Основные понятия математической статистики

Тишков Артем Валерьевич, к.ф.-м.н., доцент
Микрюкова Надежда Николаевна

Основные понятия математической статистики.

Математическая статистика – это раздел математики о методах регистрации, систематизации и анализа статистических экспериментальных данных, полученных в результате наблюдения массовых случайных явлений.

Статистическая совокупность – это множество объектов, обладающих общими признаками, которые являются наиболее важными (типичными) для характеристики этих объектов.

Серия измерений какого либо признака совокупности – это совокупность значений случайной величины.

Объём совокупности N –это число членов совокупности.

Генеральная совокупность – это совокупность всех объектов, которые имеют типичную характеристику или признак. Это все возможные значения случайной величины.

Выборочная совокупность (выборка) – это отобранная тем или иным способом часть генеральной совокупности.

Из одной генеральной совокупности можно отбирать сколь угодно много выборок, главное, чтобы выборка была репрезентативной (представительной), а для этого элементы выборки должны отбираться случайным образом.

Варианта – это числовое значение изучаемого признака(отдельные значения случайной величины).

Основные задачи, которые стоят перед математической статистикой:

1. Определение закона распределения случайной величины по имеющимся статистическим данным (по выборке – закон распределения для всей генеральной совокупности).

2. Определение неизвестных параметров распределения (по выборке оценить параметры генеральной совокупности).

3. Задача проверки правдоподобия выдвигаемых статистических гипотез.

Схема предварительной обработки экспериментальных данных.

1) Сбор экспериментальных данных.

Чтобы определить закон распределения случайной величины, нужно провести серию измерений или подсчётов для интересующей нас случайной величины (признака).

В результате получаем статистический ряд – это совокупность числовых данных или выборка объёмом **n**:

Затем производят упорядочивание членов выборки – эта операция называется ранжирование.

Ранжирование -- это расположение всех имеющихся вариант по возрастанию. Получаем *ранжированный статистический ряд*.

Пример:

При измерении частоты пульса у 10 пациентов получены следующие результаты:
90, 110, 65, 80, 90, 60, 70, 80, 70, 80

Ранжированный ряд имеет вид: 60, 65, 70, 70, 80, 80, 80, 90, 90, 110.

Колебания изучаемого признака называются варьирование. В нашем примере варьирование - это изменение частоты пульса.

Схема предварительной обработки экспериментальных данных.

2) Составление вариационного ряда.

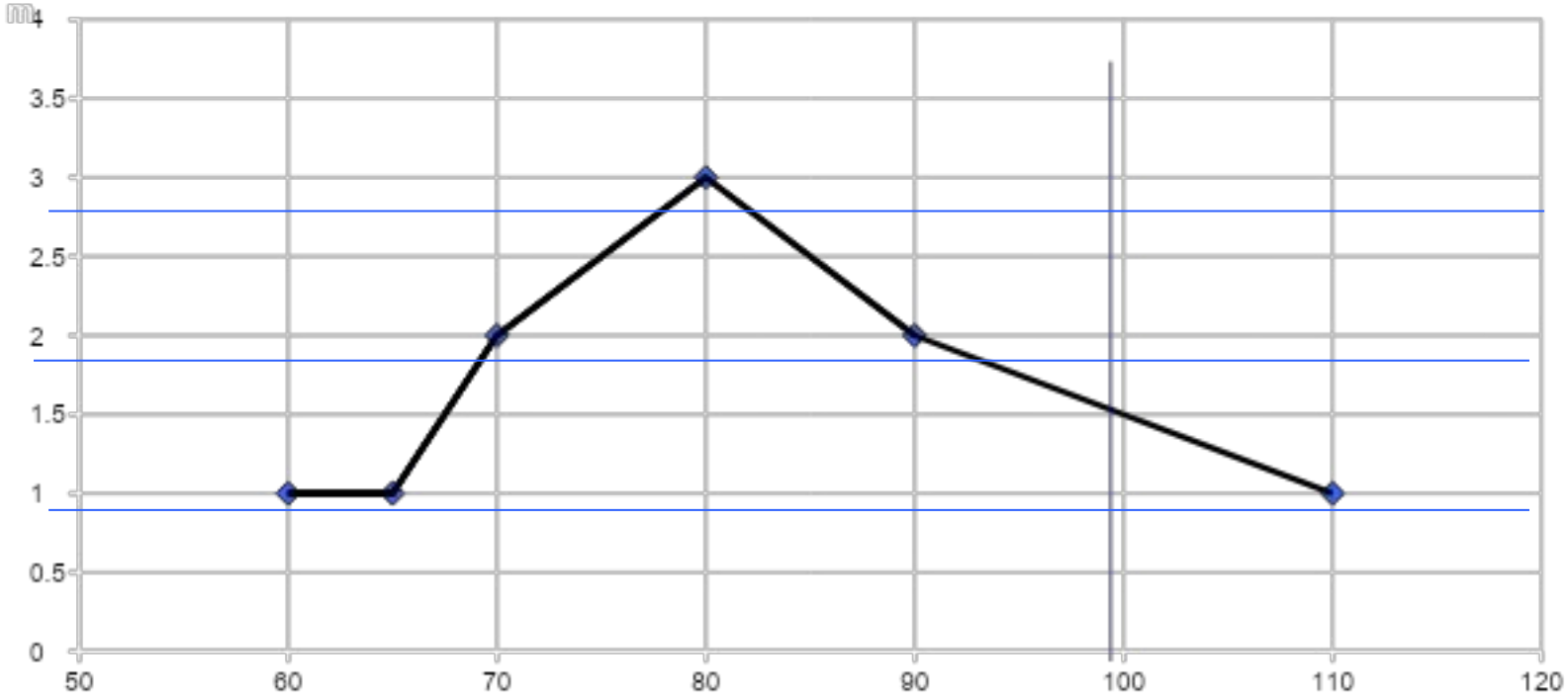
вариационный ряд (статистическое распределение) -- набор пар значение – частота, с которой это значение встретилось в выборке.

Если случайная величина изменяется дискретно, то составляем дискретный вариационный ряд.

x_i	60	65	70	80	90	110
m_i	1	1	2	3	2	1

$$\sum_{i=1}^{k} m_i = n$$

Графическое представление дискретного вариационного ряда - это полигон частот:



Если признак изменяется непрерывно, то составляется интервальный вариационный ряд: набор пар вид интервал – частота.

Для построения интервального вариационного ряда выборку разбивают на интервалы. Есть несколько рекомендаций по вычислению числа интервалов:

$k=\log_2 n+1$ (формула Стерджесса), $k=\sqrt{n}$ и др , подробнее см.

http://ami.nstu.ru/~headrd/seminar/publik_html/Z_lab_8.htm

Длина интервала ΔX рассчитывается по формуле:

$$\Delta x = \frac{x_{\max} - x_{\min}}{k}$$

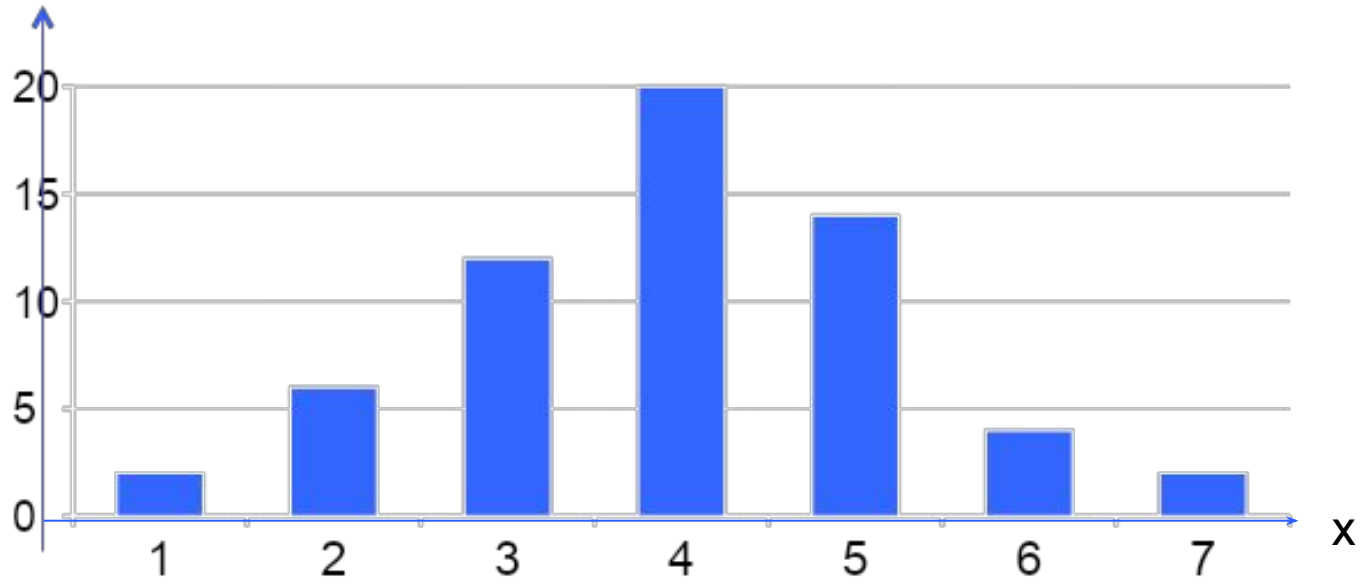
Пример. Анализ веса 60-ти новорожденных дал следующие результаты: **min вес 1,5 кг, max вес 5 кг.** Число интервалов берём **к=7**, следовательно:

Определяем границы интервалов, подсчитываем число новорожденных, вес которых попадает в каждый интервал и составляем таблицу интервальный вариационный ряд

$$\Delta кг = \frac{5_{кг} - 1,5_{кг}}{7} = 0,5$$

вес x_i (кг)	1,5-2	2-2,5	2,5-3	3-3,5	3,5-4	4-4,5	4,5-5
число новорожденных m_i	2	6	12	20	14	4	2
	0,03	0,10	0,20	0,33	0,23	0,07	0,03

Графическая характеристика непрерывного вариационного ряда - Гистограмма:



Закономерности распределения генеральной совокупности оцениваются по выборочной совокупности.

При увеличении объёма выборки ($n \to \infty$), относительные частоты стремятся к вероятностям соответствующих значений с.в., то есть *к закону распределения*.

$$\frac{m_i}{n} \to P(x_i)$$

Статистические характеристики совокупности

Характеристики генеральной совокупности

Математическое ожидание M[X]

дисперсия D[X]

среднее квадратическое отклонение σ[X]

Характеристики выборки (статистики)

$\overline{x}$ среднее арифметическое

S_n^2 - дисперсия

S_n стандартное отклонение (среднее квадратическое)

Генеральная совокупность (n→∞)

$$M[X] = \sum_{i=1}^{k} x_i \cdot P(x_i) = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$D[X] = \sum_{i=1}^{k} (x_i - M[X])^2 \cdot P(x_i) = \frac{\sum_{i=1}^{n} (x_i - M[X])^2}{n}$$

$$\sigma[X] = \sqrt{D[X]}$$

Выборка (n- конечно)

v=n-1 число степеней свободы

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

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

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

Sn стандартное отклонение

Ошибка среднего арифметического

Извлечём из генеральной совокупности N выборок, тогда их средние арифметические сами будут являться значениями случайной величины $\overline{X}\{\overline{x}_1 . \overline{x}_2, \ldots \overline{x}_N\}$

Все эти значения имеют отклонения (рассеивание) от истинного значения M[X].

Это отклонение называется ошибка среднего арифметического, она в n раз меньше отклонения каждого x_i от $\overline{x}$ для данной выборки объёмом n

$$S_{\overline{x}} = \frac{S_n}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n \cdot (n-1)}}$$

$S_{\bar{x}}$ показывает насколько выборочное среднее арифметическое близко к матожиданию M[X] генеральной совокупности.

Чем больше объём выборки n, тем ближе среднее арифметическое к M[X] генеральной совокупности (т. е., ошибка меньше, чем больше n). Этот вывод получил название Закон больших чисел.

Доверительный интервал и доверительная вероятность

Истинные значения M[X] и D[X] можно найти по генеральной совокупности, что практически невозможно. По выборке из этой совокупности мы находим лишь их точечные оценки $\bar{x}$ и S_n^2, но насколько их значения близки истинным M[X] и D[X]? Например, как велика разность $\Delta x = |\bar{x} - M[X]|$

Поэтому наряду с точечными оценками, **применяют интервальные оценки параметров генеральной совокупности по выборке**.

То есть мы хотим найти интервал ΔX, такой что:

$$\bar{x} - \Delta x \le M[X] \le \bar{x} + \Delta x \qquad M[X] - \Delta x \le \bar{x} \le M[X] \pm \Delta x$$

Если известна функция распределения, то этот интервал можно найти из соотношения:

$$\int_{M[X]-\Delta x}^{M[X]+\Delta x} f(x)dx = F(M[X]+\Delta x) - F(M[X]-\Delta x) = P(M[X]-\Delta x \le \bar{x} \le M[X]+\Delta x)$$

зная границы интервала, можно найти вероятность случайной величины $\bar{X}\{\bar{x}_1 . \bar{x}_2, ...\bar{x}_N\}$ принимать значения из данного интервала.

Но нам требуется решить обратную задачу: определить границы интервала, следовательно, для этого надо заранее задать вероятность, с которой мы этот интервал будем определять. Эту вероятность называют доверительной вероятностью $P_д$, а определённый с её помощью интервал -- доверительным интервалом $\Delta X_д$.

Доверительным интервалом какого либо параметра, называют такой интервал, о котором можно сказать, что с вероятностью $P_д$ он содержит в себе этот параметр.

Доверительную вероятность обычно берут равной $P_д=0,95$, но в особо ответственных случаях принимают $P_д=0,99$ или даже $P_д=0,999$.

С доверительной вероятностью связан уровень значимости $\alpha=1-P_д$.

Уровень значимости α --это вероятность того, что значение исследуемого параметра не попадёт в доверительный интервал.

Основная масса случайных величин в биологии и медицине распределена по нормальному закону распределения, следовательно, задав доверительную вероятность можно определить доверительный интервал:

$$P_D = \Phi\left(\frac{M[X]+\Delta X_D - M[X]}{\sigma[\bar{x}]}\right) - \Phi\left(\frac{M[X]-\Delta X_D - M[X]}{\sigma[\bar{x}]}\right) = \Phi\left(\frac{\Delta X_D}{\sigma[\bar{x}]}\right) - \Phi\left(\frac{-\Delta X_D}{\sigma[\bar{x}]}\right) =$$

$$= \Phi\left(\frac{\Delta X_D}{\sigma[\bar{x}]}\right) - \left(1 - \Phi\left(\frac{\Delta X_D}{\sigma[\bar{x}]}\right)\right) = 2\cdot\Phi\left(\frac{\Delta X_D}{\sigma[\bar{x}]}\right) - 1 \quad \Rightarrow \quad \Phi\left(\frac{\Delta X_D}{\sigma[\bar{x}]}\right) = \frac{P_D + 1}{2}$$

Например, при $P_Д$=0,95

$$\Phi\left(\frac{\Delta X_D}{\sigma[\bar{x}]}\right) = \frac{0{,}95+1}{2} = 0{,}975 \quad \Rightarrow \quad \frac{\Delta X_D}{\sigma[\bar{x}]} = 1{,}96 \quad \Rightarrow \Delta X_D = 1{,}96\cdot\sigma[\bar{x}]$$

Где $\sigma\left[\bar{x}\right]-$ стандартное отклонение для случайной величины $\bar{X}\left\{\bar{x}_1.\,\bar{x}_2,\ldots\bar{x}_N\right\}$

Но для малых выборок (n<30) распределение может значительно отличаться от нормального.

В 1908 г английский математик и химик Уильям Госсет под псевдонимом Стьюдент предложил распределение случайной величины для малых выборок.

Распределение Стьюдента

Нормированная случайная величина вычисляется по формуле:

$$t = \frac{\bar{x} - M[X]}{S_{\bar{x}}}$$

Плотность вероятности случайной величины:

$$S(t_{St}, n) = B_n \cdot \left(1 + \frac{t_{St}^{\,2}}{n-1}\right)^{\frac{-n}{2}}$$

Где B_n -- параметр , зависит от n.

По мере увеличения объёма выборок n, распределение Стьюдента довольно быстро приближается к нормальному распределению Гаусса и при n>30 практически не отличается от него.

Практическим следствием этого открытия явилась возможность определять границы доверительного интервала для M[X] с заданной доверительной вероятностью $P_Д$:

$$\Delta X_D = t_{St}(P_D, n) \cdot S_{\bar{x}} = t_{St}(P_D, n) \cdot \sqrt{\frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n \cdot (n-1)}}$$

$t_{St}(P_D, n) = t_{St}$ – коэффициент Стьюдента, находим в таблице для заданной $P_Д$ и известного n.

Таким образом, определив доверительный интервал, можно записать:

$$M[X] = \bar{x} \pm \Delta X_D$$

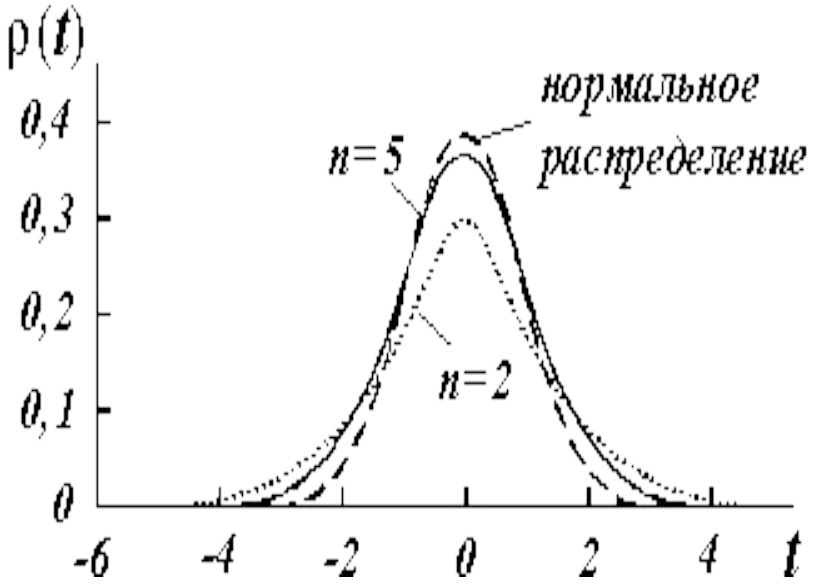


Таблица 1.2

Таблица коэффициентов Стьюдента

Число измерений	Надежность, α							
	0,6	0,7	0,8	0,9	0,95	0,98	0,99	0,999
2	1,38	2,0	3,1	6,3	12,7	31,8	62,7	53,7
3	1,06	1,3	1,9	2,9	4,3	7,0	9,9	31,6
4	0,98	1,3	1,6	2,4	3,2	4,5	5,8	12,9
5	0,94	1,2	1,5	2,1	2,8	3,7	4,6	8,6
6	0,92	1,2	1,5	2,0	2,6	3,4	4,0	6,9
7	0,91	1,1	1,4	1,9	2,4	3,1	3,7	6,0
8	0,9	1,1	1,4	1,9	2,4	3,0	3,5	5,4
9	0,89	1,1	1,4	1,9	2,3	2,9	3,4	5,0
10	0,88	1,1	1,4	1,8	2,3	2,8	3,3	4,8
11	0,88	1,1	1,4	1,8	2,2	2,7	3,2	4,6
12	0,88	1,1	1,4	1,8	2,2	2,7	3,1	4,5
13	0,87	1,1	1,4	1,8	2,2	2,7	3,1	4,3
14	0,87	1,1	1,4	1,8	2,2	2,7	3,0	4,2
15	0,87	1,1	1,3	1,8	2,1	2,6	3,0	4,1
16	0,87	1,1	1,3	1,8	2,1	2,6	3,0	4,1
17	0,87	1,1	1,3	1,7	2,1	2,6	2,9	4,0
18	0,86	1,1	1,3	1,7	2,1	2,6	2,9	4,0
19	0,86	1,1	1,3	1,7	2,1	2,6	2,9	3,9
20	0,86	1,1	1,3	1,7	2,1	2,5	2,9	3,9
∞	0,84	1,0	1,3	1,6	2,0	2,3	2,6	3,3

Пример:

При определении концентрации белка в растворе были получены следующие результаты (в мг/л):110, 112, 115, 113, 114. Найти среднее значение, стандартное отклонение и доверительный интервал для Рд=0.95.

$$\bar{x} = \frac{110+112+115+113+114}{5} = 112,8$$

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

$$= \sqrt{\frac{(110=112,8)^2 + (112+112,8)^2 + (115-112,8)^2 + (113-112,8)^2 + (114-112,8)^2}{5-1}}$$

$$= \sqrt{\frac{2,8^2 + 0,8^2 + 2,2^2 + 0,2^2 + 1,2^2}{4}} = \sqrt{\frac{7,84+0,64+4,84+0,04+1,44}{4}} = \sqrt{\frac{14,8}{4}} = 1,92$$

$$S_{\bar{x}} = \frac{S_n}{\sqrt{n}} = \frac{1,92}{\sqrt{5}} = \frac{1,92}{2,24} = 0,86$$

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

$$\sqrt{\frac{(110-112,8)^2 + (112-112,8)^2 + (115-112,8)^2 + (113-112,8)^2 + (114-112,8)^2}{5(5-1)}} =$$

$$= \sqrt{\frac{2,8^2 + 0,8^2 + 2,2^2 + 0,2^2 + 1,2^2}{5 \cdot 4}} = \sqrt{\frac{7,84 + 0,64 + 4,84 + 0,04 + 1,44}{5 \cdot 4}} =$$

$$\sqrt{\frac{14,8}{20}} = \sqrt{0,74} = 0,86$$

$$\Delta x_D = t_{st} \cdot S_{\bar{x}} = 2,8 \cdot 0,86 = 2,4$$

$M[X] = 112,8 \pm 2,4$ (/ P_D =0,95

Алгоритм обработки результатов прямых измерений

1.Провести серию измерений, не менее трех $\{x_1, x_2, \ldots x_N\}$, $N \geq 3$.

2.Найти среднее арифметическое $x = \dfrac{x_1 + x_2 + x_3 + \ldots + x_N}{N}$.

3.Вычислить доверительный интервал (случайную ошибку).
для заданной доверительной вероятности, например, $P_D = 0{,}95$.

$$\Delta x_{сл} = t_{St} \cdot \sqrt{\frac{\sum_{i=1}^{N} (x_i - x)^2}{N \cdot (N-1)}}$$

4.Найти систематическую ошибку.

а). если указан класс точности прибора:

$$Кл.т. = \frac{\Delta x_{сист}}{x_{шкалы}} \cdot 100\% \quad \Rightarrow \quad \Delta x_{сист} = \frac{Кл.т. \cdot x_{шкалы}}{100\%}$$

где X шкалы – это предел шкалы (максимальное значение на шкале)

б). если класс точности не указан (например линейка или термометр $\Delta x_{сист} = \frac{цена\_деления}{2}$

5. Вычислить общую ошибку: $\Delta x_{общ} = \sqrt{\Delta x^2_{сл.} + \Delta x^2_{сист.}}$

Эту ошибку называют еще абсолютной ошибкой.

6. Записать окончательный результат:

$$x = x \pm \Delta x_{общ}, \text{ для } P_D = 0{,}95$$

7. Кроме абсолютной ошибки желательно также найти коэффициент вариации (или относительную ошибку, выраженную в процентах $\omega\,\% = \frac{\Delta x}{x} \cdot 100\%.$

Контрольные вопросы.

- 1.Равномерный закон распределения непрерывной случайной величины.
- 2.Нормальный закон распределения непрерывной случайной величины.
- 3.Основные понятия математической статистики.
- 4.Схема предварительной обработки экспериментальных данных.
- 5.Статистические характеристики совокупности.
- 6.Ошибка среднего арифметического.
- 7.Доверительный интервал и доверительная вероятность.
- 8.Распределение Стьюдента.