


# ИНСТРУКЦИЯ ПО ВЕДЕНИЮ НАУЧНОГО ИССЛЕДОВАНИЯ

Статистический анализ  
количественных данных





Не так давно «посчитать» статистику означало умение работать с формулами. Владение статистикой в XXI веке – это ваша способность использовать статистическое программное обеспечение, а не просто способность вычислять средние значения, моды, медианы и стандартные отклонения, а потом искать значения  $p$  на обороте книги.

Безусловно, эти программы требуют базового понимания языка и логики статистики. Но сосредоточиться на формулах – все равно что учиться ездить на велосипеде, учась его делать. Конечно, способность выполнять механику может помочь вам понять логику, но если ваша основная цель, как и у большинства студентов, занимающихся проектом, заключается в том, чтобы иметь возможность провести сравнительно простой статистический анализ, вам главное будет нужно знать:

1. Как управлять вашими данными;
2. Природу переменных;
3. Роль и функции как описательной, так и выводной статистики;
4. Надлежащее использование статистических тестов;
5. Эффективное представление данных.

## Управление данными и определение переменных

Существует два важных шага между необработанными количественными данными, которые вам удалось собрать, и вашей способностью проводить статистический анализ. Это эффективно управлять вашими данными, так что вы можете создать полную базу данных, и определять ваши данные как переменные по отношению как к причине и следствию, так и к шкалам измерения.

### Управление данными

Данные могут собираться довольно быстро, и вы можете быть удивлены тем количеством, которое вам удалось собрать. Задача заключается в применении строгого и систематического подхода к управлению данными, который позволит вам создавать набор данных, которыми можно управлять и использовать на протяжении всего процесса анализа. Есть пять шагов, которые я считаю важными для эффективного управления вашими данными.

1. Ознакомьтесь с соответствующим программным обеспечением
  - Статистика IBM SPSS – комплексная и удобная для пользователя ([www.spss.com](http://www.spss.com));
  - SAS – часто институциональный стандарт, но некоторые считают, что он не такой удобный для пользователя, как SPSS ([www.sas.com](http://www.sas.com));
  - Minitab – более вводный, удобный для учащихся и небольших наборов данных ([www.minitab.com](http://www.minitab.com));
  - Excel – хотя и не является специализированной статистической программой, она может справиться с основами и легко доступна на большинстве ПК (продукт Microsoft Office);
  - R – свободная программная среда для статистических вычислений и графики ([www.r-project.org](http://www.r-project.org)).
2. Ведите учет ваших данных
3. Проверьте ваши данные на любые потенциальные проблемы
4. Введите данные – на самом деле ввод данных происходит в два этапа. Во-первых, определение ваших переменных. Второй шаг – систематически вводить ваши данные в базу данных.
5. Очистите данные – очистка данных включает в себя их прочесывание, чтобы убедиться, что все ошибки ввода обнаружены, и что набор данных выглядит по порядку.

### Переменные – причина и следствие

Основной способ разделения переменных – это на переменные причины и следствия. Это означает способность четко определять и различать ваши зависимые и независимые переменные.

**Переменные** – это концепции, имеющие более одного значения; переменные могут быть «жесткими» (например, пол, рост, доход) или «мягкими» (например, чувство собственного достоинства, ценность, политические убеждения).

**Зависимые переменные** – это то, что вы пытаетесь изучить или что вы пытаетесь измерить. Например, вам может быть интересно узнать, какие факторы вызывают хронические головные боли, сильный поток доходов или уровни успеваемости в средней школе – головные боли, доход и достижения будут зависимыми переменными.



**Независимые переменные** – те переменные, которые могут оказывать влияние на вещи, которые вы пытаетесь понять. Например, чтение может вызвать головную боль; пол может играть роль в определении дохода; родительское влияние может повлиять на уровень достижений. Независимыми переменными здесь являются чтение, пол и влияние родителей.

Хотя понимание теоретических различий между зависимыми и независимыми переменными не слишком сложно, возможность легко идентифицировать каждый тип приходит с практикой. Один из способов сделать это – просто спросить, что зависит от того, что: достижение зависит от влияния родителей; доход зависит от пола. Как я хотел бы сказать своим студентам, не имеет смысла говорить, что пол зависит от дохода, если только вы не откладываете деньги на операцию по смене пола!

## Переменные – шкалы измерений

Шкалы измерений относятся к характеру различий, которые вы пытаетесь зафиксировать в определенной переменной. Существуют четыре основные шкалы измерения, которые становятся соответственно более точными: номинальная, порядковая, интервальная и относительная. Точность каждого напрямую связана со статистическими тестами, которые могут быть выполнены на них. Чем точнее шкала измерения, тем более сложный статистический анализ вы можете сделать.

Номинальная Шкала измерения, в которой числа произвольно назначаются для представления категорий. Поскольку они произвольны и не имеют числового значения, их нельзя использовать для выполнения математических расчетов. Например, в случае пола вы бы использовали одно число для женщин, скажем, 1, а другое для мужчин, 2. В нашем примере, переменная «планы после выпуска» также является номинальной, с произвольными числовыми значениями назначается как 1 = профессионально-техническое обучение, 2 = университет, 3 = рабочая сила, 4 = выезд за границу, 5 = нерешенные и 6 = другие. При номинальном измерении коды не должны пересекаться (они должны быть взаимоисключающими) и вместе должны охватывать все возможности (быть коллективно исчерпывающими). Основная функция номинальных данных состоит в том, чтобы позволить исследователям подсчитывать ответы, чтобы понять распределение населения.

Порядковая Шкала измерения, которая упорядочивает категории каким-либо значимым образом. Однако величины различий не указаны. Например, социально-экономический статус можно классифицировать как низший, средний или высший класс. Низший класс может означать меньший статус, чем два других класса, но величина разницы не определена. Другими примерами являются авиаперелеты (экономичный, бизнес, первый класс) и предметы, где респондентов просят ранжировать выбранные варианты (самые большие экологические проблемы, стоящие перед развитыми странами). Шкалы Лайкерта, в которых респондентов просят выбрать ответ по балльной шкале (например, «мне нравится ходить на работу»: 1 = категорически не согласен, 2 = не согласен, 3 = нейтрально, 4 = согласен, 5 = полностью согласен), являются порядковыми, поскольку точная разница в величине не может быть определена. Однако многие исследователи рассматривают шкалы Лайкерта как интервалы, поскольку это позволяет им выполнять более точные статистические тесты. В большинстве небольших исследований это обычно не считается проблематичным.

	Номинальная	Порядковая	Интервальная	Относительная
Классифицировать	**	**	**	**
Упорядочить		**	**	**
Равноудаленные единицы			**	**
Абсолютный ноль				**

## Измерение центральной тенденции

Один из самых основных вопросов, которые вы можете задать относительно своих данных, касается центральной тенденции. Например: «Каков был средний балл по тесту?» «Большинство людей склоняются влево или вправо по вопросу об аборте?» Или «Что большинство людей считает главной проблемой нашей системы здравоохранения?» В статистике, есть три способа измерения центральной тенденции: среднее



значение, медиана и мода – и приведенные выше примеры вопросов соответственно относятся к этим трем показателям. Хотя показатели центральной тенденции можно рассчитать вручную, все статистические программы могут автоматически рассчитывать эти цифры.

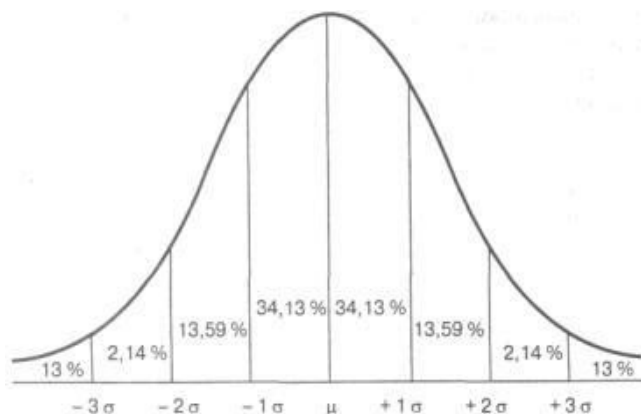
## Измерение дисперсии

Хотя показатели центральной тенденции являются стандартной и весьма полезной формой описания и упрощения данных, их необходимо дополнять информацией о различии ответов. Предположим, у вас была группа студентов с IQ 100, 100, 95 и 105 и другая группа студентов с IQ 60, 140, 65 и 135. Тогда центральной тенденцией, в данном случае средней для обеих групп будет 100. Однако разброс вокруг среднего значения потребует от вас разработки учебного плана и совершенно разных занятий с каждой группой. Есть несколько способов понять разброс, которые подходят для разных типов переменных. Как и в случае с центральной тенденцией, статистические программы могут автоматически генерировать эти цифры.

Центральная тенденция измерения, которые указывают на середину или центр распределения: среднее значение, медиана и мода.

## Измерение формы данных

Чтобы полностью понять набор данных, необходимо рассмотреть центральную тенденцию и разброс в свете формы данных или того, как данные распределяются. Как показано на рисунке, нормальная кривая имеет «колоколообразную форму»; распределение данных симметрично, причем среднее значение, медиана и мода совпадают в самой высокой точке кривой. Если распределение данных не симметрично, оно считается искаженным. В искаженных данных среднее, медиана и мода попадают в разные точки.



Куртоз характеризует, насколько пиковое или плоское распределение по сравнению с «нормальным». Положительный эксцесс указывает на относительно пиковое распределение, в то время как негативный эксцесс указывает на более плоское распределение.

Значение в понимании формы распределения заключается в статистических выводах, которые можно сделать. Как показано на рисунке, нормальное распределение подчиняется определенному набору правил, касающихся значимости стандартного отклонения, а именно:

- 68,3% случаев будут попадать в одно стандартное отклонение от среднего значения;
- 95,4% случаев попадут в два стандартных отклонения от среднего значения;
- 99,7% случаев будут находиться в пределах трех стандартных отклонений от среднего.

Таким образом, если бы у нас была нормальная кривая для данных с вашей выборки, относящихся к «возрасту участников» (среднее = 12,11, SD = 2,22;), 68,3% участников были бы в возрасте от 9,89 до 14,33 (12,11 - 2,22 и 12,11 + 2,22).

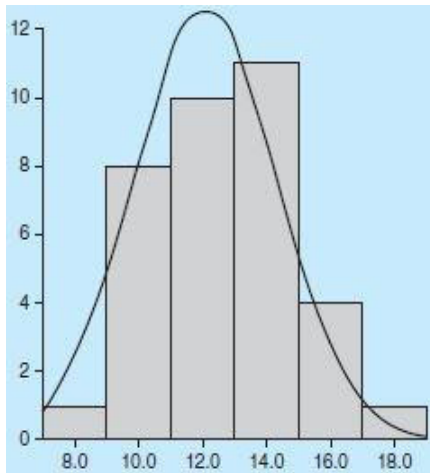


Таблица показывает фактическую кривую, асимметрию и эксцесс нашего выборочного набора данных.

Эти правила нормальной кривой позволяют использовать довольно мощные статистические тесты и обычно используются с интервальными и относительными данными (иногда их называют параметрическими тестами). Для данных, которые не соответствуют допущениям нормальной кривой (номинальные и порядковые данные), исследователь должен использовать непараметрические статистические тесты для выполнения выводов.