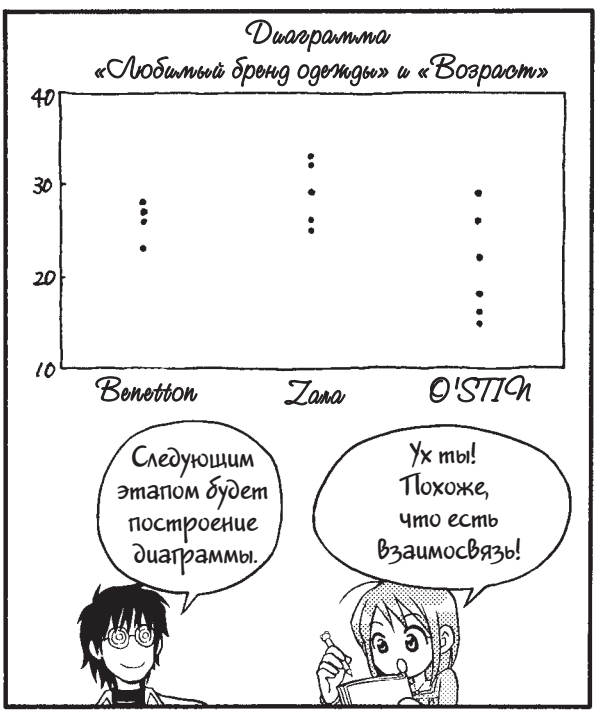
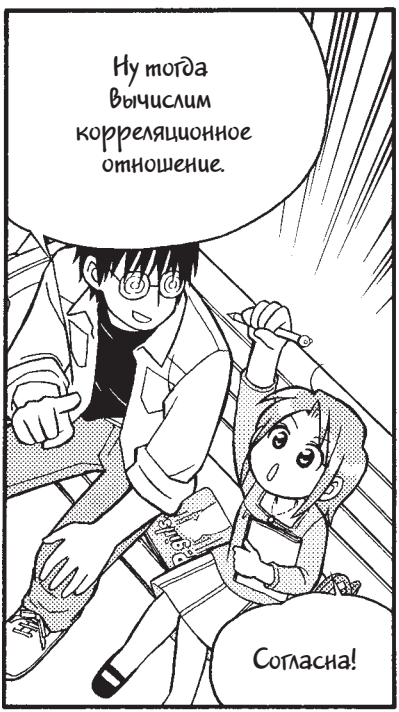


Диаграмма
«Любимый бренд одежды» и «Возраст»
40
30
20
10
Benetton
Zara
O'STIN
Следующим этапом будет построение диаграммы.
Ух ты! Похоже, что есть взаимосвязь!


Ну тогда вычислим корреляционное отношение.
Согласна!

Вычисление корреляционного отношения предусматривает выполнение нижеследующих шагов.

Шаг 1

Вычислить суммы стандартных отклонений для каждого столбца таблицы:

	(значение — среднее значение в столбце Benetton)^2	(значение – среднее значение в столбце Zara)^2	(значение — среднее значение в столбце O'STIN)^2
	$(23-26)^2=(-3)^2=9$ $(26-26)^2=0^2=0$ $(27-26)^2=1^2=1$ $(28-26)^2=2^2=4$	$(25-29)^2=(-4)^2=16$ $(26-29)^2=(-3)^2=9$ $(29-29)^2=0^2=0$ $(32-29)^2=3^2=9$ $(33-29)^2=4^2=16$	$(15-21)^2=(-6)^2=36$ $(16-21)^2=(-5)^2=25$ $(18-21)^2=(-3)^2=9$ $(22-21)^2=1^2=1$ $(26-21)^2=5^2=25$ $(29-21)^2=8^2=64$
Сумма	14	50	160
	↓ S_{BB}	↓ S_{ZZ}	↓ S_{OO}

Шаг 2

Вычислить внутригрупповую дисперсию, показывающую как сильно отличаются данные в каждой категории, как сумму $S_{TT} + S_{CC} + S_{BB}$:

$$S_{BB} + S_{ZZ} + S_{OO} = 14 + 50 + 160 = 224$$

Шаг 3

Рассчитать межгрупповую дисперсию — меру разброса данных между категориями. Для этого надо найти стандартные отклонения средних значений от общего среднего значения, умножить их на количество данных в соответствующей графе и вычислить сумму полученных произведений, т. е.:

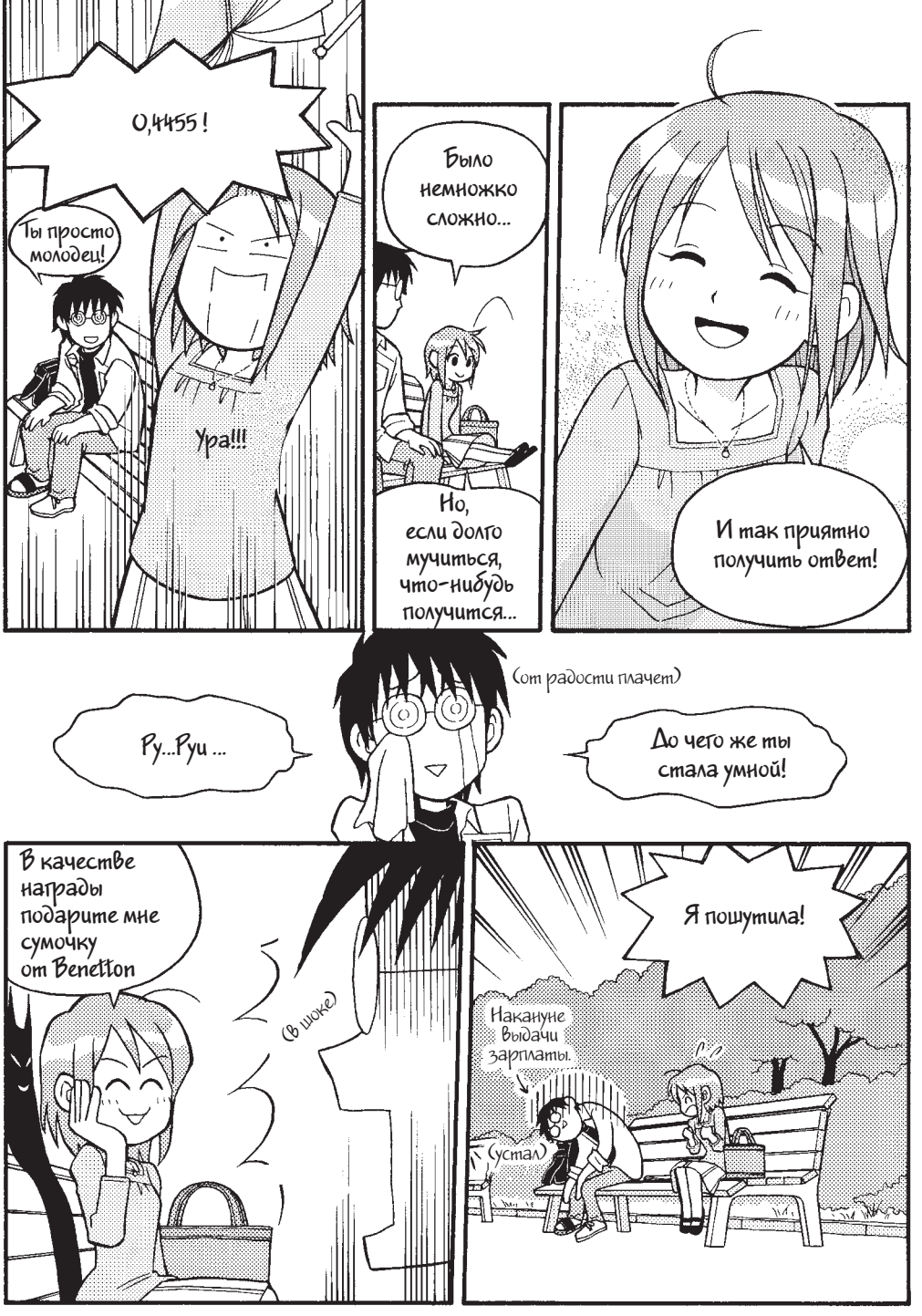
Кол-во значений Benetton × (Ср. знач. Benetton − общее ср. знач.)^2 +
+ Кол-во значений Zara × (Ср. знач. Zara − общее ср. знач.)^2 +
+ Кол-во значений O'STIN × (Ср. знач. O'STIN − общее ср. знач.)^2

$$
\begin{aligned}
&4 \times (26 - 25)^2 + 5 \times (29 - 25)^2 + 6 \times (21 - 25)^2 = \\
&= 4 \times 1 + 5 \times 16 + 6 \times 16 = \\
&= 4 + 80 + 96 = \\
&= 180
\end{aligned}
$$

Шаг 4

Вычислить корреляционное отношение по формуле:

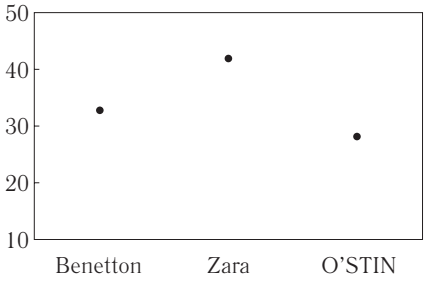
$$\frac{180}{224 + 180} = \frac{180}{404} = 0{,}4455$$


0,4455 !
Ура!!!
Ты просто молодец!
Было немножко сложно...
Но, если долго мучиться, что-нибудь получится...
И так приятно получить ответ!
(от радости плачет)
Ру...Руи ...
До чего же ты стала умной!
В качестве награды подарите мне сумочку от Benetton
(в шоке)
Я пошутила!
Накануне выдачи зарплаты.
(устал)

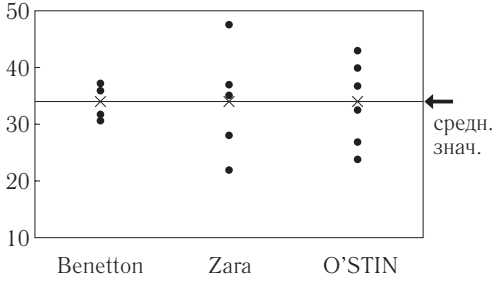
Как я уже говорил, корреляционное отношение может принимать значения от 0 до 1. Чем теснее взаимосвязь между двумя переменными, тем значение ближе к 1, чем слабее взаимосвязь, тем ближе к 0.

Диаграмма «Любимый бренд одежды» и «Возраст»

Корреляционное отношение = 1



Корреляционное отношение = 0



Корреляционное отношение = 1

⇕

Данные внутри каждой группы одинаковы

⇕

Внутригрупповая дисперсия = 0

Корреляционное отношение = 0

⇕

Средние значения всех групп одинаковы

⇕

Межгрупповая дисперсия = 0

К сожалению, в статистике нет кого-то определённого значения корреляционного отношения, свыше которого переменные считаются тесно связанными. Для справки приведём таблицу соответствия корреляционного отношения и степени взаимосвязи.

Критерии величины корреляционного отношения

Корреляционное отношение		Вывод о степени взаимосвязи	Вывод о наличии взаимосвязи
1,0—0,8	⇒	Очень тесная	Есть
0,8—0,5	⇒	Достаточно тесная	
0,5—0,25	⇒	Слабая	
< 0,25	⇒	Очень слабая	Нет

В нашем примере корреляционное отношение было равно 0,4455: значит связь слабая!

3. Коэффициент корреляции Крамера

Таблица взаимной сопряжённости «пола» и «способа признания в любви»

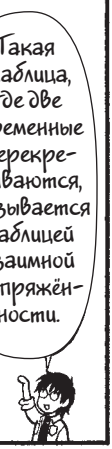
Пол респондента	Способ признания в любви, люди			Итого
	по телефону	по SMS	при встрече	
женский	34	61	53	148
мужской	38	40	74	152
Итого:	72	101	127	300

Это значит, что из 152 опрошенных молодых людей 74 хотели бы, чтобы им признавались в любви при встрече.

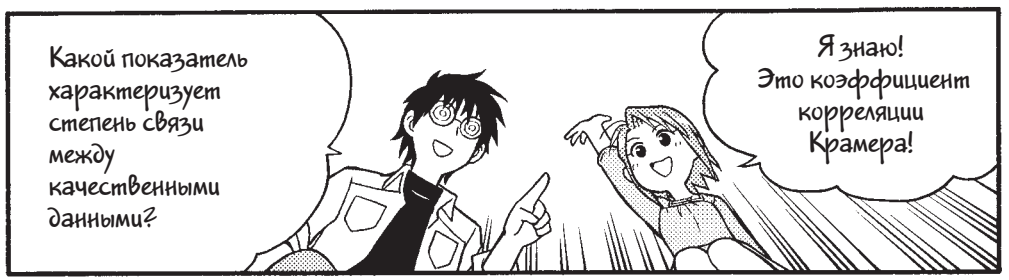
Таблица взаимной сопряжённости в %

Пол респондента	Способ признания в любви, %			Итого
	по телефону	по SMS	при встрече	
женский	23	41	36	100
мужской	25	26	49	100
Итого:	24	34	42	100

Это значит, что 49% опрошенных молодых людей хотели бы, чтобы им признавались в любви при встрече.




Похоже, что юноши и девушки предпочитают разные способы признания в любви.
Другими словами, есть взаимосвязь между «полом» и «способом признания».
Значит, лучше признаться в любви Игараси-сан при встрече
(бормочет)
Руи, ты меня слушаешь?
Да. Конечно!!


Какой показатель характеризует степень связи между качественными данными?
Я знаю! Это коэффициент корреляции Крамера!


Этот коэффициент называют либо коэффициент корреляции Крамера, либо коэффициент V Крамера, либо коэффициент независимости.
У меня голова кр...угом идёт.
Да нет, очень здорово у тебя получилось кр...угом.
(смутилась)
Не говори ничего...

Вычисление коэффициента корреляции Крамера предусматривает выполнение следующих шагов:

Шаг 1

Построить таблицу взаимной сопряжённости. Величины в клетках, обведённых жирной чертой, называются **эмпирическими частотами**.

Пол респондента	Способ признания в любви			Итого
	по телефону	по SMS	при встрече	
женский	34	61	53	148
мужской	38	40	74	152
Итого:	72	101	127	300

Шаг 2

Выполнить вычисления, как показано в следующей таблице. Величины в обведённых клетках называются **теоретическими частотами**.

Пол респондента	Способ признания в любви			Итого
	по телефону	по SMS	при встрече	
женский	$\frac{148 \times 72}{300}$	$\frac{148 \times 101}{300}$	$\frac{148 \times 127}{300}$	148
мужской	$\frac{152 \times 72}{300}$	$\frac{152 \times 101}{300}$	$\frac{152 \times 127}{300}$	152
Итого:	72	101	127	300

$$\frac{\text{Число опрошенных юношей} \times \text{Число ответов «при встрече»}}{\text{Общее число опрошенных}}$$

Если между «полом» и «способом признания в любви» полностью отсутствует какая-либо связь, то отношение эмпирических частот в строках «женский» и «мужской» будет таким-же, как отношение значений в строке «Итого» (см. Шаг 1).

Теоретические частоты, вычисленные на Шаге 2, отражают, каково было бы число соответствующих респондентов в случае полного отсутствия какой-либо связи между «полом» и «способом признания в любви».

Шаг 3

Вычислить:

$$\frac{(\textit{Эмпирическая частота} - \textit{Теоретическая частота})^2}{\textit{Теоретическая частота}}$$

Пол респондента	Способ признания в любви			Итого
	по телефону	по SMS	при встрече	
женский	$\dfrac{\left(34 - \dfrac{148 \times 72}{300}\right)^2}{\dfrac{148 \times 72}{300}}$	$\dfrac{\left(61 - \dfrac{148 \times 101}{300}\right)^2}{\dfrac{148 \times 01}{300}}$	$\dfrac{\left(53 - \dfrac{148 \times 127}{300}\right)^2}{\dfrac{148 \times 127}{300}}$	148
мужской	$\dfrac{\left(38 - \dfrac{152 \times 72}{300}\right)^2}{\dfrac{152 \times 72}{300}}$	$\dfrac{\left(40 - \dfrac{152 \times 101}{300}\right)^2}{\dfrac{152 \times 101}{300}}$	$\dfrac{\left(74 - \dfrac{152 \times 127}{300}\right)^2}{\dfrac{152 \times 127}{300}}$	152
Итого:	72	101	127	300

Чем больше разница между эмпирическими и теоретическими частотами, т.е. чем теснее связь между «полом» и «способом признания в любви», тем больше значения величин в клетках таблицы.

Шаг 4

Вычислить сумму величин в клетках таблицы на ***Шаге 3***.
Иначе говоря, вычислить значение критерия согласия Пирсона (χ_0^2).

$$\chi_0^2 = \frac{\left(34 - \frac{148 \times 72}{300}\right)^2}{\frac{148 \times 72}{300}} + \frac{\left(61 - \frac{148 \times 101}{300}\right)^2}{\frac{148 \times 101}{300}} + \frac{\left(53 - \frac{148 \times 127}{300}\right)^2}{\frac{148 \times 127}{300}} +$$

$$+ \frac{\left(38 - \frac{152 \times 72}{300}\right)^2}{\frac{152 \times 72}{300}} + \frac{\left(40 - \frac{152 \times 101}{300}\right)^2}{\frac{152 \times 101}{300}} + \frac{\left(74 - \frac{152 \times 127}{300}\right)^2}{\frac{152 \times 127}{300}} =$$

$$= 8{,}0091$$

Согласно Шагу 3, чем больше разница между эмпирическими и теоретическими частотами, т.е. чем теснее связь между «полом» и «способом признания», тем больше значение критерия согласия Пирсона (χ_0^2).

Шаг 5

Величина коэффициента корреляции Крамера вычисляется по формуле:

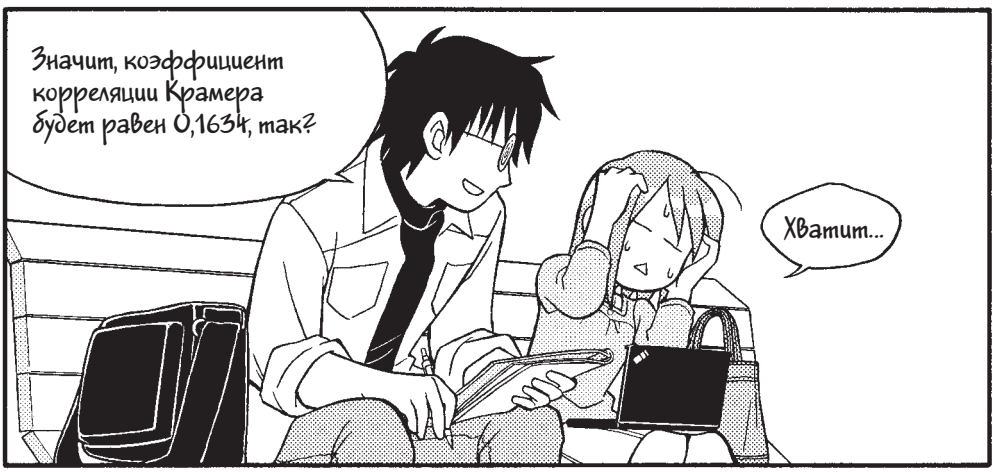
$$\sqrt{\frac{\chi_0^2}{n \times (\,\mathbf{min}\,\{\textit{кол-во строк в таблице}; \textit{кол-во столбцов в таблице}\} - 1)}},$$

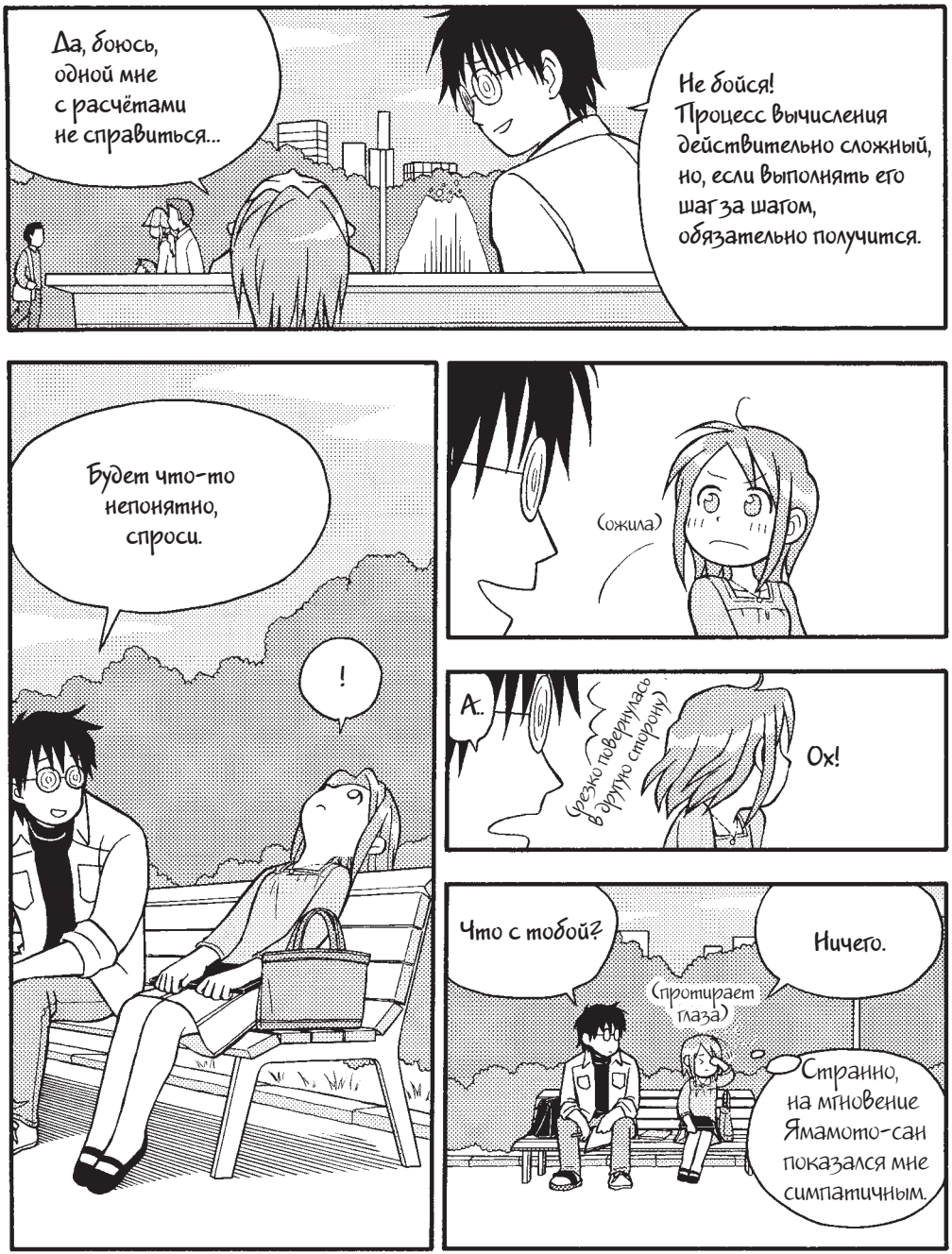
где

n — общее число единиц совокупности,

min $\{a;b\}$ — из величин a и b надо взять меньшую.

$$\sqrt{\frac{8{,}0091}{300 \times (\mathbf{min}\{2,\,3\} - 1)}} = \sqrt{\frac{8{,}0091}{300 \times (2-1)}} = \sqrt{\frac{8{,}0091}{300}} = 0{,}1634$$




Да, боюсь, одной мне с расчётами не справиться...
Не бойся! Процесс вычисления действительно сложный, но, если выполнять его шаг за шагом, обязательно получится.
Будет что-то непонятно, спроси.
!
(ожила)
А..
(резко повернулась в другую сторону)
Ох!
Что с тобой?
Ничего.
(протирает глаза)
Странно, на мгновение Ямамото-сан показался мне симпатичным.

Как я уже говорил, коэффициент корреляции Крамера может принимать значение от 0 до 1. Чем теснее связаны две переменные, тем ближе коэффициент Крамера к 1, а чем слабее связь, тем ближе к 0.

Таблицы взаимной сопряжённости «Пола» и «Способа признания в любви»

Величина коэффициента Крамера равна 1

Пол респондента	Способ признания в любви, %			Итого
	по телефону	по SMS	при встрече	
женский	17	83	0	100
мужской	0	0	100	100

Величина коэффициента Крамера равна 1 ⟺ Предпочтения девушек и юношей совершенно различаются.

Величина коэффициента Крамера равна 0

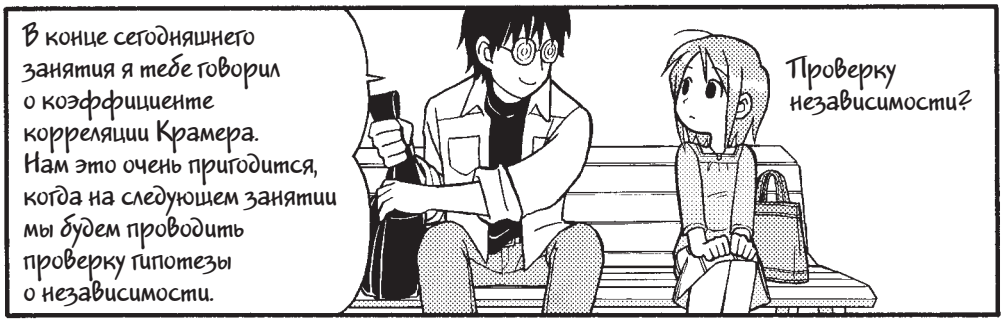
Пол респондента	Способ признания в любви, %			Итого
	по телефону	по SMS	при встрече	
женский	17	48	35	100
мужской	17	48	35	100

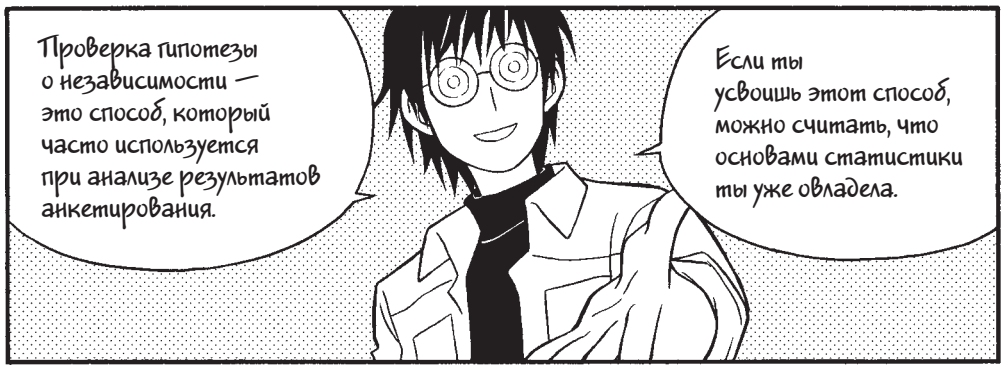
Величина коэффициента Крамера равна 0 ⟺ Предпочтения девушек и юношей одинаковы

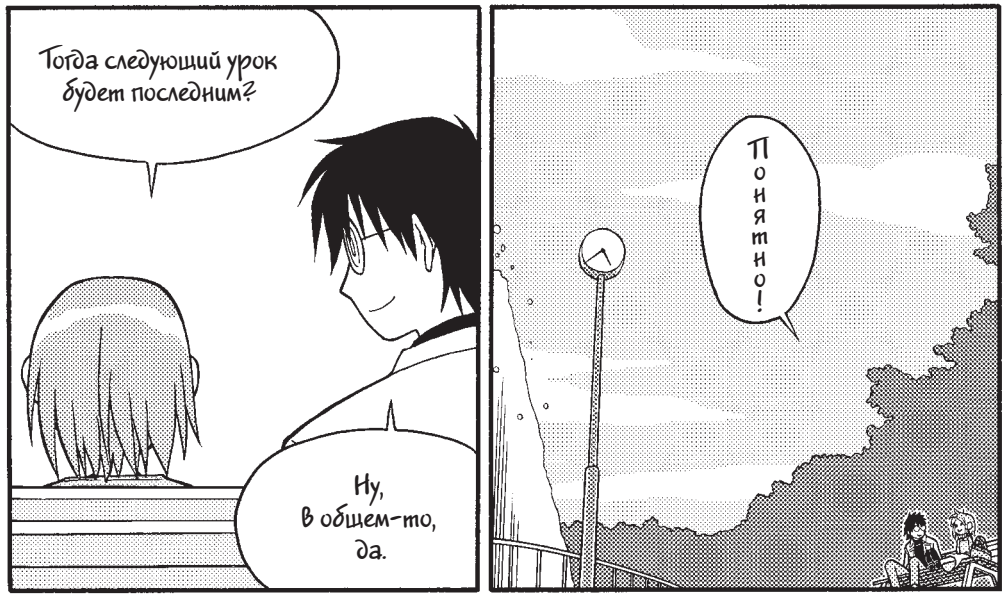
К сожалению, в статистике нет кого-то определённого значения коэффициента корреляции Крамера, свыше которого переменные считаются тесно связанными.
Для справки приведём таблицу соответствия коэффициента Крамера и степени взаимосвязи.

Критерии величины коэффициента корреляции Крамера

Коэффициент Крамера		Вывод о степени взаимосвязи	Вывод о наличии взаимосвязи
1,0—0,8	⇒	Очень тесная	Есть
0,8—0,5	⇒	Достаточно тесная	
0,5—0,25	⇒	Слабая	
< 0,25	⇒	Очень слабая	Нет


В конце сегодняшнего занятия я тебе говорил о коэффициенте корреляции Крамера. Нам это очень пригодится, когда на следующем занятии мы будем проводить проверку гипотезы о независимости.
Проверку независимости?


Проверка гипотезы о независимости — это способ, который часто используется при анализе результатов анкетирования.
Если ты усвоишь этот способ, можно считать, что основами статистики ты уже овладела.


Тогда следующий урок будет последним?
Ну, в общем-то, да.
Понятно!

Упражнение

В последнее время положение дел в компании X, владеющей сетью семейных ресторанов, трудно назвать благополучным. Чтобы узнать пожелания клиентов, компания провела анкетирование, объектами которого стали произвольно выбранные жители Японии в возрасте старше 20 лет.
Результаты анкетирования представлены в виде следующей таблицы:

Респондент	**Какую кухню Вы обычно заказываете в семейном ресторане?**	**Если после обеда будет предлагаться бесплатный напиток, Вы предпочтёте чай или кофе?**
1	китайская	кофе
2	европейская	кофе
...	...	...
250	японская	чай

На основании этой таблицы был составлен один из вариантов таблицы взаимной сопряжённости:

Обычно заказываемая кухня	**Предпочитаемый напиток**		**Итого**
	Кофе	**Чай**	
Японская	43	33	76
Европейская	51	53	104
Китайская	29	41	70
Итого:	123	127	250

Вычислите коэффициент корреляции Крамера, чтобы установить взаимосвязь между «обычно заказываемой кухней» и «предпочитаемым напитком».

Решение

Шаг 1

Составим таблицу взаимной сопряжённости.

Обычно заказываемая кухня	Предпочитаемый напиток		Итого
	Кофе	Чай	
Японская	43	33	76
Европейская	51	53	104
Китайская	29	41	70
Итого:	123	127	250

Шаг 2

Вычислим теоретические частоты.

Обычно заказываемая кухня	Предпочитаемый напиток		Итого
	Кофе	Чай	
Японская	$\frac{76 \times 123}{250}$	$\frac{76 \times 127}{250}$	76
Европейская	$\frac{104 \times 123}{250}$	$\frac{104 \times 127}{250}$	104
Китайская	$\frac{70 \times 123}{250}$	$\frac{70 \times 127}{250}$	70
Итого:	123	127	250

Шаг 3

Вычислим:

$$\frac{\textit{Эмпирическая частота} - \textit{Теоретическая частота}}{\textit{Теоретическая частота}}$$

Обычно заказываемая кухня	Предпочитаемый напиток		Итого
	Кофе	Чай	
Японская	$\dfrac{\left(43 - \dfrac{76 \times 123}{250}\right)^2}{\dfrac{76 \times 123}{250}}$	$\dfrac{\left(33 - \dfrac{76 \times 127}{250}\right)^2}{\dfrac{76 \times 127}{250}}$	76
Европейская	$\dfrac{\left(51 - \dfrac{104 \times 123}{250}\right)^2}{\dfrac{104 \times 123}{250}}$	$\dfrac{\left(53 - \dfrac{104 \times 127}{250}\right)^2}{\dfrac{104 \times 127}{250}}$	104
Китайская	$\dfrac{\left(29 - \dfrac{70 \times 123}{250}\right)^2}{\dfrac{70 \times 123}{250}}$	$\dfrac{\left(41 - \dfrac{70 \times 127}{250}\right)^2}{\dfrac{70 \times 127}{250}}$	70
Итого:	123	127	250

Шаг 4

Вычислим сумму величин в обведённых жирной чертой клетках таблицы на **Шаге 3**.
Иначе говоря,
вычислим значение критерия согласия Пирсона.

$$\chi_0^2 = \frac{\left(43 - \frac{76\times 123}{250}\right)^2}{\frac{76\times 123}{250}} + \frac{\left(33 - \frac{76\times 127}{250}\right)^2}{\frac{76\times 127}{250}}$$
$$+ \frac{\left(51 - \frac{104\times 123}{250}\right)^2}{\frac{104\times 123}{250}} + \frac{\left(53 - \frac{104\times 127}{250}\right)^2}{\frac{104\times 127}{250}}$$
$$+ \frac{\left(29 - \frac{70\times 123}{250}\right)^2}{\frac{70\times 123}{250}} + \frac{\left(41 - \frac{70\times 127}{250}\right)^2}{\frac{70\times 127}{250}}$$
$$= 3{,}3483$$

Шаг 5

Вычисляем коэффициент корреляции Крамера по формуле:

$$\sqrt{\frac{\chi_0^2}{n\times(\min\{\textit{кол-во строк в таблице};\ \textit{кол-во столбцов в таблице}\}-1)}} =$$

$$= \sqrt{\frac{3{,}3483}{250\times(\min\{3,2\}-1)}} = \sqrt{\frac{3{,}3483}{250\times(2-1)}} = \sqrt{\frac{3{,}3483}{250}} = 0{,}1157$$

Выводы

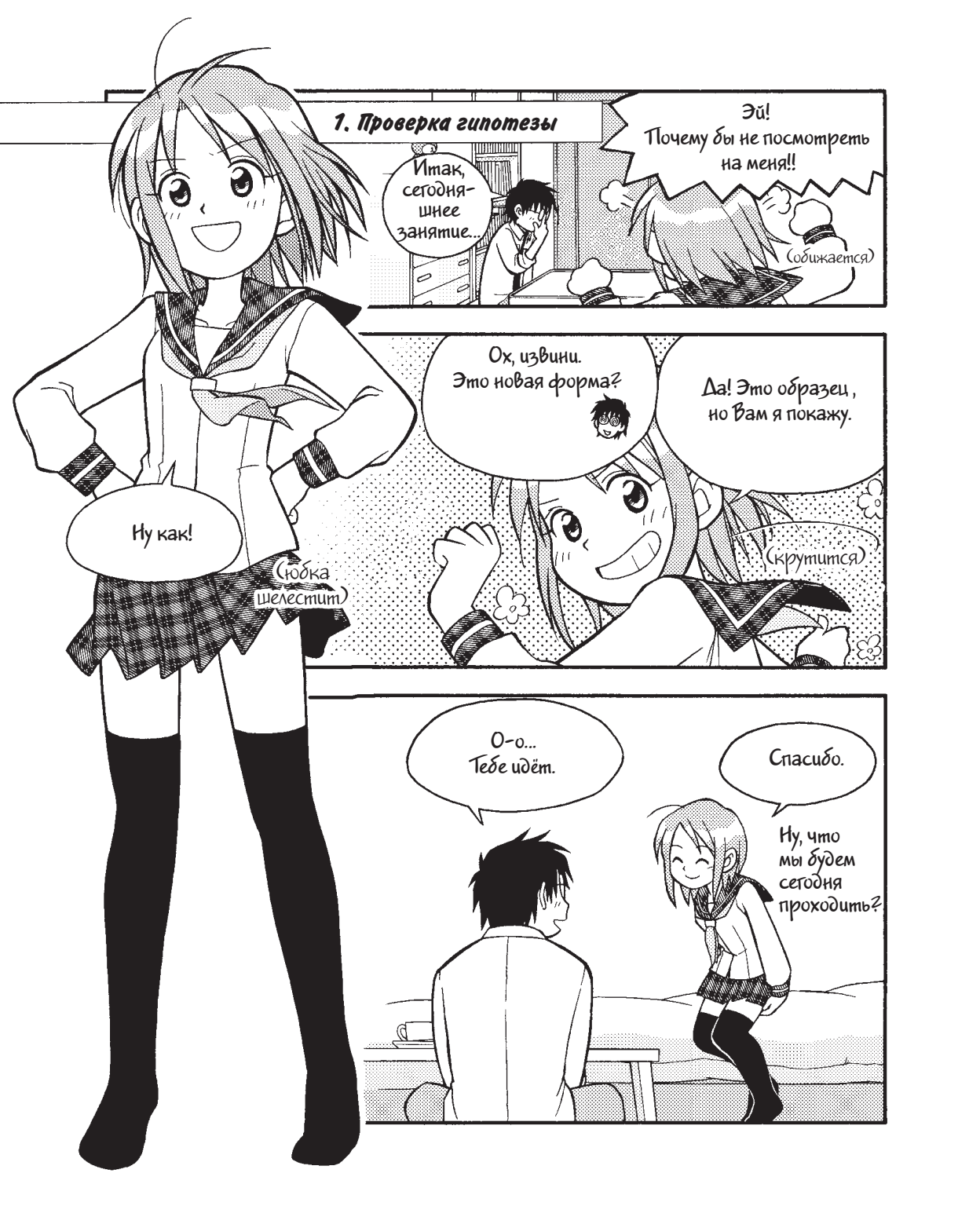
- **Коэффициент линейной корреляции является показателем тесноты связи между двумя количественными данными.**
- **Корреляционное отношение является показателем тесноты связи между количественными и качественными данными.**
- **Коэффициент корреляции Крамера является показателем тесноты связи между качественными данными.**
- **Перечисленные коэффициенты корреляции имеют следующие свойства:**

Вид функции распределения вероятностей	Значение коэффициента корреляции		Отсутствие какой-либо связи между двумя переменными величинами	Переменные величины максимально тесно связаны
	max	min		
Коэффициент линейной корреляции	−1	1	0	−1 или 1
Корреляционное отношение	0	1	0	1
Коэффициент корреляции Крамера	0	1	0	1

- **В статистике не существует какого-то определённого значения указанных коэффициентов корреляции, свыше которого переменные считаются тесно связанными.**

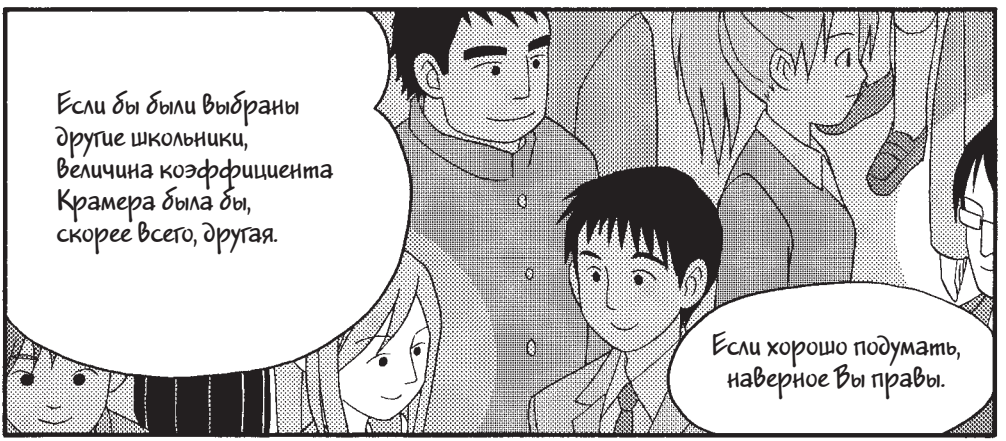
Глава 7

А что это за проверка гипотезы о независимости?

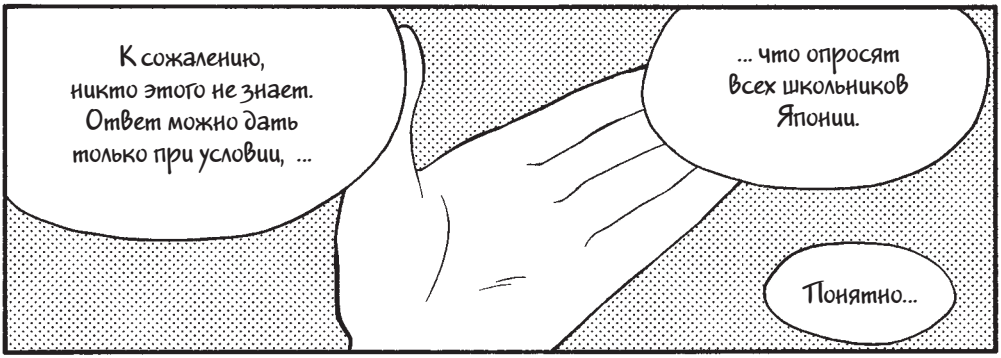

1. Проверка гипотезы
Итак, сегодняшнее занятие...
Эй! Почему бы не посмотреть на меня!!
(обижается)
Ох, извини. Это новая форма?
Да! Это образец, но Вам я покажу.
(крутится)
Ну как!
(юбка шелестит)
О-о... Тебе идёт.
Спасибо.
Ну, что мы будем сегодня проходить?


На прошлом уроке мы изучали коэффициент корреляции Крамера, помнишь?
Опросили 300 школьников! Какой способ признания в любви вы предпочитаете?
А..., про признание в любви, да?


Коэффициент корреляции Крамера в том примере был равен 0,1634. Мы сделали вывод, что «связь очень слабая», так?
Да, так и было.
Руц, теперь хорошенько подумай.
Результат этого анкетирования основан всего лишь на данных, ...
... полученных от 300 школьников, произвольно выбранных из всех школьников Японии.


Если бы были выбраны другие школьники, величина коэффициента Крамера была бы, скорее всего, другая.
Если хорошо подумать, наверное Вы правы.


Как ты думаешь, какое будет значение коэффициента корреляции Крамера для генеральной совокупности, в которую входят «все школьники Японии»?
М-м-м... Откуда мне знать?...

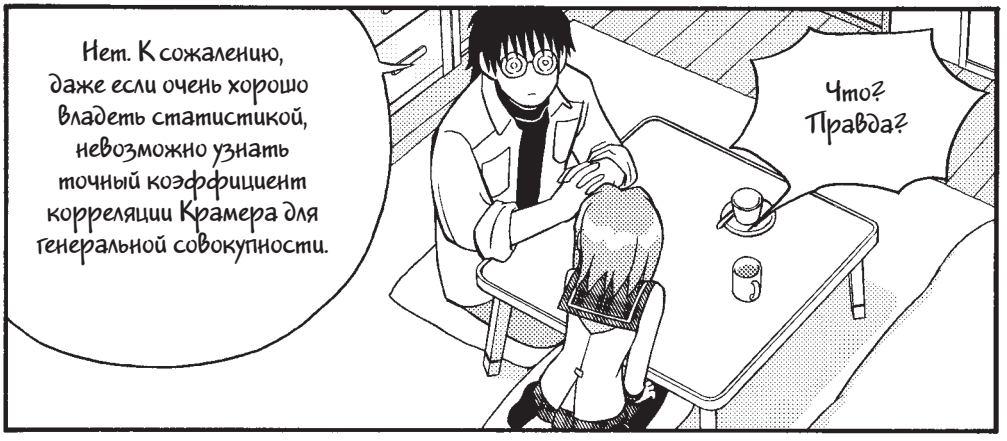

К сожалению, никто этого не знает. Ответ можно дать только при условии, ...
... что опросят всех школьников Японии.
Понятно...

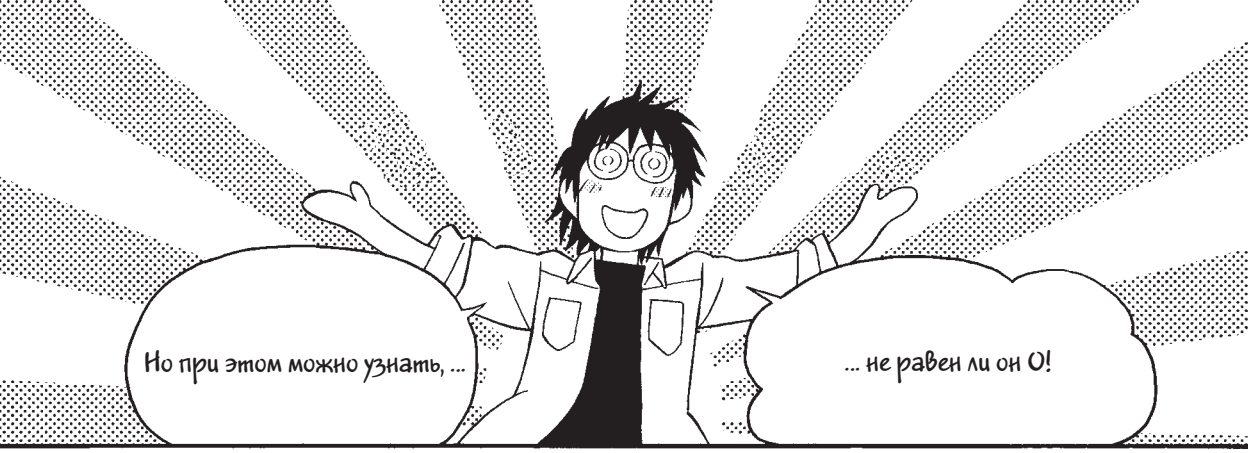

Это касается не только данного примера. Практически никогда не удаётся определить коэффициент корреляции Крамера для всей генеральной совокупности.
Поэтому, ...


...зная, что коэффициент корреляции Крамера, вычисленный на основе ответов, полученных от 300 школьников, произвольно выбранных из генеральной совокупности, равен 0,1634, считаем, что это значение соответствует коэффициенту корреляции Крамера для всей генеральной совокупности.
Нам придётся сделать подобное субъективное заключение.


Довольно туманное...

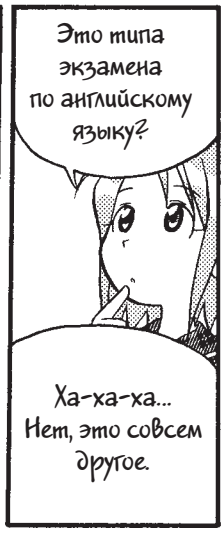

Но ведь наверняка есть какие-то статистические методы?


Нет. К сожалению, даже если очень хорошо владеть статистикой, невозможно узнать точный коэффициент корреляции Крамера для генеральной совокупности.
Что? Правда?


Но при этом можно узнать, ...
... не равен ли он 0!


Это что, выход?
?
Равен 0?
Конечно! Поскольку можно получить объективную информацию о генеральной совокупности.

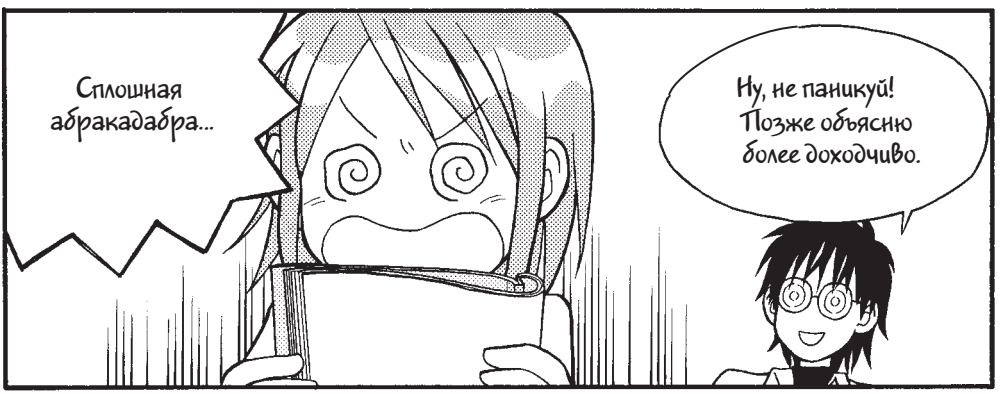

И как это делается?
Используют метод анализа, который называется «проверка гипотезы о независимости». Я тебе уже говорил об этом.


Это типа экзамена по английскому языку?
Ха-ха-ха... Нет, это совсем другое.


Проверка гипотезы о независимости — это один из способов анализа, называемых в статистике проверкой.
проверка гипотезы о независимости
Проверка
В первую очередь я тебе объясню, что такое проверка.
проверка корреляционного отношения
проверка гипотезы о равенстве долей в (двух) совокупностях
проверка гипотезы о равенстве средних величин в (двух) совокупностей
проверка гипотезы об отсутствии корреляции
Ладно.

Примеры проверяемых гипотез	
Гипотезы	**Примеры применения**
О независимости	Проверяют, равняется ли нулю коэффициент корреляции Крамера между полом и способом признания в любви для генеральной совокупности.
О корреляционном отношении	Проверяют, равняется ли нулю корреляционное отношение между любимым брэндом одежды и возрастом для генеральной совокупности.
Об отсутствии корреляции	Проверяют, равняется ли нулю коэффициент линейной корреляции между расходами в месяц на косметику и расходами в месяц на одежду для генеральной совокупности.
О равенстве средних величин (двух) совокупностей	Проверяют, получают ли школьницы Токио и школьницы Осаки одну и ту же или разную сумму на карманные расходы. (Будьте внимательны: предполагаются две генеральные совокупности).
О равенстве долей в (двух) совокупностях	Проверяют, различается ли рейтинг кабинета министров среди избирателей, проживающих в городах, и избирателей из сельской местности. (Будьте внимательны: предполагаются две генеральные совокупности).

Порядок статистической проверки гипотез	
Шаг 1.	Определить генеральную совокупность.
Шаг 2.	Сформулировать нулевую и альтернативную гипотезы.
Шаг 3.	Выбрать вид статистической проверки гипотезы.
Шаг 4.	Определить уровень значимости.
Шаг 5.	Вычислить фактическое значение выбранного статистического критерия на основе данных выборочной совокупности.
Шаг 6.	Проверить, входит ли вычисленное на Шаге 5 значение статистического критерия в критическую область.
Шаг 7.	Если фактическое значение выбранного статистического критерия (Шаг 6) входит в критическую область, делают вывод о том, что верна альтернативная гипотеза. В противном случае полагают, что нет оснований считать нулевую гипотезу ошибочной.



2. Проверка гипотезы о независимости

Объясню суть проверки гипотезы о независимости. Сегодня — это главная тема.

Проверка гипотезы о независимости — один из способов анализа, проводимый с целью выяснить, не равен ли нулю коэффициент корреляции Крамера для генеральной совокупности.

Понятно.

Другими словами, это способ анализа, предполагающий наличие взаимосвязи между двумя переменными в таблице взаимной сопряжённости.

Поня-я-тно...

	34	61	53	148
	38	40	74	152
	72	101	127	

Проверка гипотезы о независимости называется также **проверкой критерия согласия Пирсона χ^2**.

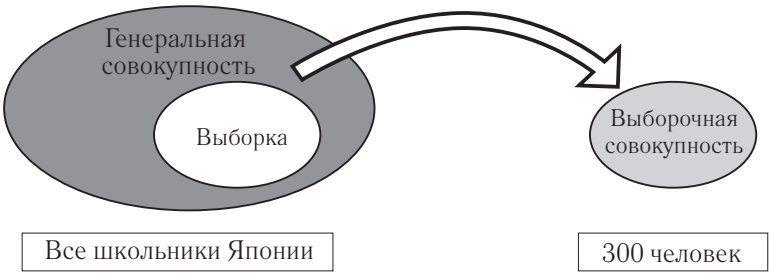
Опять этот χ^2! Ужасно...

Объяснение Критерий согласия Пирсона и распределение хи-квадрат

Перед тем, как объяснить проверку гипотезы о независимости на конкретном примере, расскажу о том, что лежит в основе такой проверки. Предположим, что провели следующее исследование (хотя в действительности это невозможно):

Шаг 1

Из генеральной совокупности, которой являются «школьники Японии», произвольно выбрали 300 человек.



Шаг 2

Среди выбранных 300 школьников провели анкетирование (см. стр. 127), а затем рассчитали величину критерия согласия Пирсона.

Шаг 3

Выбранных 300 школьников «вернули» в генеральную совокупность.

Шаг 3

Шаг 1 — Шаг 3 повторяли множество раз.

Функцией распределения критерия согласия Пирсона, полученной в этом исследовании, является распределение хи-квадрат с числом степеней свободы, равным 2, при условии, что коэффициент корреляции Крамера для генеральной совокупности «все школьники Японии» = 0. Другими словами, если коэффициент корреляции Крамера для генеральной совокупности «все школьники Японии» = 0, то критерий согласия Пирсона (χ_0^2) имеет хи-квадрат-распределение с числом степеней свободы, равным 2.

1. Способ вычисления критерия согласия Пирсона (χ_0^2) см. на стр. 130—133.
2. О распределении хи-квадрат с числом степеней свободы 2 см. на стр. 100.

Описанное выше исследование было проведено на самом деле, но с учётом следующих ограничений:

- Поскольку невозможно провести исследование среди всех школьников, живущих в Японии, считается, что генеральная совокупность «все школьники Японии» — это совокупность, состоящая из 10 000 человек, как указано в **Табл. 7.1**.
- Было сделано предположение, что коэффициент корреляции Крамера для совокупности «все школьники Японии» = 0. Другими словами, предположили, что ответы юношей и девушек на вопрос «Какой способ признания в любви вы предпочитаете?» (по телефону : по SMS : при встрече) одинаковы (см. стр. 135). Предположили также, что на основе **Табл. 7.1** была создана таблица взаимной сопряжённости.
- Исследования (Шаг 1 — Шаг 3) провели 20 тысяч раз.

Таблица 7.1. Какой способ признания в любви вы предпочитаете? (все школьники Японии)

Респондент	Пол респондента	Способ признания в любви
1	ж	при встрече
2	ж	по телефону
...	...	...
10000	м	по SMS

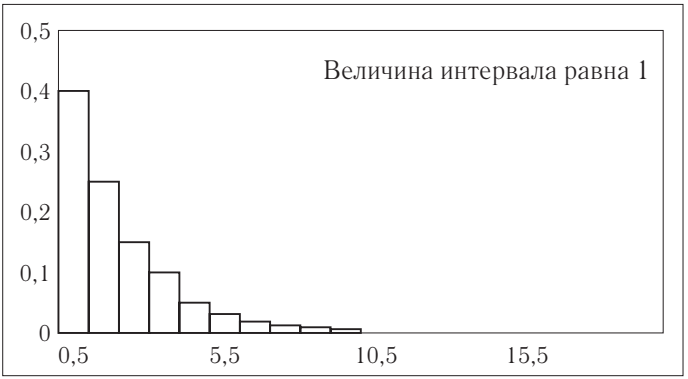
Таблица 7.2. Взаимная сопряжённость пола и способа признания в любви

Пол респондента	Способ признания в любви			Итого
	по телефону	по SMS	при встрече	
женский	400	1600	2000	4000
мужской	600	2400	3000	6000
Итого:	1000	4000	5000	10000

Таблица 7.3. Результаты исследования

Номер исследования	Критерий согласия Пирсона (χ_0^2)
1	0,8598
2	0,7557
...	...
20000	2,7953

Рис. 7.1. Гистограмма на основе данных **Табл. 7.3**



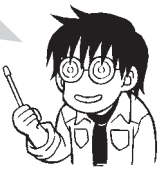
Действительно, Рис. 7.1 очень похож на график функции распределения вероятностей, когда число степеней свободы равно 2 (см. стр. 100). Нет сомнений, что величина критерия согласия Пирсона (χ_0^2) имеет хи-квадрат-распределение с числом степеней свободы 2. Это не имеет непосредственного отношения к исследованию, но я скажу вам одну очень важную вещь: число степеней свободы, равное 2, получается из выражения:

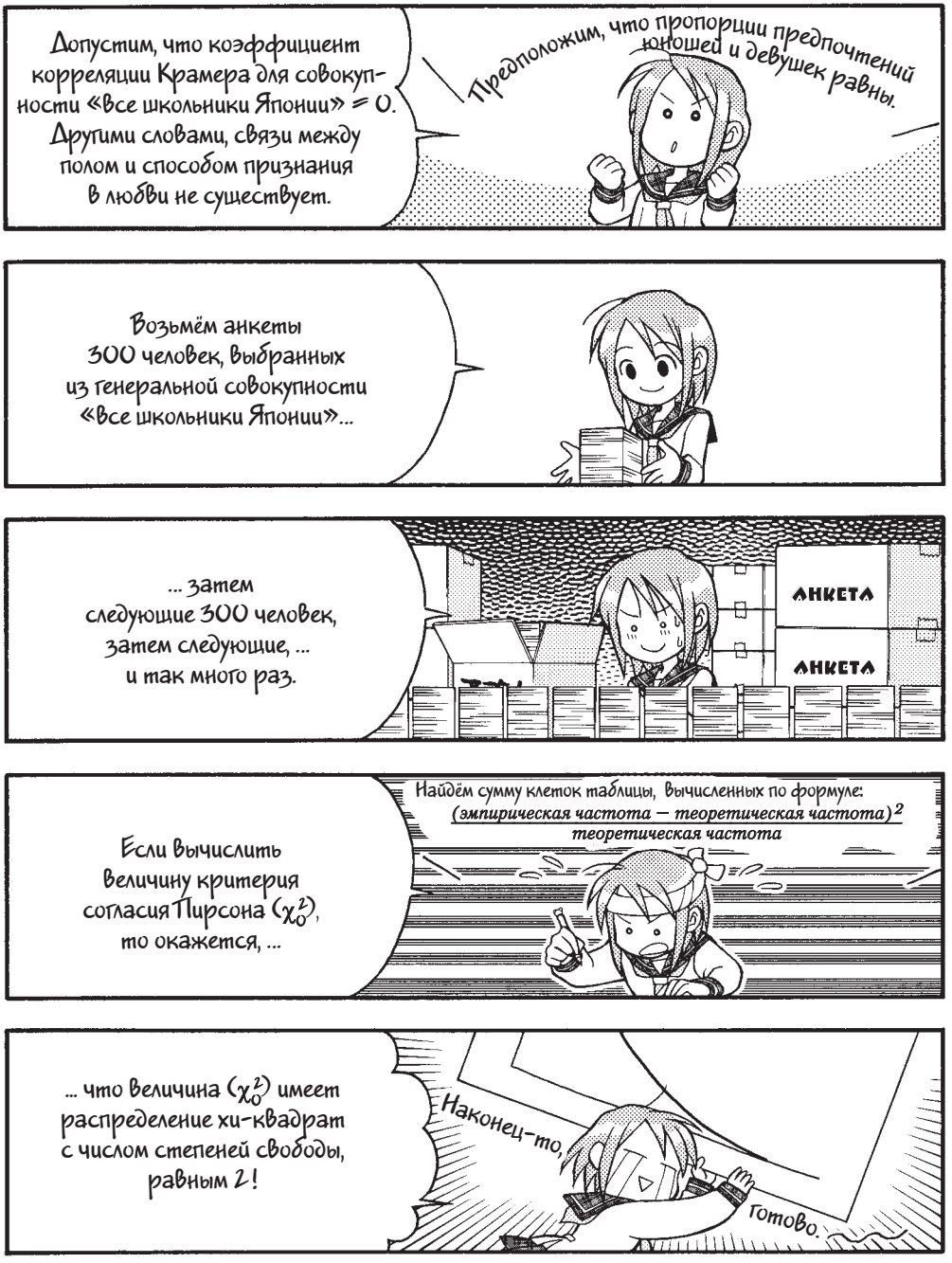
$$(2-1)\times(3-1)=1\times 2=2$$

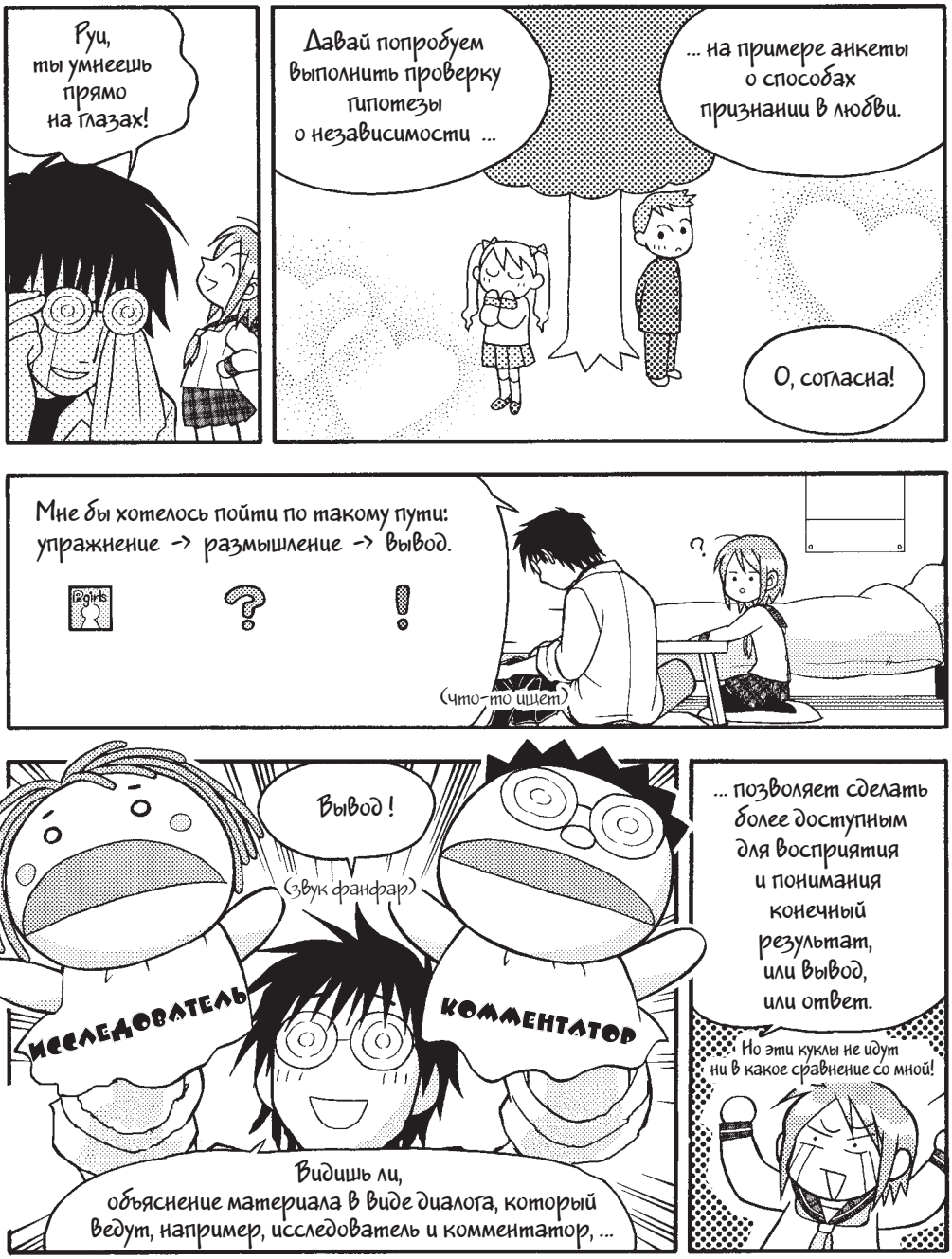
два варианта: девушки, юноши

три варианта: по телефону, по SMS, при встрече

Почему такой странный расчет? Ответ на этот вопрос выходит за рамки данной книги. Поэтому не беспокойтесь, если вы не до конца поняли этот способ расчёта.




Допустим, что коэффициент корреляции Крамера для совокупности «все школьники Японии» = 0. Другими словами, связи между полом и способом признания в любви не существует.
Предположим, что пропорции предпочтений юношей и девушек равны.
Возьмём анкеты 300 человек, выбранных из генеральной совокупности «все школьники Японии»...
... затем следующие 300 человек, затем следующие, ... и так много раз.
АНКЕТА
АНКЕТА
Если вычислить величину критерия согласия Пирсона (χ_0^2), то окажется, ...
Найдём сумму клеток таблицы, вычисленных по формуле:
$\frac{(\text{эмпирическая частота} - \text{теоретическая частота})^2}{\text{теоретическая частота}}$
... что величина (χ_0^2) имеет распределение хи-квадрат с числом степеней свободы, равным 2!
Наконец-то,
готово.


Руи, ты умнеешь прямо на глазах!
Давай попробуем выполнить проверку гипотезы о независимости ...
... на примере анкеты о способах признании в любви.
О, согласна!
Мне бы хотелось пойти по такому пути: упражнение -> размышление -> вывод.
Руи
?
!
?
(что-то ищет)
Вывод!
(звук фанфар)
ИССЛЕДОВАТЕЛЬ
КОММЕНТАТОР
Видишь ли, объяснение материала в виде диалога, который ведут, например, исследователь и комментатор, ...
... позволяет сделать более доступным для восприятия и понимания конечный результат, или вывод, или ответ.
Но эти куклы не идут ни в какое сравнение со мной!

Упражнение

Женский журнал «P-girls» решил провести опрос среди школьников по вопросу: «Какой способ признания в любви Вы предпочитаете?». Для этого кореспондент из всех школьников Японии произвольно выбрал 300 человек и провёл анкетирование. Результаты анкетирования представлены в виде следующей таблицы:

Респондент	Способ признания в любви	Возраст респондента	Пол респондента
1	при встрече	17	ж
2	по телефону	15	ж
...	...	...	..
300	по SMS	18	м

Таблица взаимной сопряжённости «пола» и «способа признания в любви»

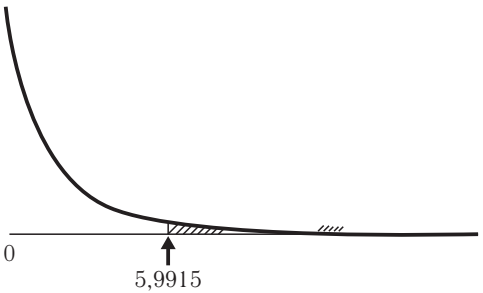
Пол респондента	Способ признания в любви			Итого
	по телефону	по SMS	при встрече	
женский	34	61	53	148
мужской	38	40	74	152
Итого:	72	101	127	300

Выясните путём проверки гипотезы о независимости, больше ли 0 коэффициент корреляции Крамера между полом и способом признания в любви для генеральной совокупности «все школьники Японии». Другими словами, надо выяснить, есть ли взаимосвязь между полом и способом признания в любви. Предположим, что уровень значимости (объясню позже) равен 0,05.

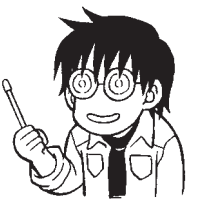
Размышление

Как уже было сказано (стр. 152—154), если коэффициент корреляции Крамера для генеральной совокупности «все школьники Японии» = 0, критерий согласия Пирсона (χ_0^2) имеет распределение хи-квадрат с числом степеней свободы, равным 2. Следовательно, если коэффициент корреляции Крамера = 0 для генеральной совокупности «все школьники Японии», вероятность того, что величина, полученная на основе данных, поступивших от случайно отобранных 300 школьников, будет, например, > 5,9915, равна 0,05. (см. Таблицу распределения хи-квадрат на стр. 103).

Рис. 7.2. Вероятность того, что $\chi_0^2 > 5{,}9915$.



Величина уже была вычислена и равняется 8,0091 (см. стр. 132).
Ну, как, не кажется ли вам, что величина слишком большая, хотя и вычислена на основе ответов, полученных от 300 человек, случайно выбранных из генеральной совокупности. Если поразмыслить, учитывая комментарий, данный на стр. 132, не естественно ли, что величина коэффициента корреляции Крамера для генеральной совокупности «все школьники Японии» >0?



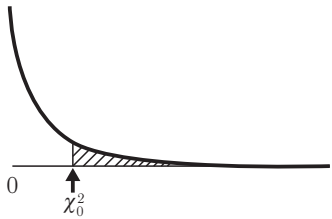
Не только в этом примере, но всегда при проверке гипотезы о независимости следует придерживаться следующей последовательности действий:

1) Сначала делают предположение, что коэффициент корреляции Крамера для генеральной совокупности = 0.
2) Затем рассчитывают величину χ_0^2 для выборочной совокупности.
3) Если χ_0^2 очень большой, делают вывод, что коэффициент корреляции Крамера для генеральной совокупности > 0.

Это надо запомнить.

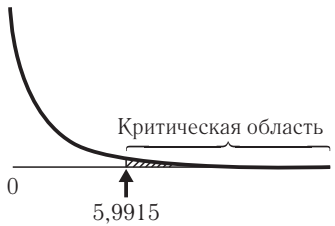
Добавлю кое-что к пункту 3. Чем больше величина χ_0^2, тем меньше вероятность, т.е. площадь заштрихованной области на **Рис. 7.3.**

Рис. 7.3. Вероятность при определённом значении



Если при проверке гипотезы о независимости оказывается, что значение вероятности заштрихованной области меньше так называемого уровня значимости, делают вывод, что величина коэффициента корреляции Крамера для генеральной совокупности > 0. Обычно в статистических исследованиях используют такой уровень значимости, как 0,05 или 0,01; выбор конкретного значения — прерогатива исследователя. Предположим, что выбрали значение уровня значимости, равное 0,05. В действительности, уровень значимости — это заштрихованная область на графике, показаном на **Рис. 7.3**.
При этом область, указанная на **Рис. 7.4**, называется критической областью.

Рис. 7.4. Критическая область при значении уровня значимости, равном 0,05



Вывод **Порядок выполнения проверки**

Шаг 1

Определяем генеральную совокупность.

Генеральная совокупность — это:

Все школьники Японии

В этом упражнении изначально определено, что генеральной совокупностью являются «все школьники Японии». Поэтому нет необходимости выполнять Шаг 1.

Пример.
При проведении проверки гипотезы о равенстве долей в (двух) совокупностях (см. стр. 149) предполагалось, что генеральными совокупностями являются «избиратели, проживающие в городах и в сельской местности». При этом необходимо уточнить, что подразумевается под городом – Токио и Осака? Или это столицы префектур? Решение этой проблемы остаётся за исследователем, поскольку это его прерогатива — определить, что будет представлять собой генеральная совокупность при проверке. Если чётко не выделить генеральную совокупность, при проведении любой проверки можно оказаться в ситуации: «Ой! Что же я хотел исследовать?!». Подобная ситуация наблюдается довольно часто, поэтому следует быть предельно внимательными.

Шаг 2

Выстраиваем нулевую и альтернативную гипотезы.

Нулевая гипотеза:

Коэффициент корреляции Крамера для генеральной совокупности = 0. «Пол» и «способ признания в любви» не связаны.

Альтернативная гипотеза:

Коэффициент корреляции Крамера для генеральной совокупности > 0. «Пол» и «способ признания в любви» связаны.

О нулевой и альтернативной гипотезах расскажу позже.

Шаг 3

Выбирают вид гипотизы для статистической проверки.

Проведём проверку гипотезы о независимости.

В этом примере изначально решено было проводить проверку гипотезы о независимости. Поэтому нет необходимости выполнять Шаг 3. В действительности, когда проводят проверку, исследователь сам должен выбрать вид гипотезы с учётом целей анализа.

Шаг 4

Определяют уровень значимости.

Пусть уровень значимости будет равен 0,05.

Поскольку будет использоваться уровень значимости, равный 0,05, нет необходимости выполнять Шаг 4. В действительности при проведении проверки исследователь должен сам выбрать уровень значимости. Обычно это 0,05 или 0,01. Уровень значимости обозначается α (альфа).

Шаг 5

Вычисляют фактическое значение выбранного статистического критерия на основе данных выборочной совокупности.

Я собираюсь провести проверку гипотезы о независимости. Поэтому в качестве статистического критерия будет выступать критерий согласия Пирсона χ^2_0.
Для данного примера значение критерия Пирсона χ^2_0 уже вычислено и равно 8,0091 (см. стр. 132).

Выбранный статистический критерий — это формула, которая преобразует данные выборочной совокупности в одну величину.
В зависимости от вида проверяемых гипотез выбираются разные критерии.
В случае, когда проводят проверку гипотезы о независимости, критерием является величина, указанная выше, а в случае проведения проверки об отсутствии корреляции (см. стр. 149) критерием будет величина, рассчитываемая по формуле:

$$\frac{\textit{линейный коэффициент корреляции}^2 \times \sqrt{\textit{общее число единиц совокупности} - 2}}{\sqrt{1 - \textit{линейный коэффициент корреляции}^2}}$$

Шаг 6

Выясняют, входит ли фактическое значение выбранного статистического критерия, вычисленное на Шаге 5, в критическую область.

Значение критерия согласия Пирсона χ_0^2, который в данном примере является статистическим критерием, равняется 8,0091.
Так как уровень значимости равен 0,05, из Таблицы хи-квадрат-распределения (стр. 103) получаем, что критическая область больше, чем 5,9915.
Как следует из рисунка, значение выбранного статистического критерия входит в критическую область.

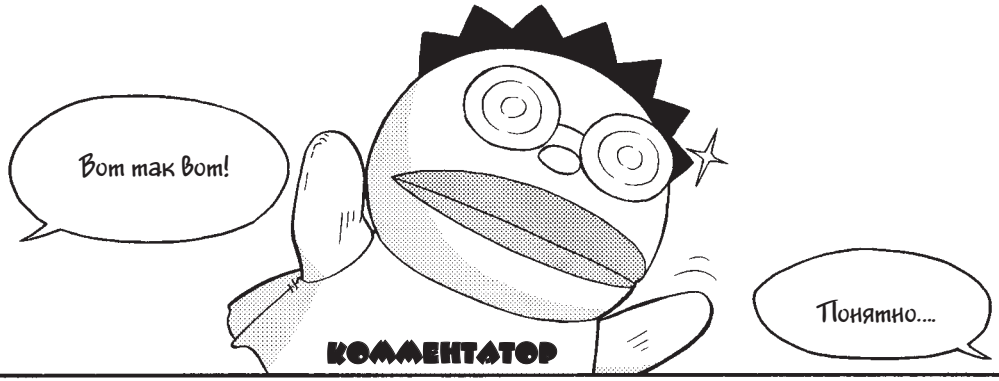
Критическая область меняется в зависимости от уровня значимости α. Если бы уровень значимости был бы равен 0,01, а не 0,05, то критическая область, как следует из Таблицы распределения хи-квадрат на стр. 103, была бы больше, чем 9,2104.

Шаг 7

Если значение статистического критерия входит в критическую область (Шаг 6), делают вывод, что «верна альтернативная гипотеза». В противном случае вывод таков: «нельзя утверждать, что нулевая гипотеза ошибочна».

Значение выбранного статистического критерия входит в критическую область. Следовательно, верна альтернативная гипотеза — Величина коэффициента корреляции Крамера для генеральной совокупности > 0. «Пол» и «способ признания в любви» связаны!

Даже если величина статистического критерия входит в критическую область, нельзя на основе проверки делать вывод, что «альтернативная гипотеза **абсолютно** верна». Можно сделать лишь такой вывод: «хотелось бы утверждать, что альтернативная гипотеза **абсолютно** верна, но существует вероятность (α x 100%) того, что верна нулевая гипотеза».


Вот так вот!
КОММЕНТАТОР
Понятно....


Однако меня беспокоит Шаг 7.
?


Мне кажется, можно сказать, что нулевая гипотеза верна, если величина статистического критерия не входит в критическую область...
критическая область
0
5.9915

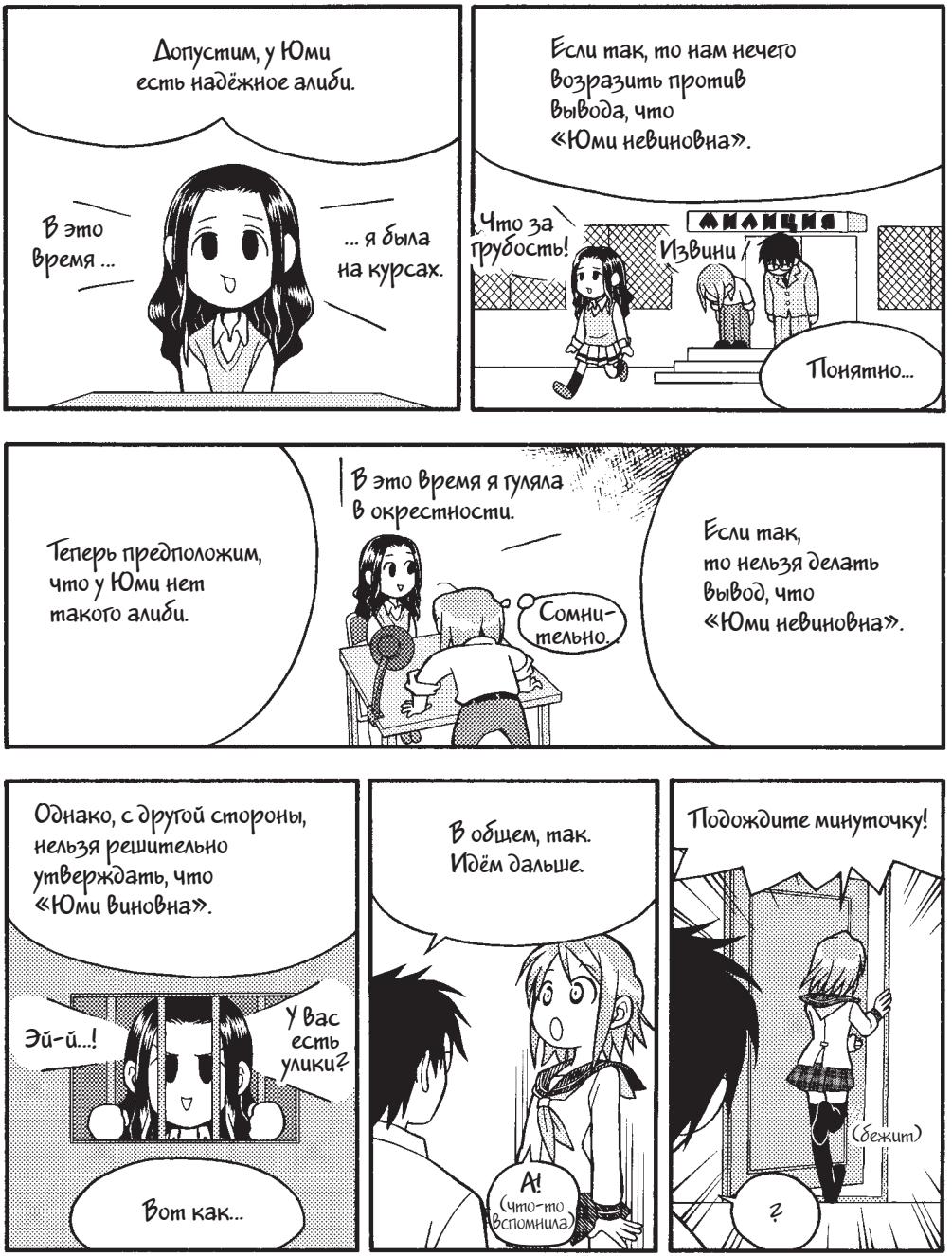

К сожалению, так сказать нельзя. Можно только сделать вывод: «нельзя утверждать, что нулевая гипотеза ошибочна».
Вот как?


Например, представь, что величина хи-квадрат в предыдущем примере была равна 2,5013.
критическая область
0
2.5013
5.9915
Она не входит в критическую область, да?..

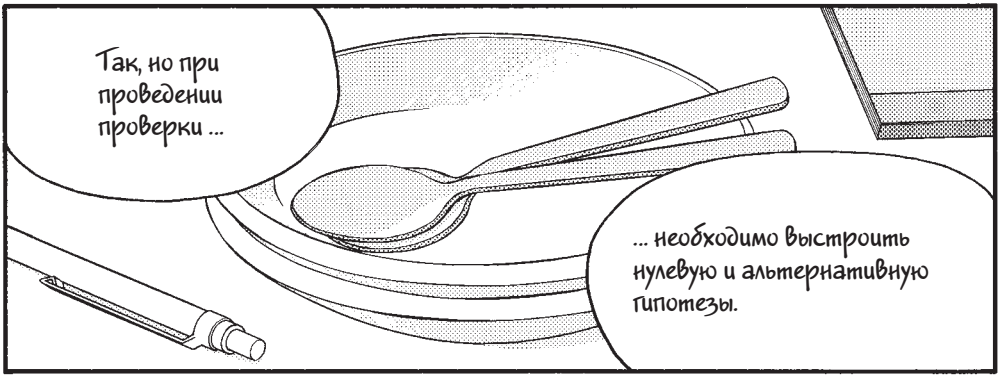

Поэтому совершенно невозможно сделать вывод, что коэффициент корреляции Крамера для генеральной совокупности > 0.
Непонятно...
Но при этом нельзя утверждать, что коэффициент корреляции Крамера для генеральной совокупности = 0.


Тогда приведу пример.
Допустим, кто-то съел пудинг, который собиралась съесть ты.
Кто посмел?!


Подозревается Юми.
Юми, это ужасно!
Я же говорю, допустим...
Пока опустим такие подробности, как виды проверяемых гипотез и значение уровня значимости...
нулевая гипотеза
Юми виновна
альтернативная
Юми невиновна
(постучал по столу стопкой бумаги)
Будем проводить проверку этих двух гипотез.


Допустим, у Юми
есть надёжное алиби.
В это
время ...
... я была
на курсах.
Если так, то нам нечего
возразить против
вывода, что
«Юми невиновна».
Что за
грубость!
Извини
Понятно...
Теперь предположим,
что у Юми нет
такого алиби.
В это время я гуляла
в окрестности.
Сомни-
тельно.
Если так,
то нельзя делать
вывод, что
«Юми невиновна».
Однако, с другой стороны,
нельзя решительно
утверждать, что
«Юми виновна».
Эй-й...!
У вас
есть
улики?
Вот как...
В общем, так.
Идём дальше.
А!
(что-то
вспомнила)
Подождите минуточку!
(бежит)
?

3. Нулевая и альтернативная гипотезы



Примеры проверяемых гипотез	
Гипотезы	**Примеры применения**
О независимости	Проверяют, равняется ли нулю коэффициент корреляции Крамера между полом и способом признания в любви для генеральной совокупности.
О корреляционном отношении	Проверяют, равняется ли нулю корреляционное отношение между любимым брэндом одежды и возрастом для генеральной совокупности.
Об отсутствии корреляции	Проверяют, равняется ли нулю коэффициент линейной корреляции между расходами в месяц на косметику и расходами в месяц на одежду для генеральной совокупности.
О равенстве средних величин (двух) совокупностей	Проверяют, получают ли школьницы Токио и школьницы Осаки одну и ту же или разную сумму на карманные расходы. (Будьте внимательны: предполагаются две генеральные совокупности).
О равенстве долей в (двух) совокупностях	Проверяют, различается ли рейтинг кабинета министров среди избирателей, проживающих в городах, и избирателей из сельской местности. (Будьте внимательны: предполагаются две генеральные совокупности).

Проверка гипотезы о независимости

Нулевая гипотеза	Коэффициент корреляции Крамера между «полом» и «способом признания в любви» для генеральной совокупности = 0.
Альтернативная гипотеза	Коэффициент корреляции Крамера между «полом» и «способом признания в любви» для генеральной совокупности > 0.

Проверка гипотезы о корреляционном отношении

Нулевая гипотеза	Корреляционное отношение между «любимым брендом одежды» и «возрастом» для генеральной совокупности = 0.
Альтернативная гипотеза	Корреляционное отношение между «любимым брендом одежды» и «возрастом» для генеральной совокупности > 0.

Проверка гипотезы об отсутствии корреляции

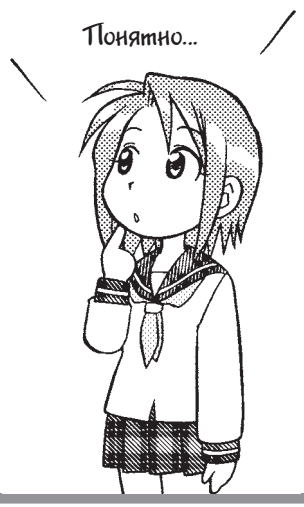
Нулевая гипотеза	Коэффициент линейной корреляции между «расходами в месяц на косметику» и «расходами в месяц на одежду» для генеральной совокупности = 0.
Альтернативная гипотеза	Коэффициент линейной корреляции между «расходами в месяц на косметику» и «расходами в месяц на одежду» для генеральной совокупности ≠ 0. или Коэффициент линейной корреляции между «расходами в месяц на косметику» и «расходами в месяц на одежду» для генеральной совокупности > 0. или Коэффициент линейной корреляции между «расходами в месяц на косметику» и «расходами в месяц на одежду» для генеральной совокупности < 0.

Проверка гипотезы о равенстве средних величин (двух) совокупностей

Нулевая гипотеза	«Суммы, получаемые на карманные расходы» школьницами Токио и школьницами Осаки, одинаковы.
Альтернативные гипотезы	«Суммы, получаемые на карманные расходы» школьницами Токио и школьницами Осаки, различны.
	«Сумма, получаемая на карманные расходы» школьницами Осаки, больше, чем сумма, получаемая школьницами Токио.
	«Сумма, получаемая на карманные расходы» школьницами Осаки меньше, чем сумма, получаемая школьницами Токио.

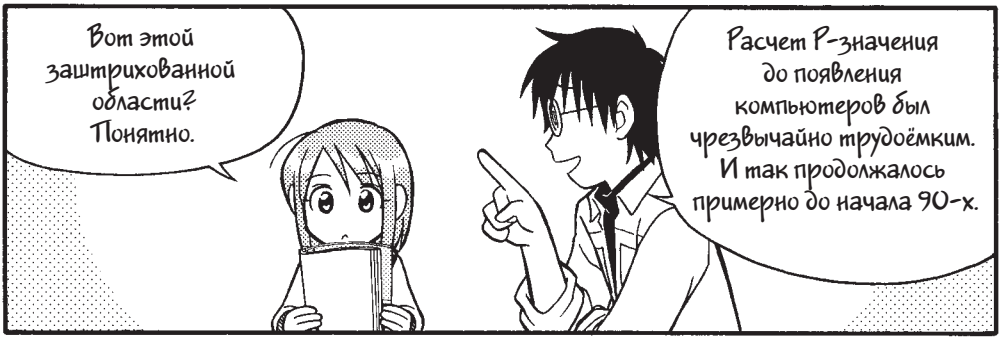
Проверка гипотезы о равенстве долей в (двух) совокупностях

Нулевая гипотеза	Рейтинг кабинета министров среди избирателей, проживающих в городах и в сельской местности, одинаков.
Альтернативные гипотезы	Рейтинг кабинета министров среди избирателей, проживающих в городах и в сельской местности, различен.
	Рейтинг кабинет министров среди избирателей, проживающих в городах, выше, чем среди избирателей, проживающих в сельской местности.
	Рейтинг кабинет министров среди избирателей, проживающих в городах, ниже, чем среди избирателей, проживающих в сельской местности.




Ну, как?
Ты поняла, что в качестве нулевых гипотез используются гипотезы, которые сложно доказать, как, например, «коэффициент корреляции Крамера для генеральной совокупности = 0», а не такие, как «коэффициент корреляции Крамера для генеральной совокупности приблизительно равен 0»?
Сплошные крайности...
А ещё, Руи, поняла ли ты, что в качестве нулевой гипотезы используются утвердительные гипотезы, содержащие слова «равно», «одинаково»,
а в качестве альтернативных используются отрицательные гипотезы со словами «не ...», «различно ...»?
...Да...
В качестве нулевых используют утвердительные гипотезы, которые сложно доказать,
а в качестве альтернативных — отрицательные гипотезы..
Мм... вот как
Достаточно, если ты поймёшь только это.

4. P-значение и порядок проверки


Вот этой заштрихованной области? Понятно.
Расчет P-значения до появления компьютеров был чрезвычайно трудоёмким. И так продолжалось примерно до начала 90-х.


Вот как...
Поэтому в большинстве случаев ...
... вывод о проверке делался на основании Гипотезы 1.


А сейчас?
Когда стало возможным вычисление P-значения с помощью Excel, выводы, как правило, делаются на основании Гипотезы 2.


Понятно...
Как я уже отмечал, порядок проведения проверки, если вывод о ней делается на основании Гипотезы 2, отличается от порядка, если вывод делается на основании Гипотезы 1.


Я хотела бы ещё раз провести анализ
(Ямамото-сан говорит женским голосом)
Исследователь
Не говорите таким странным голосом!

Шаг 6р

Выясняют, меньше ли P-значение, соответствующее величине выбранного статистического критерия, вычисленного на ***Шаге 5***, чем уровень значимости.

Уровень значимости равен 0,05.
Поскольку величина критерия согласия Пирсона равна 8,0091, P-значение равно 0,0182 и, следовательно, 0,0182 < 0,05.
Другими словами, P-значение меньше, чем выбранный статистический критерий.

Как я уже говорил, P-значение можно вычислить, используя функции Excel (но способ вычисления зависит от вида гипотезы).
Например, в Excel можно вычислить величину P-значения для проверки гипотезы о независимости.
Подробное объяснение смотрите на стр. 208.

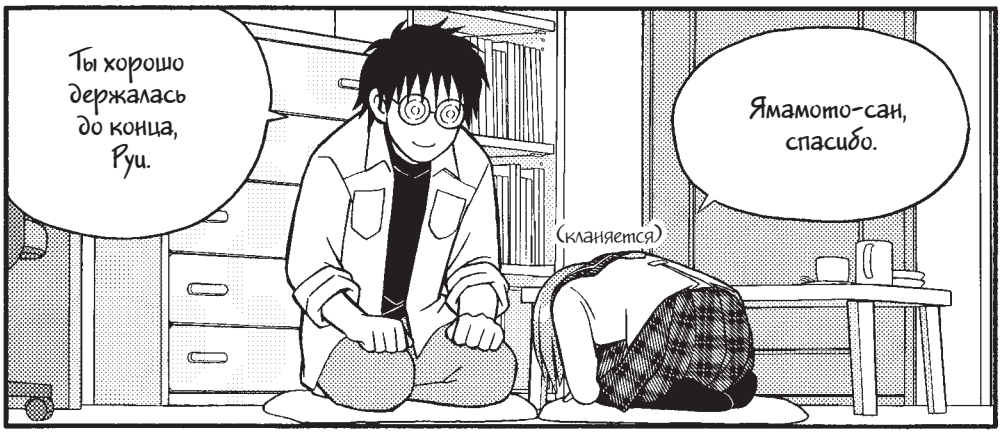
Шаг 7р

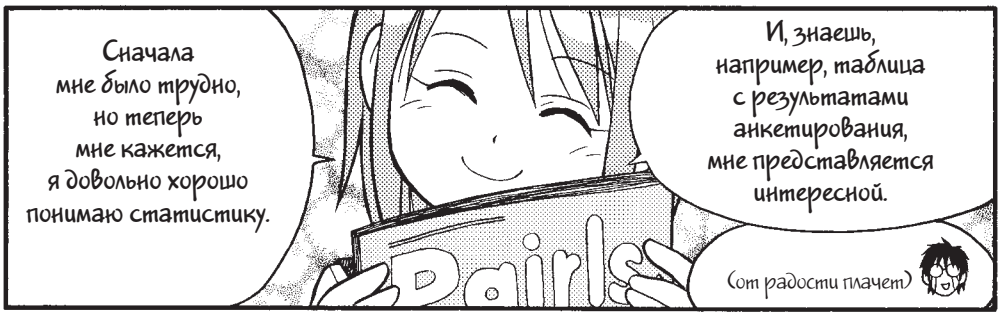
Если Р-значение меньше, чем уровень значимости (см. ***Шаг 6р***), делают вывод, что «верна альтернативная гипотеза». В противном случае вывод такой: «нельзя утверждать, что нулевая гипотеза ошибочна».

Р-значение оказалось меньше уровня значимости. Следовательно, верна альтернативная гипотеза: «величина коэффициента корреляции Крамера для генеральной совокупности > 0», т.е. «пол» и «способ признания в любви» связаны!

Даже когда Р-значение оказывается меньше уровня значимости, нельзя на основе проверки делать вывод, что «альтернативная гипотеза абсолютно верна». Можно сделать лишь такой вывод: «хотелось бы утверждать, что альтернативная гипотеза абсолютно верна, но существует вероятность ($\alpha \times 100\%$) того, что верна нулевая гипотеза».


М-ммм... Почти точно так же, как в 1-м случае.
Вроде поняла.
О! Говоришь с такой уверенностью...
(смеётся)
Ну-ка, вспоминай пример с пудингом.
Если P-значение окажется больше уровня значимости, кажется, можно сделать вывод, что «верна нулевая гипотеза», но это не так.
Можно только ...
... сделать такой вывод: «нельзя утверждать, что нулевая гипотеза ошибочна».
(улыбается)
(от радости плачет)


Ты хорошо держалась до конца, Руи.
Ямамото-сан, спасибо.
(кланяется)


Сначала мне было трудно, но теперь мне кажется, я довольно хорошо понимаю статистику.
И, знаешь, например, таблица с результатами анкетирования, мне представляется интересной.
(от радости плачет)


То, что ты говоришь, так радует меня.
Я хочу попробовать другие способы анализа!
Ха-ха-ха... почему бы и тебе, Руи, не заняться тем же, чем занимаюсь я?
В общем-то, да... хотя моей главной целью было ...
Ну, я пошел. Пока.


Решено. Я пойду на встречу с Игараси-саном.
Ой, а Игараси сейчас ...
... в отпуске, у него свадебное путешествие.
Что?????!!!!! Он женился ?
Зачем же я столько времени и сил угрохала на эту статистику...?!
Он женат....
Что? Разве твой интерес был неискренним?
Оставь меня в покое!!
(бежит в сторону Ямамото)
Ой, осторожно!
БУМ
Извините...


Все в порядке?
Я и не знала, что Ямамото-сан такой симпатичный!


Ямамото-сан, научите меня ещё чему-нибудь!
И их занятия продолжились ...
... а может и нет.

5. Проверка гипотезы о независимости и гипотезы об однородности

Проверка гипотезы об однородности очень похожа на проверку гипотезы о независимости. Ниже приводится пример такой проверки. Попробуйте разобраться, в чём заключается разница между этими двумя проверками.

Упражнение

Женский журнал «P-girls» решил провести опрос школьников на тему «Какой способ признания в любви Вы предпочитаете:
- по телефону;
- по SMS;
- при встрече»?

Журнал выдвинул такую гипотезу:

Гипотеза:
пропорции ответов «по телефону» : «по SMS» : «при встрече» зависят от пола респондентов.

Чтобы выяснить, правильна ли эта гипотеза, журналист произвольно выбрал определённое число юношей и девушек из «всех школьников Японии» и провёл анкетирование. Результаты этого анкетирования приведены в следующей таблице:

Респондент	Способ признания в любви	Возраст респондента	Пол респондента
1 ... 148	при встрече ... по SMS	17 ... 16	ж ... ж
149 ... 300	по телефону ... по SMS	15 ... 18	м ... м

Таблица взаимной сопряжённости «пола» и «способа признания в любви»

Пол респондента	Способ признания в любви			Итого
	по телефону	по SMS	при встрече	
женский	34	61	53	148
мужской	38	40	74	152
Итого:	72	101	127	300

Проверьте правильность сформулированной гипотезы путём проверки гипотезы об однородности. Предположим, что уровень значимости равен 0,05.

Решение

1	Определим генеральную совокупность	Генеральных совокупностей будет две: «все школьницы Японии» и «все школьники Японии».
2	Сформулируем основную и противоположную гипотезы	Нулевая гипотеза: «пропорции предпочтительных способов признания в любви — по телефону : по SMS : при встрече — у школьниц и школьников одинаковы». Альтернативная гипотеза: «пропорции предпочтительных способов признания в любви — по телефону : по SMS : при встрече — у школьниц и школьников различны».
3	Выбираем вид гипотезы для статистической проверки	Проведём проверку гипотезы об однородности.
4	Определим уровень значимости	Пусть уровень значимости равен 0,05.
5	Вычислим фактическое значение выбранного статистического критерия на основе данных выборочной совокупности	В этом упражнении будет проведена проверка гипотезы об однородности. Следовательно, статистическим критерием будет критерий согласия Пирсона. Величина была уже вычислена и равна 8,0091 (см. стр. 132). Если нулевая гипотеза верна, критерий согласия Пирсона имеет распределение хи-квадрат с числом степеней свободы, равным $(2-1)\times(3-1)=1\times 2=2$.
6	Проверим, входит ли значение вычисленного на Шаге 5 статистического критерия в критическую область	Величина критерия согласия Пирсона, который является статистическим критерием, равна 8,0091. Так как уровень значимости = 0,05, критическая область, как следует из таблицы распределения хи-квадрат на стр. 103, больше 5,9915. Это значит, что значение статистического критерия входит в критическую область.
7	Если значение статистического критерия входит в критическую область (Шаг 6), делают вывод: «верна альтернативная гипотеза». Если нет, — «нельзя утверждать, что нулевая гипотеза ошибочна»	Значение статистического критерия входит в критическую область. Следовательно, верна альтернативная гипотеза: «пропорции предпочтительных способов признания в любви — по телефону : по SMS : при встрече — у школьниц и школьников различны».

Как вам? И упражнение, и ответ как две капли воды похожи на проверку гипотезы о независимости. Давайте уточним, чем же отличается проверка гипотезы о независимости от проверки гипотезы об однородности.

Во-первых, определённые нами генеральные совокупности разные. В проверке гипотезы о независимости генеральная совокупность только одна — «все школьники Японии», а в проверке гипотезы об однородности генеральных совокупностей две: «все школьницы Японии» и «все школьники Японии».

Во-вторых, нулевые и альтернативные гипотезы разные. При проверке гипотезы о независимости были сформулированы следующие гипотезы:

Нулевая гипотеза	Коэффициент корреляции Крамера для генеральной совокупности = 0 и, значит, «пол» и «способ признания в любви» не связаны.
Альтернативная гипотеза	Коэффициент корреляции Крамера для генеральной совокупности > 0 и, значит «пол» и «способ признания в любви» связаны.

а при проверке гипотезы об однородности были сформулированы гипотезы:

Нулевая гипотеза	Пропорции предпочтительных способов признания в любви — по телефону : по SMS : при встрече — у школьниц и школьников одинаковы.
Альтернативная гипотеза	Пропорции предпочтительных способов признания в любви — по телефону : по SMS : при встрече — у школьниц и школьников различны.

К тому же, гипотезы формулировались в разные моменты: в случае проверки гипотезы о независимости — после сбора данных, а в случае проверки гипотезы об однородности — до сбора данных. Однако, несмотря на очевидные различия на практике часто бывает так: собираются проводить проверку гипотезы о независимости, а на деле проводят проверку гипотезы об однородности или наоборот. **Будьте внимательны!**

6. Как выразить словами вывод на основании проверки

Вывод, полученный на основании проверки, формулируется следующим образом:

> Если величина статистического критерия входит в критическую область, делают вывод о том, что верна альтернативная гипотеза. В противном случае вывод таков: «нельзя утверждать, что нулевая гипотеза ошибочна».

На самом деле такие выражения для формулирования вывода не используются. Различные выражения, которые действительно используются для формулирования вывода, сделанного на основе результатов статистической проверки, приведены в Табл. 7.4.

Таблица 7.4. Выражения, используемые для формулирования выводов на основе проверки

Случаи, когда величина статистического критерия входит в критическую область	Случаи, когда величина статистического критерия не входит в критическую область
• верна альтернативная гипотеза • альтернативная гипотеза значима • нулевая гипотеза отвергается	• нельзя утверждать, что нулевая гипотеза ошибочна • нельзя отвергнуть нулевую гипотезу • нулевая гипотеза отклонена • нельзя утверждать, что нулевая гипотеза верна • нулевая гипотеза принимается

Выражения «гипотеза значима» и «гипотеза отвергнута» используются достаточно часто. Вместе с тем я специально использовал формулировки, которые, как правило, не используются. Объясняю почему. Я заметил, что среди начинающих изучать статистическую проверку гипотез, есть такие, которые часто говорят «гипотеза значима», причём не очень хорошо понимая, какой смысл имеет данное выражение. Очевидно, они используют это выражение только потому, что уверены в величине P-значения и получили значение статистического критерия. Другими словами, они проводят статистическую проверку гипотез, не сформулировав чётко нулевую и альтернативную гипотезы. И, как мне кажется, генеральная совокупность также чётко не определена. Я раньше полагал, что не следует делать замечания начинающим постигать премудрости статистической проверки гипотез. Однако без чётко выстроенных гипотез невозможно сделать какой-либо вывод.

В связи с этим я использую такие выражения, как «верна альтернативная гипотеза» и «нельзя утверждать, что основная гипотеза ошибочна». Это позволит читателю понять и усвоить, что такое нулевая и альтернативная гипотезы.

Упражнение

Таблица взаимной сопряжённости (см. стр. 138).

Обычно заказываемая кухня	Предпочитаемый напиток		Итого
	Кофе	Чай	
Японская	43	33	76
Европейская	51	53	104
Китайская	29	41	70
Итого:	123	127	250

Выясните путём проверки гипотезы о независимости, больше ли 0 коэффициент корреляции Крамера между видом обычно заказываемой кухни и предпочитаемым напитком для генеральной совокупности «жители Японии старше 20 лет». Другими словами, есть ли взаимосвязь между видом обычно заказываемой кухни и предпочитаемым напитком? Пусть уровень значимости равен 0,01.

Ответ

1	Определим генеральную совокупность	Генеральной совокупностью будут «жители Японии старше 20 лет»
2	Сформулируем нулевую и альтернативную гипотезы	Основная гипотеза: «вид обычно заказываемой кухни и предпочитаемый напиток не связаны». Альтернативная гипотеза: «вид обычно заказываемой кухни и предпочитаемый напиток взаимосвязаны».
3	Выбираем вид гипотезы	Проведём проверку гипотезы о независимости.
4	Определим уровень значимости	Пусть уровень значимости равен 0,01.
5	Вычислим фактическое значение выбранного статистического критерия на основе данных выборочной совокупности	В этом упражнении будет проведена проверка гипотезы о независимости. Следовательно, статистическим критерием будет являться критерий согласия Пирсона. Величина была вычислена ранее и равна 3,34839 (см. стр. 141)
6	Проверим, входит ли значение вычисленного на Шаге 5 статистического критерия в критическую область	Величина критерия согласия Пирсона, являющегося статистическим критерием, равна 3,3483. Так как уровень значимости $\alpha = 0{,}01$, критическая область, как следует из таблицы распределения хи-квадрат на стр. 103, больше 9,2104. Это значит, что значение статистического критерия не входит в критическую область.
7	Если значение статистического критерия входит в критическую область (Шаг 6), делают вывод: «верна альтернативная гипотеза». Если нет — «нельзя утверждать, что нулевая гипотеза ошибочна».	Значение статистического критерия не входит в критическую область. Следовательно, нельзя утверждать, что нулевая гипотеза: «вид обычно заказываемой кухни и предпочитаемый напиток не связаны» ошибочна.

Выводы

- **Проверка — один из способов анализа, который позволяет установить, правильна ли гипотеза, сделанная исследователем о генеральной совокупности на основе данных выборочной совокупности.**
- **Такая проверка называется статистической проверкой гипотез.**
- **Статистический критерий — формула, с помощью которой преобразуются данные выборочной совокупности.**
- **Обычно используются такие значения уровня значимости, как 0,05 или 0,01.**
- **Критическая область — область значений, соответствующая определённому уровню значимости.**
- **Проверка гипотезы о независимости — один из способов анализа, который позволяет выяснить, не равен ли 0 коэффициент корреляции Крамера для генеральной совокупности. Можно также сказать, что такой анализ позволяет выяснить, есть ли связь между двумя переменными в таблице взаимной сопряжённости.**
- **Если коэффициент корреляции Крамера для генеральной совокупности равен 0, величина критерия согласия Пирсона имеет распределение хи-квадрат.**
- **P-значение в случае проверки гипотезы о независимости при условии, что нулевая гипотеза верна, — вероятность того, что критерий согласия Пирсона χ_0^2 больше или равен наблюдаемой величины.**
- **Сделать вывод о проверке можно на основании того, что:**
  1) значение статистического критерия входит в критическую область, либо
  2) P-значение меньше уровня значимости.
- **Будь то проверка гипотезы о независимости или любая другая проверка, порядок проведения анализа один и тот же, при этом он предусматривает выполнение следующих действий:**

Шаг 1	Определить генеральную совокупность.
Шаг 2	Сформулировать нулевую и альтернативную гипотезы.
Шаг 3	Выбрать вид статистической проверки гипотезы.
Шаг 4	Определить уровень значимости.
Шаг 5	Вычислить фактическое значение выбранного статистического критерия на основе данных выборочной совокупности.
Шаг 6	Проверить, входит ли вычисленное на Шаге 5 значение статистического критерия в критическую область.
Шаг 7	Если P-значение меньше, чем уровень значимости (см. Шаг 6р), то делают вывод о том, что верна альтернативная гипотеза. Если нет, то вывод «нельзя утверждать, что нулевая гипотеза ошибочна».
Шаг 6р	Проверить, меньше ли P-значение, соответствующее величине выбранного статистического критерия, вычисленного на Шаге 5, чем уровень значимости.

Приложение

Попробуем вычислить с помощью Excel

В этой главе объясняется, как с помощью программы Excel:

1. Построить таблицу (ряд) распределения.
2. Вычислить среднее значение, медиану и стандартное отклонение.
3. Составить простую статистическую таблицу.
4. Вычислить нормированное отклонение и рассчитать T-показатель.
5. Вычислить вероятность стандартного нормального распределения.
6. Вычислить значение x при распределении хи-квадрат.
7. Вычислить коэффициент линейной корреляции.
8. Проверить гипотезу о независимости.

Файлы Excel можно загрузить по адресу *http://www.dodeca.ru/books/33081.php*. Читателю, не имеющему опыта работы в Excel, рекомендуется сначала попробовать вычислить среднее значение, медиану и стандартное отклонение (см. стр. 195).

1. Построение таблиц распределения

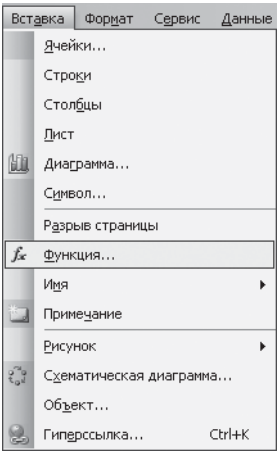
Используются данные со стр. 33.

Выберите ячейку J3.

	A	B	C	D	E	F	G	H	I	J
1	**Ресторан**	**Цена, йены**		**Ресторан**	**Цена, йены**					
2	ресторан 1	700		ресторан 26	780		**от**	**до**	**меньше**	**частота**
3	ресторан 2	850		ресторан 27	590		500	600	599	
4	ресторан 3	600		ресторан 28	650		600	700	699	
5	ресторан 4	650		ресторан 29	580		700	800	799	
6	ресторан 5	980		ресторан 30	750		800	900	899	
7	ресторан 6	750		ресторан 31	800		900	1000	999	
8	ресторан 7	500		ресторан 32	550					
9	ресторан 8	890		ресторан 33	750					
10	ресторан 9	880		ресторан 34	700					
11	ресторан 10	700		ресторан 35	600					
12	ресторан 11	890		ресторан 36	800					
13	ресторан 12	720		ресторан 37	800					
14	ресторан 13	680		ресторан 38	880					
15	ресторан 14	650		ресторан 39	790					
16	ресторан 15	790		ресторан 40	790					
17	ресторан 16	670		ресторан 41	780					
18	ресторан 17	680		ресторан 42	600					
19	ресторан 18	900		ресторан 43	670					
20	ресторан 19	880		ресторан 44	680					
21	ресторан 20	720		ресторан 45	650					
22	ресторан 21	850		ресторан 46	890					
23	ресторан 22	700		ресторан 47	930					
24	ресторан 23	780		ресторан 48	650					
25	ресторан 24	850		ресторан 49	777					
26	ресторан 25	750		ресторан 50	700					

Шаг 2

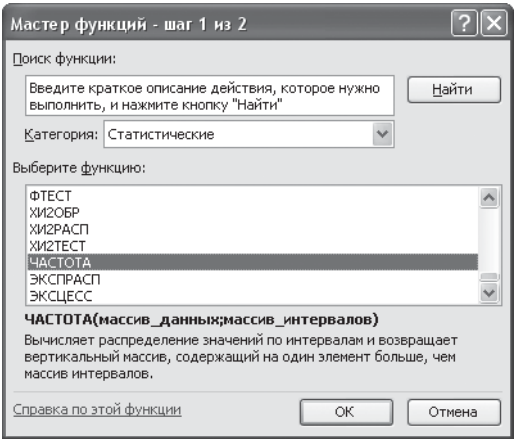
Выберите **Вставка ▸ Функция**.



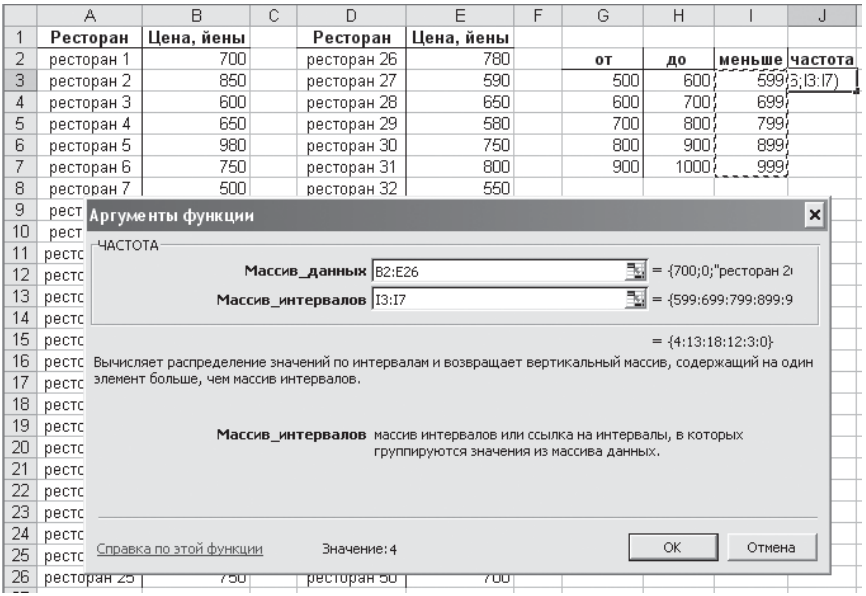
Шаг 3

Выберите **Статистические** в строке **Категория**, а затем **ЧАСТОТА*** в графе **Выберите функцию**.

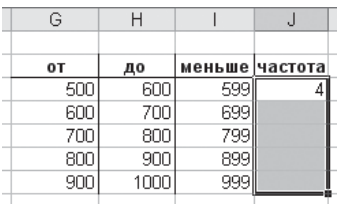
* FREQUENCY в английской версии Excel.



Шаг 4 Выделите нижеуказанную область и нажмите кнопку **OK**.

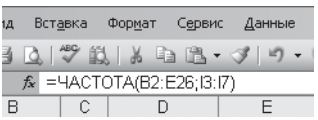


Шаг 5 Выделите ячейки, начиная с ячейки J3 и до ячейки J7.



от	до	меньше	частота
500	600	599	4
600	700	699	
700	800	799	
800	900	899	
900	1000	999	

Шаг 6 Щелкните мышью на области в строке формул.



Шаг 7 Нажмите комбинацию клавиш **Shift + Ctrl** и, удерживая ее в нажатом состоянии, нажмите клавишу **Enter**.

Шаг 8 Вычисление закончено!

G	H	I	J
от	до	меньше	частота
500	600	599	4
600	700	699	13
700	800	799	18
800	900	899	12
900	1000	999	3

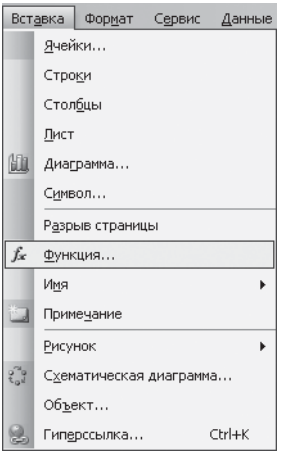
2. Вычисление среднего значения, медианы и стандартного отклонения

Используются данные со стр. 41.

Шаг 1 Выберите ячейку B10.

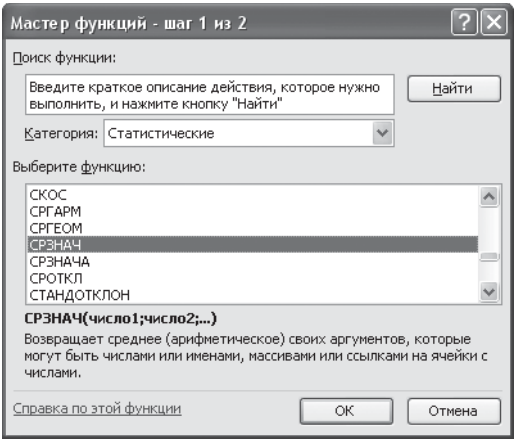
	A	B
1		Команда A
2	Руи-Руи	86
3	Дзюн	73
4	Юми	124
5	Сизука	111
6	Токо	90
7	Каздэ	38
8		
9		
10	среднее	
11	медиана	
12	отклонение	

Шаг 2 Выберите **Вставка** ▸ **Функция**.



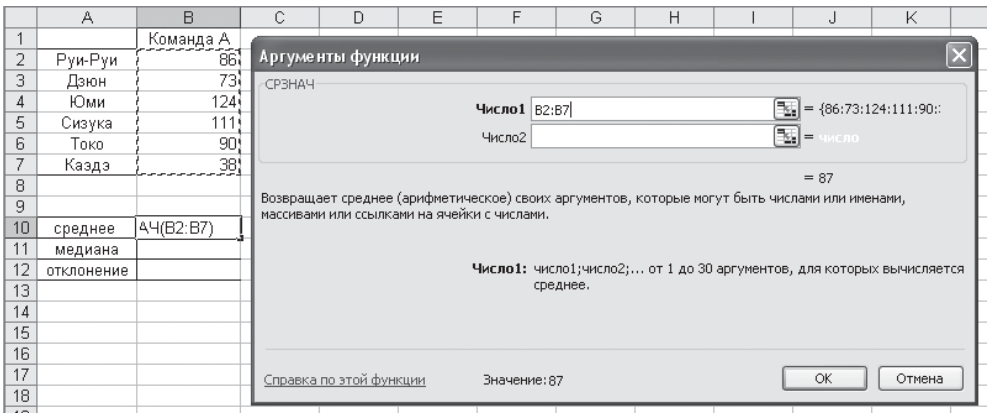
Шаг 3

Выберите **Статистические** в строке **Категория**, и **СРЗНАЧ*** в графе **Выберите функцию**.



* AVERAGE в английской версии Excel.

Шаг 4 Выделите нижеуказанную область и нажмите кнопку **OK**.



	A	B
1		Команда А
2	Руи-Руи	86
3	Дзюн	73
4	Юми	124
5	Сизука	111
6	Токо	90
7	Каэдэ	38
8		
9		
10	среднее	АЧ(B2:B7)
11	медиана	
12	отклонение	

Шаг 5

Вычисление закончено!

	A	B
1		Команда А
2	Руи-Руи	86
3	Дзюн	73
4	Юми	124
5	Сизука	111
6	Токо	90
7	Каэдэ	38
8		
9		
10	среднее	87
11	медиана	
12	отклонение	

Шаг 6 Вычислите медиану и стандартное отклонение, выполняя последовательно Шаг 1 — Шаг 5. При вычислении медианы выберите **МЕДИАНА*** в графе **Выберите функцию**, а при вычислении стандартного отклонения — **СТАНДОТКЛОНП****.

* MEDIAN в английской версии Excel.
** STDEVP в английской версии Excel.

3. Построение простой статистической таблицы

Используются данные со стр. 61.

Шаг 1 Выберите ячейку F20.

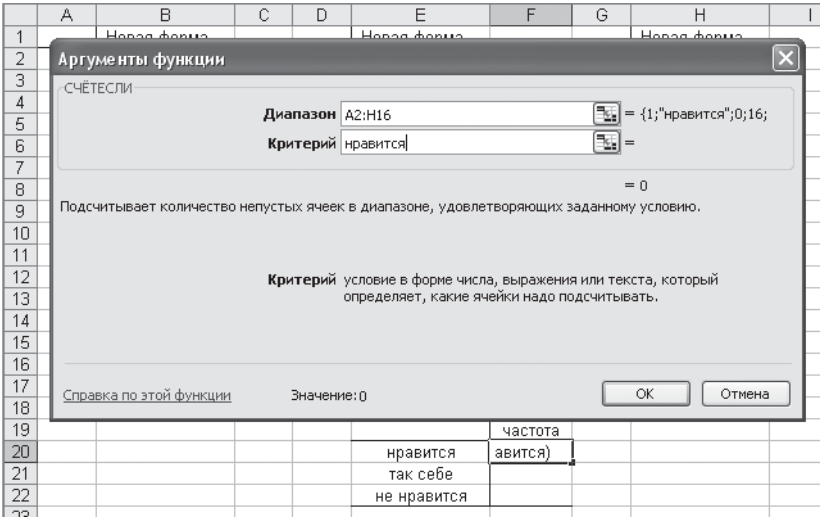
	A	B	C	D	E	F	G	H
1		Новая форма...			Новая форма...			Новая форма...
2	1	нравится		16	так себе		31	так себе
3	2	так себе		17	нравится		32	так себе
4	3	нравится		18	нравится		33	нравится
5	4	так себе		19	нравится		34	не нравится
6	5	не нравится		20	нравится		35	нравится
7	6	нравится		21	нравится		36	нравится
8	7	нравится		22	нравится		37	нравится
9	8	нравится		23	не нравится		38	нравится
10	9	нравится		24	так себе		39	так себе
11	10	нравится		25	нравится		40	нравится
12	11	нравится		26	нравится			
13	12	нравится		27	не нравится			
14	13	так себе		28	нравится			
15	14	нравится		29	нравится			
16	15	нравится		30	нравится			
17								
18								
19						частота		
20					нравится			
21					так себе			
22					не нравится			

Шаг 2 Выберите пункт **Вставка ▸ Функция**.

Шаг 3 Выберите **Статистические** в строке **Категория**, а затем **СЧЕТЕСЛИ*** в графе **Выберите функцию**.

* COUNTIF в английской версии Excel.

Шаг 4 Выделите указанную ниже область, напишите **нравится** в графе **Критерий** и нажмите кнопку **OK**.



Шаг 5

Вычисление закончено!

	A	B	C	D	E	F	G	H
1		Новая форма...			Новая форма...			Новая форма...
2	1	нравится		16	так себе		31	так себе
3	2	так себе		17	нравится		32	так себе
4	3	нравится		18	нравится		33	нравится
5	4	так себе		19	нравится		34	не нравится
6	5	не нравится		20	нравится		35	нравится
7	6	нравится		21	нравится		36	нравится
8	7	нравится		22	нравится		37	нравится
9	8	нравится		23	не нравится		38	нравится
10	9	нравится		24	так себе		39	так себе
11	10	нравится		25	нравится		40	нравится
12	11	нравится		26	нравится			
13	12	нравится		27	не нравится			
14	13	так себе		28	нравится			
15	14	нравится		29	нравится			
16	15	нравится		30	нравится			
17								
18								
19						частота		
20					нравится	28		
21					так себе			
22					не нравится			

Шаг 6 Выполняя последовательно Шаг 1 — Шаг 5, подсчитайте частоту ответов «так себе» и «не нравится».

4. Вычисление нормированного отклонения и рейтинга успеваемости

Используются данные со стр. 72.

Функция для расчета нормированного отклонения в Excel есть, а функции для расчета рейтинга успеваемости (Т-показателя) нет.
Несмотря на это, можно довольно быстро вычислить рейтинг успеваемости, используя результат вычисления нормированного отклонения.
Поэтому в данной книге считается, что рейтинг успеваемости можно рассчитать в Excel.

4.1. Вычисление нормированного отклонения

Шаг 1 Выберите ячейку E2.

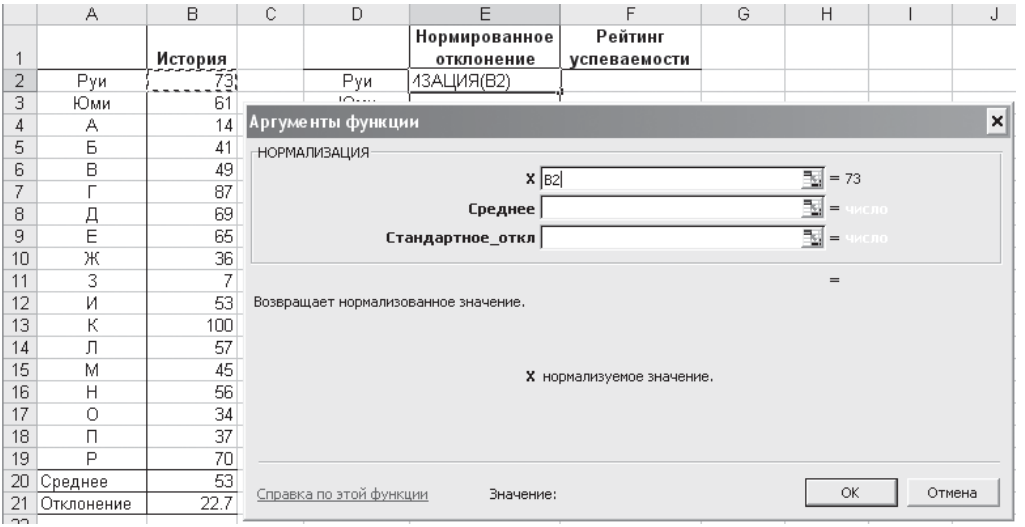
		История			Нормированное отклонение	Рейтинг успеваемости
2	Руи	73		Руи		
3	Юми	61		Юми		
4	А	14		А		
5	Б	41		Б		
6	В	49		В		
7	Г	87		Г		
8	Д	69		Д		
9	Е	65		Е		
10	Ж	36		Ж		
11	З	7		З		
12	И	53		И		
13	К	100		К		
14	Л	57		Л		
15	М	45		М		
16	Н	56		Н		
17	О	34		О		
18	П	37		П		
19	Р	70		Р		
20	Среднее	53				
21	Отклонение	22.7				

Шаг 2 Выберите **Вставка ▸ Функция**.

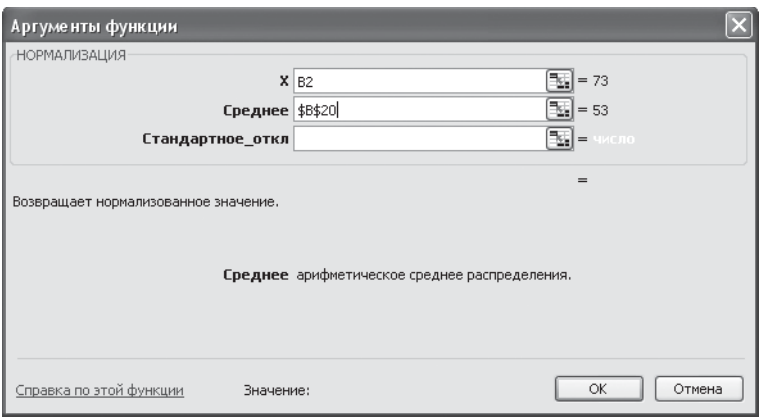
Шаг 3 Выберите **Статистические** в строке **Категория**, а затем **НОРМАЛИЗАЦИЯ*** в графе **Выберите функцию**.

* STANDARDIZE в английской версии Excel.

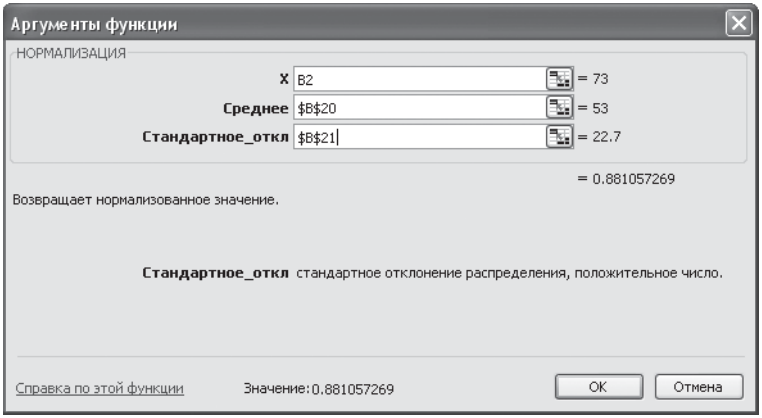
Шаг 4 Выберите ячейку B2.



Шаг 5 Выберите ячейку **B20** в графе **Среднее** и нажмите клавишу **F4**. Убедитесь, что **B20** в графе **Среднее** превратилось в **B20**.



Шаг 6 Выберите ячейку **B21** в графе **Стандартное_откл** и нажмите клавишу **F4**. Убедитесь, что **B21** в графе **Стандартное_откл** превратилось в **B21** и нажмите кнопку **OK**.



Шаг 7 Убедитесь, что было вычислено нормированное отклонение Руи.

	A	B	C	D	E	F
1		**История**			**Нормированное отклонение**	**Рейтинг успеваемости**
2	Руи	73		Руи	0.88	
3	Юми	61		Юми		
4	А	14		А		
5	Б	41		Б		
6	В	49		В		
7	Г	87		Г		
8	Д	69		Д		
9	Е	65		Е		
10	Ж	36		Ж		
11	З	7		З		
12	И	53		И		
13	К	100		К		
14	Л	57		Л		
15	М	45		М		
16	Н	56		Н		
17	О	34		О		
18	П	37		П		
19	Р	70		Р		
20	Среднее	53				
21	Отклонение	22.7				

Подведите указатель мыши к правому углу ячейки **E2**, и как только он превратится в чёрный крестик, нажмите клавишу мыши и, удерживая её в нажатом состоянии, растяните область до ячейки **E19**; отпустите клавишу мыщи.

D	E	F
	Нормированное отклонение	**Рейтинг успеваемости**
Руи	0.88	
Юми		
А		
Б		
В		
Г		
Д		
Е		
Ж		
З		
И		
К		
Л		
М		
Н		
О		
П		
Р		

Вычисление нормированного отклонения закончено!

D	E	F
	Нормированное отклонение	**Рейтинг успеваемости**
Руи	0.88	
Юми	0.35	
А	-1.71	
Б	-0.53	
В	-0.18	
Г	1.49	
Д	0.70	
Е	0.53	
Ж	-0.75	
З	-2.02	
И	0.00	
К	2.07	
Л	0.18	
М	-0.35	
Н	0.13	
О	-0.84	
П	-0.70	
Р	0.75	

4.2. Вычисление рейтинга успеваемости

Шаг 10 Выберите ячейку **F2** и «напишите» точно так же, как пишете текст в Word, выражение **=E2*10+50**, а затем нажмите клавишу **Enter**.

D	E	F
	Нормированное отклонение	**Рейтинг успеваемости**
Руи	0.88	=E2*10+50
Юми	0.35	
А	-1.71	
Б	-0.53	
В	-0.18	
Г	1.49	
Д	0.70	
Е	0.53	
Ж	-0.75	
З	-2.02	
И	0.00	
К	2.07	
Л	0.18	
М	-0.35	
Н	0.13	
О	-0.84	
П	-0.70	
Р	0.75	

Шаг 11 Повторите Шаг 8.

Шаг 12 Вычисление рейтинга успеваемости закончено!

D	E	F
	Нормированное отклонение	**Рейтинг успеваемости**
Руи	0.88	58.79
Юми	0.35	53.52
А	-1.71	32.85
Б	-0.53	44.72
В	-0.18	48.24
Г	1.49	64.95
Д	0.70	57.04
Е	0.53	55.28
Ж	-0.75	42.53
З	-2.02	29.77
И	0.00	50.00
К	2.07	70.67
Л	0.18	51.76
М	-0.35	46.48
Н	0.13	51.32
О	-0.84	41.65
П	-0.70	42.96
Р	0.75	57.47

5. Вычисление вероятности стандартного нормального распределения

Используются данные со стр. 93.

Шаг 1 Выберите ячейку **B2**.

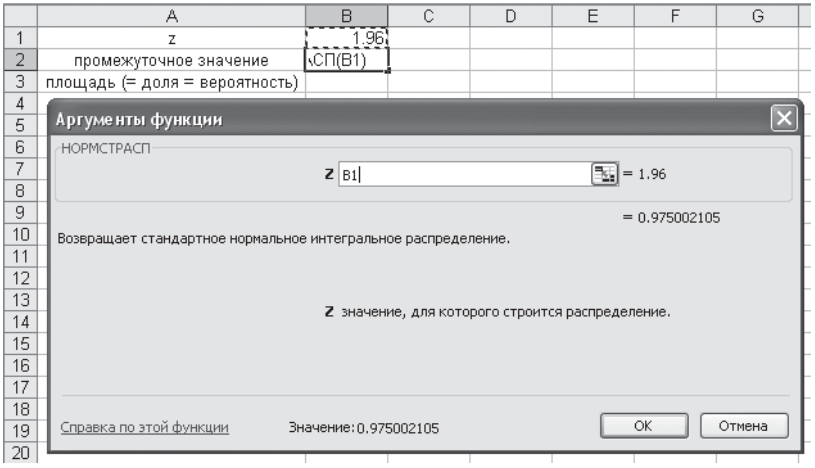
	A	B
1	z	1.96
2	промежуточное значение	
3	площадь (= доля = вероятность)	

Шаг 2 Выберите **Вставка ▸ Функция**.

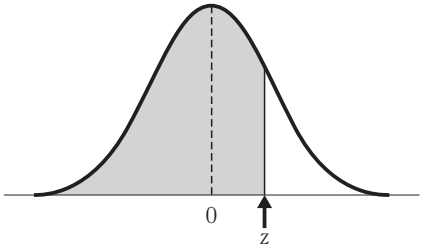
Шаг 3 Выберите **Статистические** в строке **Категория**, а затем **НОРМСТРАСП*** в графе **Выберите функцию**.

* NORMDIST в английской версии Excel.

Шаг 4 Выберите ячейку **B1** и нажмите кнопку **OK**.



Шаг 5 На самом деле функция **НОРМСТРАСП** предназначена для вычисления вероятности, указанной ниже на рисунке.



Поэтому «напишите» в ячейке B3
точно так же, как пишете в Word, выражение [=B2-0.5].

	A	B
1	z	1.96
2	промежуточное значение	0.975002
3	площадь (= доля = вероятность)	=B2-0.5

Шаг 6 Вычисление закончено!

	A	B
1	z	1.96
2	промежуточное значение	0.975002
3	площадь (= доля = вероятность)	0.475002

6. Вычисление значения x при распределении хи-квадрат

Используются данные со стр. 104.

Шаг 1 Выберите ячейку **B3**.

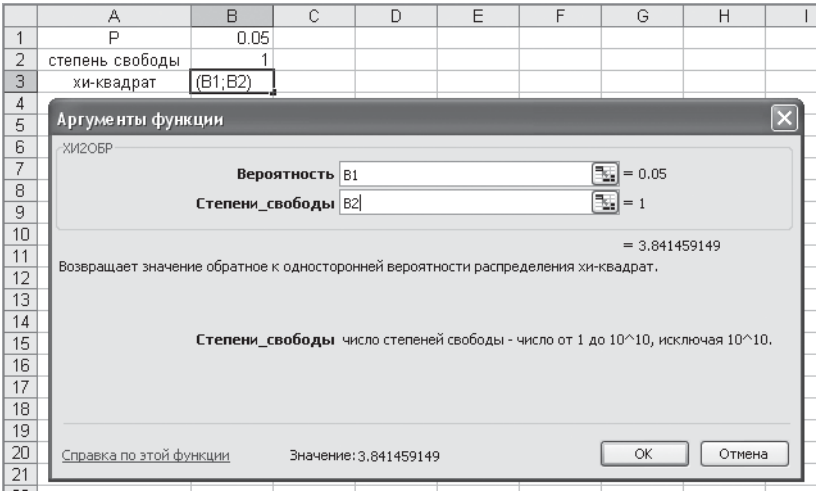
	A	B
1	P	0.05
2	степень свободы	1
3	хи-квадрат	

Шаг 2 Выберите **Вставка ▸ Функция**.

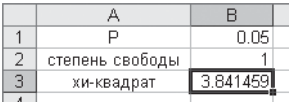
Шаг 3 Выберите **Статистические** в строке **Категория**, а затем **ХИ2ОБР*** в графе **Выберите функцию**.

* CHIINV в английской версии Excel.

Шаг 4 Выберите ячейки B1 и B2 и нажмите кнопку **OK**.



Шаг 5 Вычисление закончено!



7. Вычисление коэффициента линейной корреляции

Используются данные со стр. 116

Шаг 1 Выберите ячейку **B14**.

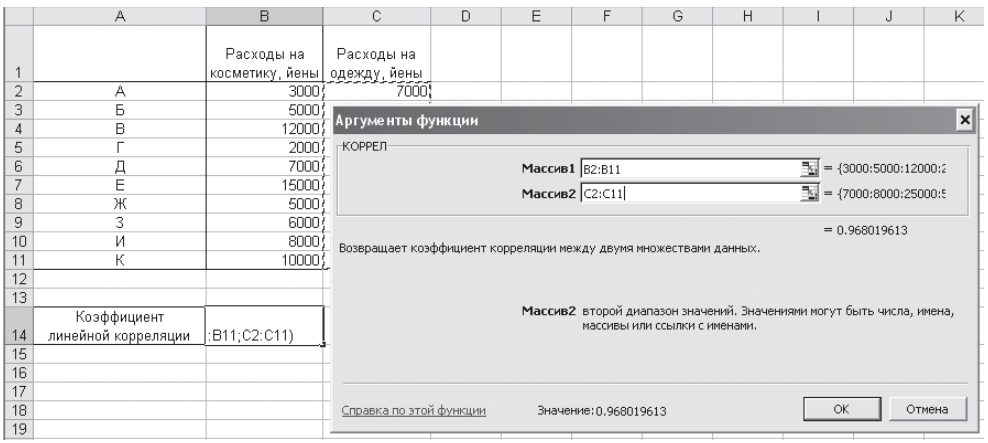
	A	B	C
1		Расходы на косметику, йены	Расходы на одежду, йены
2	А	3000	7000
3	Б	5000	8000
4	В	12000	25000
5	Г	2000	5000
6	Д	7000	12000
7	Е	15000	30000
8	Ж	5000	10000
9	З	6000	15000
10	И	8000	20000
11	К	10000	18000
12			
13			
14	Коэффициент линейной корреляции		

Шаг 2 Выберите **Вставка ▸ Функция**.

Шаг 3 Выберите **Статистические** в строке **Категория**, а затем **КОРРЕЛ*** в графе **Выберите функцию**.

* CORREL в английской версии Excel.

Шаг 4 Выделите указанную ниже область и нажмите кнопку **ОК**.



Шаг 5 Вычисление закончено!

	A	B	C
1		Расходы на косметику, йены	Расходы на одежду, йены
2	А	3000	7000
3	Б	5000	8000
4	В	12000	25000
5	Г	2000	5000
6	Д	7000	12000
7	Е	15000	30000
8	Ж	5000	10000
9	З	6000	15000
10	И	8000	20000
11	К	10000	18000
12			
13			
14	Коэффициент линейной корреляции	0.968019613	

8. Проверка гипотезы о независимости

Используются данные со стр. 157.

Шаг 1 Выберите ячейку **B8**.

	A	B	C	D	E
1		по телефону	по SMS	при встрече	Итого
2	ж	34	61	53	148
3	м	38	40	74	152
4	Итого	72	101	127	300
5					
6					
7		по телефону	по SMS	при встрече	
8	ж				
9	м				
10					
11					
12	P-значение				

Шаг 2 Напишите в ячейке B8 точно так же, как пишете в Word, выражение: =E2*B4/E4 и пока не нажимайте клавишу Enter.

	A	B	C	D	E
1		по телефону	по SMS	при встрече	Итого
2	ж	34	61	53	148
3	м	38	40	74	152
4	Итого	72	101	127	300
5					
6					
7		по телефону	по SMS	при встрече	
8	ж	=E2*B4/E4			
9	м				
10					
11					
12	P-значение				

Шаг 3 Подведите курсор к надписи E2 в ячейке B8 и трижды нажмите кнопку F4. Убедитесь, что E2 превратилось в $E2, и пока не нажимайте клавишу Enter.

	A	B	C	D	E
1		по телефону	по SMS	при встрече	Итого
2	ж	34	61	53	148
3	м	38	40	74	152
4	Итого	72	101	127	300
5					
6					
7		по телефону	по SMS	при встрече	
8	ж	=$E2*B4/E4			
9	м				
10					
11					
12	P-значение				

Шаг 4 Подведите курсор к надписи B4 в ячейке B8, дважды нажмите клавишу F4 и убедитесь, что B4 превратилось в B$4. Затем подведите курсор к надписи E4 в ячейке B8, нажмите клавишу F4 и проверьте, превратилось ли E4 в E4.

	A	B	C	D	E
1		по телефону	по SMS	при встрече	Итого
2	ж	34	61	53	148
3	м	38	40	74	152
4	Итого	72	101	127	300
5					
6					
7		по телефону	по SMS	при встрече	
8	ж	=$E2*B$4/E4			
9	м				
10					
11					
12	P-значение				

После этого нажмите клавишу Enter.

Шаг 5 Выберите ячейку B8, подведите указатель мыши к правому углу ячейки B8 и как только указатель примет вид черного крестика нажмите клавишу мыши. Удерживая клавишу мыши в нажатом состоянии, растяните область до ячейки D8, а затем отпустите клавишу мыши.

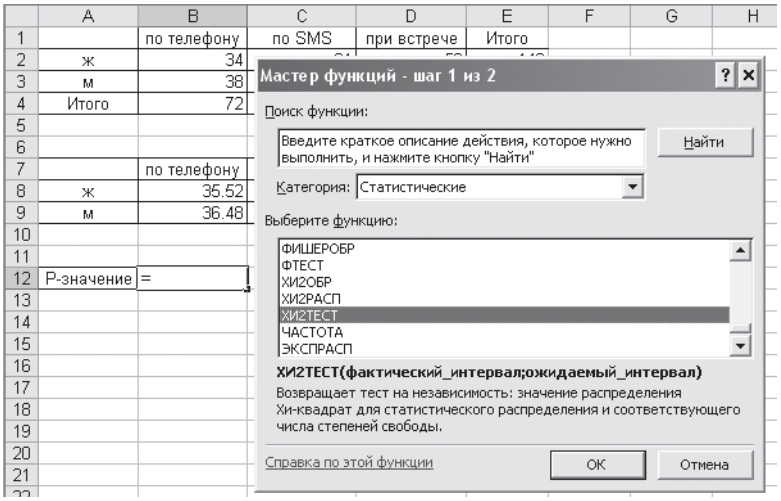
	A	B	C	D	E
1		по телефону	по SMS	при встрече	Итого
2	ж	34	61	53	148
3	м	38	40	74	152
4	Итого	72	101	127	300
5					
6					
7		по телефону	по SMS	при встрече	
8	ж	35,52			
9	м				
10					
11					
12	P-значение				

Шаг 6 Выделите ячейки B8—D8, подведите указатель мыши к правому углу ячейки D8 и как только он превратится в черный крестик нажмите клавишу мыши. Удерживая клавишу мыши в нажатом состоянии, растяните область до ячейки D9, а затем отпустите клавишу мыши.

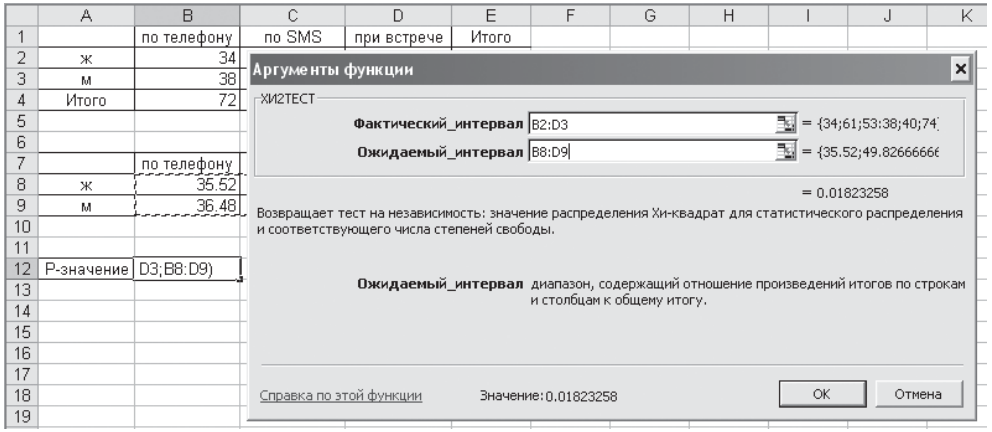
	A	B	C	D	E
1		по телефону	по SMS	при встрече	Итого
2	ж	34	61	53	148
3	м	38	40	74	152
4	Итого	72	101	127	300
5					
6					
7		по телефону	по SMS	при встрече	
8	ж	35.52	49.82666667	62.65333333	
9	м				
10					
11					
12	P-значение				

Шаг 7 Выберите ячейку B12. После этого выберите **Вставка ▸ Функция**, а затем **Статистические** в строке **Категория** и **ХИ2ТЕСТ*** в графе **Выберите функцию**.

* CHITEST в английской версии Excel.



Шаг 8 Выделите указанную ниже область и нажмите кнопку **OK**.



	A	B	C	D	E
1		по телефону	по SMS	при встрече	Итого
2	ж	34			
3	м	38			
4	Итого	72			
5					
6					
7		по телефону			
8	ж	35.52			
9	м	36.48			
10					
11					
12	P-значение	D3;B8:D9)			
13					
14					
15					
16					
17					
18					
19					

Шаг 9 Вычисление закончено!

Проверьте, соответствует ли полученная величина P-значению на стр. 177).

	A	B	C	D	E
1		по телефону	по SMS	при встрече	Итого
2	ж	34	61	53	148
3	м	38	40	74	152
4	Итого	72	101	127	300
5					
6					
7		по телефону	по SMS	при встрече	
8	ж	35.52	49.82666667	62.65333333	
9	м	36.48	51.17333333	64.34666667	
10					
11					
12	P-значение	0.01823258			

Предметный указатель

Книги Издательского дома «Додэка-XXI» можно заказать в торгово-издательском холдинге «АЛЬЯНС-КНИГА» наложенным платежом, выслав открытку или письмо по почтовому адресу: **123242, Москва, а/я 20** или по электронному адресу: **orders@alians-kniga.ru**.

При оформлении заказа следует указать адрес (полностью), по которому должны быть высланы книги; фамилию, имя и отчество получателя. Желательно также указать свой телефон и электронный адрес.

Эти книги вы можете заказать и в Интернет-магазине: **www.alians-kniga.ru**.

Оптовые закупки: тел. **(495) 258-91-94, 258-91-95**; электронный адрес **books@alians-kniga.ru**.

Син Такахаси

Занимательная статистика
Манга

Подписано в печать 15.12.2009. Формат 70x90/16. Бумага офсетная.
Гарнитура «LiteraturnayaC», «JacobC». Печать офсетная.
Объём 14,0 п. л. Усл. п. л. 16,3.
Тираж 1500 экз. Код ОНМ03.

Издательский дом «Додэка-XXI»

105318 Москва, а/я 70
Тел./факс: (495) 366-04-56, 366-11-55
E-mail: red@dodeca.ru
Web-сайт издательства: www.dodeca.ru
Интернет-магазин: www.alians-kniga.ru

Отпечатано с готовых диапозитивов в ОАО «Щербинская типография»
117623, Москва, ул. Типографская, д. 10.