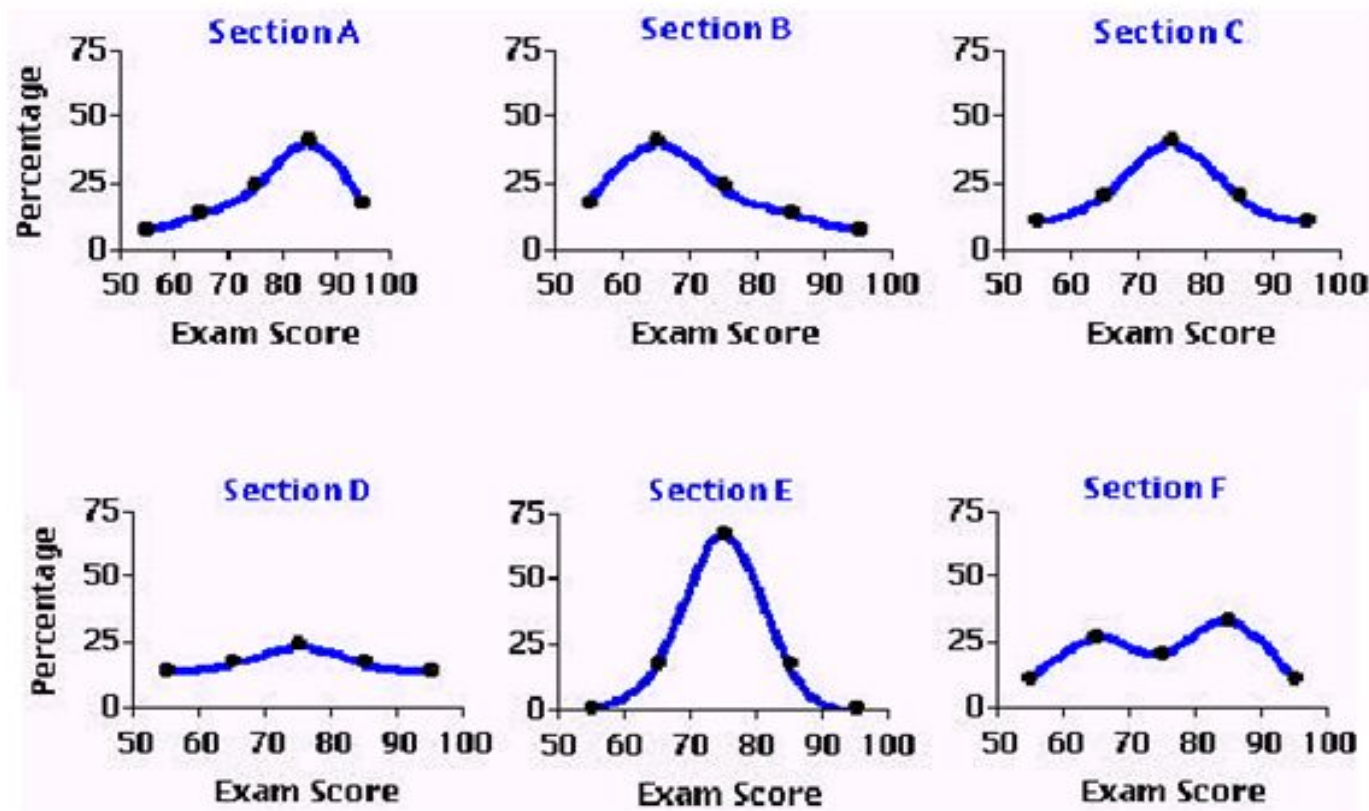


Описательная статистика

Параметры распределения

Асимметрия, эксцесс, модальность



Распределение оценок студентов по разным разделам дисциплины:

А – отрицательная асимметрия, В – положительная асимметрия, С – симметричное распределение, D – отрицательный эксцесс. E – положительный эксцесс. F –

Параметры главной тенденции:

*«Каково типичное значение
признака для данного
распределения?»*

- Среднее значение
- Мода
- Медиана

Среднее значение

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Значения	x_1	x_2	...	x_k
Частоты	p_1	p_2	...	p_k

$$\bar{X} = \frac{\sum_{i=1}^k x_i p_i}{n}$$

Медиана (Me)

Для нахождения медианы необходимо **упорядочить** выборку по возрастанию и найти элемент, стоящий посередине вариационного ряда

Если n – нечетное число, то медианой будет элемент с номером $i = (n+1)/2$ в упорядоченном по возрастанию ряду. Например, в выборке объемом 7 медианой будет 4 элемент вариационного ряда:

3,1 3,8 4,2 5,7 6,3 7,2 7,9 $Me = x_4 = 5,7$

Если n – четное число, то медианой будет среднее значение двух элементов вариационного ряда с номерами $i = n/2$ и $j = n/2 + 1$.

Например, при $n=10$ медианой будет среднее арифметическое 5 и 6 элементов вариационного ряда:

3,1 3,8 4,2 5,7 6,3 7,5 7,9 8,4 8,5 9,2

$$Me = (x_5 + x_6)/2 = (6,3 + 7,5)/2 = 6,9$$

Параметры разброса

Определяют различия в значениях признака у разных объектов

- Размах вариации
- Дисперсия
- Стандартное отклонение
- Коэффициент вариации

Дисперсия

Выборочная дисперсия:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$$

$$s^2 = \frac{\sum_{i=1}^k p_i (x_i - \bar{X})^2}{n-1}$$

Дисперсия генеральной совокупности:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

- Стандартное отклонение

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}}$$

- Коэффициент вариации

$$V = \frac{s}{\bar{X}} \cdot 100\%$$

- $V < 33\%$  выборка однородная

Стандартная ошибка среднего

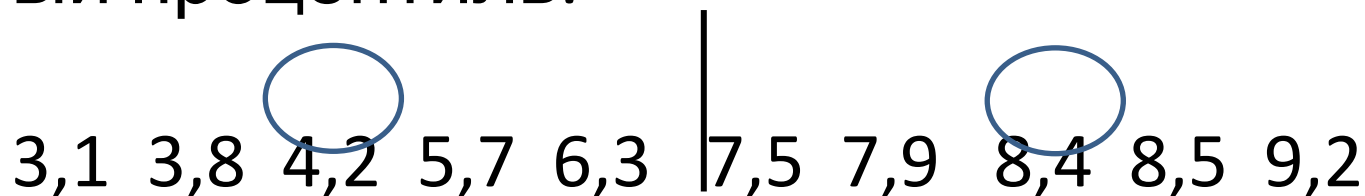
Разные выборки дают разные оценки параметров распределения. Для характеристики точности выборочных оценок используют *стандартную ошибку среднего*:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Не является параметром разброса, только показывает точность оценки среднего. Чем больше выборка, тем меньше ошибка и выше точность

Процентили

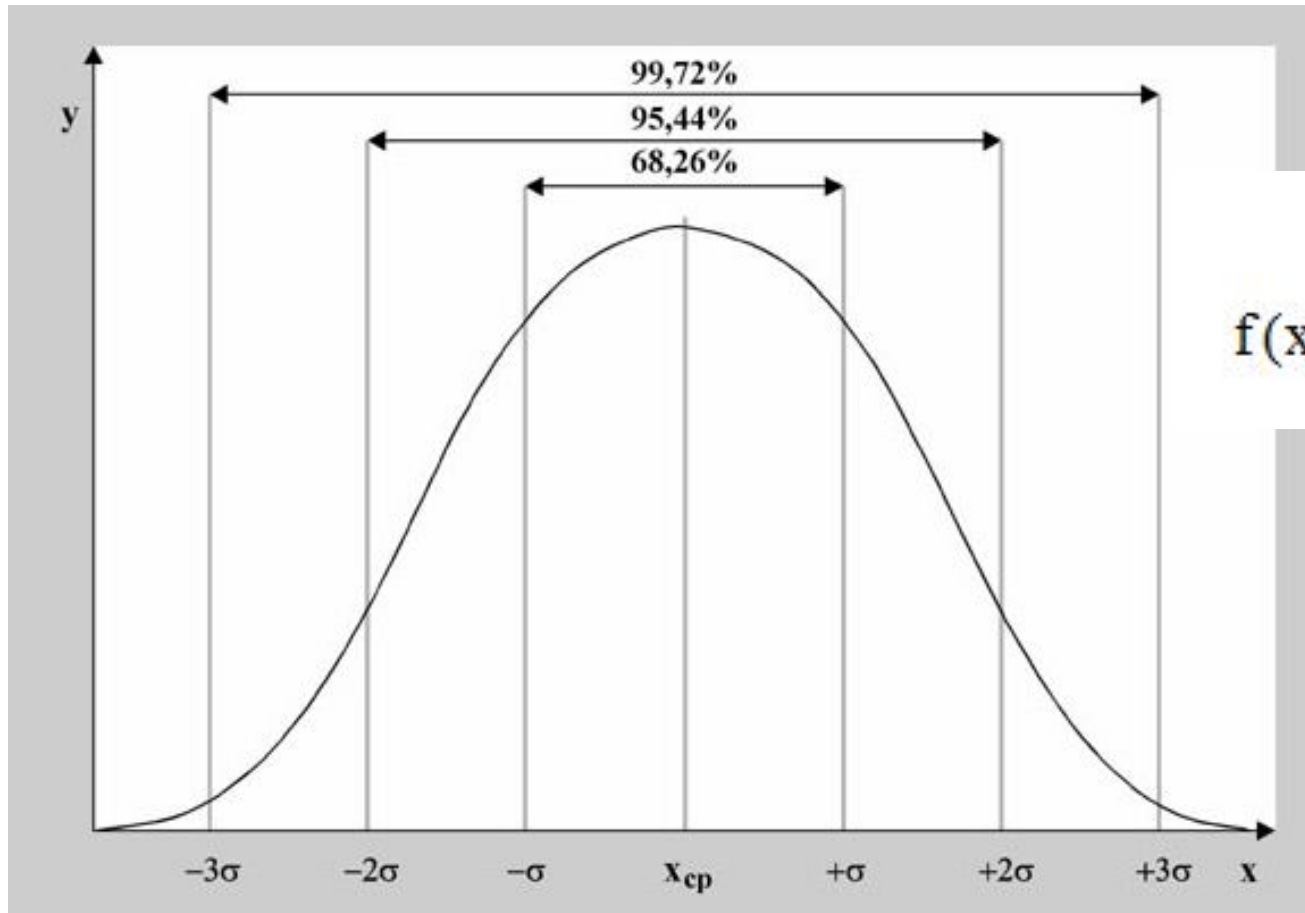
25-ый и 75-ый процентили (квартили) отсекают от распределения по четверти, т.е. одна четверть значений распределения будет не больше 25-го процентиля, а одна четверть – больше 75-го процентиля. Медиана – это 50-ый процентиль.



$$25\% = 4,2$$

$$75\% = 8,4$$

Нормальное распределение

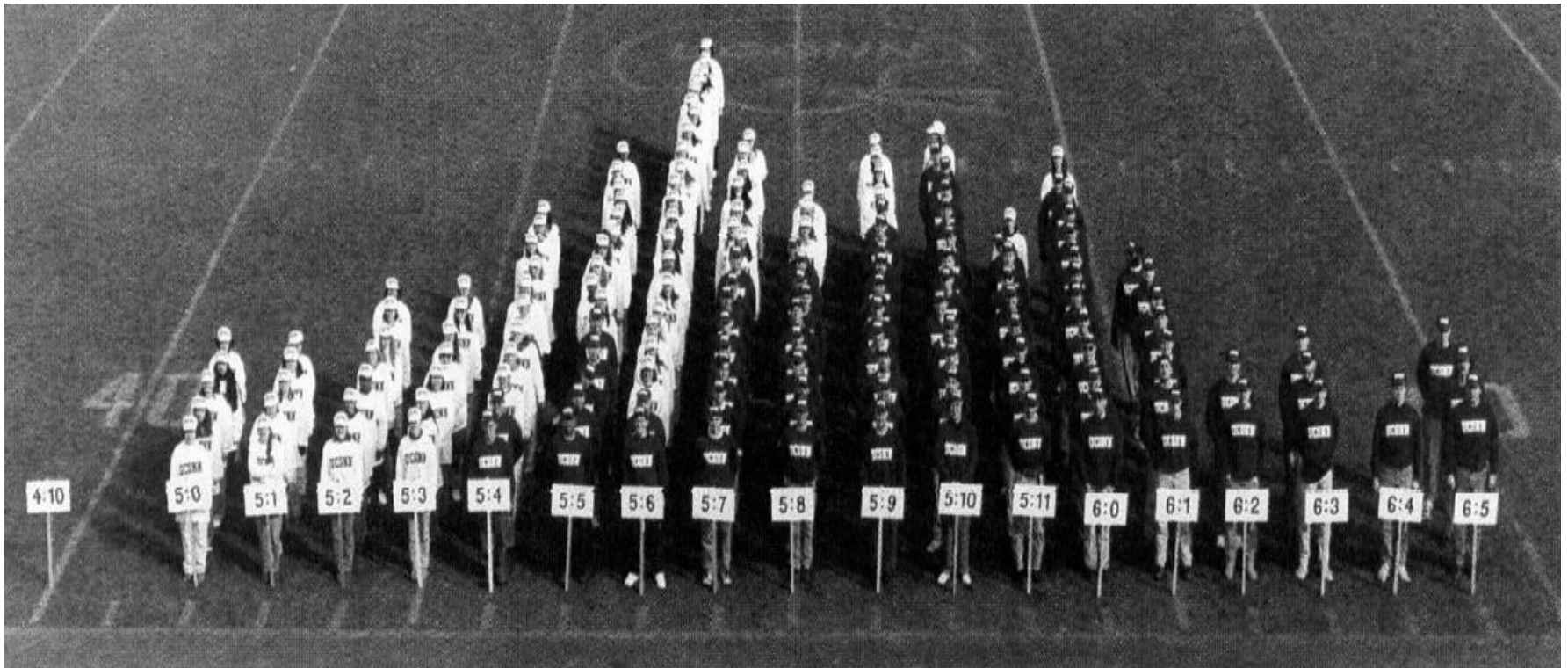


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

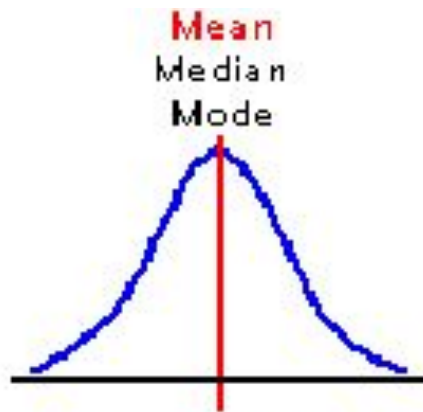
Свойства нормального распределения

- Полностью определяется средним значением и стандартным отклонением
- Мода, медиана и среднее значение совпадают
- Среднее значение характеризует положение кривой распределения и место ее максимума
- Стандартное отклонение характеризует форму кривой
- Зная среднее и стандартное отклонение, ориентировочно можно указать интервал практически всех значений изучаемой величины.

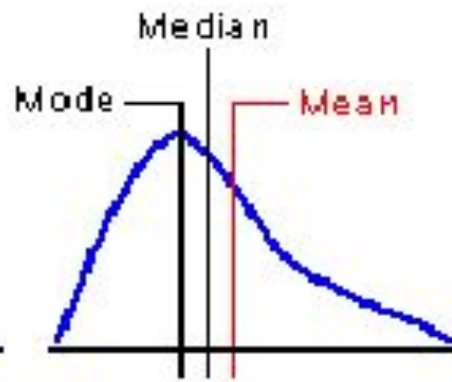
Распределение по росту



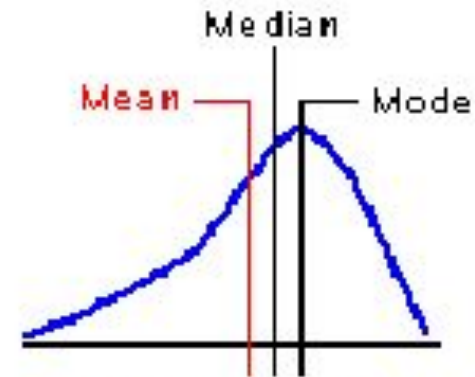
Симметричное и асимметричные распределения



Нормальное
распределение



Положительное
смещение



Отрицательное
смещение

Способы проверки соответствия распределения нормальному закону

1) Способы, основанные на визуальной оценке близости распределения признака к нормальному:

- построение гистограммы распределения признака
- построение графика функции распределения признака

2) Вычисление коэффициентов асимметрии и эксцесса. Для нормального распределения эти показатели равны 0.

3) Вычисление среднего, моды, медианы и процентилей

4) Статистические критерии для проверки нормальности распределения (Пирсона, Колмогорова-Смирнова,

Проверка соответствия распределения нормальному закону

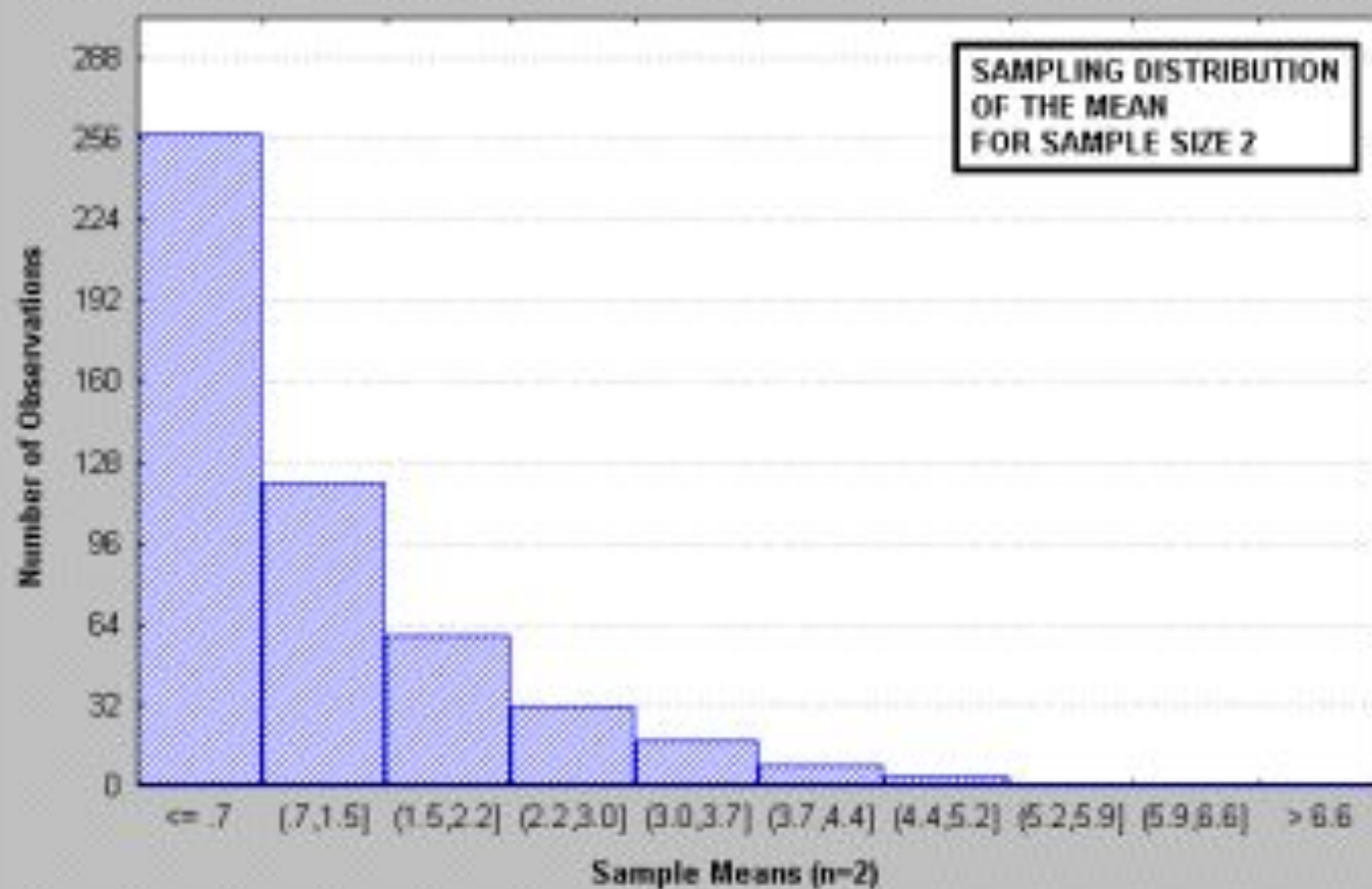
- 1) выборочные среднее, медиана и мода должны быть близки по значению и находиться примерно посередине между 25 и 75 процентилями;
- 2) интервал среднее $\pm$ два стандартных отклонения должен включать примерно 95% значений выборки и не должен содержать много значений, которых не может быть в данном распределении (*например, отрицательных, если речь идет о данных, которые могут принимать только положительные значения*).

Часто ли встречается нормальное распределение?

- Можно сказать, что из всех распределений в природе чаще всего встречается именно нормальное распределение – отсюда и произошло его название.
- Но для данных биомедицинских исследований это не всегда верно. Нормальное распределение встречается в биомедицинских признаках примерно в 20-25% (???).
- До тех пор пока выборка достаточно большая (например, 30 (100) или больше наблюдений), можно считать, что выборочное распределение нормально (???).

Sample Size of 2.5TG: Histogram

THE SAMPLING DISTRIBUTION BECOMES NORMAL AS THE SAMPLE SIZE INCREASES



Как правильно использовать параметры распределения для описания данных?

- Купе № 1: пассажиры возраста 19, 20, 21 год
- Купе №2: пассажиры возраста 54, 2 и 4 года

***Каков средний возраст пассажиров
каждого купе?***

[illegible]

Среднее = 7,

Медиана = 1,

Мода = 1,

25 процентиль = 1,

75 процентиль = 1.

Описание количественных данных в зависимости от вида их

распределения

- Для описания выборочного **нормального распределения** количественных признаков необходимо указывать: число наблюдений, среднее значение, стандартное отклонение.
- Для описания выборочного распределения количественных признаков, которое **отличается от нормального**, рекомендуется указывать: число наблюдений, медиану, 25 и 75 процентиля (нижний и верхний квартили).

1: 21, 22, 22, 23, 23, 24, 24, 24, 25, 25, 25, 25, 26, 26, 26, 26,
27, 27, 28, 29, 30

2: 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 21, 21, 21, 21, 21, 33,
34, 34, 36, 37, 42

$$n_1 = n_2 = 21$$

Среднее $_1 = 25,14$;

Ст. отклон. $_1 = 2,31$;

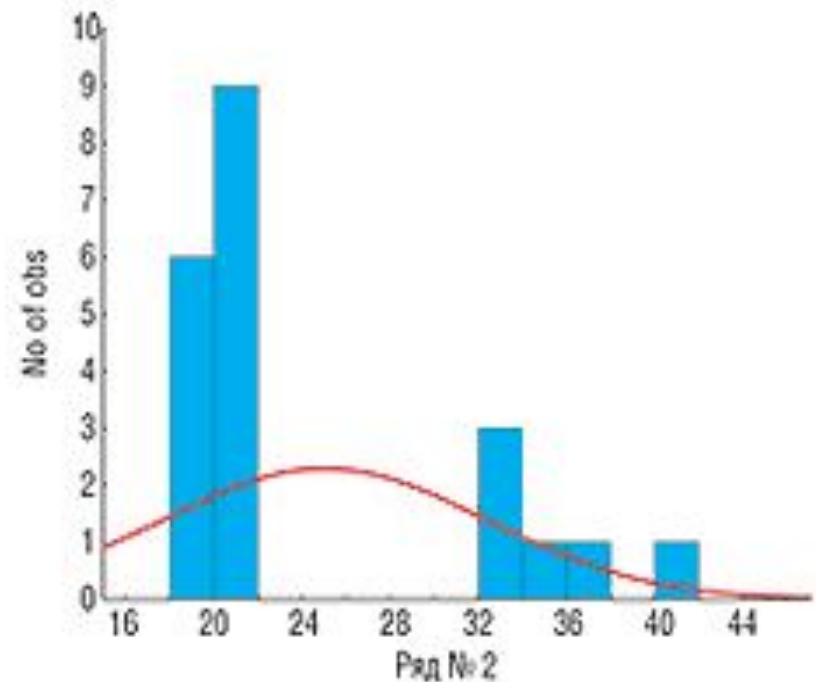
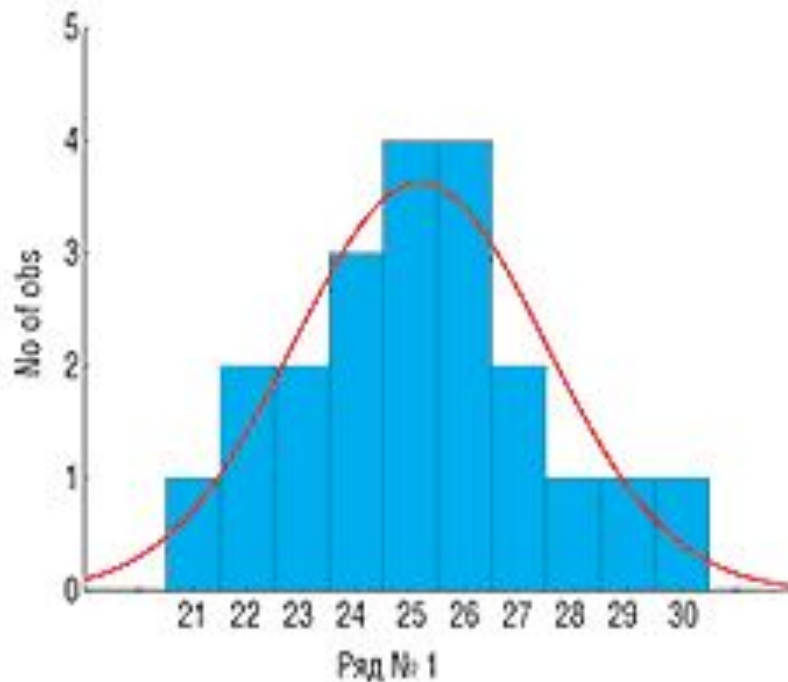
Медиана = 25; Мода=25 и 26

Среднее $_2 = 25,00$;

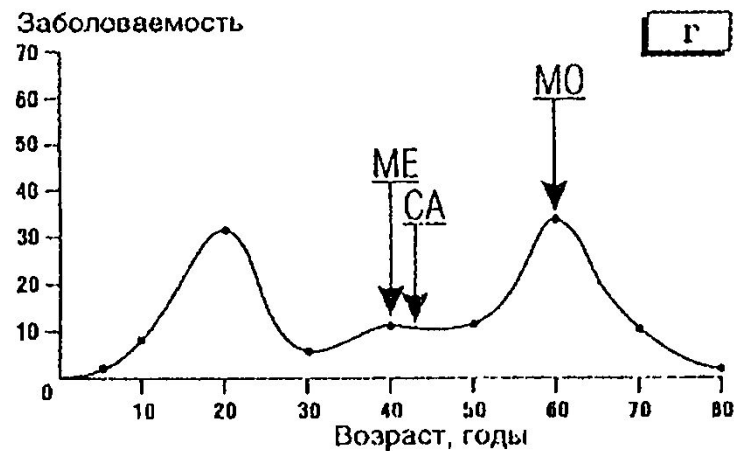
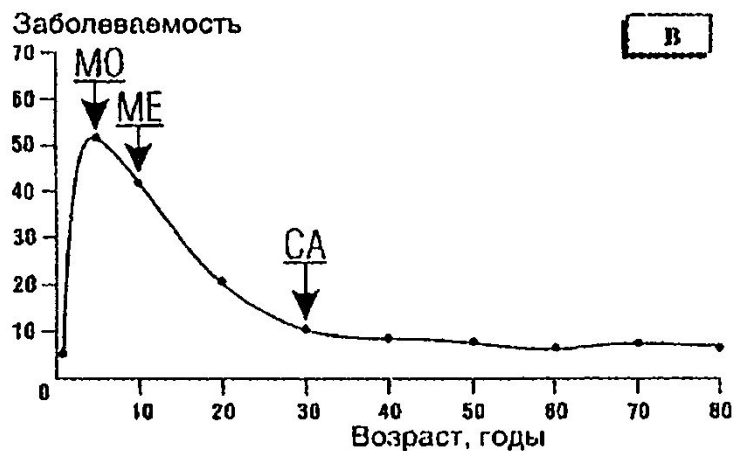
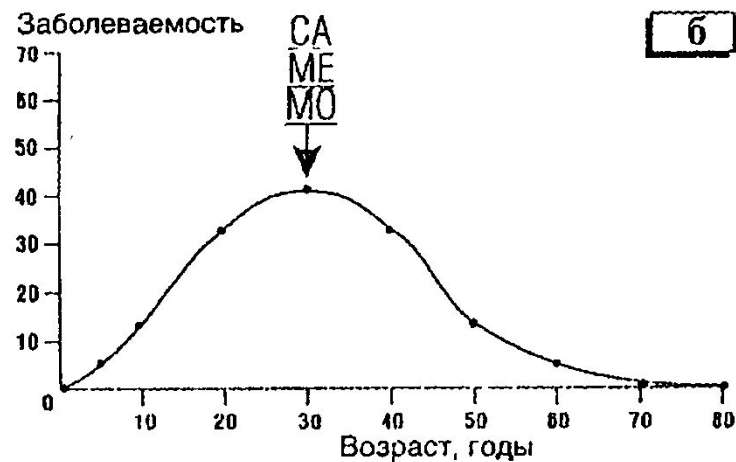
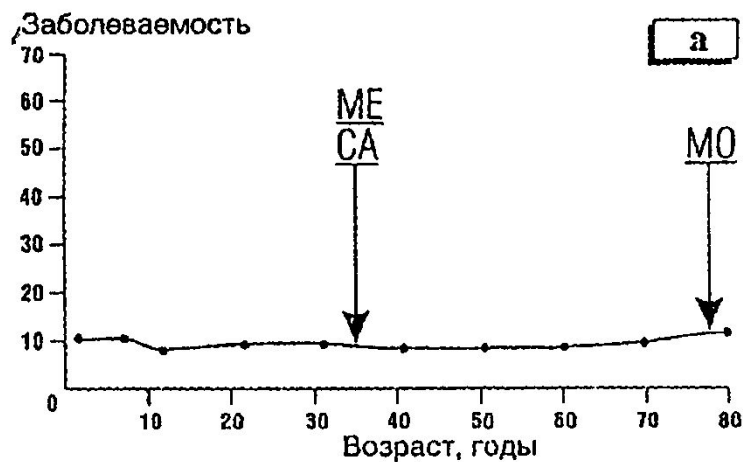
Ст. отклон. $_2 = 7,32$;

Медиана = 21; Мода 21

Визуальное представление 1 и 2 распределения



Примеры взаимного расположения параметров для разных видов распределений



Пример

- Найти параметры следующего выборочного распределения (клинические оценки тяжести серповидноклеточной анемии):
- 0 0 0 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 4 4 5 5
5 5 6 7 9 10 11
- Можно ли считать, что выборка извлечена из совокупности с нормальным распределением?

Таблица для расчета параметров распределения

Значе- ния x_i	Частоты p_i	Накоплен- ные частоты	$x_i \cdot p_i$	$x_i - \bar{X}$	$(x_i - \bar{X})^2$	$p_i \cdot (x_i - \bar{X})^2$
0	3	3	0	-3,09	9,55	28,65
1	11	14	11	-2,09	4,37	48,07
2	4	18	8	-1,09	1,19	4,76
3	4	22	12	-0,09	0,01	0,04
4	2	24	8	0,91	0,83	1,66
5	4	28	20	1,91	3,65	14,6
6	1	29	6	2,91	8,47	8,47
7	1	30	7	3,91	15,29	15,29
9	1	31	9	5,91	34,93	34,93
10	1	32	10	6,91	47,75	47,75
11	1	33	11	7,91	62,57	62,57
	$n = \sum p_i = 33$		$\sum = 102$			$\sum = 266,8$

$$n = 33$$

$$Mo = 1 \quad (p = 11)$$

$$Me = x_{(33+1)/2} = x_{17} = 2$$

$$n/4 = 33/4 = 8,25 \approx 8$$

$$25\% = x_8 = 1$$

$$3/4 = 3 \cdot 33/4 = 24,75 \approx 25$$

$$75\% = x_{25} = 5$$

$$\bar{X} \pm 2 \cdot s$$

$$3,09 - 2 \cdot 2,89 = -2,69;$$

$$3,09 + 2 \cdot 2,89 = 8,87$$

Интервал: $-2,69 \div 8,87$

$$\bar{X} = \frac{\sum x_i p_i}{n} = \frac{102}{33} = 3,09;$$

$$s^2 = \frac{\sum_{i=1}^k p_i (x_i - \bar{X})^2}{n - 1} = \frac{266,8}{32} = 8,34;$$

$$s = \sqrt{s^2} = \sqrt{8,34} = 2,89;$$

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{2,89}{\sqrt{33}} = 0,5;$$

$$V = \frac{s}{\bar{X}} \cdot 100\% = \frac{2,89}{3,09} \cdot 100\% = 93,5\%$$

Проверка нормальности

- 1) Среднее, медиана и мода не совпадают, не находятся посередине между 25 и 75-м процентилями
- 2) Около четверти значений интервала $\text{среднее} \pm \text{два стандартных отклонения}$ имеют отрицательный знак, а в исходной выборке по самой природе изучаемого признака не может быть отрицательных значений



Выборка вряд ли извлечена из совокупности с нормальным законом распределения

