Санкт-Петербургский
государственный
университет
www.spbu.ru

Анализ данных

Описательная статистика

Графеева Н.Г.

2018

Анализ данных. Описательная статистика

- Задача **описательных статистик** — первичная систематизация данных, полученных экспериментально или в ходе наблюдений и их наглядное представление. В бизнесе статистика используются повсеместно, от расчета зарплат сотрудникам до анализа популярности бренда на рынке.
- Рассмотрим основные описательные статистики и их практическое применение.

Анализ данных. Описательная статистика

Центральная тенденция

Измерение центральной тенденции (measure of central tendency) состоит в выборе одного числа, которое наилучшим
образом описывает все значения признака из набора данных.
Такое число называют **центром, типическим значением** для
набора данных, мерой центральной тенденции.

Анализ данных. Описательная статистика

Плюсы и минусы центральной тенденции

Плюсы:

- Получение информацию о распределении признака в сжатой форме.
- Можно сравнивать между собой два набора данных (две выборки).

Минус:

- Выбор центра ведет к потере информации по сравнению с распределением частот.

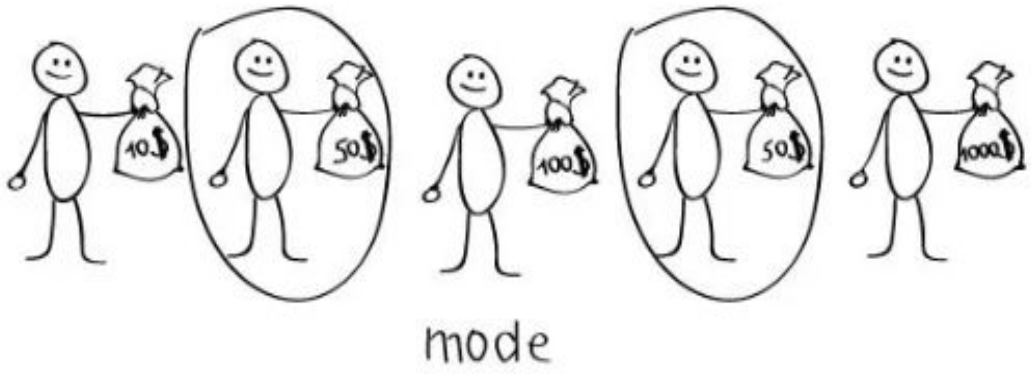
Центральная тенденция

- Мода
- Медиана
- Среднее значение
- Средневзвешенное значение

Анализ данных. Описательная статистика

Мода

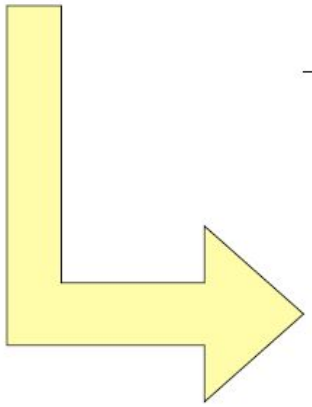
Мода – наиболее часто встречающееся значение в выборке, наборе данных. Обозначается ***Mo.***



Анализ данных. Описательная статистика

Пример (вычисление моды)

Выборка: 5 4 1 2 4 3 1 2 4 8 3 6 4 1



Варианты	Частоты	
1	3	
2	2	
3	2	
4	4	Мода=4
5	1	
6	1	
8	1	

Наиболее часто встречающееся значение

Анализ данных. Описательная статистика

Пример (вычисление моды для таксиста и светофоров)

Выборка, набор данных:

На какой свет чаще всего таксист проезжает перекрестки?

Анализ данных. Описательная статистика

Пример (вычисление моды при подсчете)

Для данных, расположенных в таблице частот, мода определяется как значение, имеющее наибольшую частоту.

КАТЕГОРИИ	*f*
Демократы	41
Коммунисты	23
Либералы	22
Любители пива	5
Зеленые	12
Всего	103

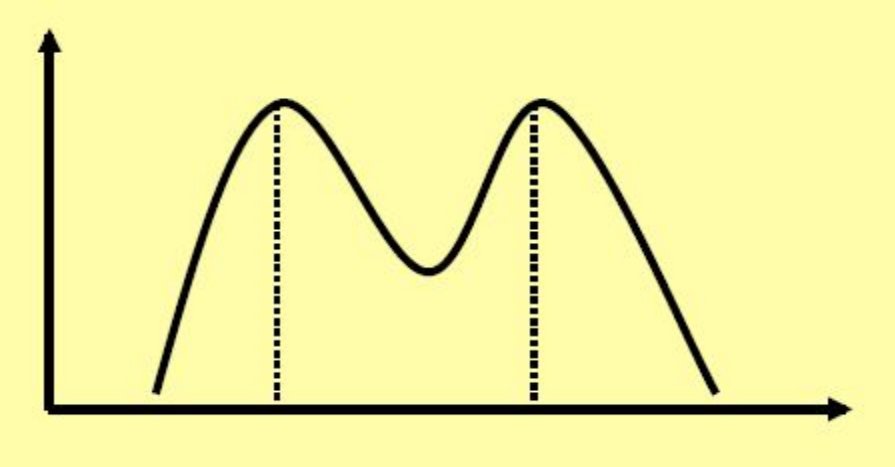
Мода «Демократы»

Анализ данных. Описательная статистика

Бимодальное распределение

Если наибольшую частоту имеют два значения выборки,

выборочное распределение называется **бимодал**

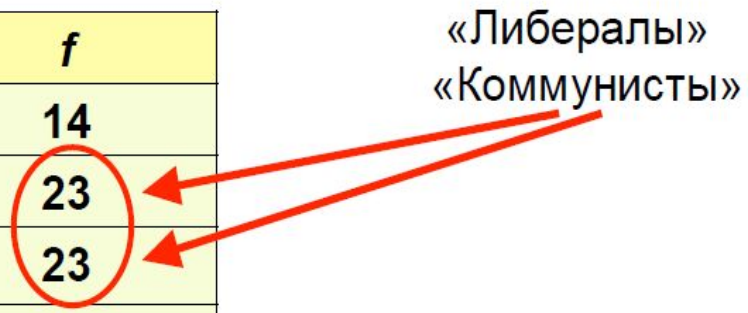


Анализ данных. Описательная статистика

Пример (бимодальное распределение)

Два значения имеют наибольшую частоту, равную 23.

КАТЕГОРИИ	*f*
Демократы	14
Коммунисты	23
Либералы	23
Любители пива	8
Зеленые	12
Всего	80



Две моды !!!

Анализ данных. Описательная статистика

Пример (бимодальное распределение на гистограмме)

Два значения имеют наибольшую частоту, равную 23.

Анализ данных. Описательная статистика

А если моды вообще нет или больше двух?

Если наибольшую частоту имеет более двух значений выборки, выборочное распределение называется **мультимодальным.** Если ни одно из значений не повторяется, мода отсутствует.

Анализ данных. Описательная статистика

Свойства моды

- Наличие одного или двух крайних значений, сильно отличающихся от остальных, не влияет на значение моды.
- Мода совпадает с точкой наибольшей плотности данных.
- Мода может иметь несколько значений.
- Мода может существовать для всех типов данных.
- Мода - единственная мера центральной тенденции, которая работает в номинальной шкале!

Анализ данных. Описательная статистика

Медиана

Еще одна характеристика центральной тенденции - **медиана**. **Медиана** основывается на понятии **вариационного ряда**.

Анализ данных. Описательная статистика

Вариационный ряд

Вариационный ряд – это упорядоченные данные, расположенные в порядке возрастания значения признака, либо в порядке убывания.

Назван так, поскольку содержит **варианты** значений признака.

Анализ данных. Описательная статистика

Пример (вариационный ряд)

Набор данных:

6 1 3 7 1 7 3

После упорядочения (в порядке возрастания) получим вариационный ряд:

1 1 3 3 6 7 7

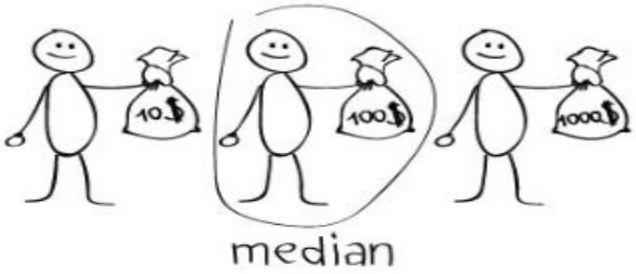
В порядке убывания получим другой вариационный ряд:

7 7 6 3 3 1 1

Анализ данных. Описательная статистика

Медиана (Median)

- **Медиана** есть значение серединного элемента для вариационного ряда.
- Обозначается ***Me.***
- Для нахождения **медианы** требуется набор данных превратить в вариационный ряд, то есть расположить все значения признака в порядке возрастания или убывания, а затем найти средний элемент. Он и есть **медиана**.



Анализ данных. Описательная статистика

Вычисление медианы

Для набора из **n** значений, если **n** нечетно, средний элемент имеет номер **(N + 1)/ 2**.

Если **n** четно, медиана находится как среднее арифметическое
двух соседних серединных элементов с номерами **N/2** и **N/2 +1**.

Анализ данных. Описательная статистика

Пример (вычисление медианы)

Для набора данных из семи чисел:

6 1 3 7 1 7 3

После упорядочения получим вариационный ряд:

1 1 3 3 6 7 7

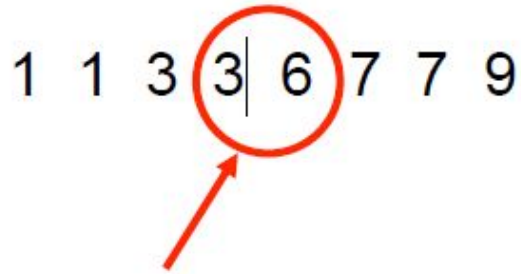
Медиана есть средний элемент.

Его номер четвертый.

Анализ данных. Описательная статистика

Пример (вычисление медианы)

Если набор данных включает восемь чисел:



Тогда медиана равна (3+6)/2=4,5

Анализ данных. Описательная статистика

Свойства медианы

- Сильно отличающиеся от остальных данных крайние значения не влияют на величину медианы.
- Значение медианы является единственным для каждого набора данных.
- Медиана может быть определена не из полного набора данных. Достаточно иметь информацию об упорядоченности, общее число элементов в наборе и несколько значений, расположенных в середине вариационного ряда.
- Медиана может быть определена для числовых и порядковых данных.

Анализ данных. Описательная статистика

Среднее (Mean)

Выборочным **средним** будем называть среднее арифметическое выборки, то есть сумму всех значений выборки, деленную на ее объем выб

$$\bar{x} = \frac{\sum x}{n}$$

где $\sum x$ = сумма всех значений выборки

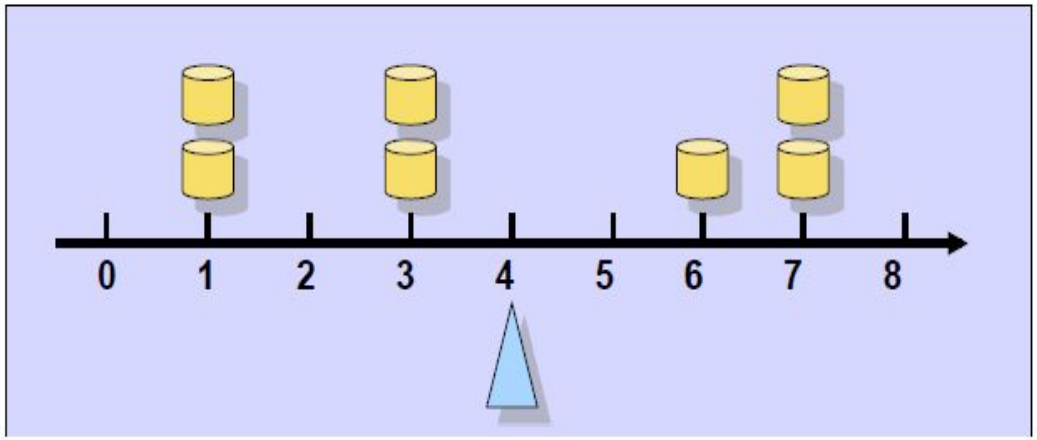
n = объем выборки

Анализ данных. Описательная статистика

Пример (вычисление среднего)

Вычислим среднее для выборки из семи значений: 1 1 3 3 6 7 7

Получим: $\bar{x} = \frac{1+1+3+3+6+7+7}{7} = \frac{28}{7} = 4$



Среднее значение является «точкой равновесия».

Анализ данных. Описательная статистика

Свойства среднего

- Вычисляется только в числовых шкалах.
- При вычислении необходимо использовать все данные.
- Для каждого набора данных имеется только одно среднее.
- Среднее есть единственная мера центральной тенденции, для которого сумма отклонений каждого значения от среднего равна нулю:

$$\sum (x - \bar{x}) = 0$$

Анализ данных. Описательная статистика

Взвешенное среднее

ГРУППА	СРЕДНЕЕ ПО ГРУППЕ	ОБЪЕМ ГРУППЫ
A	87	65
B	92	110
C	89	85
D	96	200
E	84	60
Всего		520

Среднее взвешенное

Среднее взвешенное вычисляется по формуле:

$$\overline{X} = \frac{\sum(\overline{x} \cdot n)}{N}$$

где $\sum(\overline{x} \cdot n)$ = сумма произведений средних в группе на количество элементов в этой группе,

N = общее число наблюдений во всех группах

Анализ данных. Описательная статистика

Пример (вычисление среднего взвешенного)

ГРУППА	СРЕДНЕЕ ПО ГРУППЕ	ОБЪЕМ ГРУППЫ	Произведение
A	87	65	5 655
B	92	110	10 120
C	89	85	7 565
D	96	200	19 200
E	84	60	5 040
Всего		520	47 580

$$\overline{X} = \frac{47580}{520} = 91{,}5$$

Анализ данных. Описательная статистика

Пример: где особенно уместно использовать средневзвешенное значение

Среднее для дихотомической шкалы

Среднее может также применяться и для переменной, измеренной в дихотомической шкале. Если два значения признака кодируются 0 и 1, то среднее указывает долю (относительную частоту) единиц в выборке.

Пример: 1, 0, 0, 0, 1, 1, 1, 1, 1, 0

Среднее равно 0,6. То есть 60% значений выборки принимают значение, равное единице.

Среднее – не значит лучшее

Пример. В деревне 50 жителей. Среди них 49 человек – крестьяне с месячным доходом в 1 тыс.рублей, а один житель – зажиточный владелец строительной фирмы, с месячным доходом 451 тыс.рублей. Среднее равно 10 тыс. рублей. Однако, вряд ли можно утверждать, что это число адекватно представляет доход жителей деревни. В этом случае, более разумно взять в качестве меры центральной тенденции **моду** или **медиану** (обе равны 1 тыс. рублей).

Какое типическое значение наилучшее?

В зависимости от данных каждое из трех значений может стать наилучшим! Абсолютных рекомендаций нет.

Анализ данных. Описательная статистика

Меры и шкалы

Шкала, по которой измеряется переменная, накладывает ограничения на выбор меры центральной тенденции.

Типическое значение	Номинальные данные	Порядковые данные	Дихотомические данные	Интервальные данные	Относительные данные
Мода	V	V	V	V	V
Медиана		V		V	V
Среднее			V	V	V

Анализ данных. Описательная статистика

Мера центральной тенденции – всего лишь одно число, которое не всегда достаточно емко может описать данные. Именно поэтому были придумано понятие **размаха** и **квартильного размаха**, как логическое продолжение мер центральной тенденции.

Пример (три выборки)

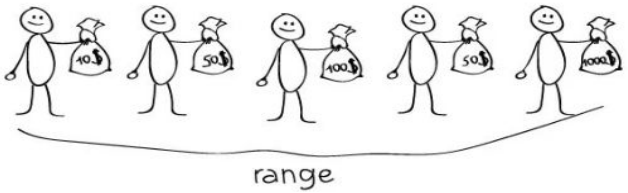
Рассмотрим три выборки:

- 999 1000 1001
- 900 1000 1100
- 1 1000 1999

Во всех трёх случаях среднее равно 1000. Однако это значение никаким образом не отражает особенности этих выборок.

Анализ данных. Описательная статистика

Размах (Range)



Размах – разность между наибольшим значением набора данных и наименьшим.

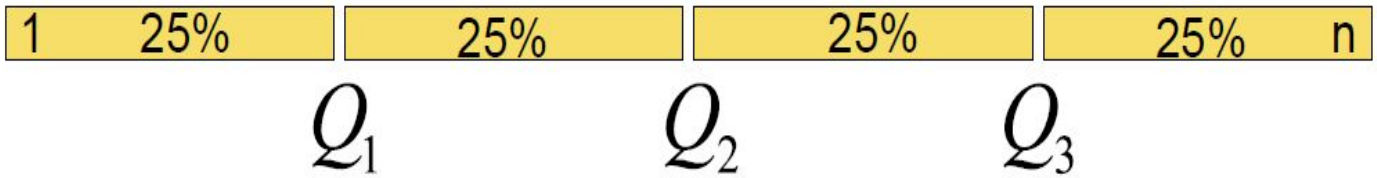
$$R = x_{\max} - x_{\min}$$

Пример: Для набора данных **27, 3, 26, 19, 12, 10, 8, 6** размах равен **R = 27 – 3 = 24**.

Размах – уже неплохо, чтобы расширить наше представление о выборке но можно пойти и дальше.

Квартили (Quartile)

- Под **квартилями** понимаются значения Q_1, Q_2, Q_3 которые делят вариационный ряд на четыре равные части.
- Второй квартиль Q_2 совпадает с **медианой**.
- Q_1 - это медиана для значений, которые левее Q_2.
- Q_3 - это медиана для значений, которые правее Q_2.



Анализ данных. Описательная статистика

Проблемы с границами при определении квартилей

Есть разные способы определения Q1 и Q3. В некоторых сама медиана (Q2), полученная на предыдущем шаге учитывается при определении Q1, Q3, в других – нет (в литературе описывают по крайней мере 9 вариантов). Рассмотрим, как это делает EXCEL и ORACLE.

Анализ данных. Описательная статистика

Применение функции КВАРТИЛЬ в EXCEL

Нечетное количество чисел

1	3	это	Q1
2			
3	5	это	Q2
4			
5			
6	7	это	Q3
7			
8			
9			

Четное количество чисел

1	2.75	это	Q1
2			
3			
4	4.5	это	Q2
5			
6			
7	6.25	это	Q3
8			

Анализ данных. Описательная статистика

Вычисление квартилей в ORACLE

Нечетное количество чисел

1		3	это Q1
2			
3		5	это Q2
4			
5			
6		7	это Q3
7			
8			
9			

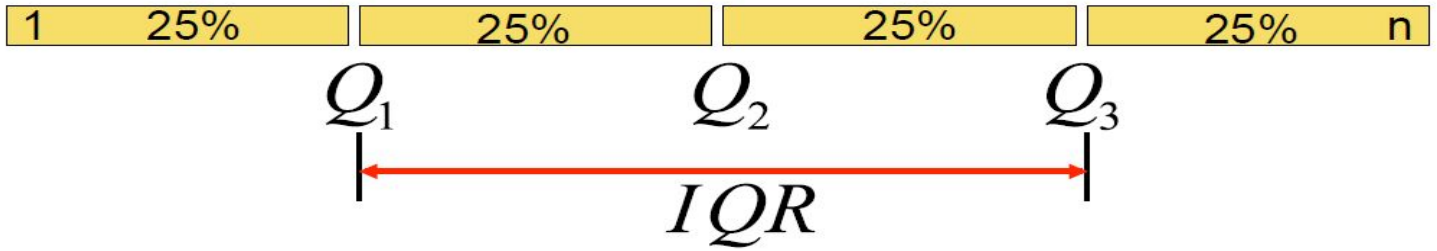
Четное количество чисел

1		2.75	это Q1
2			
3			
4		4.5	это Q2
5			
6			
7		6.25	это Q3
8			

Размах квартилей (Inter Quartile Range)

Размах квартилей - это разница между третьим и первым квартилем и вычисляется по формуле:

$$IQR = Q_3 - Q_1$$



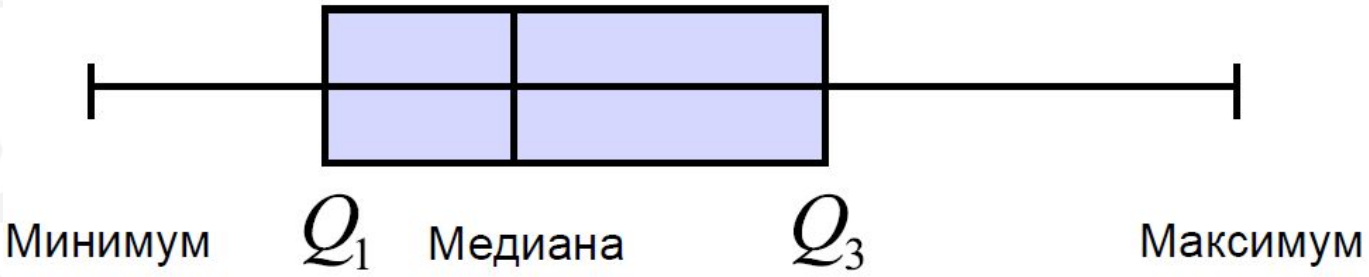
Анализ данных. Описательная статистика

Сравнение размаха и квартильного размаха

- При вычислении **размаха** используются только наибольшее и наименьшее значения признака. Распределение данных между ними полностью игнорируется.
- **Размах** – очень простая мера вариации, но очень «грубая».
- При вычислении **квартильного размаха** игнорируются только крайние значения, расположенные за пределами первого и третьего квартилей.

Коробковая диаграмма (Box plot)

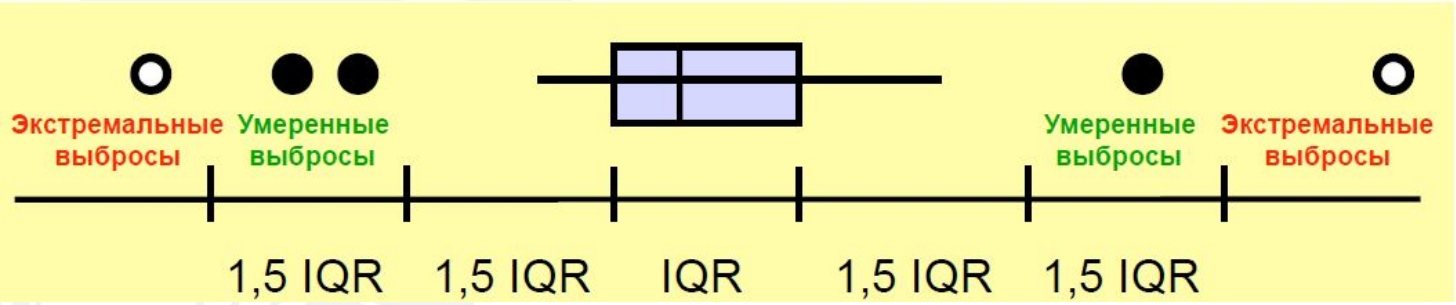
Диаграмма, основанная на пяти важных числах. Удобна для анализа данных и широко используется для представления основных характеристик выборки.



Анализ данных. Описательная статистика

Еще один способ для определения выбросов

- **Умеренные выбросы** удалены ниже первой квартили или выше третьей от 1,5 IQR, но не более 3 IQR.
- **Экстремальные выбросы** удалены ниже первой квартили или выше третьей более 3 IQR.



Анализ данных. Описательная статистика

Пример (актеры и актрисы)

Имеются данные о возрасте актеров и актрис, в котором они были удостоены Оскара. Актеры:

32	37	36	32	51	53	33	61	35
45	55	39	76	37	42	40	32	60
38	56	48	48	40	43	62	43	42	44
41	56	39	46	31	47	45	60	46
40	36							

Актрисы:

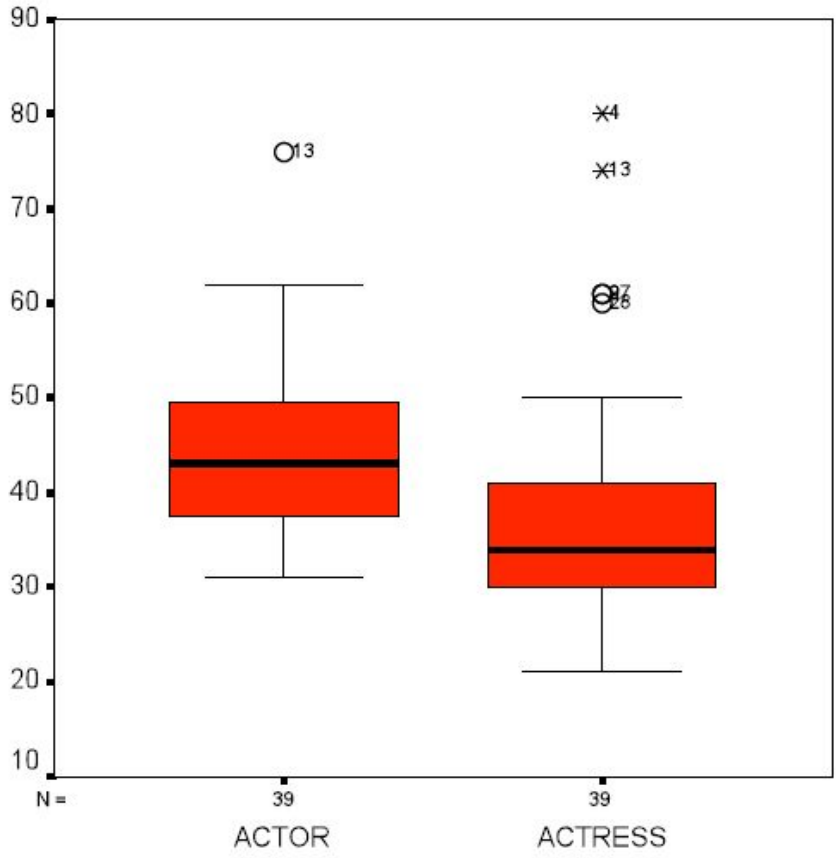
50	44	35	80	26	28	41	21	61
38	49	33	74	30	33	41	31	35
41	42	37	26	34	34	35	26	61	60
34	24	30	37	31	27	39	34	26
25	33							

Анализ данных. Описательная статистика

Пример (Box plot с расширением)

Несколько значений оказалось выбросами. Например, актер 76 лет - умеренный выброс.

Поскольку для актрис размах квартилей меньше, 80 и 74 года составили экстремальный выброс.
60 и 61 – умеренные выбросы.
Для оставшихся значений заново пересчитали статистики.



Задание 4

На сайте Москвы найдите открытые данные о том, как называли младенцев в 2015 – 2018 годах. На основании этих данных постройте три Box Plot диаграммы для своего имени и своих родителей (или братьев – сестер). Определите, были ли выбросы (умеренные или экстремальные за этот период).

Примечание: Срок сдачи: 2 недели с момента выдачи. Задание отправлять по адресу: N.Grafeeva@spbu.ru.

Topic: DataMining_2018_job4

Анализ данных