Формирование базовых компетенций студентов технического университета

М.А.Вигура, О.А.Кеда, А.Ф.Рыбалко, Н.М.Рыбалко, А.Б.Соболев

Теория вероятностей

Лекция 8

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

УГТУ-УПИ
2008г.

Цель лекции:

1. Овладеть соответствующим математическим аппаратом для дальнейшего изучения курса математики, демонстрировать и использовать математические методы в ходе изучения специальных дисциплин для будущей профессиональной деятельности.

2. Ознакомиться с основными понятиями математической статистики, со способами представления данных выборки и с вычислением численных характеристик выборки.

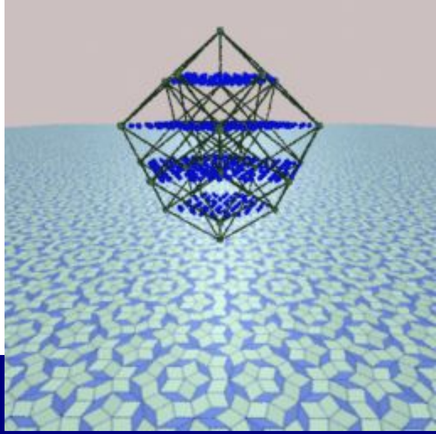
Формируемые компетенции по ФГОС:

ОНК1: способность и готовность использовать фундаментальные математические законы в профессиональной деятельности, применять методы математического анализа и теоретического исследования.

ИК1: способность использовать современные средства вычислительной техники, коммуникаций и связи.

ИК4: готовность работать с информацией из различных источников (сбор, обработка, анализ, систематизация, представление).

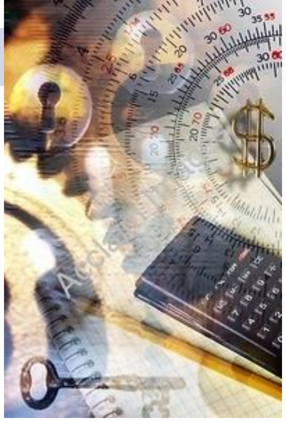
СЛК3: способность самостоятельно приобретать новые знания, используя современные образовательные и информационные технологии.



Лекция 8

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Задачи математической статистики



- <u>Математическая статистика</u> – раздел математики, изучающий *методы сбора и анализа результатов наблюдений массовых случайных явлений* с целью выявления существующих закономерностей.

Типичная задача теории вероятности –

по известным вероятностям простых случайных событий вычислить вероятность более сложного события.

Рассмотрим простую модель – ловля рыбы на удочку. Пусть известна вероятность выловить окуня, скажем, p=0.2 . Какова вероятность события “среди 20 пойманных рыб оказалось 5 окуней”?

Типичная задача математической статистики –

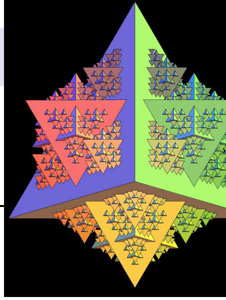
на основании результатов наблюдений оценить вероятность случайного события или характеристики случайной величины.

В только что упомянутой модели – ловля рыбы – задача статистики может быть сформулирована так: среди 20 пойманных рыб оказалось 5 окуней. Что можно сказать о вероятности поймать окуня и насколько этой оценке можно доверять?

Источники информации

- **результаты наблюдений (экспериментов)**
- ***процесс наблюдений может корректироваться на основании предварительных результатов***
- ***(последовательный анализ)***

- **априорная (доопытная) информация о свойствах изучаемого объекта, накопленная к текущему моменту.**
- ***Эта информация отражается в статистической модели, выбираемой при решении задачи.***



- Следует заметить, что *степень обоснованности применения априорной информации* зависит от компетентности и добросовестности конкретного исследователя и неверные исходные допущения могут **существенно исказить** результат статистического анализа.

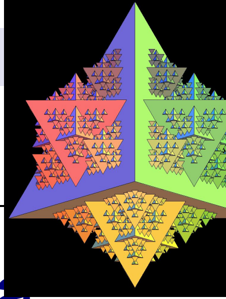
Частые задачи математической статистики

Предварительная обработка данных

Оценка неизвестной величины

Проверка статистических гипотез

Установление формы и степени связи между случайными величинами.



- Современная математическая статистика может быть определена как **теория принятия решений** *в условиях неопределенности.*

- Она включает в себя также *методы определения числа наблюдений*, необходимых для достаточно надежной оценки,

до начала исследований (планирование эксперимента)

или в процессе исследований (последовательный анализ),

что позволяет уже на этапе сбора информации *уменьшить объем собираемых данных без снижения надежности оценок*.

Генеральная и выборочная совокупности. Способы отбора

- Если нужно изучить, *как в совокупности однородных объектов распределен некоторый признак*, характеризующий эти объекты, не всегда возможно исследовать каждый объект (объектов может быть слишком много, при проверке объект может быть уничтожен, и т.п.).
- **В этих случаях отбирают часть объектов и по свойствам отобранных объектов судят о свойствах всех объектов.**

Основные определения

- ***Выборкой или выборочной совокупностью*** называют совокупность случайно отобранных объектов.
- ***Генеральной совокупностью*** называют исходное множество объектов, из которого производится выборка.
- ***Объем совокупности*** (выборочной или генеральной) – число элементов данного множества.

Для упрощения вычислений *при очень большом объеме генеральной совокупности* часто принимают, что *ее объем бесконечен*. Подобное допущение основано на законе больших чисел, погрешность, им вносимая, практически не сказывается на характеристиках выборки.

- Выборка называется **повторной**, если случайно отобранный для обследования объект ***<u>возвращается</u>*** в генеральную совокупность перед отбором следующего объекта.

- В противном случае выборка называется **бесповторной** .

- Чтобы по данным выборки можно было судить о всей совокупности, необходимо, *чтобы члены выборки представляли ее достаточно правильно*.

  <u>Такая выборка называется</u> **репрезентативной** (представительной).

Для того, чтобы выборка была репрезентативной, необходимы:

1) **случайный отбор** элементов совокупности,

2) **равновероятность** попадания в выборку любого элемента генеральной совокупности,

3) достаточно **большой объем выборки**

- Если элементы извлекаются по одному из генеральной совокупности, говорят о **простом случайном отборе** (*может быть повторным и бесповторным*).
- Если из генеральной совокупности элементы разбиваются на группы, “*серии*”, серия отбирается случайно и подвергается сплошной проверке, отбор называется **серийным**.

- Типический отбор
- Механический отбор осуществляется через регулярные интервалы Для обеспечения репрезентативности при механическом отборе необходим контроль периодичности.

Возможны также произвольные комбинации способов отбора.

Статистическое распределение выборки

Пусть *из генеральной совокупности извлечена выборка* и производятся наблюдения за случайной величиной X, причем значение x_1 наблюдалось n_1 раз, значение x_2 – n_2 раз,…, значение x_k – n_k раз, $\sum_{i=1}^{k} n_i = n$ – объем выборки.

Возможные значения случайной величины x_1, x_2,…, x_k принято называть **вариантами**, а последовательность вариант, записанную в порядке возрастания – **вариационным рядом**. Числа n_1, n_2,…, n_k называются **частотами**,

а $w_i = \dfrac{n_i}{n}$ – **относительными частотами** $\left(\sum_{i=1}^{k} w_i = 1\right)$.

статистическим распределением выборки или статистическим рядом

Называется перечень вариант и соответствующих им частот

(или относительных частот)

X	x_1	x_2	$\dots$	x_i	$\dots$	x_k
n_i	n_1	n_2	$\dots$	n_i	$\dots$	n_k

X	x_1	x_2	$\dots$	x_i	$\dots$	x_k
w_i	w_1	w_2	$\dots$	w_i	$\dots$	w_k

пример

- При 100 подбрасываниях игральной кости на верхней грани единица выпала 22 раза, двойка -16 , тройка - 13, четверка -24 , пятерка -12 и, наконец, шестерка – 13 раз. Считая число выпавших очков случайной величиной, построить для нее статистический ряд.

X	1	2	3	4	5	6
n_i	22	16	13	24	12	13
w_i	0,22	0,16	0,13	0,24	0,12	0,13

- В том случае, *если число значений случайной величины X велико*, или есть основания полагать, что случайная величина *является непрерывной* и может принять любое значение из некоторого промежутка, строят **интервальный статистический ряд**.

- Значения вариант группируют по промежуткам (*обычно одинаковой длины*), в первой строке указывается промежуток, во второй – число наблюдений, попавших в данный промежуток.

интервальный статистический ряд

Значения вариант группируют по промежуткам (обычно одинаковой длины),

в первой строке указывается промежуток, во второй – число наблюдений, попавших в данный промежуток.

- Для определения оптимальной длины частичного промежутка можно использовать **формулу Стерджеса**. Пусть значения случайной величины *X* располагаются на *отрезке* $[a,b]$, *объем выборки – n*.

Длина частичного интервала $\Delta = \dfrac{b-a}{1+log_2 n}$,

число интервалов $k = 1 + log_2 n$ (берется

ближайшее к $log_2 n$ целому), первый интервал начинается в точке $x_{min} = a - \dfrac{\Delta}{2}$

пример

- Пусть измерен рост 50 случайно выбранных человек с точностью до 1 см (результаты приведены ниже).

  175, 179, 170, 163, 159, 171, 170, 152, 168, 172, 160, 167, 165, 167, 156, 170, 181, 153, 163, 167, 179, 172, 170, 186, 180, 187, 178, 175, 168, 168, 171, 173, 178, 170, 183, 181, 180, 160, 165, 158, 173, 160, 167, 172, 180, 169, 168, 170, 188, 176.

Рост является *непрерывной случайной величиной*, но в силу ограниченной точности измерений любые значения этой величины будут принадлежать некоторому *дискретному множеству*.

- Значения роста в выборке изменяются от 152 см до 188 см, т.е., принимают 37 значений, объем выборки – 50 человек.
- Нахождение статистических характеристик данной выборки в таком виде представляет заметные вычислительные трудности.

Упорядочим данные выборки по возрастанию (ранжируем выборку):

- 152, 153, 156, 158, 159, 160, 160, 160, 163, 163, 165, 165, 167, 167, 167, 167, 168, 168, 168, 168, 169, 170, 170, 170, 170, 170, 170, 171, 171, 172, 172, 172, 173, 173, 175, 175, 176, 178, 178, 179, 179, 180, 180, 180, 181, 181, 183, 186, 187, 188.

Построим интервальный статистический ряд.

$$\Delta = \frac{188-152}{1+log_2 50} \approx \frac{36}{6,644} \approx 6,36 \approx 6 \qquad k \approx 6,644 \approx 7$$

$$x_{min} = 152 - 3 = 149$$

X	$[149,155)$	$[155,161)$	$[161,167)$	$[167,173)$	$[173,179)$	$[179,185)$	$[185,191]$
n_i	**2**	**6**	**4**	**20**	**7**	**8**	**3**
w_i	**0,04**	**0,12**	**0,08**	**0,4**	**0,14**	**0,16**	**0,06**

Полигон и гистограмма

Для наглядности часто используют *графические изображения* статистических рядов:

для дискретного ряда - <u>полигон</u>,

для интервального ряда - <u>гистограмму</u>.

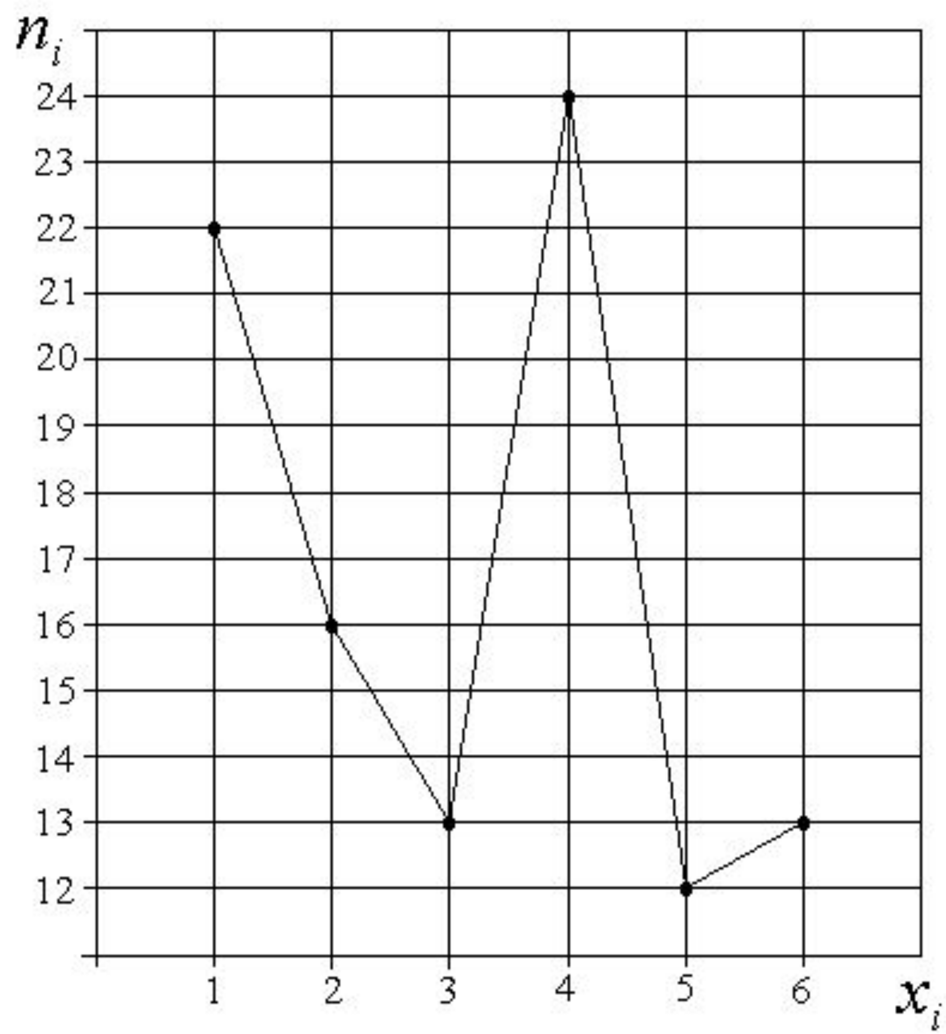
Полигон частот (относительных частот)

есть ломаная, отрезки которой соединяют точки

(x_1, n_1), (x_2, n_2),…, (x_i, n_i) ,…

или

(x_1, w_1), (x_2, w_2),…,(x_i, w_i),… .

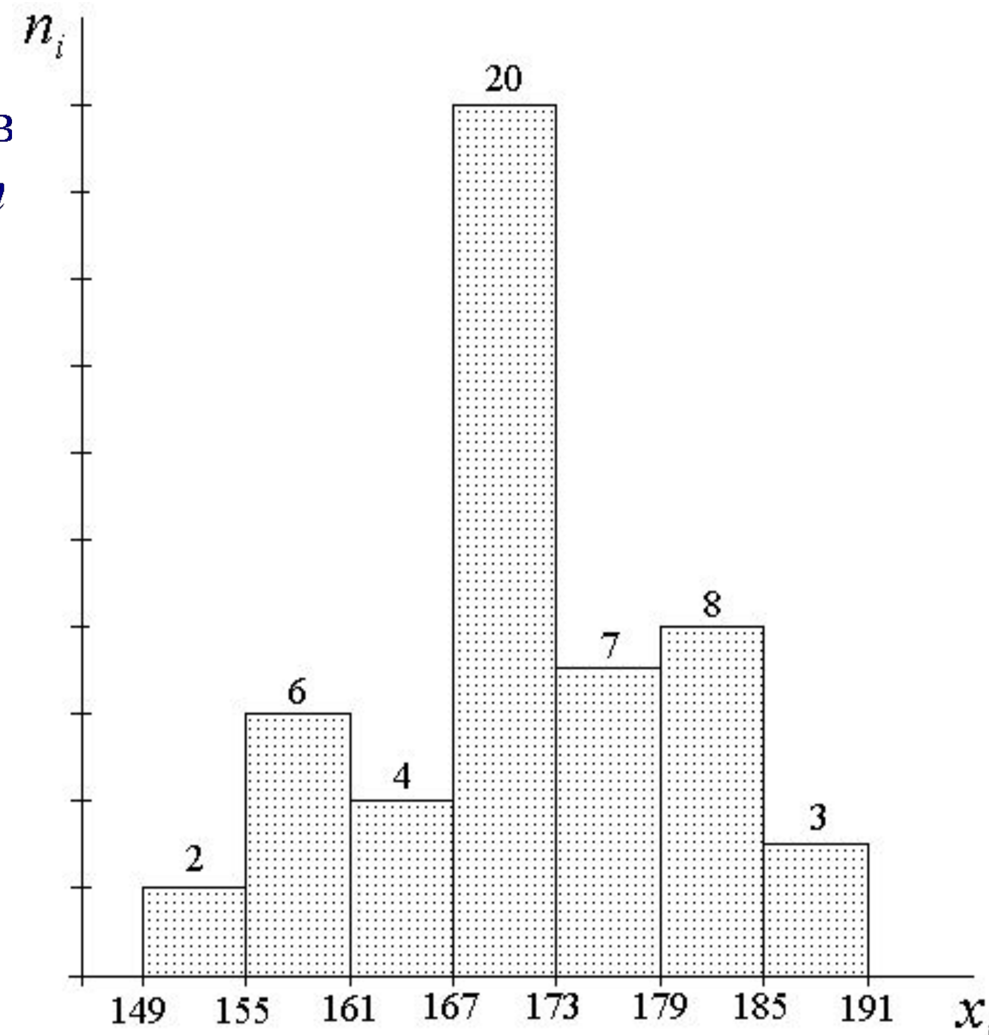


Гистограмма частот (относительных частот)

есть *ступенчатая фигура*, состоящая из прямоугольников *основаниями которых служат частичные интервалы длиной* $\Delta x_i = x_{i+1} - x_i$ *и высотами*

$$h_i = \frac{n_i}{\Delta x_i}, \quad \left(h_i = \frac{w_i}{\Delta x_i} \right).$$

Площадь всей гистограммы частот равна *n (объему выборки)*, а площадь всей гистограммы относительных частот равна 1.



Эмпирическая функция распределения

Пусть задано *статистическое распределение* случайной величины X.

- Обозначим через n_x число вариант, меньших x, n – общее число наблюдений (объем выборки).
- Относительная частота события $\{X < x\}$ равна $\dfrac{n_x}{n}$.

При изменении x меняется и относительная частота, т.е., $\dfrac{n_x}{n}$ есть функция от x. Поскольку эта функция строится по данным опыта, ее называют **эмпирической (опытной).**

- **Эмпирической функцией распределения** (функцией распределения выборки) $F^*(x)$ называется **относительная частота** события $\{X < x\}$

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

где n_x – число вариант, меньших x , n – объем выборки.

Теоретической функцией распределения

называется функция распределения $F(x)$

случайной величины *X*, вычисленная по

генеральной совокупности, т.е.,

вероятность события $\{X<x\}$.

При возрастании объема выборки различия между $F^*(x)$ и $F(x)$ уменьшаются.

теорема (Гливенко)

При *неограниченном возрастании объема выборки* эмпирическая функция распределения $F^*(x)$ **сходится по вероятности** к теоретической функции распределения $F(x)$

$$\forall \varepsilon > 0 \quad \forall x \in R \quad \lim_{n \to \infty} P\left(\left|F^*(x) - F(x)\right| < \varepsilon\right) = 1$$

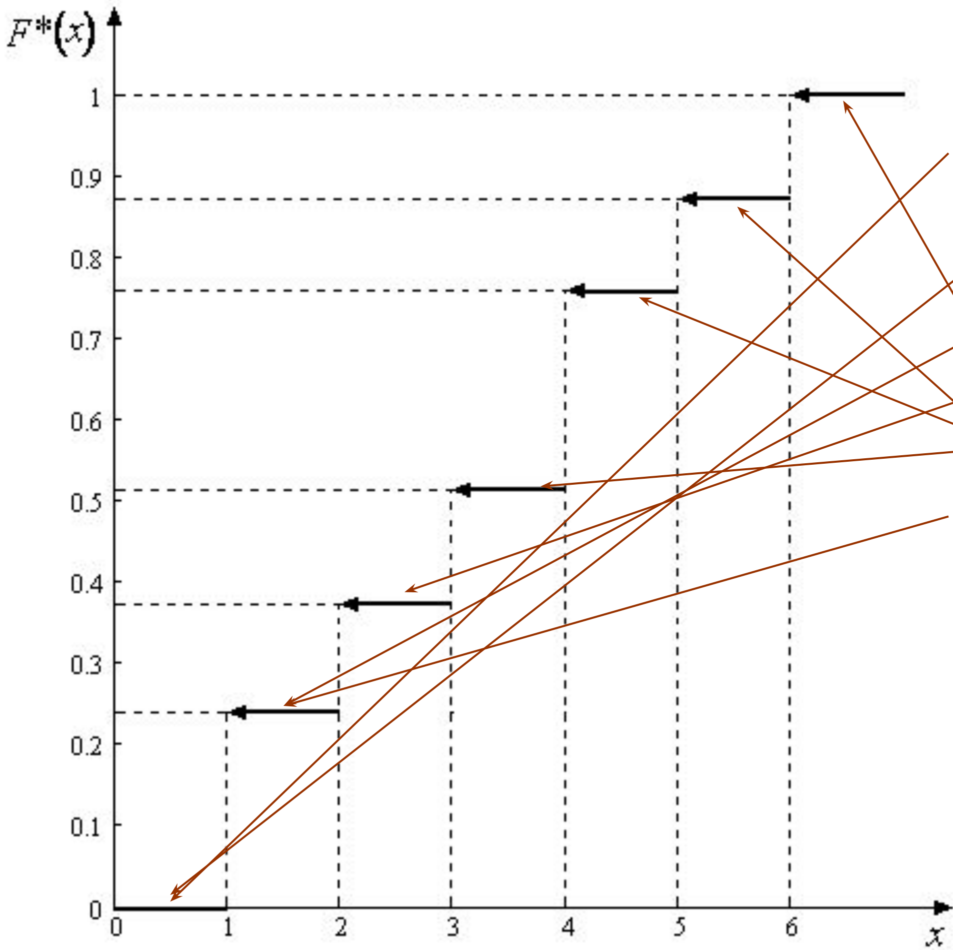
пример

Построим **эмпирическую функцию распределения** для ранее рассмотренного примера (подбрасывание кости). Распределение приведено ниже.

X	1	2	3	4	5	6
n_i	22	16	13	24	12	13
w_i	0,22	0,16	0,13	0,24	0,12	0,13

Построим

<u>эмпирическую функцию распределения</u>



$$F^*(x)=\begin{cases}0, & x\le 1,\\ 0,22, & 1<x\le 2,\\ 0,38, & 2<x\le 3,\\ 0,51, & 3<x\le 4,\\ 0,75, & 4<x\le 5,\\ 0,87, & 5<x\le 6,\\ 1, & x>6.\end{cases}$$

Для $x \le 1$ условие $\{X<x\}$ не может быть выполнено, т.е., $F^*(x)=0$ при $x\le 1$;

Для $1<x\le 2$ условие $\{X<x\}$ выполняется в 22 случаях из 100, т.е., $F^*(x)=0,22$ при $1<x\le 2$

Числовые характеристики статистического распределения выборки

Пусть имеется генеральная совокупность объема N, из которой ***<u>сделана выборка объема n</u>***. Статистический ряд, в котором присутствуют значения случайной величины X и относительные частоты их появления, можно рассматривать как ***закон распределения новой случайной величины*** XВ (исходную величину переобозначим как XГ). ***Очевидно, законы распределения этих величин в какой-то мере близки, но не совпадают.***

Замечание

- Каждой числовой характеристике случайной величины *ХГ* соответствует ее **выборочный аналог** – характеристика случайной величины *ХВ*.
- ***При возрастании объема выборки*** числовые характеристики *ХВ будут сходиться по вероятности к соответствующим характеристикам ХГ.*

числовые характеристики выборки

- ***Выборочное среднее*** *– среднее арифметическое значений выборки*

$$\overline{x}_B = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot x_i = \sum_{i=1}^{k} w_i \cdot x_i$$

числовые характеристики выборки

- **Выборочное среднее** – среднее арифметическое значений выборки

$$\overline{x}_B = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot x_i = \sum_{i=1}^{k} w_i \cdot x_i$$

числовые характеристики выборки

- **Выборочная мода** Mo^* – наиболее вероятное значение в выборке (*варианта с наибольшей частотой*).
- **Выборочная медиана** Me^* – значение случайной величины, *приходящееся на середину вариационного ряда*. Если объем выборки

четен, *n=2m*, то, $Me^* = \dfrac{x_m + x_{m+1}}{2}$

если нечетен, *n=2m+1*, то $Me^* = x_{m+1}$

числовые характеристики выборки

- ***Выборочная дисперсия*** *– среднее значение квадрата отклонения* $x_i - \overline{x}_B$

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot \left(x_i - \overline{x}_B\right)^2 = \sum_{i=1}^{k} w_i \cdot \left(x_i - \overline{x}_B\right)^2$$

Эта формула может быть преобразована к виду

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot x_i^{\,2} - \left(\overline{x}_B\right)^2 = \sum_{i=1}^{k} w_i \cdot x_i^{\,2} - \left(\overline{x}_B\right)^2 = \overline{x^2} - \left(\overline{x}\right)^2$$

$$\overline{x^2} = M^*\left(x^2\right) \qquad \overline{x} = M^*\left(x\right)$$

числовые характеристики выборки

- ***Выборочное среднее квадратическое отклонение*** $\sigma_B = \sqrt{D_B}$
- ***Исправленная выборочная дисперсия***

$$s^2 = \frac{n}{n-1} D_B = \frac{1}{n-1} \sum_{i=1}^{k} n_i \cdot \left(x_i - \overline{x}_B \right)^2$$

- ***Исправленное выборочное среднее квадратическое отклонение***

$$s = \sqrt{s^2}$$

Числовые характеристики генеральной совокупности

- ***Генеральное среднее*** *– среднее арифметическое значений признака X генеральной совокупности:*

$$\bar{x}_Г = \frac{x_1 + x_2 + \ldots + x_N}{N}.$$

Если в генеральной совокупности значения признака $x_1, x_2, \ldots, x_k$ *имеют частоты* $N_1, N_2, \ldots, N_k$*, причем* $N_1 + N_2 + \ldots + N_k = N$*, то*

$$\bar{x}_Г = \frac{1}{N}\sum_{i=1}^{k} N_i \cdot x_i = \sum_{i=1}^{k} p_i \cdot x_i = M(X),$$

где $p_i = \frac{N_i}{N}$ – вероятности появления значений признака $x_1, x_2, \ldots, x_k$.

Числовые характеристики генеральной совокупности

- ***Генеральная дисперсия*** *– среднее по генеральной совокупности значение квадрата отклонения* $x_Г - \overline{x}$

$$D_Г = \frac{1}{N}\sum_{i=1}^{k} N_i \cdot \left(x_i - \overline{x}_Г\right)^2 = \sum_{i=1}^{k} p_i \cdot \left(x_i - \overline{x}_Г\right)^2$$

Числовые характеристики генеральной совокупности

- ***Генеральное среднее квадратическое отклонение (стандартное отклонение)***

$$\sigma_Г = \sqrt{D_Г}$$

По мере увеличения объема выборки $(n \to N)$ ***числов характеристики*** X_B ***будут приближаться*** соответствующим характеристикам $X_Г$, следовательно, и п произвольном объеме выборки ***значения выборочн характеристик в какой-то мере служат оценка генеральных характеристик.***

В результате студент должен:

владеть основными понятиями математической статистики;

уметь преобразовывать выборочные данные к виду, удобному для дальнейшей обработки;

уметь вычислять численные характеристики выборки.

Перечень источников, список дополнительной литературы по теме.

1. Сборник задач по математике: Учеб. пособие для втузов : В 4 ч. Ч. 4: Теория вероятностей. Математическая статистика / Э. А. Вуколов, А. В. Ефимов, В. Н. Земсков, А. С. Поспелов; Под общ. ред. А. В. Ефимова, А. С. Поспелова. - 3-е изд., перераб. и доп. - М.: Физматлит, 2004. - 432 с.: ил.; 21 см. - Библиогр.: с. 431 (16 назв.). - ISBN 5-94052-037-5.

2. Чудесенко, Валерий Федорович. Сборник заданий по специальным курсам высшей математики (типовые расчеты): Учеб. пособие для вузов / В. Ф. Чудесенко. - 2-е изд., перераб. - М.: Высшая школа, 1999. - 126 с. - ISBN 5-06-003065-2.

3. Гмурман, Владимир Ефимович. Руководство к решению задач по теории вероятностей и математической статистике: Учеб. пособие для вузов. - 5-е изд., стер. - М.: Высш. шк., 1999. - 400 с. - ISBN 5-06-003465-8.

4. Вентцель, Елена Сергеевна. Задачи и упражнения по теории вероятностей: учеб. пособие для студентов втузов / Е. С. Вентцель, Л. А. Овчаров. - 5-е изд., испр. - М.: Академия, 2004. - 448 с.: ил.; 21 см. - (Высшее образование). - Библиогр.: с. 240 (12 назв.). - ISBN 5-7695-1054-4.

5. Агапов, Георгий Иванович. Задачник по теории вероятностей: Учеб. пособие для втузов. - 2-е изд., доп. - М.: Высш.шк., 1994. - 112с. - ISBN 5-06-002664-7.