Кафедра медицинской и биологической физики

Задачи и методы математической статистики. Выборочный метод

Лекция №1

для студентов 2 курса,
обучающихся по специальности 060609 –
Медицинская кибернетика

доц. Шапиро Л.А.

Красноярск, 2015 г.

План лекции:

1. Задачи и методы математической статистики.
2. Основные понятия выборочного метода.
3. Статистическое распределение выборки. Эмпирическая функция распределения, гистограмма.
4. Статистические оценки параметров распределения.
5. Свойства выборочных характеристик.

Актуальность темы

- Основные понятия и методы математической статистики необходимы для обработки результатов измерений в медицине и биологии

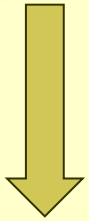
- Теория вероятностей занимается построением и изучением вероятностных моделей случайных явлений. Эти модели строятся на основе аналитических исследований изучаемых случайных явлений. По вероятностным моделям мы можем рассчитать вероятность любого события изучаемого случайного явления.

- Предмет математической статистики составляет разработка методов регистрации, описания и анализа статистических экспериментальных данных, получаемых в результате наблюдения массовых случайных явлений

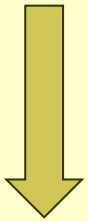
Задачи математической статистики:

- По результатам случайных экспериментов (выборкам) сделать содержательные выводы о вероятностных моделях, адекватно отражающих закономерности изменения замеряемых признаков в изучаемых процессах, явлениях

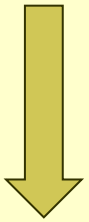
Математическая статистика (числовые данные)



Статистика случайных величин (одномерная статистика)



Многомерная статистика (факторный анализ)



Временные ряды

Задачи одномерной статистики

- Описательная статистика **(представление экспериментальных данных, определение точечных и интервальных оценок)**
- Проверка статистических гипотез **(о законе распределения, параметрах распределения)**

Основные понятия выборочного метода

- Наиболее общую совокупность, подлежащих изучению объектов называют **генеральной**
- **Выборочной** совокупностью или просто выборкой называют часть генеральной совокупности, случайным образом отобранной для наблюдений
- **Обьемом** совокупности называется число объектов этой совокупности (генеральной или выборочной)

Выборочные совокупности

- **n<30 -малые**
- **30<n<100 - средние**
- **n>100 –большие**
- **Цель:** С помощью статистических методов по свойствам выборки сделать вывод о свойствах генеральной совокупности.

  Выборка должна быть **репрезентативна** (представительна), то есть организована таким образом, чтобы отражать, по-возможности, все интересующие нас свойства генеральной совокупности.
- Выборка считается репрезентативной, если каждый объект выборки отобран случайно из генеральной совокупности, то есть все объекты имеют одинаковую вероятность попасть в выборку.

Выборки

Повторные

Объекты возвращаются в генеральную совокупность

Бесповторные

Объекты не возвращаются в генеральную совокупность

Способы отбора

Отбор, не требующий разделения

2. Простой случайный повторный отбор

Отбор, не требующий разделения генеральной совокупности на части:

1. **Простой случайный бесповторный отбор**
2. **Простой случайный повторный отбор**

Отбор, при котором генеральная совокупность разбивается на части:

1. **Типический отбор**
2. **Механический отбор**
3. **Серийный отбор**

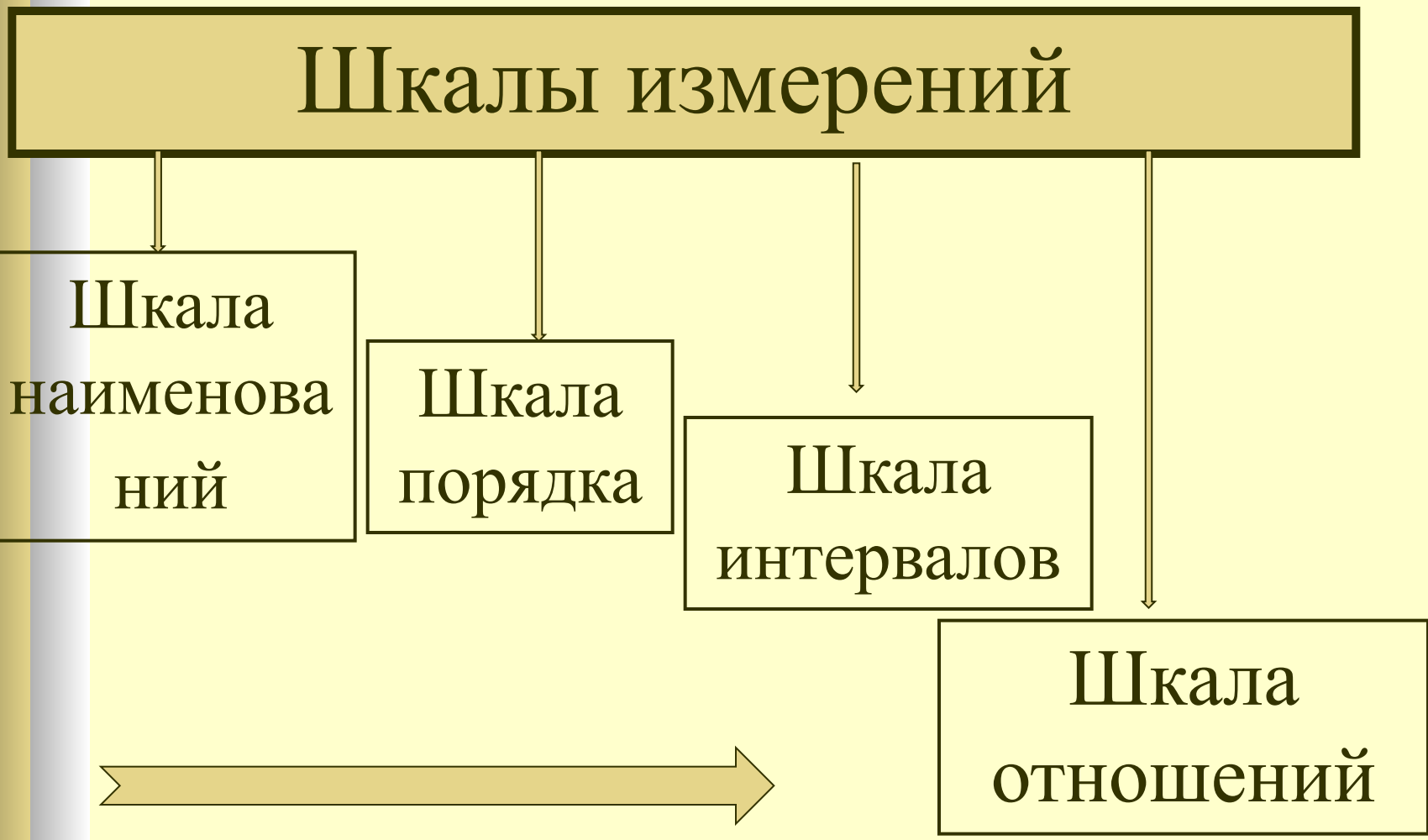
Типический отбор – объекты отбираются не из всей генеральной совокупности, а из каждой ее «типической» части

Механический отбор – генеральная совокупность делится на столько групп, сколько объектов должно войти в выборки и из каждой группы отбирается по одному объекту

Серийный отбор - объекты отбираются из генеральной совокупности не по одному, а сериями

На практике часто используются комбинированные методы


Типы данных
количественные
качественные
порядковые
(полуколичественные)
номинальные
бинарные


Шкалы измерений
Шкала наименований
Шкала порядка
Шкала интервалов
Шкала отношений
Мощность шкалы

Шкалы и допустимые преобразования

Шкала	Допустимое преобразование
Наименований	Взаимно-однозначное
Порядковая	Строго возрастающее
Интервальная	Линейное
Отношений	Подобия

- Значения изучаемого признака называются **вариантами**
- Последовательность вариант, расположенных в возрастающем порядке называется **вариационным рядом**

Например: 172, 179, 158, 186, 164

Вариационный ряд:

158, 164, 172, 179, 186

Вариационные ряды

дискретные **непрерывные**

Статистическим рядом распределения **называется набор вариант и соответствующих им абсолютных и относительных частот**

Статистический ряд распределения

X	X_1	X_2	…	X_n
m	m_1	m_2	…	m_n
m/n	m_1/n	m_2/n	…	m_n

Дискретный ряд распределения (индекс КПУ)

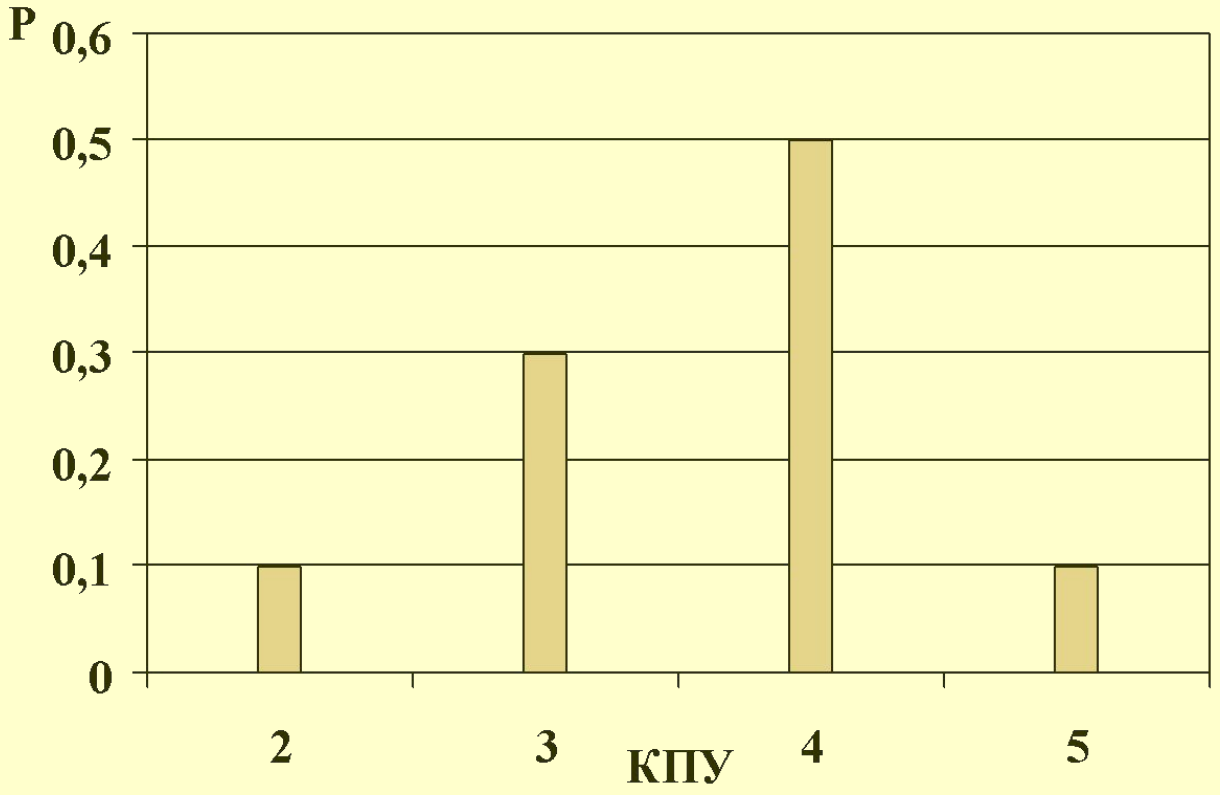
2 3 4 4 2 5 4 2 3 3 3 3 5 5 4 4
3 3 4 4 4 4 4 3 3 4 4 4 4 4

n=30

КПУ	2	3	4	5
m	3	9	15	3
m/n	3/30= 0,1	9/30= 0,3	15/30=0 ,5	3/30= 0,1

$$\sum_{i=1}^{n} P(x_i) = 1$$ — условие нормировки

Дискретный ряд распределения (график)



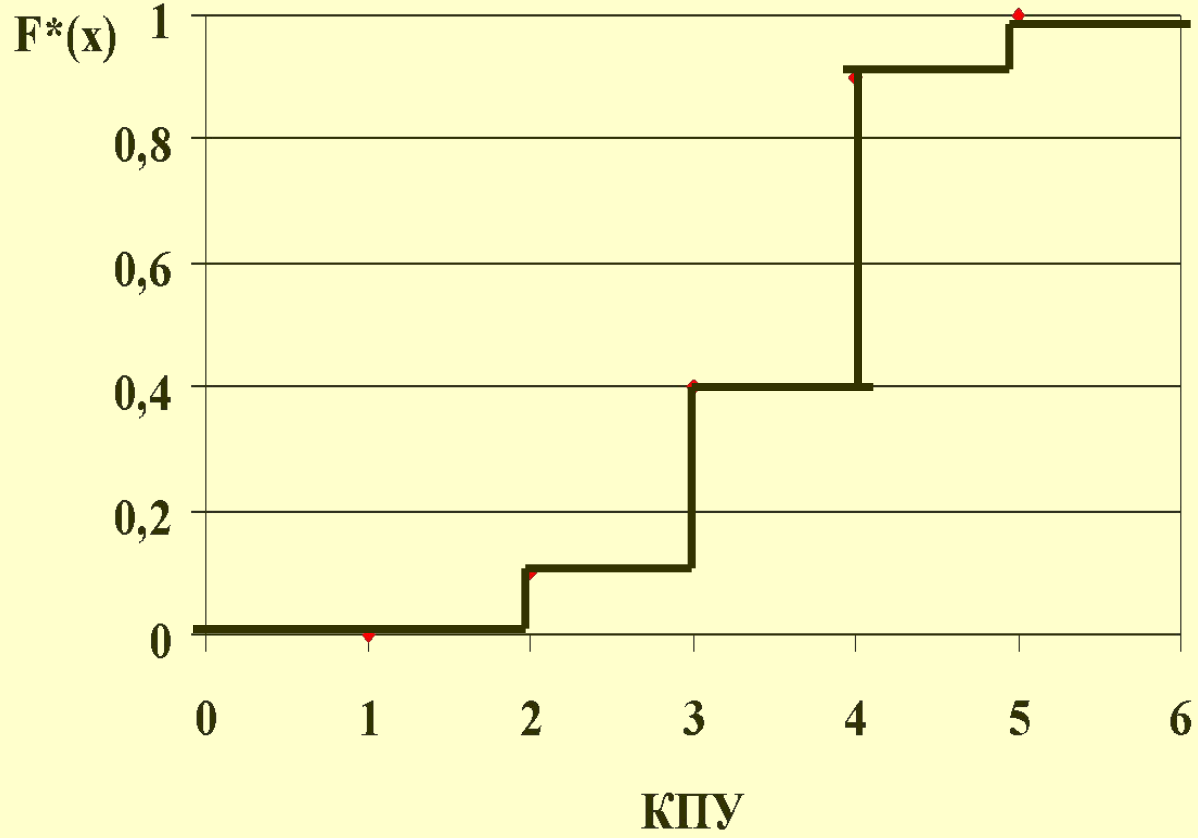
Статистическая функция распределения

- Пусть $\{x_1,\ldots,x_n\}$ - выборка наблюдений случайной величины X с функцией распределения F(x). Необходимо по выборке оценить функцию распределения.
- Определение. Статистической (иногда – эмпирической) функцией распределения случайной величины X называется частота события $X<x$ в данном статистическом материале:
- $F^*(x) = m/n$,
- где m – число X_i, таких, что $X_i<x$.

Эмпирическая функция распределения имеет скачки в точках выборки (вариационного ряда), величина скачка в точке x_i равна m/n, где m–количество элементов выборки, совпадающих с x_i. Эмпирическая функция распределения по вариационному ряду строится так:

$$F_n^*(x) = \begin{cases} 0 & \text{если } x \le x_1 \\ \dfrac{k}{n} & \text{если } x_k < x \le x_{(k+1)} \\ 1 & \text{при } x > x_n \end{cases}$$

Функция распределения вероятностей для дискретной случайной величины F*(x)



КПУ	<2	<3	<4	<5	>=5
F*(X)	0	0,1	0,4	0,9	1

Эмпирическая функция распределения

$$F_n^*(x) = \begin{cases} 0 \ \ если \ \ x \le 2 \\ 0{,}1 \ если \ \ 2 < x \le 3 \\ 0{,}4 \ если \ \ 3 < x \le 4 \\ 0{,}9 \ если \ \ 4 < x \le 5 \\ 1 \quad при \quad x > 5 \end{cases}$$

Интервальные ряды распределения

Ряд распределения студентов по росту

148	158	149	162	170	156	189
151	161	152	171	165	174	157
172	172	177	166	157	149	159
154	164	167	173	176	145	163
185	164	161	153	168	162	184
162	169	154	167	163	166	172
158	155	165	179	165	160	159
169						

На практике ряд распределения (вариационный ряд) составляют следующим образом:

- Из имеющихся значений признака x выбирают наименьшее (X_{min}), наибольшее (X_{max}), определяют размах распределения
- ($X_{max} – X_{min}$).

189-145=44

- Определяют число классов группировки. Для определения числа классов можно воспользоваться формулой: k=1+3,32·lg n, где n – число измерений. Величину k округляют до целых чисел (формула Стерджесса). Например, при n=50:

k=1+3,32·lg 50=1+3,32·1,7=6,64≈7

$$k = \sqrt[3]{n} = \sqrt[3]{50} = 4{,}6$$

Интервальные ряды распределения

- Определяют оптимальную величину класса (интервала группировки)

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

- Эту величину также можно округлять соответственно точности значений x.

$$\Delta X_i = 44/4{,}6 = 9{,}5 \approx 10$$

- Выбирают границы классов. Границы первого класса следует выбрать так, чтобы он содержал наименьшее значение, но не начинался с него, например, класс может начинаться с величины $(X_{min} - \frac{\Delta x_i}{2})$.
- Последующие классы образуются добавлением величины интервала ΔX_i. Если нижняя граница класса совпадает с верхней границей предыдущего класса, это значение следует отнести к данному классу. Например, [1–2), [2–3) и т.д.

Статистический ряд распределения студентов по росту

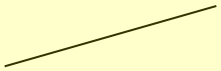
X	140-150	150-160	160-170	170-180	180-190
m	4	14	20	9	3
m/n	4/50 0,08	14/50 0,28	20/50 0,4	9/50 0,18	3/50 0,06
$f(x)=\frac{m}{n\cdot\Delta x}$	0,08/10 0,008	0,028	0,04	0,018	0,006

Гистограмма распределения студентов по росту (m, m/n, f(x))

Эмпирическая функция распределения вероятностей F*(x)

X	<140	<150	<160	<170	<180	>180
m	0	4	18	38	47	50
m/n F*(x)	0	4/50 0,08	18/50 0,36	38/50 0,76	47/50 0,94	50/50 1

Эмпирическая функция распределения F*(x)



Статистические оценки параметров распределения

- Задача: Изучить количественный признак генеральной совокупности.
- Если можно теоретически оценить вид распределения, то необходимо вычислить соответствующие параметры:

Нормальное распределение	M(x) и σ
Распределение Пуассона	параметр λ
Биномиальное распределение	p и т.д.

- Пусть для изучения признака в генеральной совокупности извлечена выборка объемом n:

$$x_1, x_2, x_3, \ldots, x_n$$

- Статистической оценкой **(статистикой)** неизвестного параметра теоретического распределения называют функцию от наблюдаемых случайных величин $\Theta_n(x_1,\ldots,x_n)$

Истинные моменты	Оценки для истинных моментов
M(X)=a	$\bar{x} = \frac{\sum_{i=1}^{n} m_i x_i}{n}$
$D(X)=\sigma^2$	$D(x) = s^2 = \frac{\sum_{i=1}^{n} m(x_i - \bar{x})^2}{n}$
α_k Начальные моменты	$\bar{x}^k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \cdot m_i$
μ_k Центральные моменты	$\bar{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k \cdot m_i$

- В качестве оценки M(X) используется выборочное среднее:

1. Если значения признака $x_1, x_2, x_3, \ldots, x_n$ имеют соответственно частоты $m_1, m_2, m_3, \ldots, m_n$, причем $m_1+m_2+m_3+ \ldots + m_n=n$

$$\overline{x} = \frac{x_1 m_1 + x_2 m_2 + \ldots + x_n m_n}{n} = \frac{\sum_{i=1}^{n} m_i x_i}{n}$$

2. Если все значения признака различны, $m_i = 1$:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Средняя арифметическая есть средняя взвешенная значений признака с весами, равными соответствующим частотам

- Отклонением называют разность между значением признака и его средней арифметической

$$(x - \bar{x})$$

- Сумма произведений отклонений на соответствующие частоты равна 0:

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

- Среднее значение отклонений равно 0:

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

- Оценкой D(X) служит выборочная дисперсия:

- 1. $$D(x) = s^2 = \frac{\sum_{i=1}^{n} m_i (x_i - \overline{x})^2}{n}$$

- 2. $$D(x) = s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

- Среднее квадратическое отклонение:

$$s = \sqrt{D(x)}$$

Асимметрия-скошенность распределения

Эксцесс-островершинность распределения

Обычно рассматривают безразмерные коэффициенты асимметрии и эксцесса:

$$Sk = \frac{\mu_3}{\sigma^3} \qquad Ex = \frac{\mu_4}{\sigma^4} - 3$$

$$Sk = A = \frac{\sum (x_i - \bar{x})^3 m_i}{n \cdot s^3}$$

$$E = \frac{\sum (x_i - \bar{x})^4 m_i}{n \cdot s^4} - 3$$

Коэффициент вариации

- Характеризует относительное значение среднего квадратического отклонения и служит для сравнения разброса несоизмеримых показателей

$$V = \frac{s}{\overline{x}} \cdot 100\%$$

Числовые характеристики интервального ряда

$x_{i+1}-x_i$	$<x_i>$	m_i	$<x_i>m$	$(<x_i>-\bar{x})$	$(<x_i>-\bar{x})^2$	$(<x_i>-\bar{x})^2 \cdot m_i$
140;150	145	4	580	-18,6	345,96	1383,84
150;160	155	14	2170	-8,6	73,96	1035,44
160;170	165	20	3300	1,4	1,96	39,2
170;180	175	9	1575	11,4	129,96	1169,64
180;190	185	3	555	21,4	457,96	1373,88
	Σ	**50**	**8180**			**5002**

$$\bar{X} = \frac{8180}{50} = 163{,}6 \qquad D = \frac{5002}{50} = 100$$

поправка Шеппарда

При вычислении выборочной дисперсии для уменьшения ошибки, вызванной группировкой (особенно при малом числе интервалов) вычитают из вычисленной дисперсии 1/12 квадрата длины частичного интервала:

$D'_{в} = D_{в} - (1/12)h^2$

$D = 100 - (1/12) \cdot 100 = 91,67$

$s = \sqrt{91,67} = 9,6$

$$V = \frac{9,6}{163,6} \cdot 100\% = 5,9\%$$

Коэффициенты асимметрии и эксцесса:

m_i	$(< x_i > -\bar{x})^3$	$(< x_i > -\bar{x})^3 \cdot m_i$	$(< x_i > -\bar{x})^4$	$(< x_i > -\bar{x})^4 \cdot m_i$
4	-6434,86	-25739,4	119688,3	478753,3
14	-636,06	-8904,78	5470,08	76581,14
20	2,744	13333,9	3,84	76,832
9	1481,5	29401,03	16889,6	152006,4
3	9800,3	555	209727,4	629182,1
Σ=50		**8145,6**		**1336600**

$$A = \frac{8145,6}{50 \cdot 9,6^3} = 0,184 \qquad E = \frac{1336600}{50 \cdot 9,6^4} - 3 = 0,147$$

Заключение

Нами рассмотрены:

- **Основные понятия выборочного метода;**
- **Способы построения дискретных и интервальных вариационных рядов.**

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА:

Основная литература:

- Попов А.М. Теория вероятней и математическая статистика /А.М. Попов, В.Н. Сотников. – М.: ЮРАЙТ, 2011. – 440 с.
- Герасимов А. Н. Медицинская статистика: учебное пособие / А. Н. Герасимов. – М. : Мед. информ. агентство, 2007. – 480 с.
- Балдин К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин. – М. : Флинта, 2010. – 488с.

Учебно–методические пособия:

- Шапиро Л.А., Шилина Н.Г. Руководство к практическим занятиям по медицинской и биологической статистике Красноярск: ООО «Поликом». – 2003.

БЛАГОДАРЮ ЗА ВНИМАНИЕ