СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

КОГДА И ЗАЧЕМ ПРИМЕНЯЕТСЯ

При наличии большого массива данных:

- Получение усредненных данных
- Оценка связей между переменными
- Классификация
- Кластеризация
- Редукция данных

ВИДЫ ШКАЛ

- Номинативная

Категория
1 («левые»)
2 («либералы»)
3 («национал-патриоты»)
4 («центристы»)

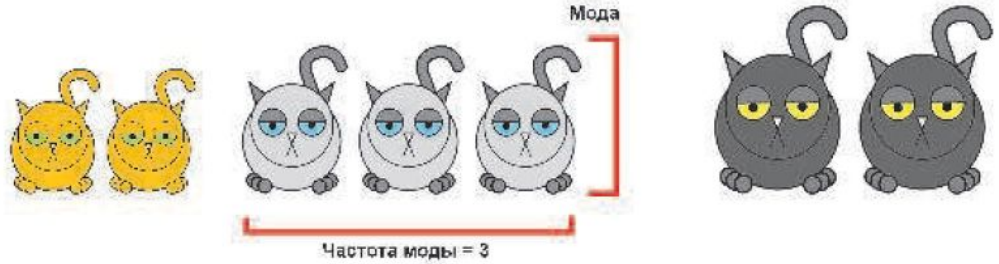
- Ранговая (порядковая)

Бегун	Ранг
A	1
B	2
C	3
D	4

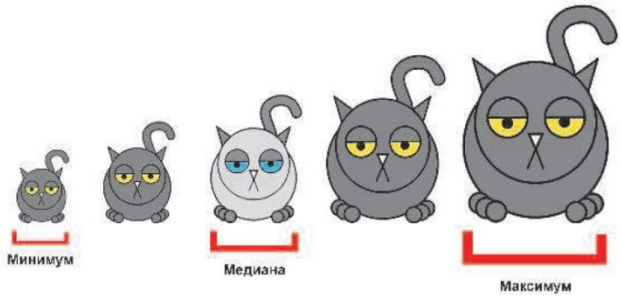
- Абсолютная (метрическая)

ЗНАЧЕНИЯ = МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

- Мода

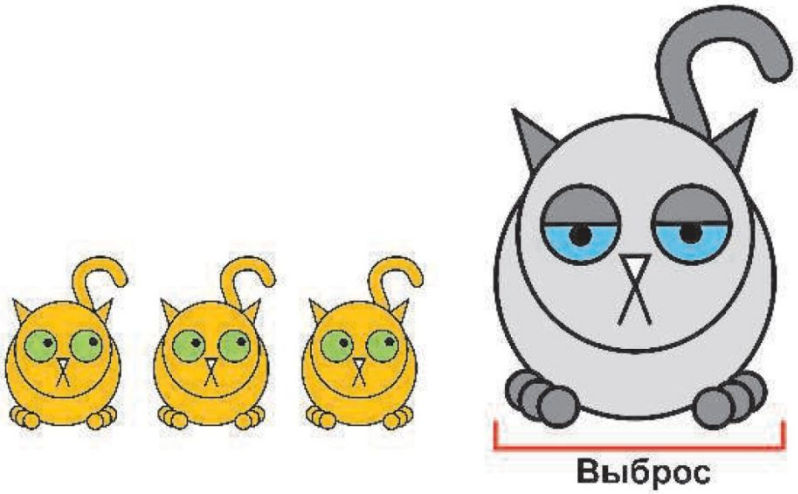


- Медиана



- Среднее арифметическое

- Выброс:



- Квантиль – точка на числовой оси, делящая всю совокупность упорядоченных измерений на две группы с известным соотношением их численности.
- Процентили – это величины (99 точек), делящие выборку данных на сто групп, содержащих (по возможности) равное количество наблюдений
- Квартили – 3 точки значения признака на числовой оси (P_{25}, P_{50}, P_{75}), делящие множество на 4 части.

МЕРЫ ИЗМЕНЧИВОСТИ

- Размах — разность между минимальным и максимальным значением: $R = X_{max} - X_{min}$

  Межквартильный размах: $R = X_{75} - X_{25}$

- Дисперсия – мера изменчивости для метрических данных, пропорциональная сумме квадратов отклонений измеренных значений от их среднеарифметического

- Стандартное отклонение - квадратный корень из дисперсии

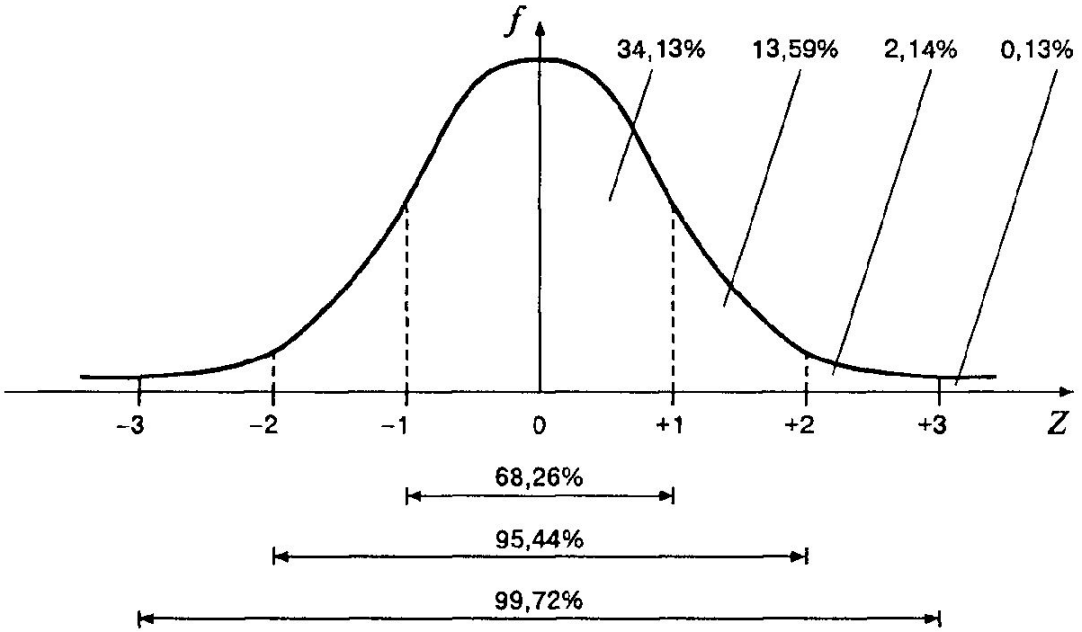
СТАНДАРТНОЕ ОТКЛОНЕНИЕ: ПРИМЕР РАСЧЕТА

№	% голосов	Средн.арифм.	Разность м/у средним и значениями	Возводим в квадрат
1	2	$\frac{2+3+2+3+90}{5}=20$	2-20=-18	324
2	3		3-20=-17	289
3	2		2-20=-18	324
4	3		3-20=-17	289
5	90		90-20=70	4900
∑	100			**6126**

Сумма квадратов /N-1	Дисперсия	Станд.отклонение
6126/(5-1)	1531,5	39,13

$$\sigma_x = \sqrt{D_x} = \sqrt{\frac{\sum_i (x_i - M_x)^2}{N-1}}$$

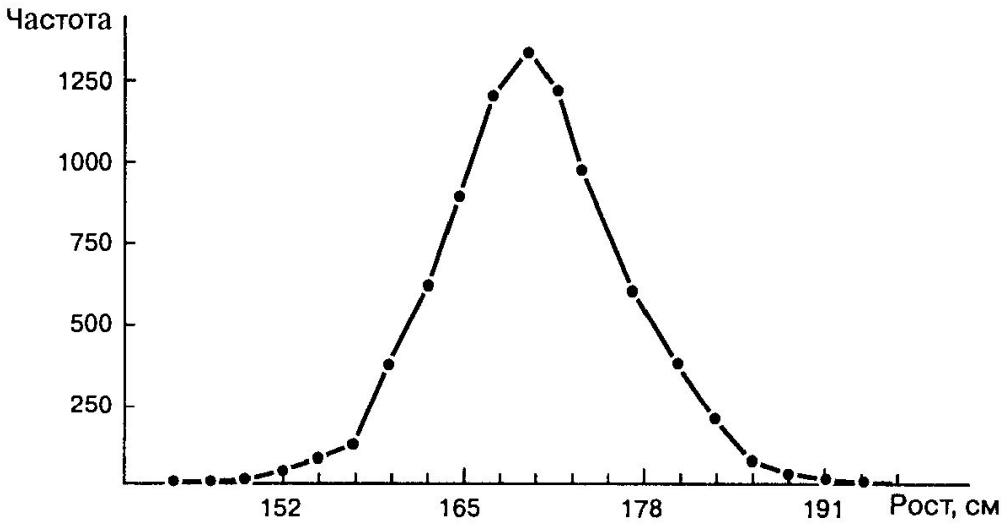
ЗАКОН НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ



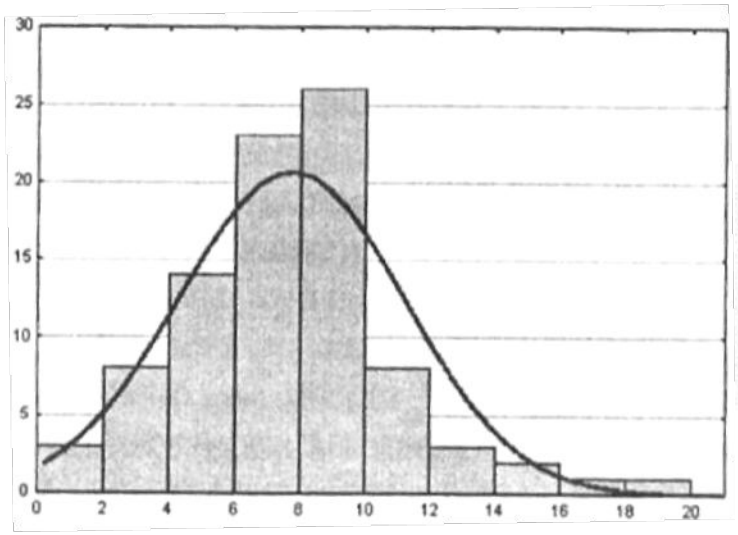
Нормальное распределение признака можно определить, если:

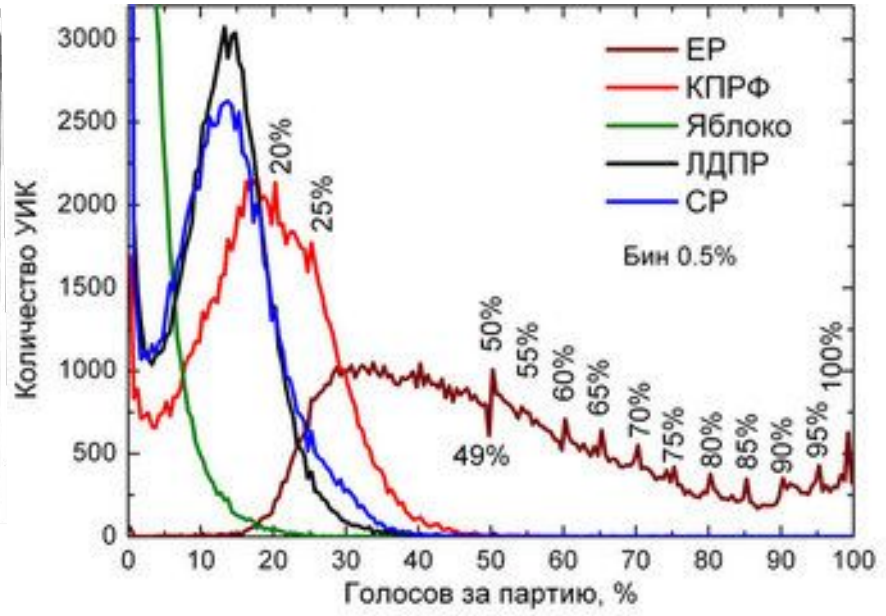
1) В ряду есть единственная мода, находящаяся в центре распределения;
2) Частоты симметрично убывают по направлениям к предельным значениям ряда;
3) Распределение признака подчиняется правилу «трех сигм»: 68,26% случаев – в пределах одного стандартного отклонения, 95,5% - в пределах двух, 99,7% - в пределах трех отклонений.

ПРИМЕРЫ



.1. Полигон частот для роста 8585 взрослых людей, родившихся в Англии в XIX в.





СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ

В гумманитарных науках устанавливается, как правило, на уровне 5% (p=0,05).

Применяется для сравнения нескольких выборок и означает, что вероятность случайного появления обнаруженных различий составляет не более 5%.

Чем меньше значение p/уровня, тем выше статистическая значимость результата исследования, подтверждающего гипотезу.

X-КВАДРАТ ПО ПИРСОНУ: НАЛИЧИЕ СВЯЗИ МЕЖДУ ПЕРЕМЕННЫМИ

- Критерий Хи-квадрат показывает, является ли отклонение реально измеренных признаков от их вероятностного распределения случайным или можно говорить о связи признаков.

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(n_{эмпир} - n_{теор}\right)^2}{n_{теор}}$$

РАСЧЕТ ХИ-КВАДРАТ

Участие в выборах				
Респонденты	Участвуют практически всегда	На выборы не ходят	Участвуют в выборах время от времени	Итого
Мужчины	80 (а)	200 (б)	200 (в)	480
Женщины	370 (г)	50 (д)	100 (е)	520
Всего	450	250	300	1000

Находим теоретические (ожидаемые) частоты:

$$n_{теор} = \frac{итого\ по\ строке \times итого\ по\ столбцу}{общее\ число\ наблюдений}$$

Респон-денты	Участие в выборах			Итого
	Участвуют практически всегда	На выборы не ходят	Участвуют в выборах время от времени	
Мужчины	216 (а)	120 (б)	144 (в)	480
Женщины	234 (г)	130 (д)	156 (е)	520
Всего	450	250	300	1000

Ячейка	Частота, n_i	Ожидаемая частота, $\tilde{n}_i$	Разность реальной и ожидаемой частот	Квадрат разности реальной и ожидаемой частот	Отношение квадрата разности реальной и ожидаемой частот к соответствующему значению ожидаемой частоты
а	80	216	−136	18496	85,63
б	200	120	80	6400	53,33
в	200	144	56	3136	21,78
г	370	234	136	18496	79,04
д	50	130	−80	6400	128,00
е	100	156	−56	3136	20,10
Σ	1000	1000			$\chi^2 = 387{,}88$

далее – сравнение с табличным критическим значением с учетом «степени свободы».

$$df = (r - 1)(c - 1)$$

где r и c - количество категорий в колонке (column) и строке (row)

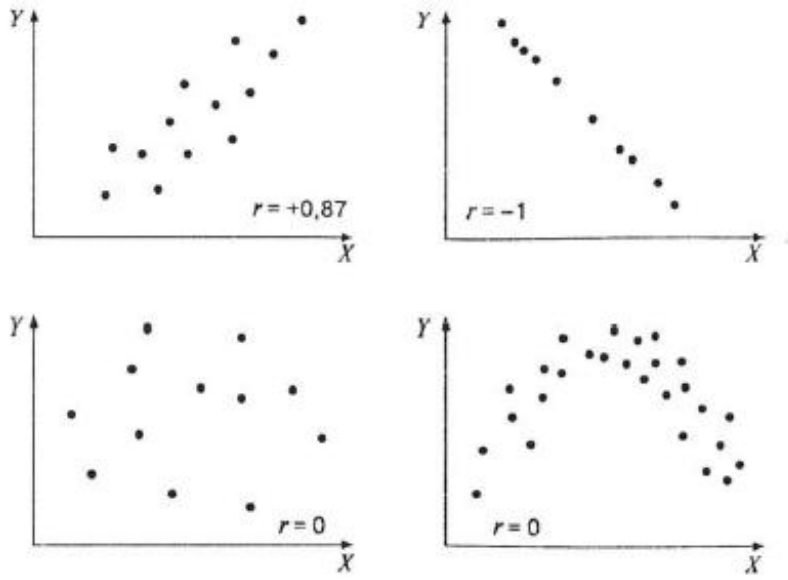
В примере: $df = (3 - 1)(2 - 1) = 2$

Число степеней свободы	Уровень значимости α					
	0,01	0,05	0,1	0,90	0,95	0,99
1	6,6	3,8	2,71	0,02	0,004	0,0002
2	9,2	6,0	4,61	0,21	0,1	0,02
3	11,3	7,8	6,25	0,58	0,35	0,12
4	13,3	9,5	7,78	1,06	0,71	0,30
5	15,1	11,1	9,24	1,61	1,15	0,55
6	16,8	12,6	10,6	2,20	1,64	0,87
7	18,5	14,1	12,0	2,83	2,17	1,24
8	20,1	15,5	13,4	3,49	2,73	1,65
9	21,7	16,9	14,7	4,17	3,33	2,09
10	23,2	18,3	16,0	4,87	3,94	2,56
11	24,7	19,7	17,3	5,58	4,57	3,05
12	26,2	21,0	18,5	6,30	5,23	3,57
13	27,7	22,4	19,8	7,04	5,89	4,11
14	29,1	23,7	21,1	7,79	6,57	4,66
15	30,6	25,0	22,3	8,5	7,26	5,23
16	32,0	26,3	23,5	9,31	7,98	5,81
17	33,4	27,6	24,8	10,1	8,67	6,41
18	34,8	28,9	26,0	10,9	9,39	7,01
19	36,2	30,1	27,2	11,7	10,1	7,63
20	37,6	31,4	28,4	12,4	10,9	8,26
21	38,9	32,7	29,6	13,2	11,6	8,90
22	40,3	33,9	30,6	14,0	12,63	9,54
23	41,6	35,2	32,0	14,8	13,1	10,2
24	43,0	36,4	33,2	15,7	13,8	10,9
25	44,3	37,7	34,4	16,5	14,6	11,5
26	45,6	38,9	35,6	17,3	15,4	12,2
27	47,0	40,1	36,7	18,1	16,2	12,9
28	48,3	41,3	37,9	18,9	16,9	13,6
29	49,6	42,6	39,1	19,8	17,7	14,3
30	50,9	43,8	40,3	20,6	18,5	15,0

ЧТО ТАКОЕ КОРРЕЛЯЦИЯ?

- Корреляция – наличие статистической взаимосвязи признаков, когда каждому определенному значению одного признака X соответствует определенное значение Y.

CORRELATION IS NOT CAUSATION



Примеры рассеивания и соответствующих коэффициентов корреляции

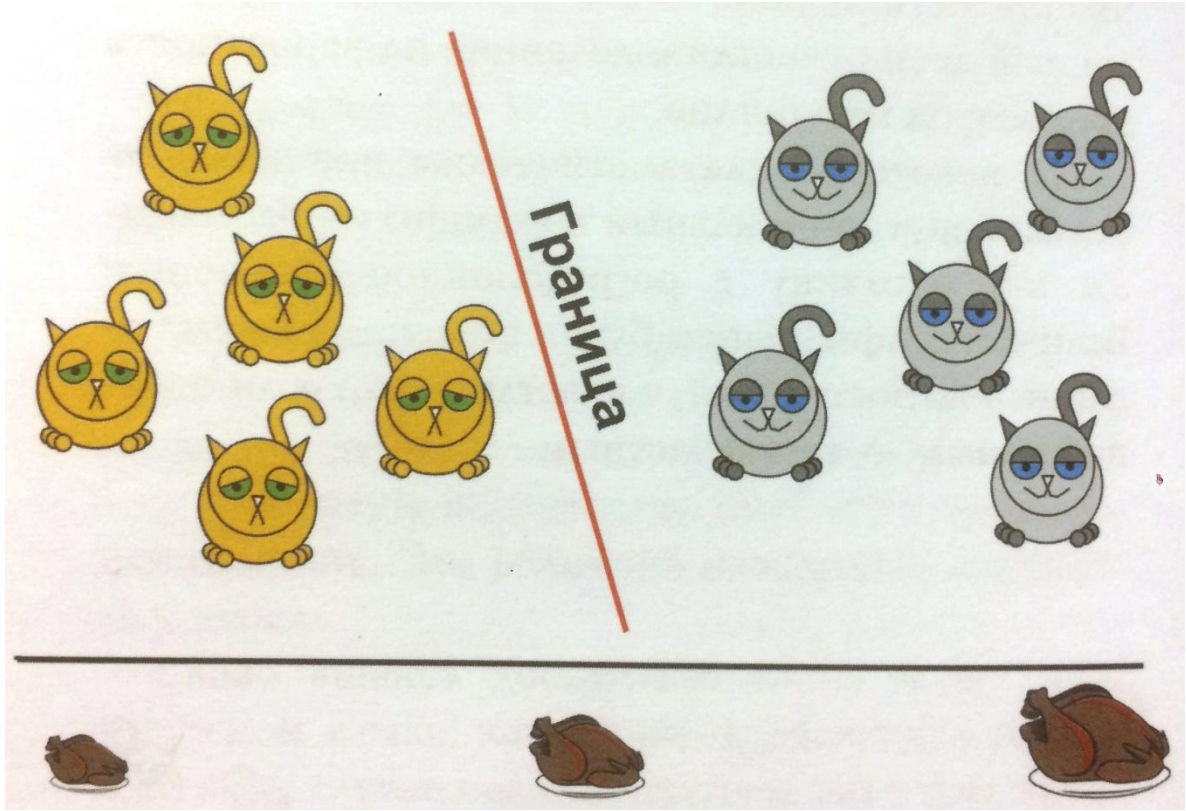
РЕГРЕССИОННЫЙ АНАЛИЗ

- Целью регрессионного анализа является измерение связи между зависимой переменной (объясняемой) и одной (парный регрессионный анализ) или несколькими (множественный) независимыми переменными (предикторы).

- Позволяет определить влияние переменных на исследуемую проблему.

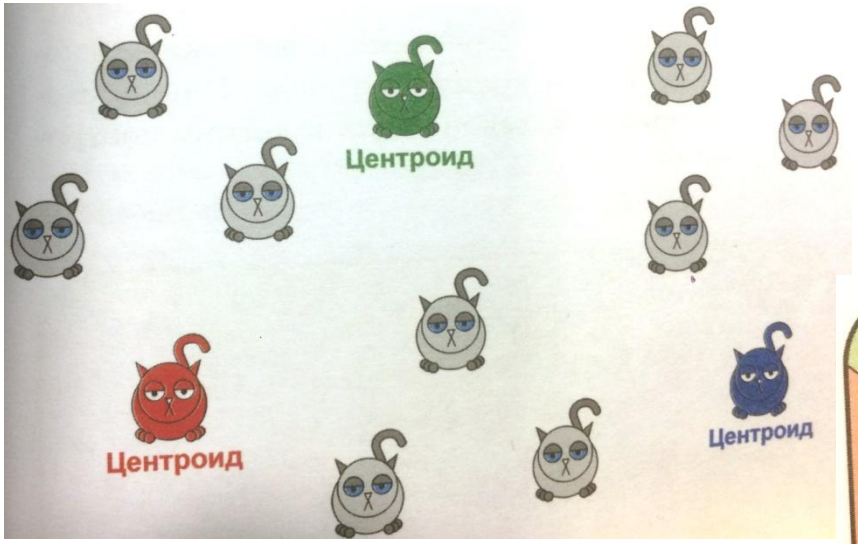
ДИСКРИМИНАНТНЫЙ АНАЛИЗ

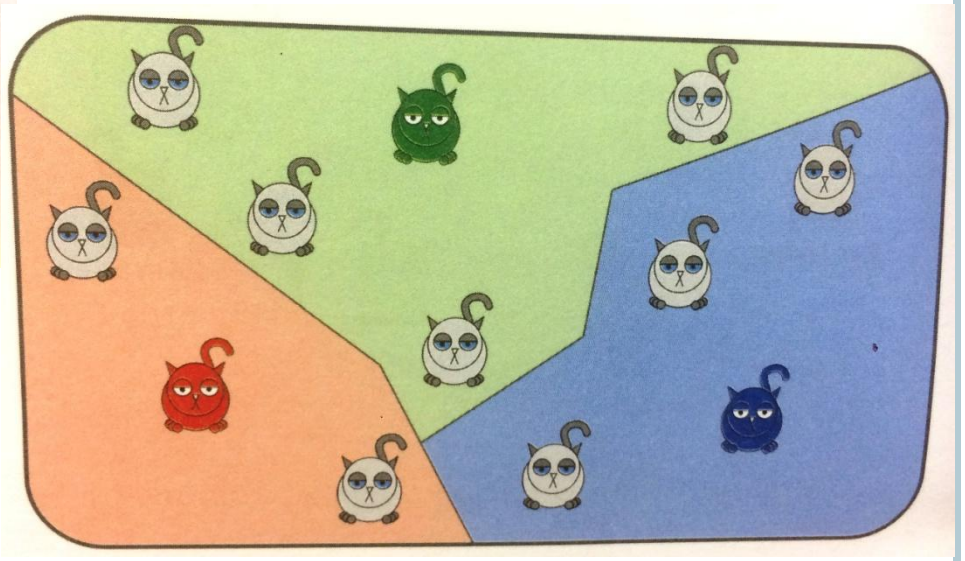
- Позволяет определить критерии для отнесения объекта измерения к тому или иному классу.



КЛАСТЕРНЫЙ АНАЛИЗ

- Позволяет разбить объекты на классы, при этом число классов может быть как известно заранее, так и нет.





ФАКТОРНЫЙ АНАЛИЗ

- Позволяет сократить количество переменных, заменив их набором факторов. Может являться предварительной процедурой перед регрессионным анализом, если ряд предикторов коррелируют между собой.