

## **1.3. Разведочный анализ данных**



## Цель, задачи

**Цель**– представить наблюдаемые данные в компактной и простой форме, позволяющей выявить имеющиеся закономерности и связи

**Разведочный анализ данных (РАД) включает:**

- преобразование данных и способы наглядного их представления
- выявление аномальных значений
- грубая оценка типа распределения
- сглаживание

# Вопросы анализа данных

- 1. Какой обработке подвергнуть наблюдения?**
- 2. Какую модель выбрать?**
- 3. Какие заключения можно сделать?**

# Пример РАД

**Разведочный анализ (*Exploratory data analysis*) – средство получения более полной информации об изучаемом явлении**

**Наблюдения  $n$  пар  $(x_1, Y_1), \dots, (x_n, Y_n)$  опишем уравнением**

$$(1) \quad \mathbf{M}(Y_i) = \beta_0 + \beta_1 x_j, \quad i = 1, \dots, n$$

**Минимальный предварительный анализ - график рассеяния точек  $(x_j, Y_j)$ .**

# Предварительная обработка данных. Оценка среднего

Оценка  $\hat{m}$  - истинного среднего  $m$  независимой случайной величины  $x$  по выборке объема  $n$

Доверительный интервал:  $\hat{m} \pm t S_{\hat{m}}$

$t$ -распределение Стьюдента:  $t = \hat{m} / S_{\hat{m}}$

95%-е доверительные интервалы

Для нормального распределения  $t = 1,96,$

Для  $t$ -распределения при числе степеней свободы  $\nu$  ( $\nu = n - 1$ ), равных 1; 3 и 12, величина  $t$ , соответственно, равна 12,7; 4,3 и 2,18.

# Причины отличия реального распределения от нормального

1. Большинство измерений проводится в конкретных единицах
2. Резкая асимметрия некоторых распределений (например,  $\chi^2$ , F) при малых выборках, обрывистые края у равномерного распределения
3. Поведение на «хвостах» распределения, которое существенно отличается от значений основного количества наблюдений

# Робастные оценки

*Робастные оценки* - robust – крепкий, здоровый,

Пример робастной оценки среднего, терпимой к отклонению хвостов распределения от нормального - *медиана* распределения

# Мера разброса

- **среднеквадратическое отклонение  $\sigma$**
- **дисперсия  $\sigma^2$**
- **размах  $R$**

**Оценки этих величин обозначают,  
соответственно,  $S, S^2, R$**

**Оценка разброса по  $S$  – в линейных  
преобразованиях типа  $Y = \beta + \alpha X$**



# Разбиение данных на три группы



# Качество результатов

- ***Простая перепроверка.*** Проверка полученной модели на данных, отличных от тех, по которым определены параметры модели
- ***Двойная перепроверка.*** Проверка на данных отличных, как от тех, по которым строилась модель, так и от тех, которые использованы для вычисления параметров модели

# Неоднородные выборки

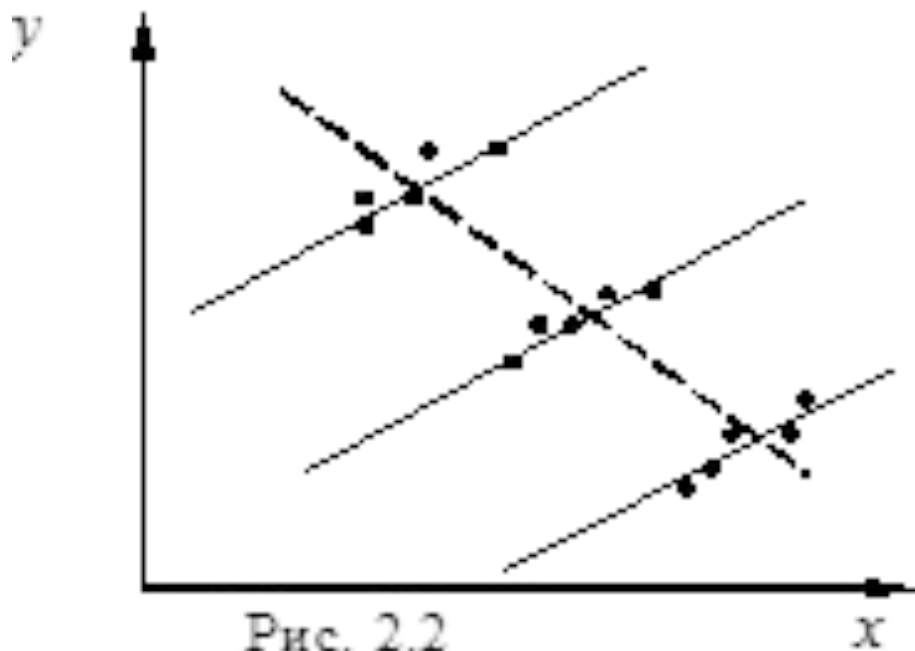


Рис. 2.2

# Разделение неоднородной совокупности на однородные

Пусть выборка изучаемой совокупности  $x_1, \dots, x_n$ , содержит элементы двух независимых случайных величин с плотностями распределений  $f(x, \theta_1)$  и  $f(x, \theta_2)$ .

Обозначим через  $A$  – множество элементов выборки, принадлежащих к первой случайной величине,  $B$  – множество элементов выборки из второй совокупности.

Требуется найти оценки неизвестных параметров  $\theta_1, \theta_2$  и множества  $A$  и  $B$ .

Для оценки этих четырех неизвестных используем метод максимума правдоподобия

# Обнаружение аномальных наблюдений

## Причины:

- грубые ошибки при регистрации измерений,
- случайные импульсные помехи,
- сбои оборудования,
- измерения в ошибочных единицах
- и др.

# Обнаружение аномальных наблюдений. Критерий проверки

Пусть наблюдения  $x_1, \dots, x_n$  являются реализациями независимых случайных величин, подчиняющихся одинаковому нормальному  $N(\mu, \sigma^2)$  распределению

Основная гипотеза  $H_0$ :  $Mx_i = \mu, Dx_i = \sigma^2, i = 1, \dots, n$ .

Альтернативная гипотеза  $H_1$  : одна или несколько величин имеют среднее  $\mu + d$

# Обнаружение аномальных наблюдений. Критерий проверки

При построении критерия возможны варианты, зависящие от степени информации о  $\mu$  и  $\sigma$ .

Рассмотрим случай, когда значения  $\mu$  и  $\sigma$  неизвестны. Критериальная статистика:

$$D_n = (x_{(n)} - \bar{x}) / S \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Распределение величины  $D_n$  получены К. Пирсоном и Н. В. Смирновым. Критические значения  $D_n'$  вычислены Н. В. Смирновым и Ф. Граббсом

$H_0 - D_n < D_a$  - наблюдение не является аномальным

$H_1 - D_n > D_a$  - наблюдение является аномальным

# Общие выводы об удалении аномальных наблюдений

1. Для данных с неправдоподобными наблюдениями использовать *робастные процедуры* оценивания
2. Существенно выделяющиеся данные обнаруживать, преобразовывать и удалять, при этом интерпретировать, привлекая знания, не относящиеся к статистической природе
3. Процедуры удаления существенно выделяющихся и подозрительно больших наблюдений с последующим оцениванием близких к робастным оценкам



# Простые числовые и графические сводки данных

## Процедура «стебель с листьями» (Stem-and-Leaf)

250 688 695 795 795 895 895 895 1099 1166 1333 1499 1693 1699  
1775 1895

Три вида записи «стебля с листьями» цен на 17 автомобилей «Шевроле»:  
а – единица = 100 \$; б – единица = 10 \$; в – единица = 100 \$

#		
2	0*	12
7	0.	667788
4	1*	0134
4	1.	6678
√17		

#		
1	1	5
1	2	5
2	6	98
2	7	99
3	8	999
1	10	9
1	11	6
1	13	3
1	14	9
2	16	99
1	17	7
1	18	9
√17		

#		
1	0*	1
1	T	2
0	F	
3	S	677
3	□	888
2	1*	01
1	T	3
1	F	4
4	S	6677
1	□	8
√17		

## Списки использованной литературы и источников:

- А.А.Большаков, Р.Н.Каримов «Методы обработки многомерных данных и временных рядов» Москва 2007 г.
- Электронный учебник StatSoft по анализу данных.