Элементы математической статистики

Математическая статистика является частью общей прикладной математической дисциплины «Теория вероятностей и математическая статистика», однако задачи, решаемые ею, носят специфический характер. Если теория вероятностей исследует явления, полностью заданные их моделью, то в математической статистике вероятностная модель определена с точностью до неизвестных параметров. Отсутствие сведений о параметрах компенсируется пробными испытаниями, на основе которых и восстанавливается недостающая информация.

Цель математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

Первая задача математической статистики состоит в указании методов сбора и группировки статистических сведений, которые получены в результате экспериментов или наблюдений.

Вторая задача — это разработка методов анализа статистических данных: оценки неизвестных вероятности события, а также функций и параметров распределения; оценка зависимости случайной величины от других случайных величин; проверка статистических гипотез о виде и величинах параметров неизвестного распределения. Рассмотрим некоторые из этих вопросов.

1. Выборочный метод

1. Выборки

На практике сплошное исследование (каждого объекта из интересующей нас совокупности) проводят крайне редко. К тому же, если эта совокупность содержит большое число объектов или исследование объекта требует нарушения его функционального стандарта, то сплошное исследование нереально. В таких случаях из всей совокупности случайно отбирают ограниченное число объектов и подвергают их исследованию.

Введем основные понятия, связанные с выборками.

Генеральной совокупностью называется совокупность объектов, из которых производится выборка.

Выборочной совокупностью (выборкой) называется совокупность случайно отобранных объектов из генеральной совокупности.

Число объектов в совокупности называется ее ***объемом.***

Пример. Пусть из 2000 изделий отобрано для обследования 100 изделий. Тогда объем генеральной совокупности $N = 2000$, а объем выборки $n = 100$.

Выборку можно осуществлять двумя способами.

Если после исследования объект из выборки возвращается в генеральную совокупность, то такая выборка называется ***повторной (возвратной).***

Если после исследования объект из выборки не возвращается в генеральную совокупность, то выборка называется ***бесповторной (безвозвратной).***

Выборка называется ***репрезентативной (представительной)***, если по ее данным можно достаточно уверенно судить об интересующем нас признаке генеральной совокупности.

2. Способы отбора

Различают два вида способов отбора:

Без расчленения генеральной совокупности на части.	***С расчленением генеральной совокупности на части.*** Этот способ отбора включает в себя следующие разновидности соответственно способам расчленения генеральной совокупности.		
К этому виду относятся простые случайные отборы (повторные либо бесповторные), когда объекты извлекают по одному из генеральной совокупности; такой отбор можно производить с использованием таблицы случайных чисел.	Отбор, при котором объекты отбираются из каждой «типической» части генеральной совокупности, называется ***типическим.*** Например, отбор деталей из продукции каждого станка, а не из их общего количества, является типическим.	Если генеральную совокупность делят на число групп, равное объему выборки, с последующим отбором из каждой группы по одному объекту, то такой отбор называется ***механическим.***	***Серийным*** называется отбор, при котором объекты отбираются не по одному, а сериями; этот способ используется, когда исследуемый признак имеет незначительные колебания в различных сериях.

На практике часто употребляется комбинирование перечисленных способов отбора. Например, генеральную совокупность разбивают на серии одинакового объема, затем случайным образом отбирают несколько серии и в завершение

случайным извлечением отдельных объектов составляют выборку. Конкретная комбинация способов отбора объектов из генеральной совокупности определяется требованием репрезентативности выборки.

3. Статистическое распределение выборки.

Пусть из генеральной совокупности извлечена выборка объема n, в которой значение x_1 некоторого исследуемого признака X наблюдалось n_1 раз, значение x_2 наблюдалось n_2 раз, значение x_k признака X наблюдалось n_k раз.

- ✧ Значения x_i называются ***вариантами,*** а их последовательность, записанная в возрастающем порядке, — ***вариационным рядом***.
- ✧ Операция расположения случайной величины по возрастанию называется ***ранжированием статистических данных.***
- ✧ Числа n_i называются ***частотами,*** а их отношения к объему выборки $W_i = \frac{n_i}{n}$ — ***относительными частотами***. При этом $\sum_{i=1}^{n} n_i = n, \quad \sum_{i=1}^{n} W_i = 1.$
- ✧ Перечень вариант и соответствующих им частот или относительных частот называют ***статистическим рядом.***
- ✧ ***Модой*** M_0 называется варианта, имеющая наибольшую частоту.
- ✧ ***Медианой*** m_e называется варианта, которая делит пополам вариационный ряд на две части с одинаковым числом вариант в каждой.
  - ! Если число вариант нечетно, т. е. $k = 2l + 1$, то $m_e = x_{l+1}$;
  - ! Если число вариант четно, т. е. $k = 2l$, то $m_e = \frac{x_l + x_{l+1}}{2}$
- ✧ ***Размахом варьирования*** называется разность между максимальной и минимальной вариантами или длина интервала, которому принадлежат все варианты выборки: $R = x_{\max} - x_{\min}$.
- ✧ Перечень вариант и соответствующих им частот называется ***статистическим распределением выборки***. (Здесь имеется аналогия с законом распределения случайной величины: в теории вероятностей — это

соответствие между возможными значениями случайной величины и их вероятностями, а в математической статистике — это соответствие между наблюдаемыми вариантами и их частотами (относительными частотами)).

Пример. Выборка задана в виде распределения частот:

x_i	4	7	8	12	17
n_i	2	4	5	6	3

Найти распределение относительных частот и основные характеристики вариационного ряда.

Решение.

1. Найдем объем выборки: $n= 2 + 4 + 5 + 6 + 3 = 20$.
2. Распределение относительных частот имеет вид:

x_i	4	7	8	12	17
W_i	0,1	0,2	0,25	0,3	0,15

   Контроль: $0.1 + 0,2 + 0,25 + 0,3 + 0,15 = 1$
3. Мода этого вариационного ряда равна 12,
4. Число вариант в данном случае нечетно, $k = 2l + 1$ *или* $5 = 2 \cdot 2 + 1 \Rightarrow l = 2$, поэтому $m_e = x_{l+1} = x_3 = 8$.
5. Размах варьирования $R = x_{\max} - x_{\min} = 17 - 4 = 13$

4. Эмпирическая функция распределения.

Пусть n_x — число наблюдений, при которых наблюдалось значение признака X, меньшее x. При объеме выборки, равном n, относительная частота события $X < x$ равна $\frac{n_x}{n}$.

Определение. Функция, определяющая для каждого значения x относительную частоту события $X < x$,

$$F^*(x) = \frac{n_x}{n}$$

называется ***эмпирической функцией распределения***, или ***функцией распределения выборки.***

В отличие от эмпирической функции распределения $F^*(x)$ выборки функция распределения $F(x)$ генеральной совокупности называется ***теоретической функцией распределения***. Различие между ними состоит в том, что функция $F(x)$ определяет вероятность события $X < x$, а $F^*(x)$ — относительную частоту этого события.

Из теоретических результатов обшей теории вероятностей следует, что при больших n вероятность отличия этих функций друг от друга близка к единице.

Нетрудно увидеть, что $F^*(x)$ обладает всеми свойствами $F(x)$, что вытекает из ее определения.

Свойства:

1) значения $F^*(x)$ принадлежат отрезку [0; 1];

2) $F^*(x)$ является неубывающей функцией;

3) пусть x_m и x_M — соответственно, минимальная и максимальная варианты. Тогда:

$$F^*(x) = 0 \text{ при } x \le x_m$$

$$F^*(x) = 1 \text{ при } x > x_M .$$

Сама же функция $F^*(x)$ служит для оценки теоретической функции распределения $F(x)$ генеральной совокупности.

Пример. Построить эмпирическую функцию по заданному распределению выборки:

x_i	*2*	*4*	*6*
n_i	*10*	*15*	*25*

Решение.

Находим объем выборки: $n = 10 + 15 + 25 = 50$.

Наименьшая варианта равна *2*, поэтому $F^*(x) = 0$ при $x \le 2$.

Значение $X < 4$ (или $x_1 = 2$) наблюдалось 10 раз, значит, $F^*(x) = \frac{10}{50} = 0{,}2$ при $2 < x < 4$.

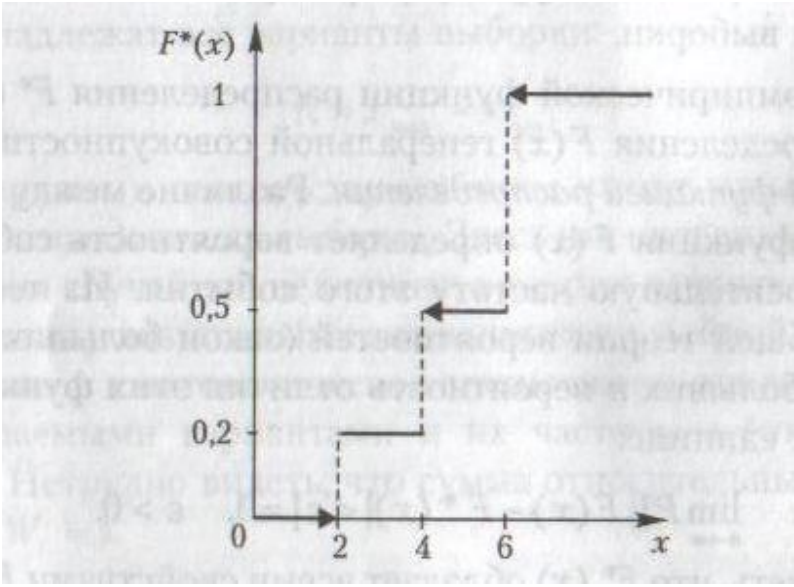
Значения $X < 6$ (а именно $x_1 = 2$ и $x_2 = 4$) наблюдались $10 + 15 = 25$ раз, значит, при $4 < x < 6$ функция $F^*(x) = \frac{25}{50} = 0{,}5$.

Поскольку $x = 6$ — максимальная варианта, то $F^*(x) = 1$ при $x > 6$.

Напишем формулу искомой эмпирической функции:

$$F^*(x) = \begin{cases} 0, & x \le 2, \\ 0{,}2, & 2 < x \le 4, \\ 0{,}5 & 4 < x \le 6, \\ 1, & x > 6. \end{cases}$$

График этой функции:



4. Интервальный статистический ряд.

В случае, когда число значений признака велико или признак является непрерывным, составляют интервальный статистический ряд. В первую строку таблицы статистического распределения вписывают промежутки:

$[x_0;x_1)$, $[x_1;x_2)$,…, $[x_{k-1};x_k)$, которые берут обычно одинаковыми по длине h.

Длина $h=\frac{x_{\max}-x_{\min}}{m}$.

m - количество промежутков можно вычислять по формуле $m=1+\log_2 n$

$x_{нач}=x_{\min}-\frac{h}{2}$ - начало первого интервала.

Пример: Измерили рост 30 наудачу отобранных студентов. Результаты измерений таковы:

178	*160*	*154*	*183*	*155*	*153*	*167*	*186*	*163*	*155*
157	*175*	*170*	*166*	*159*	*173*	*182*	*167*	*171*	*169*
179	*165*	*156*	*179*	*158*	*171*	*175*	*173*	*164*	*172*

Построить интервальный статистический ряд.

Решение:

$n=30$

$m=6$.

$h=\frac{x_{\max}-x_{\min}}{m}=\frac{186-153}{6}\approx 6$.

$x_{нач}=x_{\min}-\frac{h}{2}=153-\frac{6}{3}=150$.

Рост	*[150 – 156)*	*[156 – 162)*	*[162 – 168)*	*[168 – 174)*	*[174 – 180)*	*[180 – 186)*
Частота	*4*	*5*	*6*	*7*	*5*	*3*
Относительная частота	*0,13*	*0,17*	*0,20*	*0,23*	*0,17*	*0,10*

5. Полигон и гистограмма.

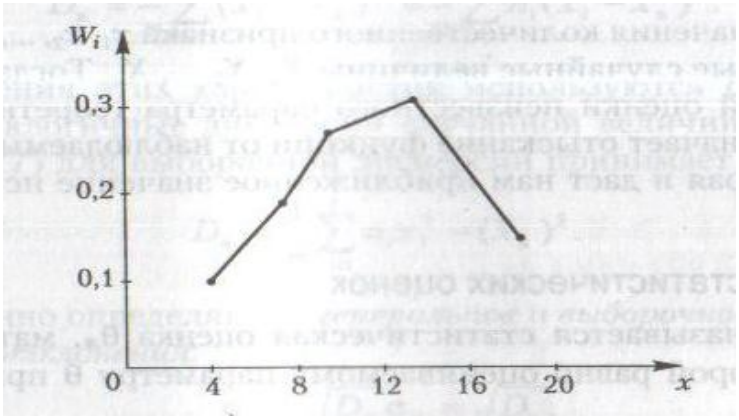
Каждую пару значений (x_i, n_i) из распределения выборки можно трактовать как точку на координатной плоскости. Точно так же можно рассматривать и пары значений (x_i, W_i) относительного распределения выборки.

Ломаная, отрезки которой соединяют точки (x_i, n_i), называется ***полигоном частот.***

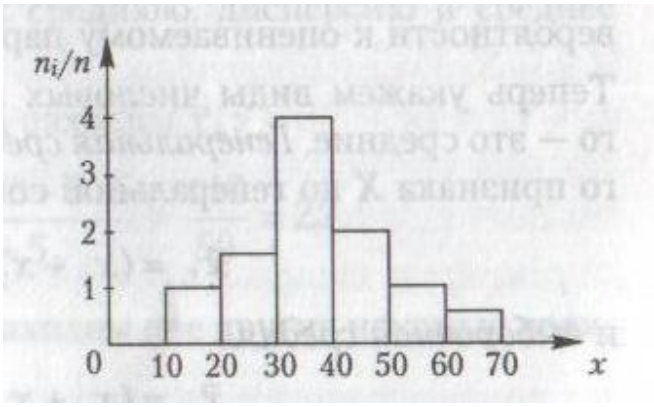
Ломаная, соединяющая на координатной плоскости точки (x_i, W_i), называется ***полигоном относительных частот.***

На рисунке показан полигон относительных частот для распределения:

x_i	4	7	8	12	17
W_i	0,1	0,2	0,25	0,3	0,15



Для случая непрерывного признака X удобно разбить интервал $(x_{\min}, x_{\max})$ его наблюдаемых значений на несколько частичных интервалов длиной h каждый и найти для каждого из этих интервалов сумму частот h_j, попавших в него. Ступенчатая фигура, состоящая из прямоугольников с основаниями длиной h и высотами $\frac{n_j}{n}$ (плотность частоты), называется ***гистограммой частот***.



Геометрический смысл гистограммы: нетрудно увидеть, что площадь ее равна сумме всех частот, или объему выборки. На рисунке изображена гистограмма выборки объема $n = 100$.

Аналогичным образом определяется и гистограмма относительных частот. Высоты прямоугольников, составляющих ступенчатую фигуру, определяются отношениями сумм относительных частот, попадающих в интервал $(x_{\min}+(j-1)h, x_{\max}+jh)$ к длине интервала h, т. е. величинами $\frac{W_j}{h}$. В этом случае площадь гистограммы относительных частот равна единице (сумме относительных частот выборки).

Упражнения.

1. В результате тестирования группа абитуриентов набрала баллы: *5, 3, 0, 1, 4, 2, 5, 4, 1, 5*. Записать полученную выборку в виде:

а) вариационного ряда;

б) статистического ряда.

Ответ: а) *0, 1, 1, 2, 3, 4, 4, 5, 5, 5*

б)

x_i	*0*	*1*	*2*	*3*	*4*	*5*
n_i	*1*	*2*	*1*	*1*	*2*	*3*
W_i	*0,1*	*0,2*	*0,1*	*0,1*	*0,2*	*0,3*

2. По результатам предыдущего примера построить эмпирическую функцию распределения и ее график.

3. По результатам предыдущего примера построить полигон частот и полигон относительных частот.

2. Статистические оценки параметров распределения

Рассмотрим значения количественного признака $x_1, x_2, ...x_n$ в выборке как независимые случайные величины $X_1, X_2, ...X_n$. Тогда нахождение статистической оценки неизвестного параметра теоретического распределения означает отыскание функции от наблюдаемых случайных величин, которая и даст нам приближенное значение искомого параметра.

1. Виды статистических оценок.

Несмещенной называется статистическая оценка θ^*, математическое ожидание которой равно оцениваемому параметру θ при любой выборке:

$$M(\theta^*) = \theta .$$

Смещенной называется оценка, при которой $M(\theta^*) \neq \theta$.

Эффективной называется оценка, которая имеет минимальную дисперсию при заданном объеме выборки n.

Состоятельной называется статистическая оценка, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру.

Теперь укажем виды числовых характеристик оценок. Прежде всего — это средние.

Генеральная средняя для изучаемого количественного признака *X* по генеральной совокупности

$$\overline{x_Г} = \frac{x_1 + x_2 + ... + x_N}{N}$$

и выборочная средняя

$$\overline{x_B} = \frac{x_1 + x_2 + ... + x_k}{n}$$

Если значения признака $x_1, x_2, ..., x_k$ в выборке имеют, соответственно, частоты $n_1, n_2, ..., n_k$, то последнюю формулу можно переписать в виде

$$\overline{x_B} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$$

Выборочная средняя является несмещенной оценкой; это аналог математического ожидания случайной величины X_B.

Введем в рассмотрение величины, характеризующие отклонение значений количественного признака X от своего среднего значения. Это генеральная дисперсия и выборочная дисперсия

$$D_Г = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x_Г})^2$$

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^{\,2} - (\overline{x_B})^2$$

Соответственно определяются генеральное и выборочное средние квадратические отклонения:

$$\sigma_Г = \sqrt{D_Г} \qquad \sigma_B = \sqrt{D_B}\ .$$

Пример. Выборка задана таблицей распределения

x_i	*1*	*2*	*3*	*5*
n_i	*15*	*20*	*10*	*5*

Найти выборочные характеристики: среднюю, дисперсию и среднее квадратнческое отклонение.

Решение.

$$\overline{x_B} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{15\cdot 1 + 20\cdot 2 + 10\cdot 3 + 5\cdot 5}{15+20+10+5} = \frac{110}{50} = 2{,}2\ .$$

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^{\,2} - (\overline{x_B})^2 = \frac{15\cdot 1 + 20\cdot 4 + 10\cdot 9 + 5\cdot 25}{15+20+10+5} - 2{,}2^2 = 6{,}2 - 4{,}84 = 1{,}36$$

$$\sigma_B = \sqrt{D_B} = \sqrt{1{,}36} \approx 1{,}167$$

2. Эмпирические моменты

Для вычисления сводных характеристик выборок используют эмпирические моменты, аналогичные соответствующим теоретическим моментам.

Обычным эмпирическим моментом порядка s называется среднее значение s-х степеней разностей $x_i - C$, где x_i — наблюдаемая варианта, C — произвольная постоянная (ложный ууль —либо мода, либо любая варианта, расположенная примерно в середине вариационного ряда):

$$M'_s = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - C)^s$$

При $C = 0$ имеем ***начальные эмпирические моменты*** порядка s; в частности, в случае $s = 1$

$$M_1 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \overline{x_B}$$

Центральным эмпирическим моментом порядка s называется обычный момент при $C = \overline{x_B}$:

$$m_s = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x_B})^s$$

В частности, центральный момент второго порядка

$$m_2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x_B})^2 = D_B$$

иными словами, это совпадает с выборочной дисперсией,

3. Асимметрия и эксцесс эмпирического распределения

Нормальное распределение является одним из самых распространенных в применениях математической статистики. Для оценки отклонения эмпирического распределения от нормального используют следующие характеристики.

Асимметрия эмпирического распределения определяется следующим равенством:

$$a_s = \frac{m_3}{\sigma_B^3}$$

Эксцесс эмпирического распределения определяется следующим равенством:

$$e_k = \frac{m_4}{\sigma_B^4} - 3$$

В эти формулы входят центральные эмпирические моменты, определяемые. а также выборочное среднее квадратическое отклонение.

Асимметрия и эксцесс служат для сравнения полигона эмпирического распределения с нормальным распределением:

- ✧ знак a_s, указывает на расположение длинной части ломаной относительно математического ожидания (справа при $a_s > 0$ и слева при $a_s < 0$);
- ✧ e_k характеризует «крутизну» ломаной (при $e_k > 0$ сравниваемая кривая более высокая и острая, при $e_k < 0$ она более низкая и плоская).

Пример. Найти асимметрию и эксцесс эмпирического распределения:

варианта	*1*	*2*	*3*	*4*	*5*	*6*	*10*
частота	*5*	*10*	*15*	*35*	*16*	*15*	*4*

Решение.

$$\overline{x_B} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{5\cdot 1+10\cdot 2+15\cdot 3+35\cdot 4+16\cdot 5+15\cdot 6+4\cdot 10}{5+10+15+35+16+15+4} = 4{,}2\,.$$

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i {x_i}^2 - (\overline{x_B})^2 = \frac{5\cdot 1+10\cdot 4+15\cdot 9+35\cdot 16+16\cdot 25+15\cdot 36+4\cdot 100}{5+10+15+35+16+15+4} - 4{,}2^2 = 3{,}56$$

$$\sigma_B = \sqrt{D_B} = \sqrt{3{,}56} \approx 1{,}887$$

$$m_3 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x_B})^3 = \frac{5\cdot(-3{,}2)^3+10\cdot(-2{,}2)^3+15\cdot(-1{,}2)^3+35\cdot(-0{,}2)^3+16\cdot 0{,}8^3+15\cdot 1{,}8^3+4\cdot 5{,}8^3}{100} = 5{,}796$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x_B})^4 = \frac{5\cdot(-3{,}2)^4+10\cdot(-2{,}2)^4+15\cdot(-1{,}2)^4+35\cdot(-0{,}2)^4+16\cdot 0{,}8^4+15\cdot 1{,}8^4+4\cdot 5{,}8^4}{100} = 54{,}8032$$

$$a_s = \frac{m_3}{\sigma_B^3} = \frac{5{,}796}{1{,}887^3} = 0{,}863$$

$$e_k = \frac{m_4}{\sigma_B^4} - 3 = \frac{54{,}8032}{1{,}887^4} - 3 = 4{,}324$$

4. Доверительный интервал

Все оценки, приведенные ранее, определяются одним числом, т. е. являются ***точечными***. При малых объемах выборки точечная оценка может приводить к большим ошибкам и значительно отличаться от оцениваемого параметра.

Более широкое применение получил метод доверительных интервалов, разработанный американским статистиком Ю. Нейманом.

Определение. Доверительным интервалом для параметра θ с надежностью оценки p называется числовой промежуток $(\theta^* - \delta,\ \theta^* + \delta)$, содержащий истинное значение данного параметра с вероятностью равной p:

$$P\ (\theta^* - \delta < \theta < \theta^* + \delta) = p\ .$$

θ^* – оценка неизвестного параметра (например, точечная оценка $\overline{x_B}$).

$\delta > 0$ – некоторое число (например, $\delta = \frac{t \cdot \sigma}{\sqrt{n}}$, где $2\Phi(t) = p$).

Обычно надежность оценки p задается числом, близким к единице. Иными словами, доверительный интервал покрывает неизвестный параметр с заданной надежностью. Число $\alpha = 1 - p$ называется уровнем значимости.

Доверительный интервал для оценки параметра θ с заданной надежностью, σ, $\overline{x_B}$, n находят по формуле:

$$(\overline{x_B} - \frac{t\sigma}{\sqrt{n}} < \theta < \overline{x_B} + \frac{t\sigma}{\sqrt{n}})$$

Параметр t находят из равенства $2\Phi(t) = p$, где $\Phi(t) = \frac{p}{2}$ определяют по таблице.

Упражнения.

1. Найти групповые средние совокупности каждой группы.

1-я группа x_i *0,1* *0,4* *0,6*

	n_i	*3*	*2*	*5*
2-я группа	x_i	*0,1*	*0,3*	*0,4*
	n_i	*10*	*4*	*6*

Ответ: *0,41; 0,23.*

2. По условиям предыдущей задачи найти общую среднюю.

Ответ: *0,29.*

3. Для распределения статистической совокупности

x_i	*4*	*7*	*10*	*15*
n_i	*10*	*15*	*20*	*5*

найти ее дисперсию.

Ответ: *9,84.*

4. Для заданных среднего квадратического отклонения σ, выборочного среднего $\overline{x_B}$ и объема выборки n найти доверительные интервалы неизвестного математического ожидания с заданной надежностью p.

а) $\sigma = 2$, $\overline{x_B} = 5{,}4$, $n = 10$, $p = 0{,}95$;

б) $\sigma = 3$, $\overline{x_B} = 20{,}12$, $n = 25$, $p = 0{,}99$;

Ответ: *а) (4,16; 6,64)* *б) (18,57; 21,67).*

Задача:

Игральную кость подбросили 20 раз. Получены результаты выпавших очков при бросании: 1, 3, 2, 3, 1, 6, 5, 4, 5, 5, 3, 1, 3. 3, 5, 4, 6, 2, 6, 4

1. Записать полученную выборку в виде:

а) вариационного ряда;

б) статистического ряда.

2. Найти распределение относительных частот и основные характеристики вариационного ряда.

3. Построить эмпирическую функцию по заданному распределению выборки:

4. Построить график этой функции.