Об авторах

Автор презентации:

- Котов Александр Ильич

Оформление презентации:

- Котова Нина Александровна

Регрессионный анализ
Условные обозначения.

$\hat{\bar{x}}$ - выборочное среднее.

$\hat{S}^2$ - выборочная дисперсия.

$\hat{S}_{\_a}^{\ 2}$ - исправленная (несмещенная) выборочная дисперсия.

$\hat{\bar{K}}_{xy}$ - выборочная ковариация.

$\hat{\bar{K}}_{xy\_a}$ - исправленная (несмещенная) выборочная ковариация.

$\hat{r}_{xy}$ - выборочный коэффициент корреляции.

Регрессионный анализ

До сих пор Вы изучали способы обработки выборочной совокупности такой, о которой можно было бы сказать: Выборочная совокупность представлена в виде результатов n экспериментов, в каждом из которых реализовывалось значение какой-то случайной величины X. В результате получалась выборка объема n: $x_1, x_2, x_3, \ldots x_n$

Пусть теперь в эксперименте получается реализация случайного вектора – системы случайных величин (X,Y). В результате n экспериментов получается выборочная совокупность объема n:

Пример №1 (общий вид):

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (x_n, y_n)$$

Пример №1(частный случай):

- **X: -1.75 -1.63 -1.86 -1.78 -1.69 -1.70 -1.72**
- **Y: 2.36 2.42 2.63 2.50 2.68 2.51 2.49**

В этом примере, очевидно, объем выборки n=7.

Заметим, что в выборке нет одинаковых пар, и, даже по отдельности значения X и Y не повторяются. Можно предположить, что мы имеем дело с системой непрерывных случайных величин.

Пример №2: Результаты наблюдений сведены в таблицу:

	i	1	2	3	4	5
j	X / Y	2	4	6	8	10
1	3	8	5	1	0	0
2	6	4	10	5	2	1
3	9	2	6	10	7	2
4	12	0	1	5	12	8

В примере 2, очевидно, объем выборки n, равный сумме частот по всей таблице: n=89 (проверьте!). Случайный вектор практически достоверно представляет собой систему дискретных случайных величин. В не закрашенной части таблицы приведены частоты m_{ij}. Например, значение случайного вектора (X=10,Y=12) – (строка j=4, столбец i=5) повторяется $m_{ij}=8$ раз в проведенных 89 наблюдениях.

Пример №2(продолжение). Вычислим относительные частоты $\omega_{ij} = m_{ij}/n$. В результате получим аналогичную таблицу относительных частот:

	i	1	2	3	4	5
j	X / Y	2	4	6	8	10
1	3	0,0899	0,0562	0,0112	0	0
2	6	0,0449	0,1124	0,0562	0,0225	0,0112
3	9	0,0225	0,0674	0,1124	0,0787	0,0225
4	12	0	0,0112	0,0562	0,1348	0,0898

Сумма относительных частот по всей таблице равна единице. Не путать с таблицей распределения систем дискретных случайных величин!

Пример №3: В случае, если мы имеем дело с системой непрерывных случайных величин, и объем выборки достаточно большой (сотни), то удобно строить интервальную таблицу. На следующем слайде приводится интервальная таблица, полученная обработкой выборки объема n=1423, аналогичной примеру №1. Размахи выборки по случайным величинам X и Y разбиты на ni=10 и nj=12 интервалов соответственно. В таблице указаны середины соответствующих интервалов.

Внимание: $n \neq ni \cdot nj$

Таблица частот для середин интервалов dx=0,2 dy=0,5

	i	1	2	3	4	5	6	7	8	9	10
j	X / Y	0,1	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,9
1	2,5	2	4	7	13	20	13	7	4	2	1
2	3,5	3	5	9	15	20	15	9	5	3	2
3	4,5	5	8	12	17	20	17	12	8	5	4
4	5,5	9	12	15	19	20	19	15	12	9	7
5	6,5	15	17	19	20	20	20	19	17	15	13
6	7,5	20	20	20	20	20	20	20	20	20	20
7	8,5	15	17	19	20	20	20	19	17	15	13
8	9,5	9	12	15	19	20	19	15	12	9	7
9	10,5	5	8	12	17	20	17	12	8	5	4
10	11,5	3	5	9	15	20	15	9	5	3	2
11	12,5	2	4	7	13	20	13	7	4	2	1
12	13,5	2	3	5	12	20	12	5	3	2	1

Вычисление статистических оценок.

Если выборка представлена в форме примера №1, то можно воспользоваться формулами: (1)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \overline{S}_1^{\,2} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad \overline{S}_2^{\,2} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$\overline{K}_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Исправленным (несмещенным) оценкам припишем индекс _a: (2)

$$\overline{S}^{\,2}{}_{1\_a} = \frac{n}{n-1}\overline{S}_1^{\,2} \qquad \overline{S}^{\,2}{}_{2\_a} = \frac{n}{n-1}\overline{S}_2^{\,2} \qquad \overline{K}_{xy\_a} = \frac{n}{n-1}\overline{K}_{xy}$$

$$\overline{S}_{1\_a} = \sqrt{\overline{S}^{\,2}_{1\_a}} \qquad \overline{S}_{2\_a} = \sqrt{\overline{S}^{\,2}_{2\_a}} \qquad \overline{r}_{xy} = \frac{\overline{K}_{xy\_a}}{\sqrt{\overline{S}_{1\_a}^{\,2}\overline{S}_{2\_a}^{\,2}}} = \frac{\overline{K}_{xy}}{\sqrt{\overline{S}_1^{\,2}\overline{S}_2^{\,2}}} = \frac{\overline{K}_{xy}}{\overline{S}_{1\_a}\overline{S}_{2\_a}}$$

Вычисление статистических оценок (продолжение).

По вышеприведенным формулам (1), (2) следует вычислять статистические оценки и в случаях иного представления выборки, если, конечно, информация в форме примера №1 не утеряна. Именно в этой форме наиболее удобно проводить вычисления средствами EXCEL. Указанные формулы легко вводить в ячейки EXCEL. Кроме того, функции СРЗНАЧ, ДИСП, КОВАР избавляют даже и от этой необходимости.

Вычисление статистических оценок (продолжение).

Для выборки, представленной в примере№2 удобно использовать такие формулы :

$$\bar{\bar{x}} = \sum_{i=1}^{ni} x_i \sum_{j=1}^{j=nj} \omega_{ij} \qquad \bar{\bar{y}} = \sum_{j=1}^{nj} y_j \sum_{i=1}^{i=ni} \omega_{ij} \qquad \bar{\bar{S}}_1^{\,2} = \sum_{i=1}^{ni} ((x_i - \bar{\bar{x}})^2 \sum_{j=1}^{j=nj} \omega_{ij})$$

$$\bar{\bar{S}}_2^{\,2} = \sum_{j=1}^{nj} ((y_j - \bar{\bar{y}})^2 \sum_{i=1}^{i=ni} \omega_{ij}) \qquad \bar{\bar{K}}_{xy} = \sum_{i=1}^{i=ni} \sum_{j=1}^{j=nj} \omega_{ij} (x_i - \bar{\bar{x}})(y_j - \bar{\bar{y}}) \qquad (3)$$

Исправленные оценки пересчитываются по выборочным оценкам по тем же формулам (2), что и ранее.

Вычисление статистических оценок (продолжение).

Внимание! Формулы (3), приведенные для примера №2 – это не другие, а **ТЕ ЖЕ САМЫЕ** формулы (1), что приведены для примера №1 Они выводятся одни из других, и дают идентичные результаты!

Вычисление статистических оценок (продолжение).

Для выборки, представленной в примере№3, если утеряна информация формы примера №1, следует использовать формулы (3), (2), предварительно создав такую же таблицу, но для относительных частот. При этом роль значений x_i, y_j играют середины соответствующих интервалов. В этом случае выборочные оценки вычисляются ПРИБЛИЖЕННО!

Формулы (1) и (3) не совпадают в этом случае!

Регрессионный анализ. Цель и задачи.

Целью регрессионного анализа является выявление характера связи случайных величин, входящих в систему случайных величин методами математической статистики. Существо причинных связей невозможно выявить статистическими методами, и это не является целью регрессионного анализа.

Регрессионный анализ. Цель и задачи (продолжение).

- **Как невозможно найти точно математическое ожидание по выборке (а только его оценку в виде выборочного среднего) так и невозможно точно найти линии регрессии по выборочной совокупности в приведенных примерах.**
- **Однако приближенное нахождение линий регрессии является задачей регрессионного анализа.**
- Зависимость условной средней одной величины от соответствующих значений другой величины называется корреляционной связью, а уравнение связи $y_k(x) = f(x)$ - уравнением регрессии y на x.

Вычисление условных средних

Если выборка случайного вектора представлена в форме примера №2 или №3, и объем выборки достаточно большой, то возможно вычислить условные средние y(xi) по формуле:

$$y(x_i) = \frac{\sum_{j=1}^{nj} y_j \omega_{ij}}{\omega_i}, \quad где \quad \omega_i = \sum_{j=1}^{nj} \omega_{ij} \qquad (4)$$

Корреляционное поле.

- **Корреляционным полем называется диаграмма, изображающая совокупность значений двух признаков.**
- **Средствами EXCEL легко получить корреляционное поле, которое по сути дела является просто точечной диаграммой.**
- **По виду корреляционного поля и, используя другую информацию о системе случайных величин (если она известна), выбирается вид уравнения регрессии (этап спецификации).**

Линейное уравнение парной регрессии.

Пример №4: В результате n=30 экспериментов получена таблица и построено корреляционное поле.

По виду корреляционного поля делаем вывод о линейной зависимости Y от x, то есть считаем, что систематическая часть y: =α+βx. Проводим вычисления в среде EXCEL:

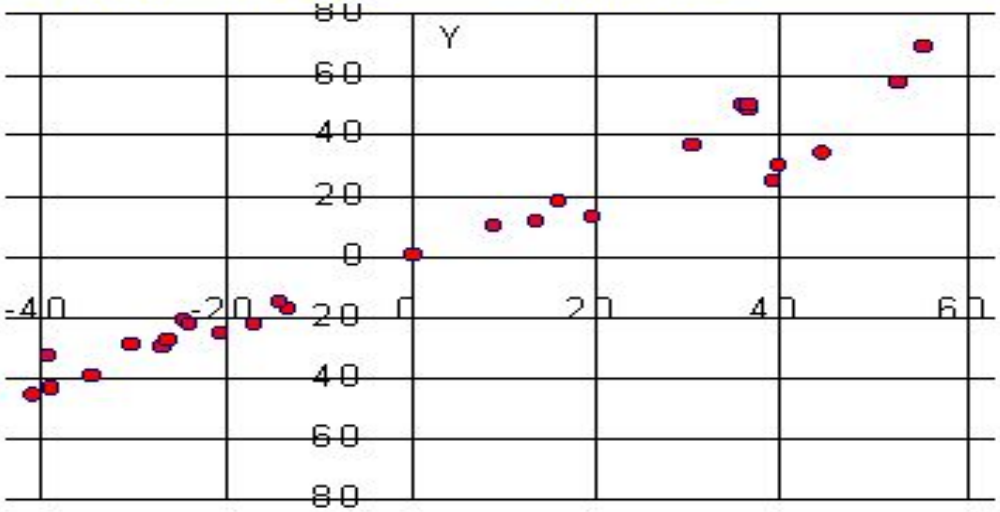
На следующих слайдах показан лист EXCEL с результатами вычислений и с формулами :

Таблица 1. Обработка наблюдений

Номер наблюдения	X	Y	$(X_i - \hat{x})^2$	$(Y_i - \hat{y})^2$	$(X_i - \hat{x})(Y_i - \hat{y})$	$\beta X_i + \alpha$
1	19,65239072	13,29660595	486,01882	247,19672	346,6154305	20,5038215
2	-16,9940358	-22,31126714	213,17729	395,42843	290,3383547	-17,61185068
3	36,43878558	48,74839325	1507,942	2618,8067	1987,211305	37,96322433
4	39,11628098	25,12693186	1723,0572	759,15771	1143,709839	40,74806713
5	-24,5646249	-20,66338782	491,56127	332,60651	404,3469818	-25,48596336
29	30,54717343	36,48341409	1085,0843	1487,1091	1270,290777	32,18164859
30	9,151298376	10,38736983	133,28107	155,42631	143,9284048	9,92796013
Сумма	-71,8033133	-62,38917623	34796,031	39586,992	36191,08995	

Таблица 2. Результаты обработки.

$\hat{\bar{x}}$	-2,393443778
$\hat{\bar{y}}$	-2,079639208
$\hat{S}^2_{1\_a}$	1199,863146
$\hat{S}^2_{2\_a}$	1365,068697
$\hat{K}_{xy\_a}$	1247,968619
$\hat{r}_{xy}$	0,975125591
β	1,04009
α	0,40976



Формулы при вводе выглядят так:

	B	C	D	E	F	G
1	Таблица 1. Обработка наблюдений					
2	Номер наблюдения	X	Y	$(X_i - \hat{x})^2$	$(Y_i - \hat{y})^2$	$(X_i - \hat{x})(Y_i - \hat{y})$
3	1	19,652	13,296	=(C3-D36)^2	=(D3-D37)^2	=(C3-D36)*(D3-D37)
4	2	-16,994	-22,311	=(C4-D36)^2	=(D4-D37)^2	=(C4-D36)*(D4-D37)
5	3	36,438	48,748	=(C5-D36)^2	=(D5-D37)^2	=(C5-D36)*(D5-D37)
6	4	39,116	25,126	=(C6-D36)^2	=(D6-D37)^2	=(C6-D36)*(D6-D37)

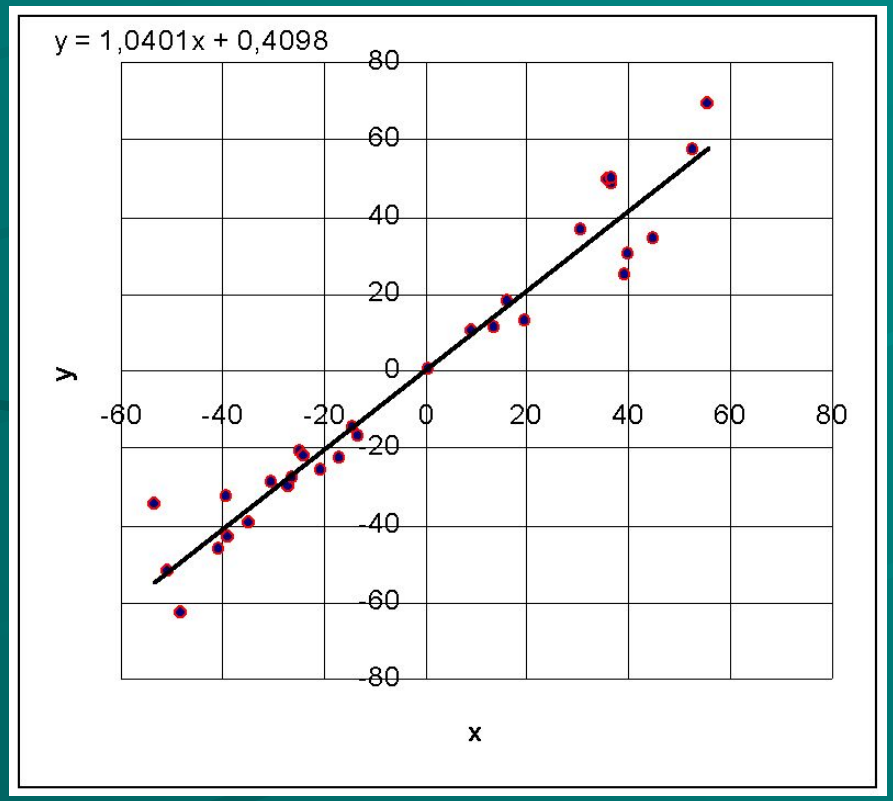
Метод наименьших квадратов дает следующие формулы для вычисления коэффициентов α и β:

$$\beta = \frac{\overline{K}_{xy}}{\overline{S}_1^{\,2}} = \frac{\overline{K}_{xy\_a}}{\overline{S}^{\,2}_{1\_a}} = \frac{\overline{r}_{xy}\,\overline{S}_{2\_a}}{\overline{S}_{1\_a}} \qquad \alpha = \overline{y} - \beta\overline{x}$$

Вычисления по этим формулам приводят к линейной регрессии Y на x:

y=0.4098+1.0401*x

Если на точечной диаграмме выделить маркеры мышкой, встав на один из них, то можно через контекстное меню добавить линию тренда. При этом следует выбрать линейный тренд и задать «показывать уравнение на диаграмме». Получится такой график:



Литература.

- 1. Вентцель Е.С. Теория вероятностей. М. Наука, 1976.
- 2. Вентцель Е.С. Овчаров Л.А. Теория вероятностей и ее инженерные приложения. М. Наука, 1988.
- 3. Гмурман В.Е. Теория вероятностей и математическая статистика. М.:Высш.шк.,2001
- 4. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. М.:Высш.шк.,2001
- 5. Вентцель Е.С. Овчаров Л.А. Задачи и упражнения по теории вероятностей М.:Высш. шк.,2002
- 6. Курзенев В.А. Основы матеметической статистики для управленцев. СпБ, СЗАГС 2002.