Статистический анализ данных. Первые шаги

Лекция 10

Это «первые шаги», а не «введение» или «основные понятия», потому что статистический анализ основывается на теории вероятностей и математической статистике, которую вы еще не проходили. В этой лекции положения статистического анализа поясняются не с помощью математической теории, а на основе здравого смысла.

Эти «первые шаги» помогут понять смысл формул для оценок параметров линейно й зависимости по методу наименьших квадратов.

Понятие выборки

- **Выборка – это последовательность наблюдений.**

Это могут быть наблюдения любой природы: некоторой физической величины (температуры, давления, напряжения) или экономические данные (стоимость какого либо объекта или заработная плата), или медицинские и т. д.

Наблюдения могут проводиться на одним объектом в последовательные моменты времени или в один момент времени над несколькими объектами.

Представим эти наблюдения как массив чисел из n элементов:

$x_1, x_2, \ldots, x_n$

n называется объемом или длиной выборки.

Значение n может быть весьма велико.

Как описать свойства выборки?

Составить о ней общее представление?

По каким параметрам можно сравнить две выборки, описывающие объекты или явления одинаковой природы?

Например, имеются оценки двух студенческих групп по какому-либо предмету. Как понять, какая группа лучше учится?

А если оценки не двух групп, а двух факультетов?

Характеристики выборки

- **Среднее значение:**

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Это наиболее распространенная характеристика центра выборки.

Обычно, когда говорят «средний», подразумевают «типичный», хотя это не всегда правильно. Например, если оценки такие: 3, 5, 3, 5, 3, 5, то среднее равно 4, хотя оценку 4 не получил ни один студент. Еще один пример нетипичности среднего значения – это «средняя по госпиталю температура». Среднее значение является хорошей характеристикой выборки, когда наблюдения более или менее равномерно заполняют интервал от x_{min} (минимального значения среди наблюдений) до x_{max} (максимального значения среди наблюдений).

Значения x . и x также являются характеристиками выборки.

Характеристики выборки

- **Медиана** (от лат. *mediāna* — середина).

Медиана — это такое число, что половина из элементов выборки больше него, а другая половина меньше.

Медиану можно найти, упорядочив элементы выборки по возрастанию или убыванию и взяв средний элемент. Например, выборка {11, 9, 3, 5, 5} после упорядочивания превращается в {3, 5, 5, 9, 11} и её медианой является число 5.

Если в выборке чётное число элементов, медиана может быть не определена однозначно: для числовых данных чаще всего используют полусумму двух соседних значений (то есть медиану набора {1, 3, 5, 7} принимают равной 4), хотя в соответствии с определением можно было взять, например, 4.5.

Медиана является важной характеристикой выборки и, так же как среднее значение, может быть использована в качестве центра выборки, в случаях сильной «неравномерности» выборки.

Характеристики выборки

- Предположим, что в одной комнате оказалось 19 бедняков и один миллионер. У каждого бедняка есть 5 рублей, а у миллионера — 1 млн рублей. В сумме получается 1 000 095рублей. Если мы разделим деньги равными долями на 20 человек, то получим 50 004,75 ₽. Это будет **среднее арифметическое** значение суммы денег, которая была у всех 20 человек в этой комнате. Такой суммы нет ни одного человека в комнате.
- Медиана в этом случае будет равна 5 рублям (полусумма десятого и одиннадцатого, *срединных* значений упорядоченного ряда). Можно интерпретировать это следующим образом. Разделив всю компанию на две равные группы по 10 человек, мы можем утверждать, что в первой группе у каждого не больше 5 рублей, во второй же — не меньше 5 рублей. В общем случае можно сказать, что медиана — это то, сколько принёс с собой «средний» человек. Наоборот, среднее арифметическое — неподходящая характеристика, так как оно значительно превышает сумму наличных, имеющуюся у среднего человека. Но рассматриваемая выборка существенно неоднородна, то есть содержит существенно различающиеся значения.

Характеристики выборки

- **Мода** — значение во множестве наблюдений, которое встречается наиболее часто. Таким образом, мода – наиболее типичное значение. Иногда в совокупности встречается более чем одна мода *(например: 6, 2, 6, 6, 8, 9, 9, 9, 0; мода — 6 и 9)*. В этом случае можно сказать, что совокупность мультимодальна.
- Мода как средняя величина употребляется чаще для данных, имеющих нечисловую природу. Среди перечисленных цветов автомобилей — *белый, чёрный, синий металлик, белый, синий металлик, белый* — мода будет равна *белому* цвету. При экспертной оценке с её помощью определяют наиболее популярные типы продукта, что учитывается при прогнозе продаж или планировании их производства.

Характеристики рассеяния, разброса, изменения выборки

- **Размах:** $R=x_{max}-x_{min}$

Не самая лучшая характеристика рассеяния выборки. Например,

Она одинакова для выборок:

-10, 10, -10, -10, -10, 10, 10, -10, 10, -10, 10

1, -1, -2, 2, -5, 5, 10, 1, -10, 3, -4, 6, 4, -6, -7, 8

Характеристики рассеяния, разброса, изменения выборки

Выборочная дисперсия:

$$D_x = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

- Дисперсия характеризует среднее отклонение (разброс, рассеяние, изменение, вариацию) наблюдений относительно их среднего значения.
- Почему суммируются не отклонения, а их квадраты? – Для того, чтобы положительные отклонения не компенсировались отрицательными. Иначе при больших отклонений можно получить маленькую сумму.
- Почему суммируются не модули отклонений, а их квадраты? - Потому, что использование модулей приводит к сложным алгебраическим выражениям. Ведь, например, взять производную от функции $y=|x|$ сложнее, чем от $y=x^2$.
- В теории вероятностей доказывается, что лучшую оценку рассеяниея можно получить, если в знаменателе использовать не n, а п-1. Почему, мы попытаемся понять позже. Но на практике используется и приведенная на этом слайде формула.

Характеристики рассеяния, разброса, изменения выборки

В теории вероятностей доказывается формула:

$$D_x = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

Если использовать обозначения, принятые в лекции 9 (слайд 9), то получим:

$$D_x = \overline{x^2} - \overline{x}^2 \qquad (1)$$

Эта формула нам пригодится для выяснения смысла формул (4) лекции 9 - оценивания коэффициентов линейной зависимости по МНК.

Характеристики рассеяния, разброса, изменения выборки

- Единицы измерения D_x – это единицы измерения x в квадрате.
- Чтобы рассеяние измерялось в тех же единицах, что и x, рассматривается характеристика:

$$s_x = \sqrt{D_x}$$

- s_x называется **выборочным средним квадратичным отклонением или стандартным отклонением.**
- Естественно:

$$D_x = s_x^2 \qquad (2)$$

Зависимость двух выборок

Пусть мы проводим наблюдения так, что в одном наблюдении определяем сразу два параметра *x* и *y* . Например, рост и вес человека. Каждое наблюдение можно изобразить точкой на плоскости в декартовой системе координат. Такая картинка (см. лекцию 9) называется полем корреляции (или полем рассеяния).

Если между *x* и *y* существует зависимость:

$$y=f(x)+\varepsilon,$$

где ε – случайная величина и значения ε не очень велики (что значит «не очень велики», определим потом), то **выборка y зависит от выборки x**.

Если при этом функция $f(x)$ линейная, то существует линейная зависимость выборки *y* от выборки *x*.

Числовые характеристики зависимости двух выборок

- **Выборочная ковариация выборок x и y:**

$$K_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Из теории вероятностей известно: если x и y независимы, то

$$Kx,y=0.$$

Поэтому выборочная ковариация считается мерой зависимости x и y.

Числовые характеристики зависимости двух выборок

В теории вероятностей доказывается формула:

$$K_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

Если использовать обозначения, принятые в лекции 9 (слайд 9), то получим:

$$K_{x,y} = \overline{xy} - \overline{x} \cdot \overline{y} \qquad (3)$$

Эта формула нам пригодится для выяснения смысла формул (4) лекции 9 - оценивания коэффициентов линейной зависимости по МНК.

Числовые характеристики зависимости двух выборок

Величина Kx,y зависит от единиц измерения x и y. Например, пусть x – рост человека, и он измеряется в метрах. Если мы будем измерять x в сантиметрах, то Kx,y увеличится в 100 раз.

Поэтому в качестве меры зависимости выборок x и y используется безразмерная величина:

$$r = \frac{K_{x,y}}{s_x s_y}$$

Величина r называется **выборочным коэффициентом корреляции**.

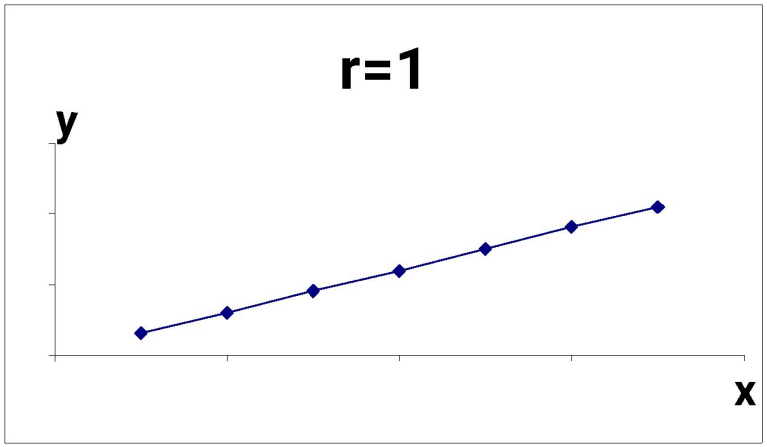
Числовые характеристики зависимости двух выборок

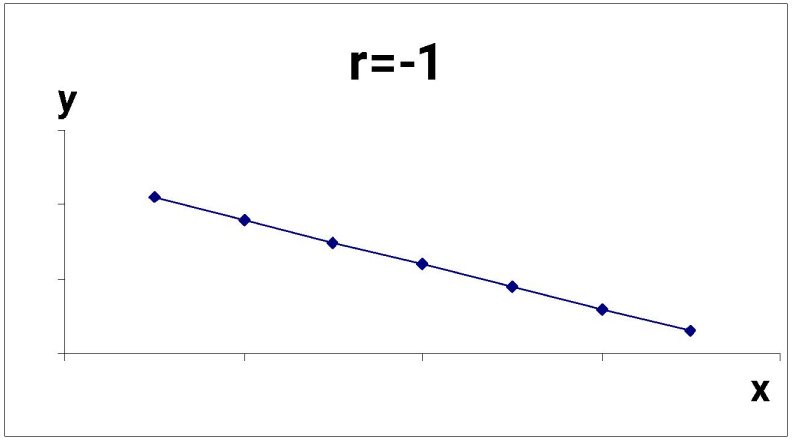
Свойства выборочного коэффициента корреляции (доказываются в теории вероятностей):

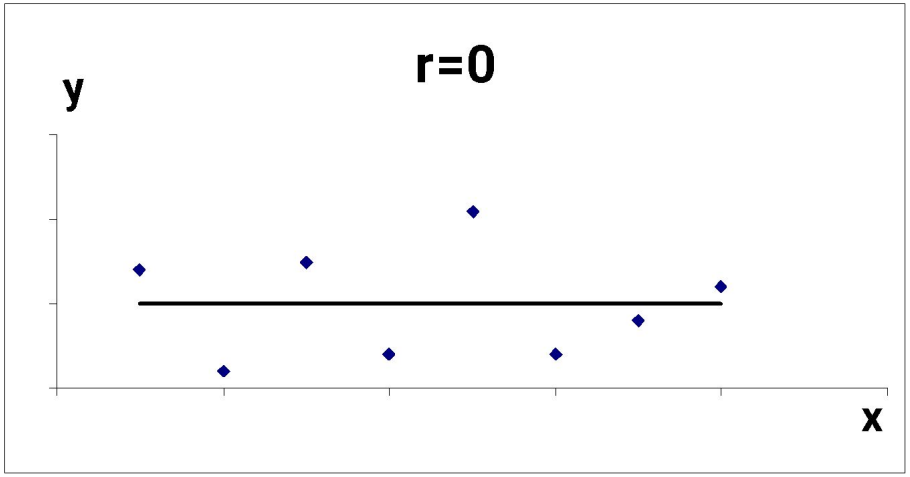
1. $-1 \le r \le 1$. Чем ближе $|r|$ к 1, тем сильнее y зависит от x.
2. При $r = \pm 1$ корреляционная связь - линейная (наблюдения располагаются на прямой)
3. При $r = 0$ связь отсутствует, линия регрессии параллельна оси OX.

Таким образом, выборочный коэффициент корреляции характеризует степень линейной зависимости y от x.

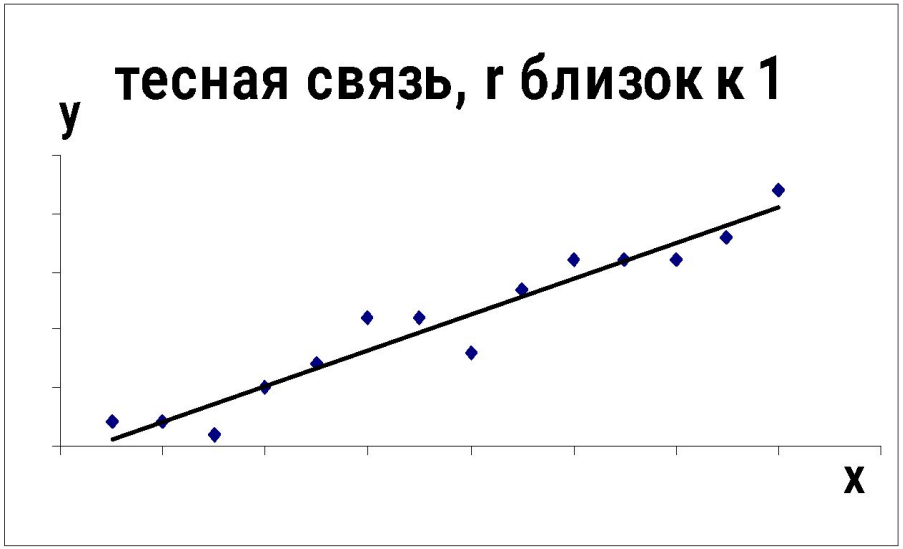
Выборочный коэффициент корреляции характеризует степень линейной зависимости y(x)

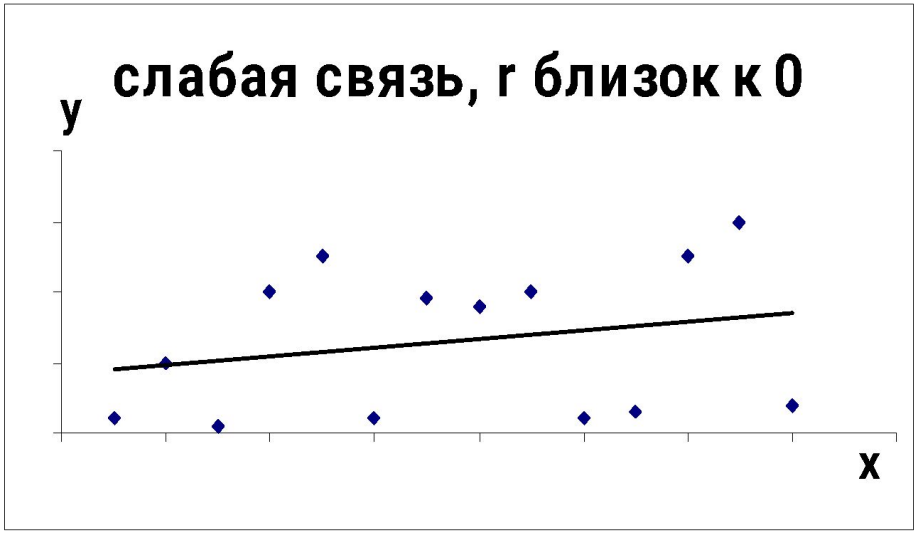






Выборочный коэффициент корреляции характеризует степень линейной зависимости y(x)





Определение параметров функции $y=mx+b$ по наблюдениям ее значений методом наименьших квадратов

Вспомним формулу для оценок параметров m и b по МНК:

$$m = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2},$$

$$b = \overline{y} - m\overline{x}.$$

Учитывая формулы (1) и (3), получаем:

$$m = \frac{K_{x,y}}{D_x} \qquad (4)$$

Таким образом, МНК-оценка коэффициента наклона прямой равна отношению ковариации выборок x и y к дисперсии x.

Определение параметров функции $y=mx+b$ по наблюдениям ее значений методом наименьших квадратов

Из формул (2), (4) и определения выборочного коэффициента корреляции, получим формулы, связывающие значения коэффициента m и коэффициента корреляции:

$$m = \frac{s_y}{s_x} r; \qquad r = \frac{s_x}{s_y} m.$$

Обратите внимание, что между коэффициентом детерминации и коэффициентом корреляции существует связь:

$$R^2=r^2$$

Встроенные функции Matlab для вычисления характеристик выборок

- mean (x) – возвращает среднее значение выборки;
- median(x) – возвращает медиану выборки;
- std(x) - возвращает среднее квадратичное отклонение выборки;
- cov(x,y) - для двух векторов x, y одинаковой длины возвращает матрицу 2×2, на главной диагонали которой стоят дисперсия x (элемент с индексами (1,1) и дисперсия y (элемент с индексами (2,2) , а вне главной диагонали – два одинаковых числа; значение ковариации x и y;
- corrcoef(x,y) – для двух векторов x, y одинаковой длины возвращает матрицу 2×2, на главной диагонали которой стоят единицы, а вне главной диагонали – два одинаковых числа; это и есть значение коэффициента корреляции; при других аргументах эта функция может возвращать попарные коэффициенты корреляции набора векторов;

Встроенные функции Mathcad для вычисления характеристик выборок

- mean (x) – возвращает среднее значение выборки;
- median(x) – возвращает медиану выборки;
- var(x) – возвращает дисперсию (вариацию) выборки;
- stdev((x) – возвращает среднее квадратичное отклонение выборки;
- cvar(x,y) - вычисляет ковариацию выборок x и y;
- corr(x,y) – вычисляет коэффициент корреляции выборок x и y.