Элементы теории статистической обработки данных

Основные определения и понятия.

Основные определения

- **Переменная** – величина, которую можно измерять, контролировать, варьировать в статистических исследованиях.
- **Эмпирические исследования** – нахождение зависимостей (корреляций) между некоторыми переменными.

1. исследование корреляций - такой вид исследования, когда зависимости троятся на основании уже имеющихся фактов, или переменных, получаемых в ходе эксперимента, на которые экспериментатор не оказывает воздействия;
2. экспериментальные исследования предполагают варьирование некоторыми переменными и измерение воздействия этих изменений на другие переменные.

- **Зависимые и независимые переменные**. Независимыми переменными называются переменные, которые варьируются исследователем, тогда как зависимые переменные - это переменные, которые измеряются или регистрируются.
- Шкалы измерений. В каждом измерении присутствует некоторая ошибка, определяющая границы "количества информации", которое можно получить в данном измерении в соответствии с измерительной шкалой. Различают следующие типы шкал:

a) номинальная,

Типичные примеры номинальных переменных - пол, национальность, цвет, город и т.д.

b) порядковая (ординальная),

Типичный пример порядковой переменной - социоэкономический статус семьи.

c) Интервальная,

Например, температура, измеренная в градусах Фаренгейта или Цельсия, образует интервальную шкалу.

d) относительная (шкала отношения).

Типичными примерами шкал отношений являются измерения времени или пространства. Например, температура по Кельвину образует шкалу отношения, и вы можете не только утверждать, что температура 200 градусов выше, чем 100 градусов, но и что она вдвое выше. Интервальные шкалы (например, шкала Цельсия) не обладают данным свойством шкалы отношения.

- **Связи между переменными**. Независимо от типа, две или более переменных связаны (зависимы) между собой, если наблюдаемые значения этих переменных распределены согласованным образом.

Другими словами, мы говорим, что переменные зависимы, если их значения систематическим образом согласованы друг с другом в имеющихся у нас наблюдениях.

Основные статистические величины

- **Описательные статистики**

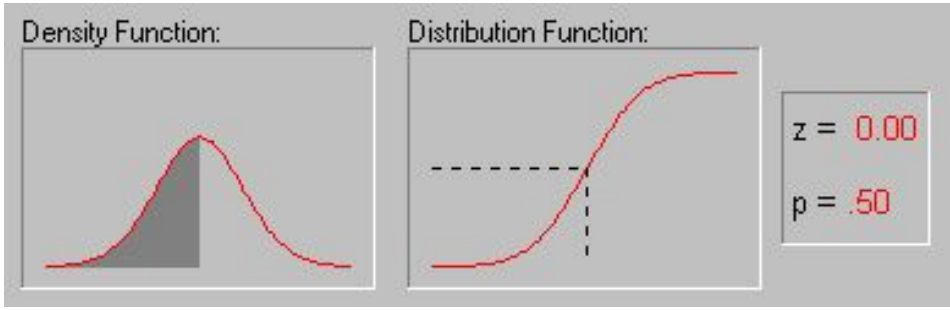
"Истинное" среднее и доверительный интервал. Вероятно, большинство из вас использовало такую важную описательную статистику, как среднее. Среднее - очень информативная мера "центрального положения" наблюдаемой переменной, особенно если сообщается ее доверительный интервал. *Доверительный интервал* для среднего представляет интервал значений вокруг оценки, где с данным уровнем доверия, находится "истинное" (неизвестное) среднее.

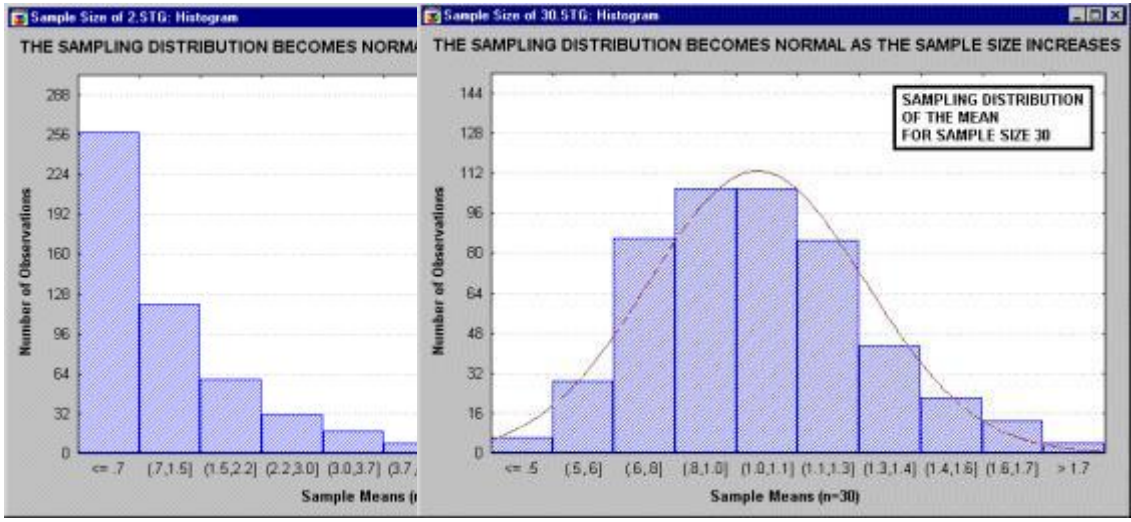
Пример. Среднее значение выборки равно 23, а нижняя и верхняя границы доверительного интервала с уровнем *p*=.95 равны 19 и 27 соответственно. Отсюда следует, что с вероятностью 95% интервал с границами 19 и 27 включает в себя среднее значение.

При установлении большего уровня доверия, интервал становится шире, поэтому возрастает вероятность, с которой он "накрывает" неизвестное среднее, и наоборот. Вычисление доверительных интервалов основывается на предположении нормальности наблюдаемых величин. Увеличение разброса наблюдаемых значений уменьшает надежность оценки.

Нормальное распределение

Форма распределения; нормальность. Важным способом "описания" переменной является форма ее распределения, которая показывает, с какой частотой значения переменной попадают в определенные интервалы. Многие наблюдаемые переменные имеют нормальное распределение, поэтому большинство статистических исследований и расчетов строится на основании нормального распределения изучаемых переменных.



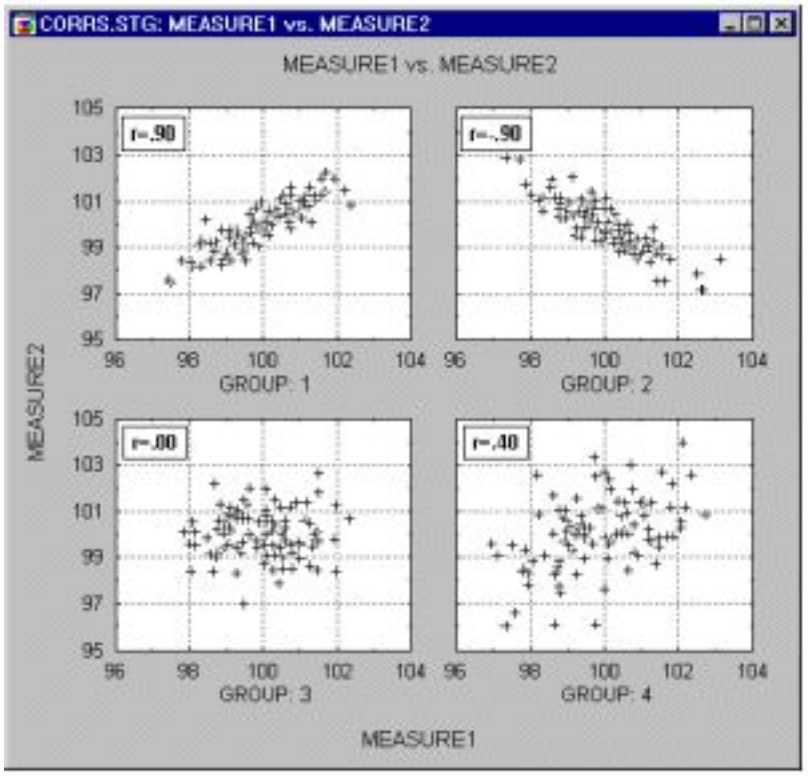


Корреляции

Определение корреляции. Корреляция представляет собой меру зависимости переменных. Наиболее часто используемый коэффициент корреляции *Пирсона r* называется также *линейной* корреляцией, т.к. измеряет степень линейных связей между переменными.

Коэффициенты корреляции изменяются в пределах от -1.00 до +1.00. Обратите внимание на крайние значения коэффициента корреляции. Значение -1.00 означает, что переменные имеют строгую *отрицательную* корреляцию. Значение +1.00 означает, что переменные имеют строгую *положительную* корреляцию.

Отметим, что значение 0.00 означает отсутствие корреляции.



Простая линейная корреляция (Пирсона r)

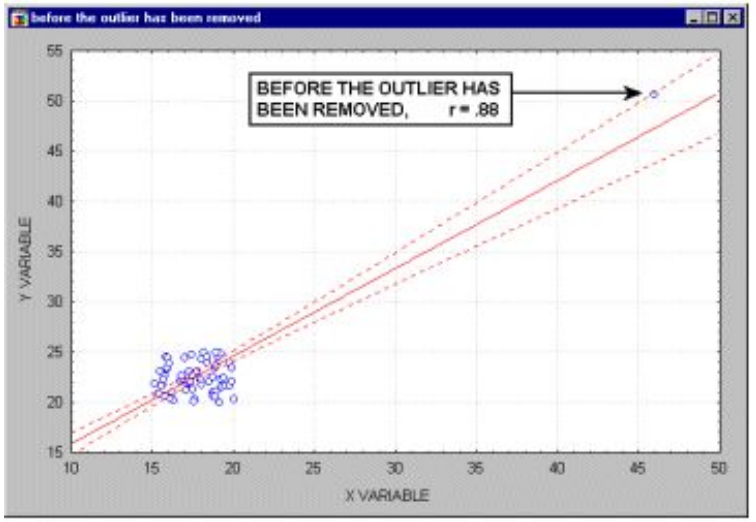
Корреляция Пирсона (далее называемая просто *корреляцией*) предполагает, что две рассматриваемые переменные измерены, по крайней мере, в интервальной шкале. Она определяет степень, с которой значения двух переменных "пропорциональны" друг другу. Важно, что значение коэффициента корреляции не зависит от масштаба измерения. Например, корреляция между ростом и весом будет одной и той же, независимо от того, проводились измерения в *дюймах* и *фунтах* или в *сантиметрах* и *килограммах*. *Пропорциональность* означает просто *линейную зависимость*. Корреляция высокая, если на графике зависимость "можно представить" прямой линией (с положительным или отрицательным углом наклона).

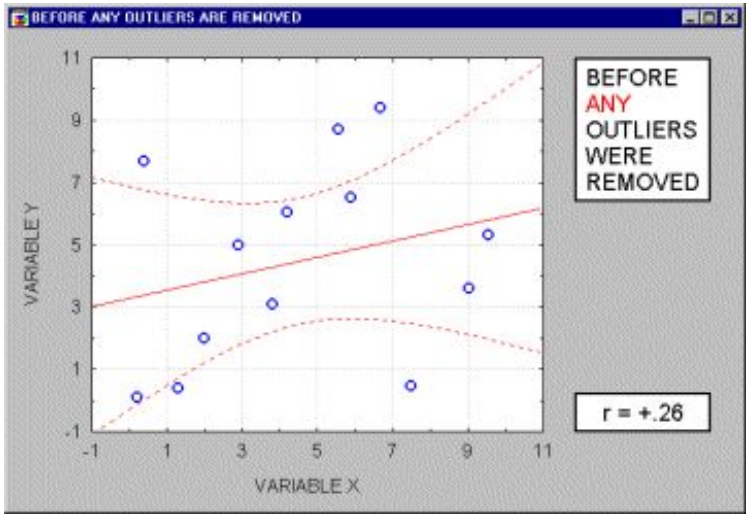
Проведенная прямая называется *прямой регрессии* или прямой, построенной *методом наименьших квадратов*. Последний термин связан с тем, что сумма *квадратов* расстояний (вычисленных по оси Y) от наблюдаемых точек до прямой является минимальной.

Коэффициент корреляции Пирсона (*r*) представляет собой меру линейной зависимости двух переменных. Если возвести его в квадрат, то полученное значение коэффициента детерминации представляет долю вариации, общую для двух переменных (иными словами, "степень" зависимости или связанности двух переменных). При этом чтобы правильно оценить зависимость между переменными, нужно знать как "величину" корреляции, так и ее *значимость*.

Уровень значимости и выбросы

- **Значимость корреляций**. Уровень значимости, вычисленный для каждой корреляции, представляет собой главный источник информации о надежности корреляции. Критерий значимости основывается на предположении, что распределение остатков (т.е. отклонений наблюдений от регрессионной прямой) для зависимой переменной *y* является нормальным (с постоянной дисперсией для всех значений независимой переменной *x*).
- **Выбросы**. По определению, выбросы являются нетипичными, резко выделяющимися наблюдениями. Так как при построении прямой регрессии используется сумма *квадратов* расстояний наблюдаемых точек до прямой, то выбросы могут существенно повлиять на наклон прямой и, следовательно, на значение коэффициента корреляции.



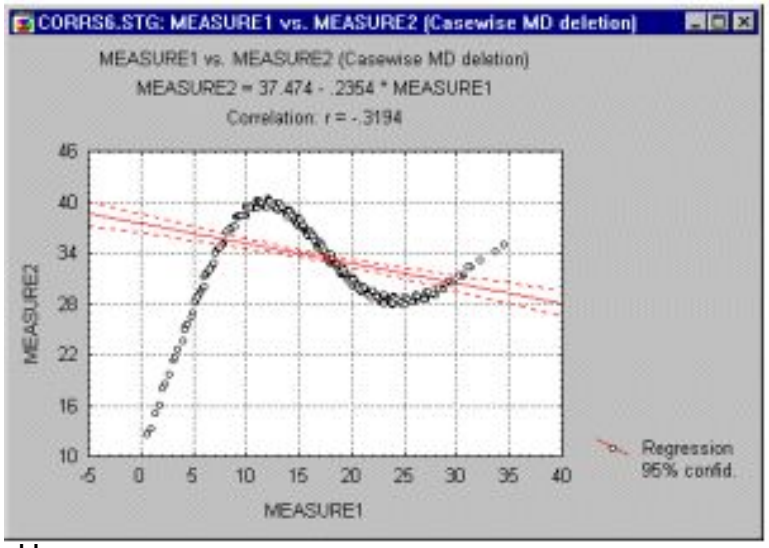


Нелинейные зависимости между переменными

Например, график справа показывает сильную корреляцию между двумя переменными, которую невозможно хорошо описать с помощью линейной функции.

Существуют два наиболее точных метода исследования нелинейных зависимостей. Оба они непросты и требуют хорошего навыка "экспериментирования" с данными. Эти методы состоят в следующем:

- Необходимо найти и математически описать функцию, которая наилучшим способом согласуется с данными. После того, как функция определена, проверяют "степень согласия" ее с исходными данными.
- Данные, разбивают некоторой переменной на группы (например, на 4 или 5 групп). Эту переменную определяют как группирующую переменную, а затем применяют дисперсионный анализ.



Что делать, если корреляция сильная, однако зависимость явно нелинейная? К сожалению, не существует простого ответа на данный вопрос, так как не имеется естественного обобщения коэффициента корреляции *Пирсона r* на случай нелинейных зависимостей. Однако, если кривая монотонна (монотонно возрастает или, напротив, монотонно убывает), то можно преобразовать одну или обе переменные, чтобы сделать зависимость линейной, а затем уже вычислить корреляцию между преобразованными величинами. Для этого часто используется логарифмическое преобразование.