Сеченовский Университет
наук о жизни

Институт лидерства и управления здравоохранением

Инструменты статистического анализа научных данных
Описательная статистика

Косаговская Ирина Игоревна
Высшая школа управления здравоохранением

16.11.2023

Неверно организованный эксперимент не спасет никакой статистический анализ!

Какой представляется исследователю обработка результатов?

Основные этапы статистического анализа

- **Описание полученного массива данных**

- **Анализ данных и проверка различных гипотез**

Описание полученного массива данных Descriptive Statistics

Прежде чем приступать к описанию признака, определите его тип.

ВНИМАНИЕ !

От типа признака зависит выбор статистического пути его описания (обобщения)!

Признаки, или переменные (variables), могут принимать различные конкретные значения (values).

Типы признаков (виды шкал)

- **Качественные, категориальные**
  (qualititative, categorical)
  - Номинальные (Nominal)
    (частный случай : бинарные, дихотомические
    (Binary – dichotomous)
  - Порядковые, ординальные, ранжируемые (Ordinal)

- **Количественные, интервальные**
  (quantitative, numerical, interval)
  - Дискретные (Discrete)
  - Непрерывные (Continuous)

Типы признаков

Качественные, категориальные или номинальные (categorical/nominal) –

не поддающиеся непосредственному измерению, например, характеристики пациента: диагноз, пол, профессия, семейное положение.

Типы признаков

Качественные данные, которые могут быть отнесены только к двум противоположным категориям (да – нет), принимающие одно из двух значений (жив – умер; курит – не курит) называются дихотомическими (dichotomous).

Типы признаков

- **Порядковые, ординальные** или ранжируемые (ordinal) – эти признаки можно расположить в естественном порядке (ранжировать), но при этом отсутствует количественная мера расстояния между величинами.

  Примеры порядковых признаков:

  - оценка тяжести состояния пациента,

  - стадия болезни,

  - самооценка состояния здоровья.

  При этом допускается, что тяжелое течение заболевания «хуже», чем среднетяжелое, а очень тяжелое – «еще хуже», однако нельзя сказать во сколько или на сколько хуже.

Типы признаков

Таким образом, порядковые данные занимают промежуточное положение между количественными и номинальными типами.

Их можно упорядочить как количественные данные, но над ними нельзя производить арифметические действия, также как и над номинальными данными.

Типы признаков

Количественные или **интервальные** (interval) – признаки, количественная мера которых четко определена; наиболее удобный для статистического анализа тип данных.

Типы признаков

Количественные признаки могут быть **непрерывными (**continuous), принимающими любое значение на непрерывной шкале (масса тела, температура, биохимические показатели крови)

Типы признаков

- **дискретными** (discrete), принимающие лишь определенные значения из диапозона измерения, обычно целые, (число рецидивов, число детей в семье, число заболеваний у одного больного, число выкуриваемых сигарет)

Type of data - Examples

- Qualitative
  - **Dichotomous - binary**
    - Gender: male or female.
    - Employment status: employment or without employment.
  - **Ordinal**
    - Socioeconomic level: high, medium, low.
    - none, light, moderate, heavy smoker
  - **Nominal**
    - Residency place: center, North, South, East, West.
    - race
    - Civil status: single, married, widowed, divorced, free union.
- Quantitative
  - **Discrete**
    - Number of offspring: 1,2,3,4.
    - days in the hospital
  - **Continuous**
    - Glucose in blood level: 110 mg/dl, 145 mg/dl.
    - blood pressure,
    - age

Редукция данных

Возраст:

Годы – количественный признак

Десятилетия – интервальный признак

Периоды – ординальный (порядковый)

Молодой-пожилой - номинальный

Перечень признаков, подлежащих изучению в ходе исследования, оформляется в виде **регистрационного документа** (анкета, бланк, карта и т.п.), включающего вопросы, которые исследователь хочет изучить в ходе эксперимента и в дальнейшем заполняется на каждую единицу наблюдения.

Регистрационный документ должен отвечать определенным требованиям:

- - **включать обязательные вопросы** (номер единицы наблюдения, дата заполнения, название учреждения, паспортная часть, подпись лица, заполнившего документ)
- - **однозначность формулировок** вопросов (предполагает унифицированность заполнения документа для любого лица)
- - **удобство для чтения и заполнения**, а также для шифровки и обработки данных (альтернативные ответы или подсказы ответов)

Фрагмент анкеты для родителей

1. Ф.И.О. ребенка_____
2. Возраст ребенка: _____ лет.
3. Пол:
   3.1. Муж.
   3.2. Жен.
4. Рост _____ см
7. Количество детей в семье:
   7.1. Один
   7.2. Двое
   7.3. 3 и более
8. Образование матери:
   8.1. Неполное среднее
   8.2. Среднее
   8.3. Специальное среднее
   8.4. Высшее
   8.5. Ученая степень
31. В чем заключается ваше общение с ребенком:
   31.1. Проверка уроков
   31.2. Совместные прогулки
   31.3. Совместные беседы, обсуждения
   31.4. Совместное посещение кинотеатров, театров, музеев, выставок и др.
   31.5. Другое (указать, что)

horse - SPSS for Windows Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window He

7:interval 82

	case	nominal	ordinal	interval
1	1.00	.00	3.00	66.00
2	2.00	.00	2.00	65.00
3	3.00	.00	6.00	73.00
4	4.00	.00	5.00	70.00
5	5.00	1.00	1.00	60.00
6	6.00	.00	4.00	67.00
7	7.00	.00	7.00	82.00

Описательная статистика

- занимается представлением и описанием данных и включает:

- Методы представления данных (таблицы, гистограммы и т.д.)
- Описание массива данных

Описание массива данных

Номинальные и порядковые (ординальные) признаки описываются (обобщаются) путем расчета **доли (пропорции, относительной частоты)**

Описание массива данных

- Единственный способ описать качественные признаки заключается в расчете доли случаев (или пропорции), которая приходится на то или иное значение, от общего числа объектов.

Доля (частота) может быть выражена в процентах:

$$P = \frac{n_1}{n_1 + n_2} \times 100\% = \frac{n_1}{n} \times 100\%$$

где n_1 и n_2 – численности групп (имеющих и не имеющих изучаемый признак), а $n=n_1+n_2$ – численность всей совокупности.

Описание массива данных

- Эти показатели чаще всего используются для характеристики структуры изучаемой совокупности или оценки частоты изучаемого явления в популяции.
- Масштабирующим коэффициентом может быть 100 (%), 1000 (‰), 10 000 (‱), 100 000 (⁗).

Пример

- Был выделен 21 кишечный паразит при обследовании детей:

Giardia lamblia	***Giardia lamblia***	***Giardia lamblia***
Entamoeba histolytica	***Entamoeba histolytica***	***Entamoeba histolytica***
Ascaris lumbricoides	***Ascaris lumbricoides***	***Ascaris lumbricoides***
Enterobius vermicularis	***Enterobius vermicularis***	***Enterobius vermicularis***
Ascaris lumbricoides	***Ascaris lumbricoides***	***Ascaris lumbricoides***
Enterobius vermicularis	***Enterobius vermicularis***	***Enterobius vermicularis***
Giardia lamblia	***Giardia lamblia***	***Giardia lamblia***

Пример

Визуальное упорядочивание

Giardia lamblia	***Ascaris lumbricoides***	***Enterobius vermicularis***
Giardia lamblia	***Ascaris lumbricoides***	***Enterobius vermicularis***
Giardia lamblia	***Ascaris lumbricoides***	***Enterobius vermicularis***
Giardia lamblia	***Ascaris lumbricoides***	***Enterobius vermicularis***
Giardia lamblia	***Ascaris lumbricoides***	***Entamoeba histolytica***
Giardia lamblia	***Enterobius vermicularis***	***Entamoeba histolytica***
Ascaris lumbricoides	***Enterobius vermicularis***	***Entamoeba histolytica***

Пример

- Частотное распределение

Parasite	n
Giardia lamblia	6
Ascaris lumbricoides	6
Enterobius vermicularis	6
Entamoeba histolytica	3
Total	21

Пример

- Распределение относительных частот (долей, пропорций)

Parásito	n	%
Giardia lamblia	6	28.57
Ascaris lumbricoides	6	28.57
Enterobius vermicularis	6	28.57
Entamoeba histolytica	3	14.29
Total	21	100.00

Описание массива данных

Описание (обобщение) количественного признака:

1. **Оценка центральной тенденции**

2. **Оценка разнообразия (разброса, рассеяния).**

ВНИМАНИЕ !

От вида распределения зависит выбор статистического пути описания (обобщения) и анализа количественного признака!

Вид распределения

Под **видом распределения** случайной величины понимают **соответствие**, устанавливаемое между всеми возможными числовыми значениями случайной величины и вероятностями их появления в совокупности.

Вид распределения

Вид (закон) распределения может быть представлен:

- аналитической зависимостью в виде формулы;
- в виде графического изображения;
- в виде таблицы

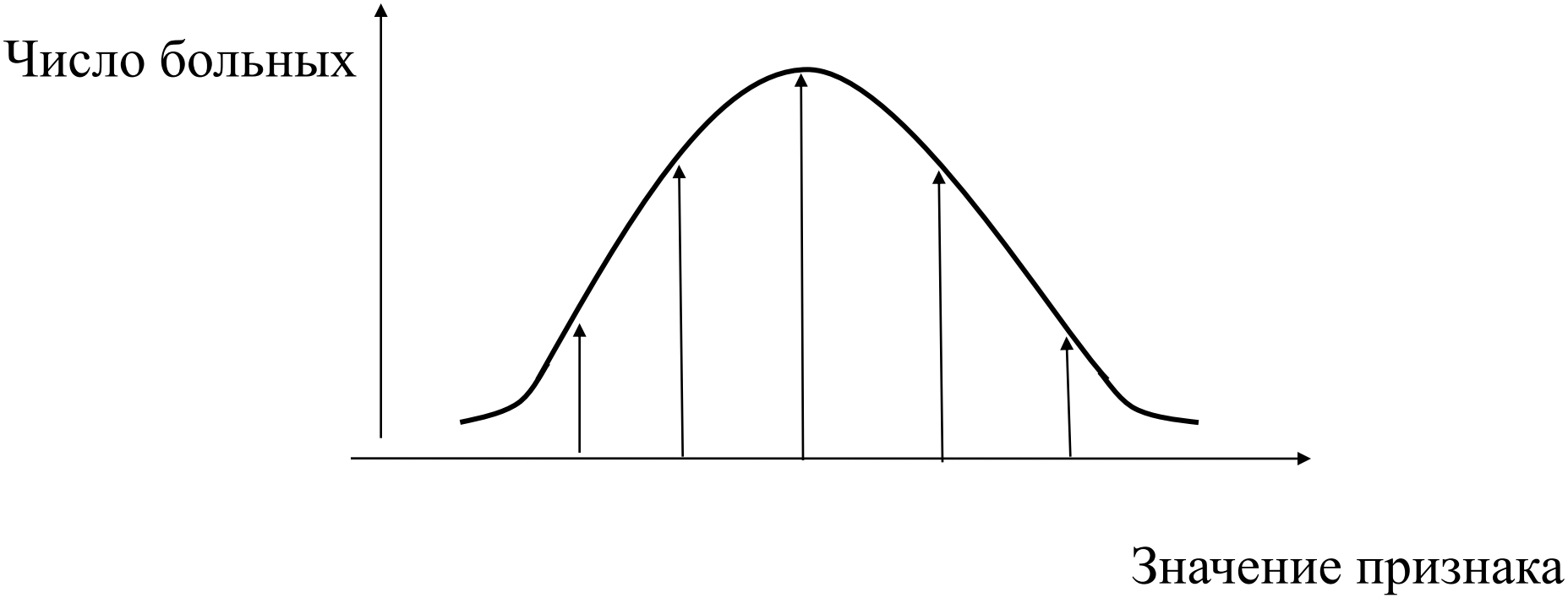
Виды распределения

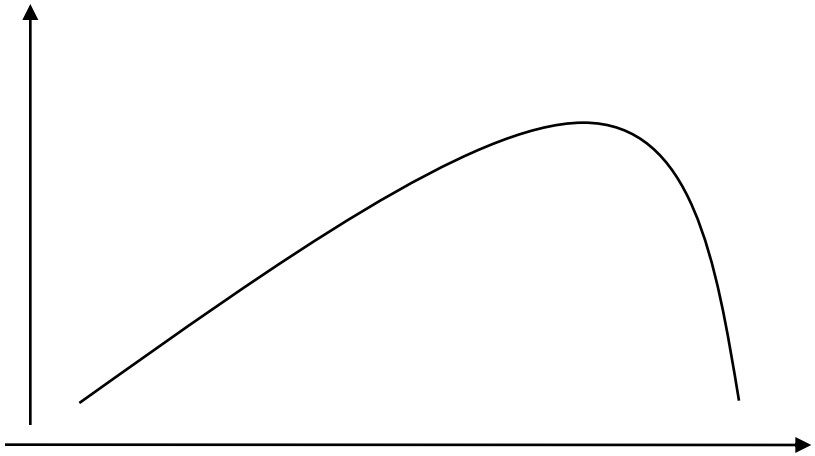
Нормальное (**гауссово**, симметричное, колоколообразное) распределение **(normal, Gaussian distribution)**– описывает совместное воздействие на изучаемое явление небольшого числа случайно сочетающихся факторов (по сравнению с общей суммой факторов), число которых неограничено велико.

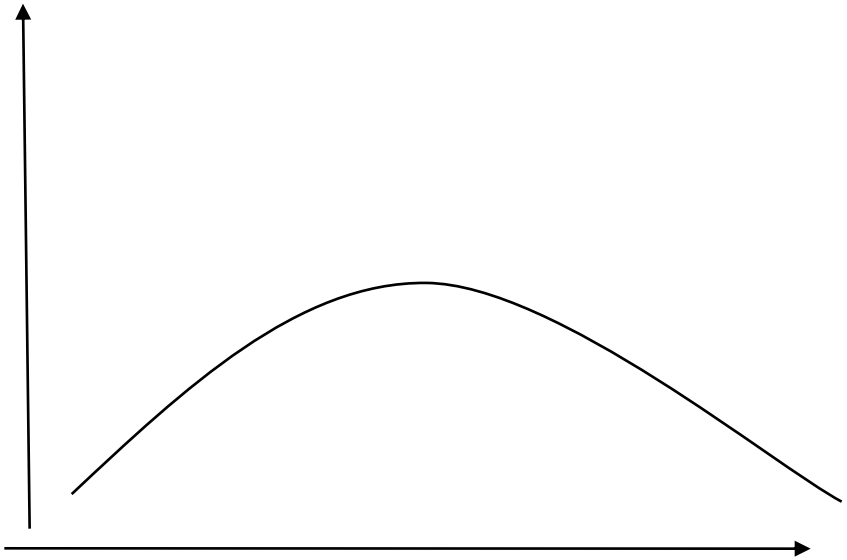
Встречается в природе наиболее часто, за что и получило название «нормального».

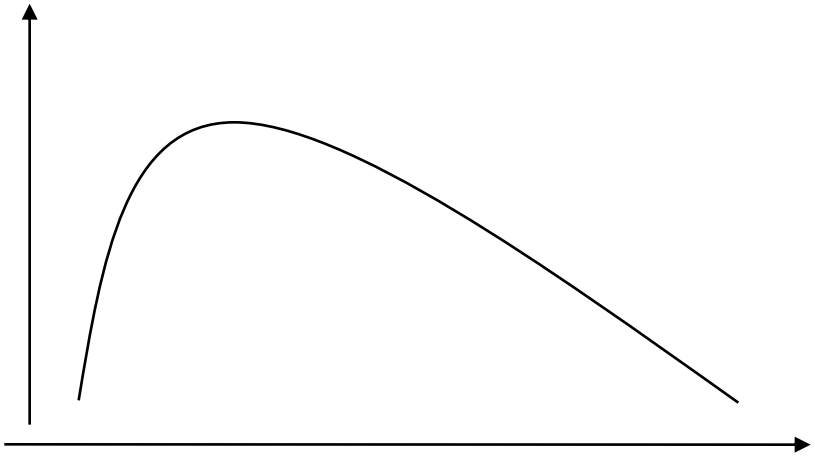
Характеризует распределение **непрерывных** случайных величин.

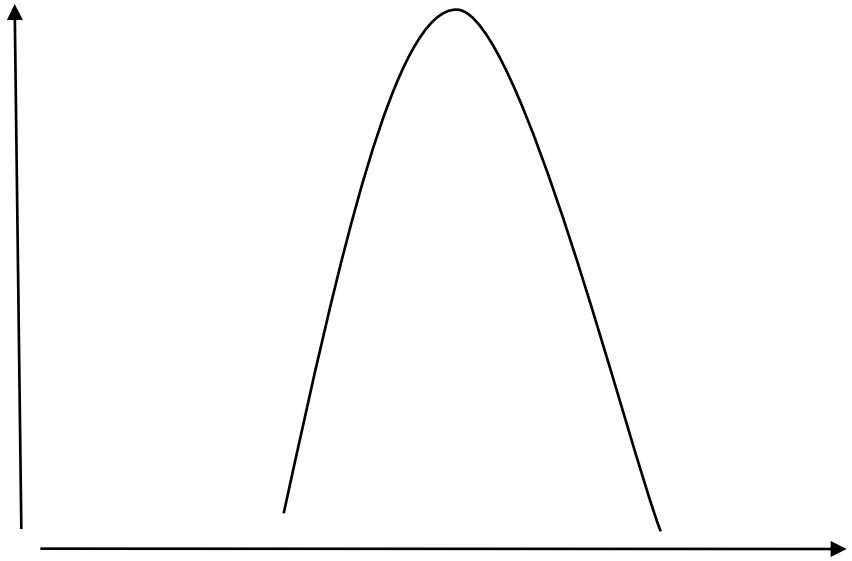
Кривая нормального распределения











Вид распределения

нормальное

отличное от нормального

параметрическая статистика

непараметрическая статистика

Вариационный ряд (frequency table)- ранжированный ряд распределения по величине какого-либо признака.

Этот признак носит название варьирующего, а его отдельные числовые значения называются вариантами и обозначаются через "x".

Число, показывающее, сколько раз данная варианта встречается в вариационном ряду, называется частотой и обозначается через "p"

Пример

Результаты измерения частоты пульса у курящих студентов-медиков в возрасте 20 лет:

68,78,75,68,80,75,70,65,70,68,78,75,70,70,75,75,68,70,
80,82,75,70,78,75,78,75,70,75,68,75,70,70,75,75,68

Построим вариационный ряд:

x	65	68	70	75	78	80	82
p	1	6	9	12	4	2	1

Вариационный ряд можно разбивать на отдельные части, которые называются **квантилями (quantile)**.

Название квантилей	Число частей, на которые разбивается ряд
Медиана	2
Терциль	3
Квартиль	4
Дециль	10
Процентиль	100

ХАРАКТЕРИСТИКИ (МЕРЫ) ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

Показатели, характеризующие **центральную тенденцию (central tendency)** :

- ***средняя величина,***
- ***медиана,***
- ***мода***

ХАРАКТЕРИСТИКИ РАЗБРОСА (РАЗНООБРАЗИЯ)

Показатели, характеризующие **разнообразие** (рассеяние, вариацию, разброс) **(spread)** признака:

- ***размах (амплитуда),***
- ***стандартное отклонение (среднее квадратическое отклонение),***
- ***дисперсия,***
- ***интерквартильный интервал,***
- ***коэффициент вариации***

Выбор характеристик центральной тенденции и разнообразия признака зависит от вида распределения!

ВИД РАСПРЕДЕЛЕНИЯ

нормальное

отличное от нормального

оценка центральной тенденции

средняя арифметическая, мода, медиана

мода, медиана

- **Средняя величина** - обобщающий коэффициент, который характеризует наиболее типичный размер определенного признака в целом для совокупности или для отдельных ее частей.

- Расчет средних величин имеет смысл только для качественно однородной совокупности, в связи с этим в одной совокупности может быть столько средних, на сколько однородных групп она может быть разбита.

Виды средних величин

Средняя арифметическая(mean) - применяется, если варианты возрастают (убывают) в арифметической прогрессии.

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{x} = \frac{\sum (x_i \times p)}{n}$$

$\bar{x}$- средняя арифметическая;

x_i – числовое значение признака (варианта);

p - частота встречаемости числового значения признака (варианты);

n - число наблюдений

Виды средних величин

- **Средняя геометрическая** - вычисляется, если варианты возрастают (убывают) в геометрической прогрессии

$$\bar{x}_g = \sqrt[n]{x_1 x_2 x_3 \ldots x_n}$$

На практике используют логарифмированную формулу:

$$\log \bar{x}_g = 1/n(\log x_1 + \log x_2 + \log x_3 + \ldots \log x_n)$$

Структурные средние

- **Мода (Mo) (mode)-** наиболее часто встречающаяся в вариационном ряду варианта

Мода используется:

- при малом числе наблюдений, когда велико влияние состава совокупности на среднюю ;
- для характеристики центральной тенденции при ассиметричных распределениях, когда велико влияние на среднюю крайних вариант;

Структурные средние

- **Медиана (Me)(median)** - варианта, которая делит вариационный ряд на две равные части.
- При нечетном количестве значений медиана всегда будет совпадать с одним из измеренных значений. При четном количестве медиана будет средним арифметическим двух соседних значений.

Медиана используется:

- при необходимости знать, какая часть вариант лежит выше и ниже срединного значения ;
- для характеристики центральной тенденции при ассиметричных распределениях .

Пример

Результаты измерения частоты пульса у некурящих студентов-медиков в возрасте 20 лет:
68,58,65,55,70,62,60,65,70,58,62,58,62,60,60,65,62,55,62,58,60,70,62,
65,60,68,65,62,68,65,60,62,60,68,65,60,62,60,65,62,68

Построим вариационный ряд (ранжированный ряд распределения по величине какого-либо признака):

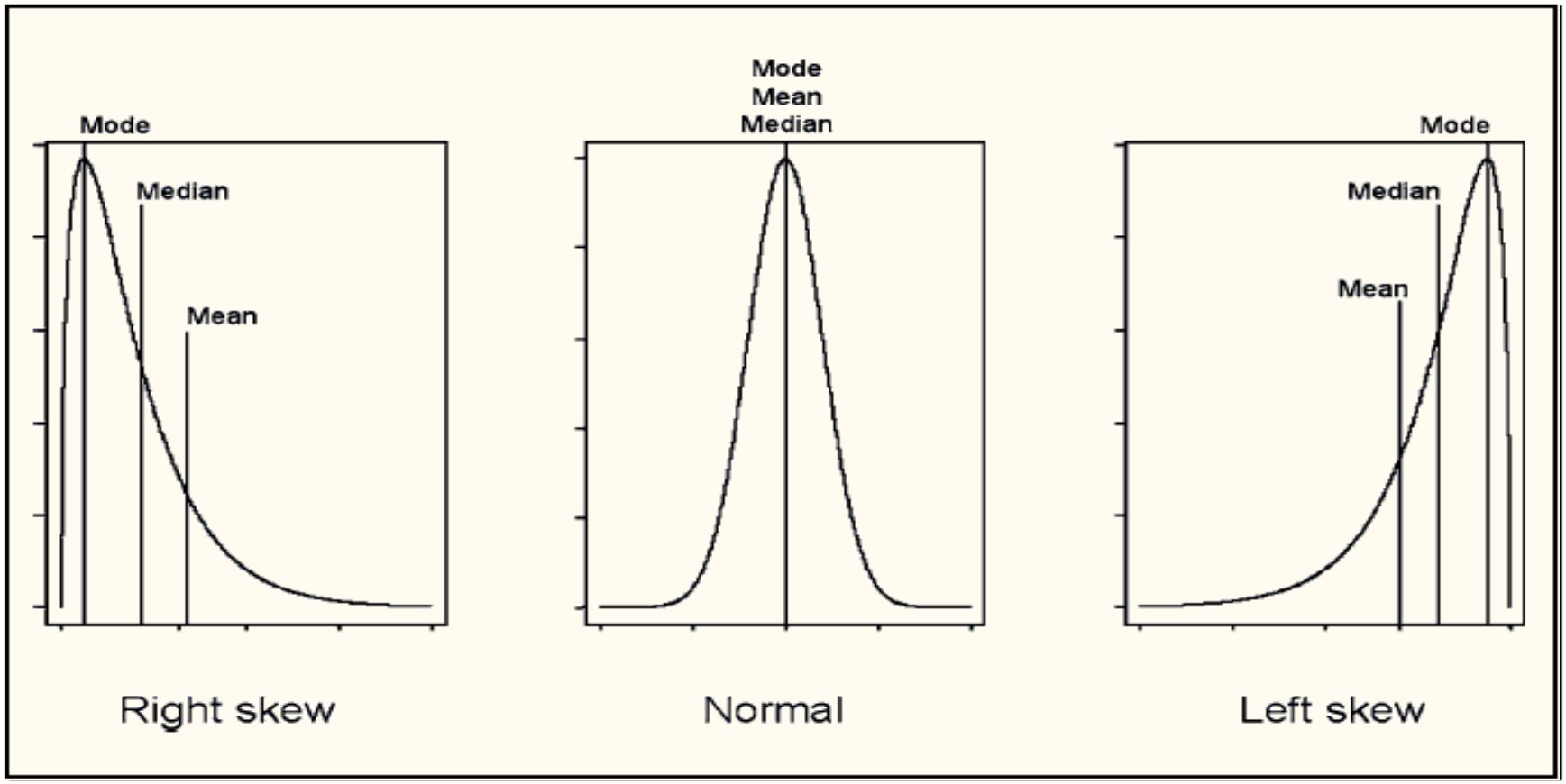
x	65	68	70	75	78	80	82
p	1	6	9	12	4	2	1

$$\bar{x} = \frac{\sum(x_i \times p)}{n}$$

Mean = 73,1 уд.в мин.

Moda = 75 уд.в мин.

Median = 75 уд.в мин.


Mode
Median
Mean
Right skew
Mode
Mean
Median
Normal
Mode
Median
Mean
Left skew

Средняя может вводить в заблуждение

- Group 1 data: 1,1,1,2,3,3,5,8,20
  - Mean: 4.9 Median: 3 Mode: 1
- Group 2 data: 1,1,1,2,3,3,5,8,10
  - Mean: 3.8 Median: 3 Mode: 1
- Когда объем совокупности небольшой, единичные значения, резко отличающиеся по своей величине от остальных, оказывают большое влияние на размер средней, при этом практически не влияя на моду и медиану.
- В этом случае мода и медиана более информативны, чем средняя.

Характеристики разнообразия вариационного ряда

Размах вариации (амплитуда) (range)

$$A = X_{max} - X_{min}$$

$$A = 82 - 65 = 17 \text{ (уд.в мин.)}$$

x	65	68	70	75	78	80	82
p	1	6	9	12	4	2	1

Характеристики разнообразия (разброса)

Стандартное отклонение

(среднее квадратическое отклонение)

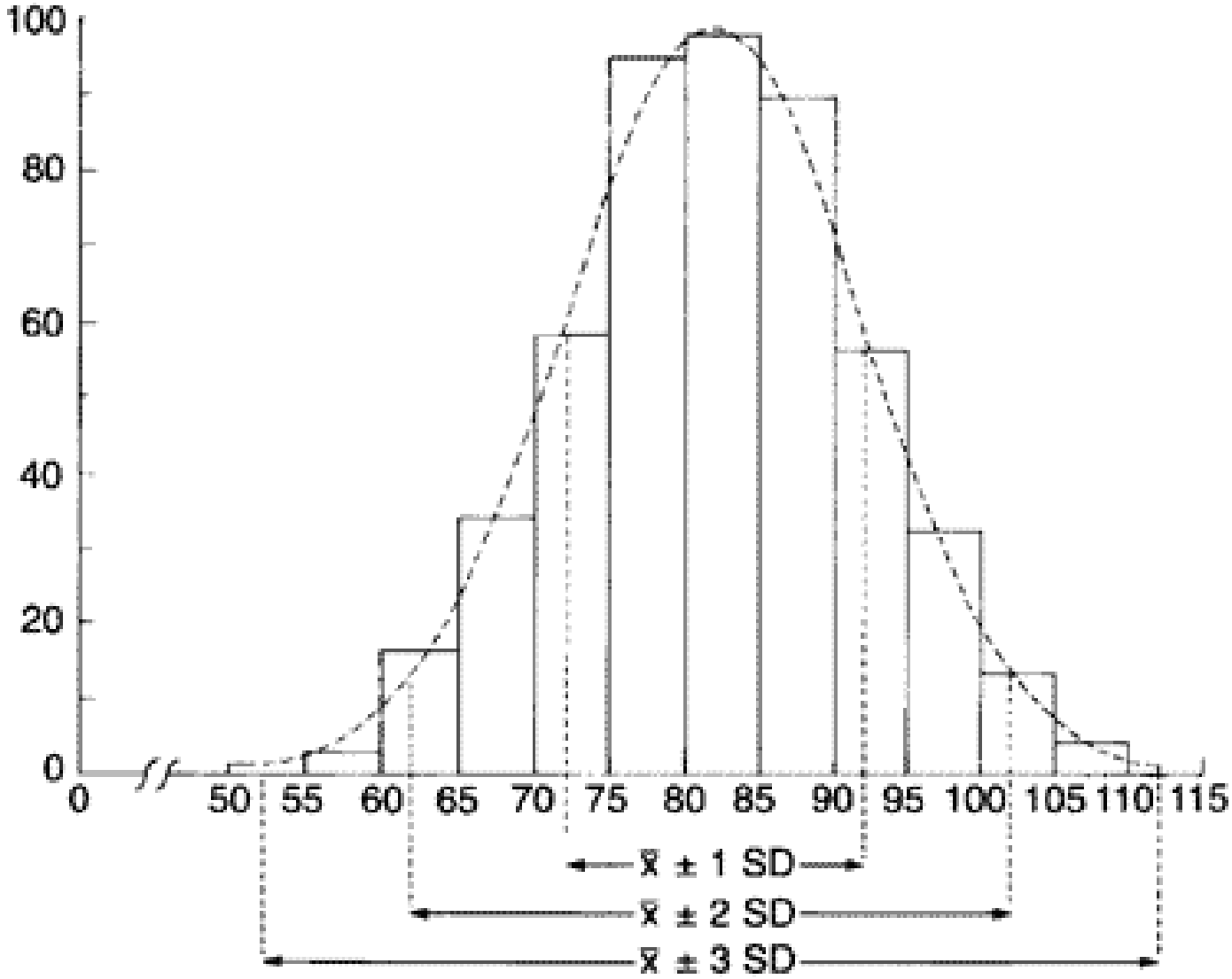
(standard deviation, SD)

$$\delta\ (SD) = \pm \sqrt{\frac{\sum (x_i - \overleftarrow{x})^2 \cdot p}{n-1}}$$

Правило трех сигм

При нормальном распределении

- 68.3 % всех вариант отклоняются от средней не более, чем на 1σ;
- 95.4% вариант находятся в пределах $\bar{x} \pm 2\sigma$;
- 99.7% вариант находятся в пределах $\bar{x} \pm 3\sigma$


100
80
60
40
20
0
0
50
55
60
65
70
75
80
85
90
95
100
105
110
115
x̄ ± 1 SD
x̄ ± 2 SD
x̄ ± 3 SD

Наиболее распространенные ошибки: средняя арифметическая используется для характеристики признаков с «анормальным» распределением или для порядковых признаков

Уровень глюкозы 8,2 ± 7,5 ммоль/л

Выраженность боли: 2,5 ± 1,2 балла

(1 – слабая, 2 – средняя, 3 – сильная)

- Возраст больных составлял от 18 до 68 лет (средний возраст - 22,8±4,2 года).

  3 сигмы=12,6 ; 10,2 $\leftrightarrow$ 35,4

- Сроки поступления больных составили от 1 до 9 дней (в среднем 2,2±1,4 дня).

  3 сигмы=4,2; -2 $\leftrightarrow$ 6,4

Характеристики разнообразия (разброса)

- **Дисперсия (варианса) (variance)**

$$\delta^2 = \frac{\sum(x_i - \bar{x})^2 \cdot p}{n-1}$$

Характеристики разнообразия (разброса)

- **Коэффициент вариации (variation coefficient)**

$$C_v = \frac{\delta}{\bar{x}} \cdot 100\ \%$$

Характеристики разнообразия (разброса)

Вариационный ряд

- считается однородным при C_v <10 % ,
- обладающим средней вариабельностью (разнообразием) при C_v =10-15%
- обладающим значительной вариабельностью при C_v >15% .

Характеристики разнообразия (разброса)

- Коэффициент вариации используется при сравнении вариационных рядов, имеющих различную размерность,

  или одной размерности, но обладающими резкими различиями в своих значениях, затрудняющими их сопоставление.

Характеристики разнообразия (разброса)

- **Интерквартильный интервал (inter-quartile range, IQR)**

Интерквартильный интервал

- Вариационный ряд разбивают на четыре интервала, получая, соответственно, 25%, 50% и 75% квантили;
- 25% и 75% квантили называют также нижним (low quartile) и верхним квартилями(high quartile).
- 50% квантиль – это медиана.
- Внутри интерквартильного интервала (между 25% и 75% квантилями) лежат 50% наиболее типичных (близких к центральному) значений.

105 112 125 132 155 158 160 165 168 170 175 185 185 235

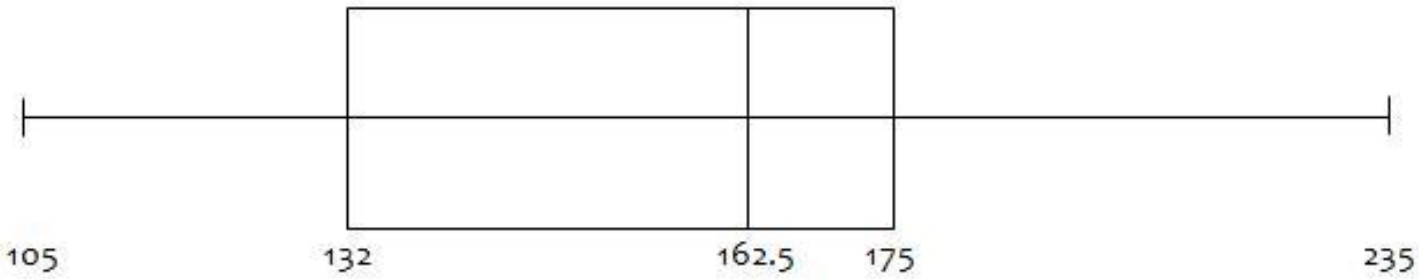
$$\text{Median} = \frac{160 + 165}{2} = 162.5$$

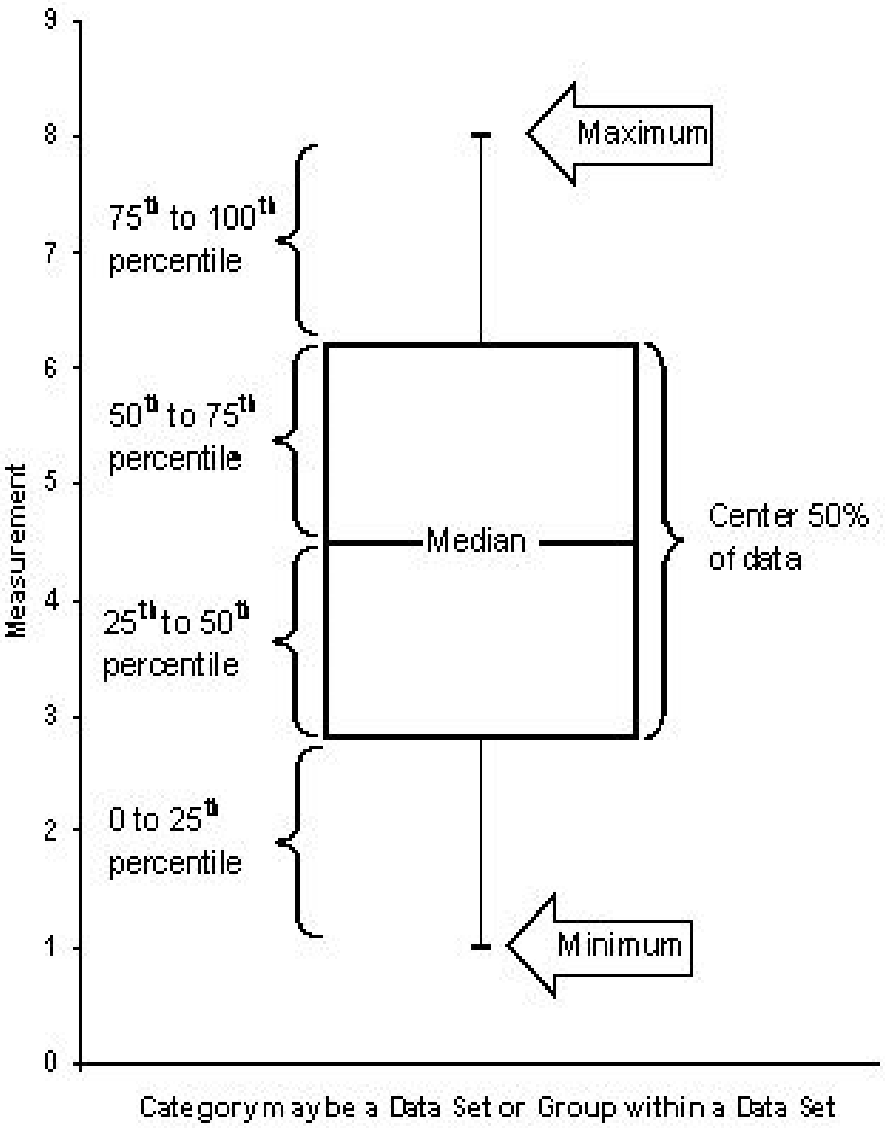
105 112 125 132 155 158 160 | 165 168 170 175 185 185 235

Quartile 1, Q_1

Quartile 3, Q_3

Interquartile range (IQR): $Q_3 - Q_1$; 175 – 132





Первый квартиль (Q_1) — это точка на шкале измеренных значений, ниже (левее) которой располагаются 25 % измеренных значений.

Второй квартиль (Q_2) — это точка, ниже (левее) которой располагаются 50 % измеренных значений. Второй квартиль также называется **медианой.**

Третий квартиль (Q_3) — это точка на шкале измеренных значений, ниже (левее) которой располагаются 75 % значений.

Пример

- Group 1 data: 1,1,1,2,3,3,5,8,20
  - Mean: 4.9 Median: 3
- Group 2 data: 1,1,1,3,3,3,5,8,10
  - Mean: 3.8 Median: 3
- SDs: group 1: 6.1 group 2: 3.2
- Interquartile range: 1 - 5

Вид распределения

нормальное

отличное от нормального

Оценка разнообразия (разброса)

стандартное отклонение SD

интерквартильный интервал IQR

В случае **нормального распределения** признак описывается **средней величиной и стандартным отклонением.**

Если **распределение неизвестно или оно отлично от нормального** центральную тенденцию и разброс можно описать с помощью **медианы, нижнего и верхнего квартиля (интерквартильным интервалом).**

Проверка нормальности распределения

- Если Me занимает срединное положение между 25-м и 75-м процентилем, то распределение близко к нормальному.

Проверка нормальности распределения

Тесты на нормальность:

- Шапиро-Вилка (Shapiro-Wilk)
- Колмогорова-Смирнова (Kolmogorov-Smirnov)
- Крамера-вон Майса (Kramer-von Mises)
- Андерсона-Дарлинга (Anderson-Darling)

ЗАДАНИЯ ДЛЯ АСПИРАНТОВ В ЧЕК-ЛИСТ ПО МЕТОДАМ СТАТИСТИЧЕСКОГО АНАЛИЗА (лекция 1)

Какими характеристиками описываются центральная тенденция и разброс в случае нормального распределения

А. Средней арифметической и стандартным отклонением
Б. Модой, медианой и интерквартильным интервалом
В. Средней геометрической, размахом, амплитудой
Г. Средней квадратической и случайной ошибкой
Д. Средней алгебраической и доверительным интервалом

Какими характеристиками описываются центральная тенденция и разброс в случае распределения, отличного от нормального

А. Средней арифметической и стандартным отклонением
Б. Модой, медианой и интерквартильным интервалом
В. Средней геометрической, размахом, амплитудой
Г. Средней квадратической и случайной ошибкой
Д. Средней алгебраической и доверительным интервалом

Приведены результаты измерения частоты пульса у некурящих студентов-медиков в возрасте 20 лет:
68,58,65,55,70,62,60,65,70,58,62,58,62,60,60,65,62,55,62,58,60,70,62,65,60,68,65,62,68,65,60,62,60,68,65,60,62,60,65,62,68
Вычислите среднюю арифметическую, моду, медиану, стандартное отклонение, интерквартильный интервал. Нарисуйте box-plot.

Спасибо!
Вопросы?