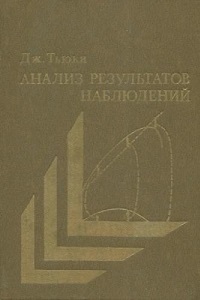
**Джон Тьюки. Анализ результатов наблюдений. Разведочный анализ**

В книге, написанной в 1977 г. известным американским специалистом по математической статистике, изложены основы разведочного анализа данных, т.е. первичной обработки результатов наблюдений, осуществляемой посредством простейших средств — карандаша, бумаги и логарифмической линейки. На многочисленных примерах автор показывает, как представление наблюдений в наглядной форме с помощью схем, таблиц и графиков облегчает выявление закономерностей и подбор способов более глубокой статистической обработки. Изложение сопровождается многочисленными упражнениями с привлечением богатого материала из практики. Живой, образный язык облегчает понимание излагаемого материала.

Джон Тьюки. Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981. – 696 с.



Автор подразделяет статистический анализ на два этапа: разведочный и подтверждающий. Первый этап включает преобразование данных наблюдений и способы их наглядного представления, позволяющие выявить внутренние закономерности, проявляющиеся в данных. На втором этапе применяются традиционные статистические методы оценки параметров и проверки гипотез. Настоящая книга посвящена разведочному анализу данных (о подтверждающем анализе см. [Фишер. Статистический вывод](http://baguzin.ru/wp/?p=15657)). Для чтения книги не требуется предварительных знаний по теории вероятностей и математической статистике.

*Прим. Багузина.* Учитывая год написания книги, автор сосредотачивается на наглядном представлении данных с помощью карандаша, линейки и бумаги (иногда миллиметровой). На мой взгляд, сегодня наглядное представление данных связано с ПК. Поэтому я попытался совместить оригинальные идеи автора и обработку в Excel. Мои комментарии набраны с отступом.

### Глава 1. КАК ЗАПИСЫВАТЬ ЧИСЛА («СТЕБЕЛЬ С ЛИСТЬЯМИ»)

График имеет наибольшую ценность тогда, когда он вынуждает нас заметить то, что мы совсем не ожидали увидеть. Представление чисел в виде стебля и листьев позволяет выявить закономерности. Например, приняв основанием стебля десятки, число 35 можно отнести к стеблю 3. Лист будет равен 5. Для числа 108 стебель – 10, лист – 8.

В качестве примера я взял 100 случайных чисел, распределенных по нормальному закону со средним 10 и стандартным отклонением 3. Чтобы получить такие числа я воспользовался формулой =НОРМ.ОБР(СЛЧИС();10;3) (рис. 1). Откройте приложенный файл Excel. Нажимая F9, вы будете генерировать новый ряд случайных чисел.

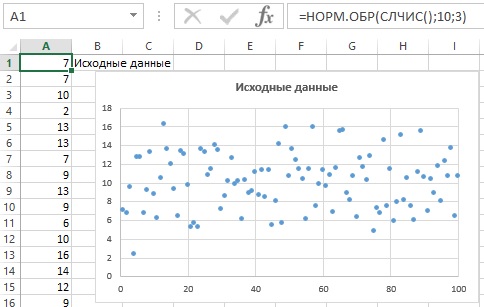


Рис. 1. 100 случайных чисел

Видно, что числа в основном распределены в диапазоне от 5 до 16. Однако заметить какую-либо интересную закономерность сложно. График «стебель и листья» (рис. 2) выявляет нормальное распределение. В качестве ствола были взяты пары соседних чисел, например, 4-5. Листья отражают число значений в этом диапазоне. В нашем примере таких значений 3.

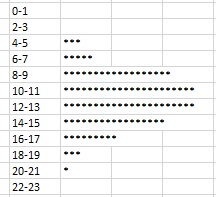


Рис. 2. График «стебель и листья»

В Excel есть две возможности, позволяющие быстро изучить частотные закономерности: функция ЧАСТОТА (рис. 3; подробнее см. [Функция массива ЧАСТОТА](http://baguzin.ru/wp/?p=14715)) и сводные таблицы (рис. 4; подробнее см. [Группировка данных сводной таблицы в Excel 2013](http://baguzin.ru/wp/?p=9739), раздел *Группировка числовых полей*).



Рис. 3. Анализ с помощью функции массива ЧАСТОТА

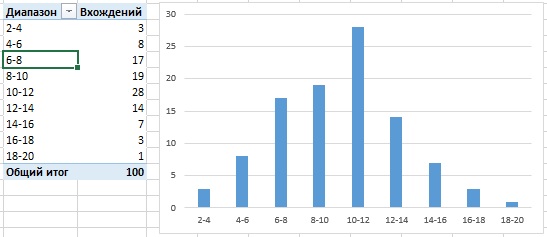


Рис. 4. Анализ с помощью сводных таблиц

Представление в виде стебля с листьями (частотное представление) позволяет выявить следующие особенности данных:

* разделение на группы;
* несимметричное спадание к концам — один «хвост» длиннее другого;
* неожиданно «популярные» и «непопулярные» значения;
* относительно какого значения «центрированы» наблюдения;
* как велик разброс данных.

### Глава 2. ПРОСТЫЕ СВОДКИ ДАННЫХ – ЧИСЛОВЫЕ И ГРАФИЧЕСКИЕ

Представление чисел в виде стебля с листьями позволяет воспринять общую картину выборки. Перед нами стоит задача научиться выражать в сжатом виде наиболее часто встречающиеся общие особенности выборок. Для этого используются сводки данных. Однако, несмотря на то, что сводки могут быть очень полезными, но они не дают всех подробностей выборки. Если этих подробностей не так много, чтобы в них запутаться, лучше всего иметь перед глазами полные данные, размещенные отчетливо удобным для нас способом. Для больших массивов данных сводки необходимы. Мы не предполагаем и не ожидаем, что они заменят полные данные. Разумеется, нередко бывает, что добавление подробностей мало что дает, но важно осознать, что иногда подробности дают многое.

Если для характеристики выборки как целого нам нужно выбрать несколько чисел, которые легко найти, то нам наверняка понадобятся:

* крайние значения — наибольшее и наименьшее, которые мы пометим символом «1» (в соответствии с их рангом или глубиной);
* какое-то срединное значение.

*Медиана* = срединное значение.

Для ряда, представленного в виде стебля с листьями, срединное значение легко найти подсчетом вглубь от любого из концов, приписывая крайнему значению ранг «1». Таким образом, каждое значение в выборке получает свой *ранг*. Счет можно начинать с любого конца. Наименьший из двух получаемых таким образом рангов, которые можно приписать одному и тому же значению, мы назовем *глубиной* (рис. 5). Глубина крайнего значения всегда 1.

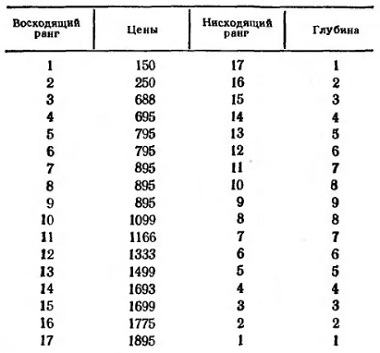
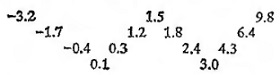


Рис. 5. Определение глубины на основе двух направлений ранжирования

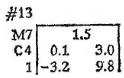
*глубина (или ранг) медианы = (1 + число значений)/2*

Если мы хотим добавить еще два числа, чтобы образовать 5-числовую сводку, то естественно определять их подсчетом до половины расстояния от каждого из концов к медиане. Процесс нахождения медианы, а затем и этих новых значений можно представить себе, как складывание листа бумаги. Поэтому эти новые значения естественно назвать *сгибами* (сейчас чаще используется термин *квартиль*).

В свернутом виде ряд из 13 значений может выглядеть, например, так:



Пять чисел для характеристики ряда в порядке возрастания будут: –3,2; 0,1; 1,5; 3,0; 9,8 — по одному в каждой точке перегиба ряда. Пять чисел (крайние значения, сгибы, медиана), из которых состоит 5-числовад сводка, мы будем изображать в виде следующей простой схемы:



где слева мы показали количество чисел (отмечено знаком #), глубину медианы (буквой М), глубину сгибов (буквой С) и глубину крайних значений (всегда 1, больше ничем отмечать не надо).

На рис. 8 показано, как изобразить 5-чиловую сводку графически. Такого типа график называется «ящик с усами».

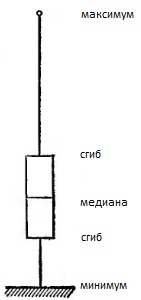


Рис. 8. Схематическая диаграмма или ящик с усами

К сожалению, Excel стандартно строит биржевые диаграммы, основанные только на трех или четырех значениях (рис. 9; как обойти это ограничение см. [Биржевая диаграмма, она же блочная, она же ящичная](http://baguzin.ru/wp/?p=4349)). Для построения 5-числовой сводки можно воспользоваться статистическим пакетом R (рис. 10; подробнее см. [Базовые графические возможности R: диаграммы размахов](http://r-analytics.blogspot.ru/2011/11/r_08.html#.WBXznfmLSUk); если вы не знакомы с пакетом R, можно начать с [Алексей Шипунов. Наглядная статистика. Используем R!](http://baguzin.ru/wp/?p=14666)). Функция boxplot() в R помимо 5 чисел отражает также выбросы (о них чуть позже).

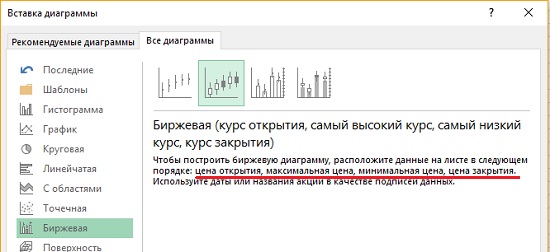


Рис. 9. Возможные типы биржевых диаграмм в Excel

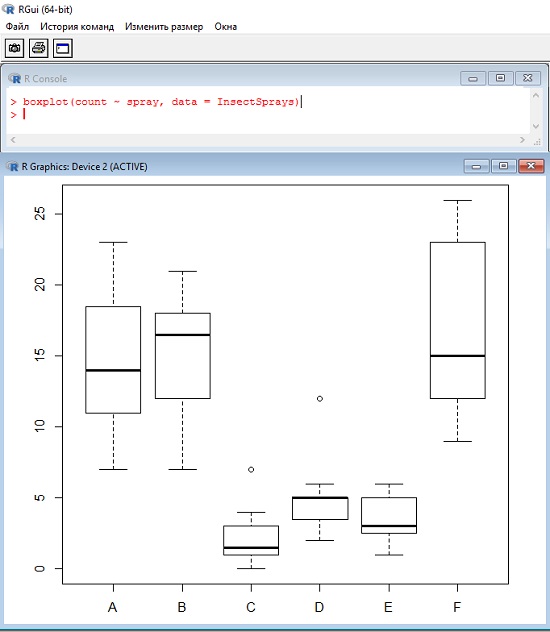


Рис. 10. Ящичная диаграмма в R; для построения такого графика достаточно выполнить команду boxplot(count ~ spray, data = InsectSprays), будут загружены данные, хранящиеся в программе, и построен представленный график

Далее автор описывает, как строить графики с помощью миллиметровки и кальки. Калька нужна, чтобы линии миллиметровки не отвлекали внимание от самого графика. В современном понимании это выражается в [принципе Эдварда Тафти минимизации количества элементов диаграммы](http://baguzin.ru/wp/?p=2071).

При построении диаграммы «ящик с усами» мы будем придерживаться следующей простой схемы:

* «С-ширина» = разность между значениями двух сгибов;
* «шаг» — величина, в полтора раза большая, чем С-ширина;
* «внутренние барьеры» находятся снаружи сгибов на расстоянии одного шага;
* «наружные барьеры» — снаружи на один шаг дальше внутренних;
* значения между внутренним и соседним наружным барьерами будут «внешними»;
* значения за наружными барьерами будем называть «отскакивающими» (или выбросы);
* «размах» = разность между крайними значениями.

Зимой 1893—1894 гг. [Рэлей](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D1%80%D0%B5%D1%82%D1%82,_%D0%94%D0%B6%D0%BE%D0%BD_%D0%A3%D0%B8%D0%BB%D1%8C%D1%8F%D0%BC_(%D0%BB%D0%BE%D1%80%D0%B4_%D0%A0%D1%8D%D0%BB%D0%B5%D0%B9)) исследовал плотность азота, полученного различными способами. На рис. 11 показаны схематическая и точечная диаграммы для 15 значений веса азота, рассматриваемых как единая выборка. Главный факт здесь — резкое разделение на две абсолютно изолированные подгруппы. Это отчетливо видно из индивидуальных значений точечной диаграммы, но почти совершенно не видно на схематической диаграмме (только почти, так как опытный исследователь должен сразу заметить, что «усы» здесь ненормально коротки по сравнению с «ящиком», и почувствовать необходимость более подробного рассмотрения).

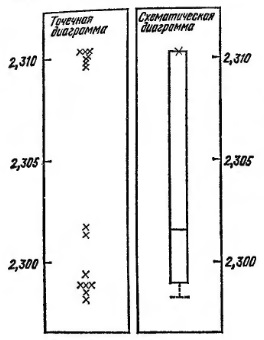


Рис. 11. Точечная и схематическая диаграммы, рассматривающие выборку из 15 измерений, как единое целое

Ясно, что схематические диаграммы не дают нам представления о том, что происходит около середины выборки, и если мы интересуемся именно этим, то придется прибегать к каким-то другим средствам. Как исправить положение? Нужно разбить 15 измерений на две выборки, и построить для них две схематические диаграммы (рис. 12).

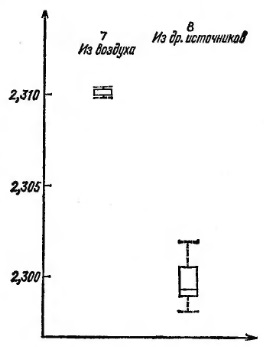


Рис. 12. Те же данные, иначе сгруппированные

Измерения Рэлея дают нам повод еще раз напомнить об одном старом положении: цели могут быть разными, и вид графика должен соответствовать цели. Мы можем пытаться выявить, либо общее поведение, либо подробности. Почти всегда нужно выбирать что-то одно. И еще, **мы фактически не видим своих результатов до тех пор, пока не представим их в графической наглядной форме.**

### Глава 3. ПРОСТЫЕ ПРЕОБРАЗОВАНИЯ

В большинстве случаев естественная схема записи позволяет получить хорошее представление о структуре чисел в выборке. Однако существует довольно много исключений. Если трудно понять числа в том виде, в каком они были первоначально записаны, нужно перейти к другому виду (не потеряв при этом ничего ценного для нас) — такому, в котором числа воспринимаются легче из-за особенностей человеческого восприятия вообще и имеющегося графического и вычислительного аппарата, в частности.

Положительные числа, которые не слишком близки друг к другу, скорее всего выиграют от преобразования с использованием логарифмов. Другой подходящий способ состоит в извлечении квадратного корня. Иногда представляется полезным вместо чисел использовать их обратные величины. Поскольку при этом меняется порядок следования чисел, используют отрицательные числа. Например, 5 преобразуют в «–1/5».

Нередко естественной измеряемой величиной оказывается время, за которое происходит какое-либо событие. Но, поскольку некоторые события так никогда и не происходят, то выручает использование обратного времени, например, «–1000/t».

Оказывается, что для наших целей преобразования наблюдений роль нулевой степени отлично исполняют логарифмы (рис. 13).

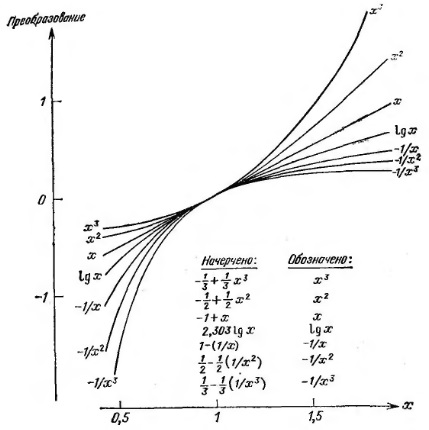


Рис. 13. Функции от *х*

**Мы рассматриваем преобразование данных как орудие анализа, которое позволяет нам лучше воспринимать данные.**

### Глава 4. ЭФФЕКТИВНОЕ СРАВНЕНИЕ, ВКЛЮЧАЯ ВЫБОР ПРЕОБРАЗОВАНИЯ

Существует две причины, по которым преобразованию одного вида отдается предпочтение перед другим:

* симметричность рассеяния значений внутри каждой отдельной выборки;
* согласие в степени рассеяния от выборки к выборке.

Из самих данных обычно бывает трудно извлечь точные указания на то, как их следует анализировать. При особо тщательном выборе методов анализа приходится полагаться на опыт работы с другими группами данных аналогичного содержания.

В обыденной речи постоянно встречаются два вида сравнений: «Билл на голову выше Джима», «Джордж весит вдвое больше своего брата Джека». В каждом из этих утверждений говорится, что нужно сделать с одним человеком, чтобы сравнять его с другим. Первое утверждение основано на знаке плюс и говорит, сколько нужно прибавить к росту Джима, чтобы он стал равен росту Билла. Второе основано на понятии «во сколько раз» и говорит, на что нужно умножить вес Джека, чтобы он стал равен весу Джорджа.

Для многих целей сложение проще умножения. Столетия назад для сведения операции умножения к сложению изобрели логарифмы, чем намного облегчили ручные вычисления. Мы можем, будем и должны использовать их, чтобы избежать сравнений путем умножения. Там, где, по всей видимости, не обойтись без умножения, применение логарифмов позволяет нам сравнивать посредством сложения.

Если мы хотим сравнить несколько выборок одновременно, нам нужно их центрировать примерно на одном уровне. Если первоначально они этому условию не удовлетворяют, придется все-таки добиться этого с помощью введения каких-то поправок. На рис. 14 мы сдвинули схематические диаграммы на различные расстояния. Мы выбрали круглые числа для расстояний, в связи с чем эти три схематические диаграммы не выровнены совершенно точно. Все же они находятся под одним и тем же «микроскопом», и мы ясно видим согласованность в ширине как от сгиба до сгиба, так и от одного крайнего значения до другого.

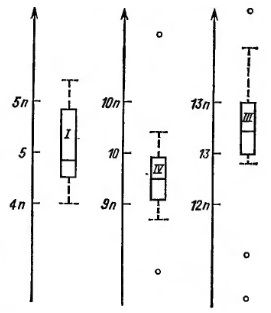


Рис. 14. Выровненные схематические диаграммы способствуют лучшему сравнению выборок

Для характеристики вариации используется понятие *остатка*.

*Остаток = данное значение — сводное значение (например, медиана)*

Мы скоро научимся использовать остатки:

* как ключ к последовательному поэтапному усовершенствованию нашего анализа;
* как ключ к исследованию адекватности анализа.

### Глава 5. ГРАФИКИ ЗАВИСИМОСТИ

После того как мы нанесли данные на график и определили их общее поведение, полезно построить новый график, который подчеркивал бы отклонения, — короче, очень часто бывает полезен график остатков. Здесь и в дальнейшем мы постоянно будем расчленять имеющиеся данные согласно ключевому соотношению:

*данное = аппроксимация ПЛЮС остаток.*

На протяжении всей книги нас будут интересовать графики, чтобы их рассматривать, а не чтобы находить из них числа. Наши графики — средство зрительного представления данных, а не хранилище количественной информации. Необходимость видеть поведение данных невозможно преувеличить. Мы должны сделать менее заметным или совсем убрать все, что может нам помешать видеть, что, собственно, происходит на графике. На каждой оси мы используем, лишь четыре-пять меток, и лишь два-три числа.

Больше всего мы будем связаны с графическим вычитанием при решении основного соотношения:

*данные = неполное описание ПЛЮС остатки = аппроксимация ПЛЮС остатки*

для нахождения из него остатков:

*остатки = данные МИНУС неполное описание* (рис. 15).

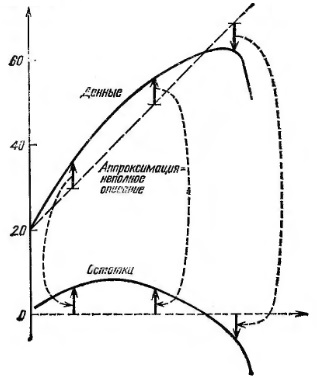


Рис. 15. Образование остатков путем вычитания неполного описания из данных

Если мы вычтем неудачную прямую и обнаружим это, взглянув на остатки, мы всегда сможем исправить дело еще одним вычитанием. График наших первых остатков будет отличаться от графика наилучших наличием некоторого наклона.

Рассмотрим пример с населением США по данным переписей через каждые 10 лет с 1800 по 1950 г. (рис. 16). Обращают на себя внимание два обстоятельства:

* более ранние годы были годами ускоренного роста, возможно на какой-то определенный постоянный процент за год;
* в дальнейшем каждые десять лет количество населения увеличивалось примерно на одно и то же число.

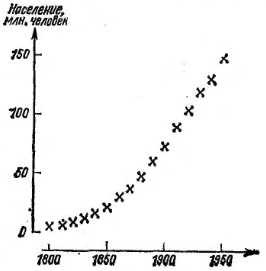


Рис. 16. Население США (в миллионах человек, масштаб линейный)

На рис. 17а показан тот же график, на котором нанесена еще прямая для сравнения. Её формула: у = 35 + 1,4 (*х* – 1870). Если вычесть ее из первоначальных данных получим остатки (рис. 17б). Аналогично можно поступить и для периода с 1800 по 1870 гг. Только теперь нужно строить логарифмические графики.

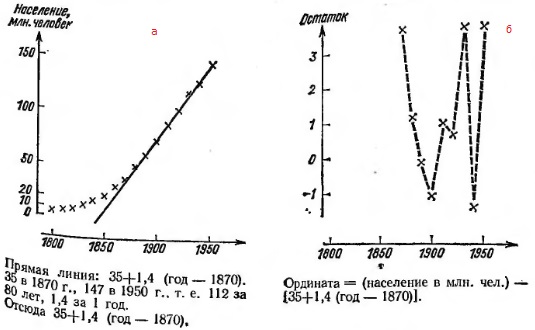


Рис. 17. Население США: (а) в миллионах человек с прямой сравнения; (б) остатки, равные отклонению от указанной прямой (масштаб различный)

Мы увидели конкретные примеры весьма общих принципов, а именно:

* приближенная линеаризация графика с помощью выбора подходящего масштаба всегда позволяет гораздо яснее видеть локальные или характерные особенности;
* выравнивание данных путем вычитания неполного описания всегда позволяет нам растянуть ось ординат я внимательнее всмотреться в оставшиеся особенности почти любого вида.

Поскольку оценка наклона нам будет нужна для выравнивания данных, мы хорошо сделаем, если приложим все усилия для получения возможно лучшей его оценки. Первый шаг в изучении искусства построения графиков — максимально возможное выпрямление зависимости или облака точек. В поисках прямолинейности, по-видимому, разумно вместо *у* наносить на график у2, √у̅, log y, –1/у и т.п. Разумно с той же целью вместо *х* использовать х2, √х̅, log х, –1/х и т.п. После выпрямления графика обычно бывает очень выгодно выровнять его, обычно с помощью изображения остатков (по отношению к неполному описанию — прямой, которую мы, возможно, еще и не нанесли на график).

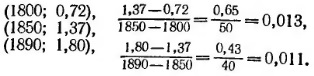
### Глава 6. ВЫПРЯМЛЕНИЕ ГРАФИКОВ (с помощью трех точек)

Если совокупность данных выглядит искривленной, можно проверить это, выбрав три характерные точки. Например, для прироста населения США в первой половине рассмотренного периода можно было бы выбрать точки, соответствующие 1800, 1850 и 1890 гг. Вот эти три точки: (1800; 5,3); (1850; 23,2); (1890; 62,9). Проверить, лежат ли какие-либо три точки на одной прямой, можно путем сравнения угловых коэффициентов прямых, проходящих соответственно через первые две и последние две точки. Имеем:



Эти значения сильно различаются. Второй коэффициент наклона больше, поэтому кривая выгнута книзу (выпуклая вниз). Сказать, что у кривой «выпуклость вверх», означает, что из трех точек на этой кривой средняя находится выше прямой, соединяющей две другие. Аналогичным образом «выпуклость вниз» означает, что средняя точка находится ниже прямой, соединяющей две другие. Если у наших трех точек выпуклость ВВЕРХ, прямолинейность нужно искать ВЫШЕ по лестнице (см. рис. 13); если у наших трех точек выпуклость ВНИЗ, прямолинейность нужно искать НИЖЕ по лестнице. Для преобразования *у* правило выглядит очень простым: двигайтесь по лестнице в ту сторону, куда указывает выпуклость кривой.

Применение этого правила к населению США в первой половине периода показывает, что мы должны спускаться по лестнице, так как выпуклость кривой направлена вниз. Возьмем log *y*. Мы видим, что log 5,3 = 0,72, log 23,2 = 1,37, log 62,9 = 1,80, откуда получаем три точки и два угловых коэффициента:



Поскольку легче произвести измерения для величин *х* = год – 1800, находим три значения функции у – 0,012 (год — 1800), а именно   
0,72 – 0,012 (0) = 0,72  
1,37—0,012 (50) = 0,77,  
1,80—0,012 (90) =0,72,  
среднее из которых с точностью до двух знаков будет 0,74.

Итоговая функция:

*население (логарифм от миллионов) = 0,74 + 0,012 (год – 1800)*

Аналогичным образом можно преобразовать только *х*. Для населения США получим наилучшую линейную аппроксимацию при странной зависимости *у ~ х7*.

Существует довольно естественный соблазн, особенно сильный в том случае, когда мы получили лишь одну из аппроксимаций, превратить «хорошую аппроксимацию» в принцип «это должно было происходить вот так», или «фундаментальный закон прироста населения». Лишь один пример хорошей аппроксимации сам по себе еще совершенно не дает оснований для таких сильных утверждений. Тот факт, что мы получили две хорошие аппроксимации совершенно различного вида, подчеркивает необходимость избегать столь поспешных выводов. И обратно — хорошей аппроксимацией можно с полным правом воспользоваться для многих целей совершенно независимо от того, является она «фундаментальным законом» или нет.

Не всегда стоит преобразовывать лишь одну из переменных. Иногда приносит успех преобразование сразу обеих, *у* и *х*. Спрямлению может способствовать также изменение начала координат. Подумать о разумном начале отсчета бывает необходимо иногда по отношению к *х*, а иногда по отношению к *у*. Естественный подход к данным такого рода заключается в вычислении

*наблюдение МИНУС фон,*

где «фон» означает константу, выбранную таким образом, чтобы учесть медленно меняющиеся компоненты. Мы должны найти этот «фон» из тех же самых данных, которые мы хотим выровнять.

На рис. 18 показаны четыре возможных случая и естественные шаги продвижения по лестницам для *х* и *у* (можно двигаться по одной из них или по обеим сразу). Преобразования следует применять на остальных точках только тогда, когда они уже эффективно подействовали на три выбранные.

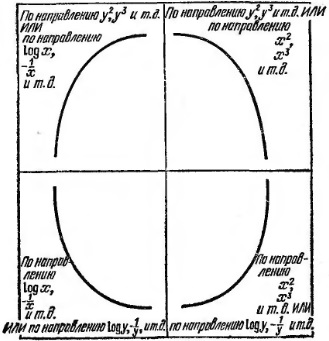


Рис. 18. Как двигаться в отдельности по каждой переменной (четыре случая формы кривых)

### Глава 7. СГЛАЖИВАНИЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

В предыдущих главах особое место было отведено прямым линиям. Но не все на свете — прямая линия. Нам нужны приемы обработки, которые помогли бы видеть, что говорят данные, даже если они не располагаются по прямой. Два основных принципа здесь — это опора на соседние значения и применение нескольких шагов сглаживания. На каждом шаге каждое значение сравнивают с несколькими соседними (нередко лишь с двумя или тремя) и затем соответственно изменяют его. Поскольку мы стремимся получить гладкие кривые, этот процесс называется сглаживанием. Будучи процессом аппроксимации, он приводит к общему соотношению:

*имеющиеся наблюдения = плавная компонента ПЛЮС неровности.*

Наибольший интерес представляют данные, для которых *х* непрерывно возрастает с одинаковым шагом, например, когда мы имеем одно значение каждый год или на каждой миле. В таких ситуациях часто говорят о «временных рядах».

**Медиана по тройкам.** Если нам дана последовательность чисел 4, 7, 9, 3, 4, 11, 12, 1304, 10, 15, 12, 13, 17, мы будем брать по три значения, переставлять их внутри каждой тройки в порядке возрастания и брать их медиану (рис. 19). Поскольку группы значений, от которых мы находим медиану, можно представить себе скользящими вдоль ряда, такие последовательности медиан часто называют скользящими медианами. Аналогичным образом можно сделать повторное сглаживание. Рано или поздно последовательность перестанет изменяться при сглаживании. Сглаженные последовательности ведут себя намного правильнее, чем первоначальные данные (рис. 20).

Excel предоставляет несколько возможностей сглаживания. Подробнее см. [Анализ временных рядов](http://baguzin.ru/wp/?p=6212) и [Прогнозирование на основе экспоненциального сглаживания](http://baguzin.ru/wp/?p=15074).

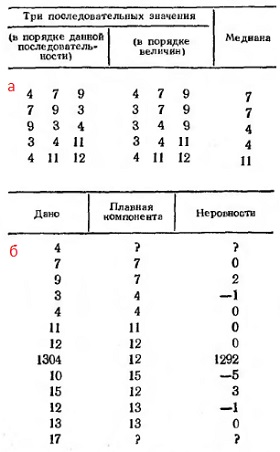


Рис. 19. Вычисление скользящей медианы: (а) подробно для части данных; (б) для всей выборки

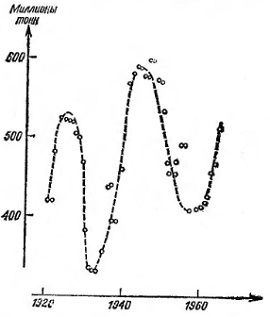


Рис. 20. Сглаженная кривая

### Глава 10. ИСПОЛЬЗОВАНИЕ ДВУХФАКТОРНОГО АНАЛИЗА

Наступило время рассмотреть двухфакторный анализ — как вследствие его важности, так и потому, что он является введением в разнообразные методы исследования. В основе двухфакторной таблицы (таблицы «откликов») лежат:

* один вид откликов;
* два фактора — и каждый из них проявляется в каждом наблюдении.

**Двухфакторная таблица остатков. Анализ «строка-плюс-столбец».** На рис. 21 приведены среднемесячные значения температуры для трех мест в Аризоне.

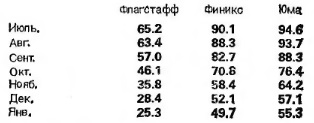


Рис. 21. Среднемесячные температуры в трех городах Аризоны, °F

Определим медиану по каждому месту, и вычтем ее из отдельных значений (рис. 22).

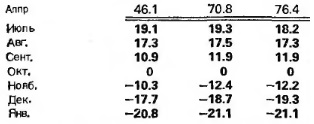


Рис. 22. Значения аппроксимации (медианы) для каждого города и остатки

Теперь определим аппроксимацию (медиану) по каждой строке, и вычтем ее из значений строки (рис. 23).

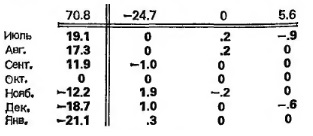


Рис. 23. Значения аппроксимации (медианы) для каждого месяца и остатки

Для рис. 23 мы вводим понятие «эффект». Число –24,7 представляет собой эффект столбца, а число 19,1 — эффект строки. Эффект показывает, как проявляется фактор или множество факторов в каждой из наблюденных величин. Если проявляющаяся часть фактора больше, чем то, что остается, то легче разглядеть и понять, что происходит с данными. Число, которое было вычтено из всех без исключения данных (здесь 70,8), называем «общее». Оно есть проявление всех факторов, общих для всех данных. Таким образом, для величин на рис. 23 справедлива формула:



Это и есть схема конкретного анализа «строка-ПЛЮС-столбец». Мы возвращаемся к нашей старой уловке — попытаться найти простое частичное описание — частичное описание, которое легче воспринимается — частичное описание, вычитание которого даст нам возможность глубже взглянуть на то, что еще не было описано.

Что нового мы сможем узнать благодаря полному двухфакторному анализу? Самый большой остаток, равный 1,9, мал по сравнению с величиной изменения эффекта от пункта к пункту и от месяца к месяцу. Во Флагстаффе приблизительно на 25°F прохладнее, чем в Финиксе, в то время как в Юме на 5–6°F теплее, чем в Финиксе. Последовательность эффектов месяцев монотонно убывает от месяца к месяцу, сначала медленно, затем быстро, затем снова медленно. Это похоже на симметрию относительно октября (такую закономерность я ранее наблюдал на примере продолжительности дня; см. [Управляйте по тенденциям, а не по событиям](http://baguzin.ru/wp/?p=2123). – *Прим. Багузина*); Мы сняли обе завесы — эффект сезона и эффект места. После этого мы смогли увидеть довольно многое из того, что ранее оставалось незамеченным.

На рис. 24 приведена **двухфакторная диаграмма**. Хотя основное на этом рисунке — это аппроксимация, мы не должны пренебрегать остатками. В четырех точках мы нарисовали короткие вертикальные черточки. Длины этих черточек равны величинам соответствующих остатков, так что координаты вторых концов представляют не значения аппроксимации, а

*Данные = аппроксимация ПЛЮС остаток.*

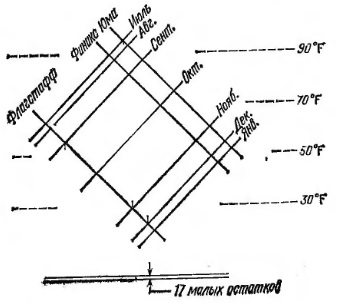


Рис. 24. Двухфакторная диаграмма

Заметим также, что свойство этой или любой другой двухфакторной диаграммы — «шкала лишь в одном направлении», задающими вертикальный размер, т.е. пунктирными горизонтальными линиями, проведенными по бокам картинки, и отсутствием какого-либо размера в горизонтальном направлении.

О возможностях Excel см. [Двухфакторный дисперсионный анализ](http://baguzin.ru/wp/?p=5919). Любопытно, что некоторые формулы, использованные в этой заметке, носят имя Тьюки

Дальнейшее изложение, на мой взгляд, стало совсем сложным…