ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Лекция 7.

Основные изучаемые вопросы:

- **1. Статистическая оценка параметров распределения.**
- **2. Вариационные ряды и их числовые характеристики.**
- **3. Ошибка выборочных наблюдений.**

СТАТИСТИЧЕСКАЯ ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

- **Понятие о статистической оценке параметров**
- Методы математической статистики используются при анализе явлений, обладающих свойством ***статистической устойчивости***. Это свойство заключается в том, что, хотя результат X отдельного опыта не может быть предсказан с достаточной точностью, значение некоторой функции $\theta^*_n = \theta^*_n(x_1, x_2, \ldots, x_n)$ от результатов наблюдений ***при неограниченном увеличении объема выборки теряет свойство случайности и сходится по вероятности к некоторой неслучайной величине X.***
- ***Генеральной совокупностью*** называют множество результатов всех наблюдений, которые могут быть сделаны при данном комплексе условий.

- В некоторых задачах генеральную совокупность рассматривают как случайную величину ***X***.
- ***Выборочной совокупностью (выборкой)*** называют множество результатов, случайно отобранных из генеральной совокупности.
- Выборка должна быть ***репрезентативной***, т.е. правильно отражать пропорции генеральной совокупности. Это достигается случайностью отбора, когда все объекты генеральной совокупности имеют одинаковую вероятность быть отобранными.
- Задачи математической статистики практически сводятся к обоснованному ***суждению об объективных свойствах генеральной совокупности по результатам случайной выборки.***
- ***Параметры генеральной совокупности есть постоянные величины, а выборочные характеристики (статистики) - случайные величины.***

- В самом общем смысле статистическое оценивание параметров распределения можно рассматривать как совокупность методов, позволяющих ***делать научно обоснованные выводы о числовых параметрах генеральной совокупности по случайной выборке из нее.***
- Всякую однозначно определенную функцию результатов наблюдений, с помощью которой судят о значении параметра X, называют ***оценкой (или статистикой)*** параметра $\overline{X}_{выб}$.
- Рассмотрим некоторое множество выборок объемом n каждая. Оценку параметра $\overline{X}$, вычисленную по ***i***-ой выборке, обозначим через $\overline{X}_{выб\ i}$. Так как состав выборки случаен, то можно сказать, что $\overline{X}_{выб\ i}$ примет неизвестное заранее числовое значение, т.е. является случайной величиной.

- Известно, что случайная величина определяется соответствующим законом распределения и числовыми характеристиками, следовательно, и ***выборочную оценку также можно описывать законом распределения и числовыми характеристиками***.
- Основная задача теории оценивания состоит в том, чтобы произвести выбор оценки $\overline{X}_{выб\,i}$ параметра X, позволяющей получить хорошее приближение оцениваемого параметра.
- Выборочные данные используются для анализа всей генеральной совокупности, но для этого требуется представить их в виде, удобном для обработки. Для этого применяются различные формы упорядочивания данных - по возрастанию, по совпадающим значениям, по интервалам и т.п. Обычно для решения проблемы наглядности и удобства обработки изучаемой совокупности используют ***вариационные ряды***.

- ***Упорядоченный в порядке возрастания или убывания ряд значений признака (вариантов) с соответствующими им весами называется вариационным рядом (рядом распределения).***
- Порядковый номер *варианта* (значения признака) называется его рангом: x_1 - 1-й вариант (1-е значение признака), x_2 - 2-й вариант (2-е значение признака), x_i - i-й вариант (i-е значение признака). Значения признака (варианты) обычно обозначаются: $x_1, x_2, \ldots, x_n$.
- ***Весами вариантов*** называют соответствующие им ***частоты*** или ***частости***.
- Под ***частотой*** i-го варианта понимают величину m_i, которая указывает, ***сколько раз встречается этот вариант*** (значение признака) в рассматриваемой статистической совокупности.

- Например, если 10 студентов имеют по экзамену оценку пять, то частота варианта $x_4 = 5$ будет иметь значение $m_4 = 10$.
- ***Сумма частот всех вариантов рассматриваемого вариационного ряда равна объему исследуемой совокупности:***

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

где n - объем исследуемой выборочной совокупности;
$к$ - количество значений признака (вариантов);
m_i - частота варианта.

- ***Частостью*** или ***относительной частотой*** называют величину ω_i, которая показывает, какая часть единиц совокупности имеет этот вариант.

- ***Частость*** рассчитывается как отношение частоты варианта к сумме всех частот ряда:

$$\omega_i = \frac{m_i}{\sum_{i=1}^{k} m_i}.$$

- Очевидно, что ***сумма всех частостей равна 1***.
- Различают дискретные и интервальные вариационные ряды.
- ***У дискретного вариационного ряда*** значения изучаемого признака отличаются друг от друга на некоторую конечную величину.

Значения x_i	x_1	x_2	...	x_n
Частоты m_i	m_1	m_2	...	m_n

- ***Интервальные вариационные ряды*** содержат не конкретные значения вариантов изучаемого признака, а интервалы, в которые попадают эти значения, если они могут отличаться друг от друга на сколь угодно малую величину.
- Общий вид интервального вариационного ряда показан в таблице

Значения признака	$a_1 - a_2$	$a_2 - a_3$	…	$a_{n-1} - a_n$
Частоты m_i	m_1	m_2	…	m_n

- В интервальных вариационных рядах в каждом интервале выделяют верхнюю и нижнюю границы интервала.
- Разность между верхней и нижней границами интервала называют ***интервальной разностью,*** или ***длиной (величиной) интервала.***

- Если интервалы в вариационном ряде имеют одинаковую длину (интервальную разность), их называют ***равновеликими***, в противном случае - ***неравновеликими***.
- Если интервалы имеют разную величину, то при построении гистограммы по оси ординат необходимо откладывать значения ***абсолютной*** или ***относительной плотности интервала***.
- ***Абсолютная плотность*** *i*-го интервала $f(a)_i$ определяется как отношение частоты интервала m_i к его длине k_i:

$$f(a)_i = \frac{m_i}{k_i}.$$

- ***Относительная плотность*** *i*-го интервала $f(o)_i$ определяется как отношение частости интервала ω_i к длине интервала k_i:

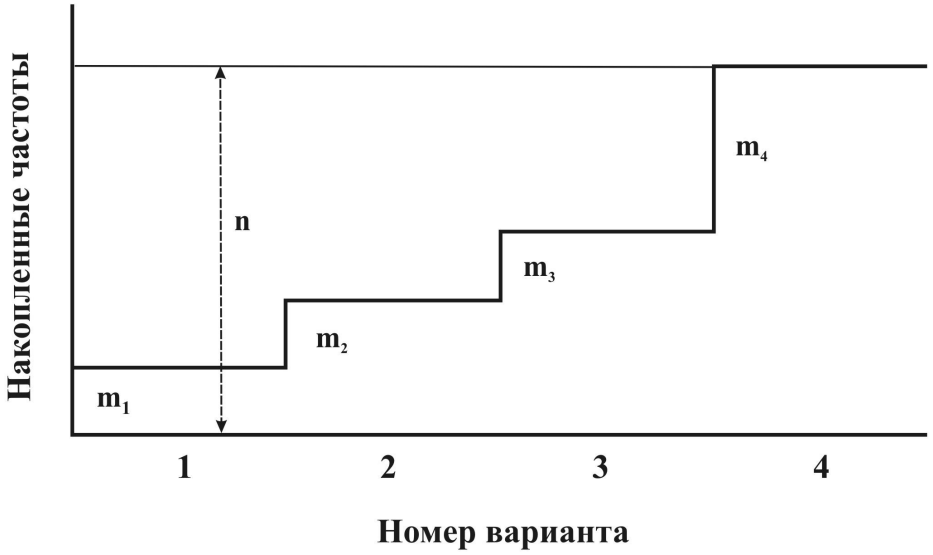
$$f(o)_i = \frac{\omega_i}{k_i}.$$

- ***Накопленные частоты (частости)*** показывают, сколько единиц совокупности (какая их часть) не превышают заданного значения (варианта) *x*.
- Накопленные частоты n_i по данным дискретного ряда можно рассчитать по следующей формуле:

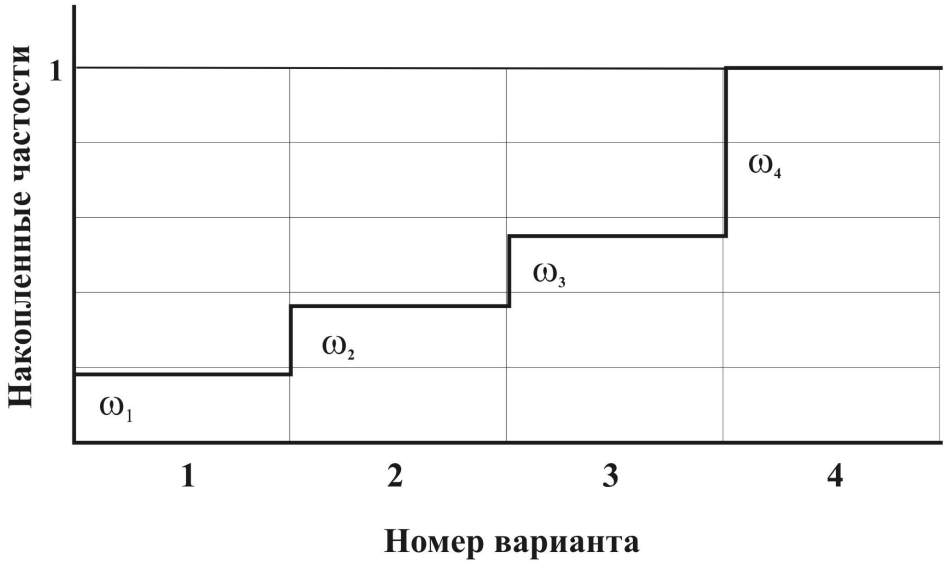
$$n_i = m_1 + m_2 + \ldots + m_i.$$

- Для интервального вариационного ряда ***накопленные частоты*** (частости) ***вычисляются как сумма частот*** (частостей) ***всех интервалов, не превышающих данный***.
- Дискретные и интервальные вариационные ряды графически можно графически представить в виде ***кумуляты***.
- ***При построении кумуляты по данным дискретного ряда по оси абсцисс откладываются значения вариантов, а по оси ординат - накопленные частоты или частости***.

Кумулята накопленных частот



Кумулята накопленных частостей



ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ ВАРИАЦИОННОГО РЯДА

- Одной из основных числовых характеристик ряда распределения (вариационного ряда) является ***средняя арифметическая.***
- ***Простую среднюю арифметическую*** обычно используют, когда все частоты равны единице или одинаковы. Она вычисляется по формуле

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

где x_i – i-тое значение признака;

n - число значений признака (вариантов).

- ***Средняя арифметическая взвешенная*** рассчитывается в том случае, когда частоты отличны друг от друга. Расчет производится по формуле

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

где m_i – частота i-го значения признака.

- Если весами вариационного ряда являются частости вариантов, то расчет средней взвешенной можно производить по формуле

$$\bar{x} = \sum_{i=1}^{n} x_i \cdot \omega_i,$$

где ω_i – частость i-го значения признака.

- ***Иногда средняя арифметическая недостаточно характеризует выборочную совокупность.*** Это происходит в тех случаях, когда колебания вариантов около средней арифметической велики. Например, если бы половина студентов получили оценку 5, а вторая половина - оценку 2, то средний показатель знаний студентов оценивался бы в 3,5 балла, что не отражало бы действительного качества знаний.
- Для того, ***чтобы оценить масштабы колебаний изучаемого признака около средней арифметической, используются различные показатели вариации***.
- К числу основных показателей вариации относятся: ***дисперсия, среднее квадратическое отклонение, коэффициент вариации.***

- По аналогии с математическим ожиданием дисперсия может быть определена с использованием формул:

  - *простая дисперсия*

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

  - *взвешенная дисперсия*

$$D(X)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 m_i}{\sum_{i=1}^{n} m_i};$$

  - *дисперсия с использованием частости вариантов*

$$D(X)=\sum_{i=1}^{n}(x_i-\bar{x})^2\omega_i.$$

- ***Среднеквадратическое отклонение*** определяется формулой

$$\sigma(X) = \sqrt{D(X)}.$$

- ***Коэффициент вариации*** рассчитывается по формуле

$$V(X) = \frac{\sigma(X)}{\overline{x}} \cdot 100\%.$$

- Принято считать, что если коэффициент вариации больше 35 %, то изучаемая статистическая совокупность ***является неоднородной и колеблемость признака высока***. Следовательно, использование средней арифметической для ее характеристики неверно - средняя арифметическая не типична для изучаемой совокупности. В таком случае необходимо использовать ***моду*** или ***медиану*** для характеристики наиболее типичного значения варианта признака рассматриваемой совокупности.

- ***Модой*** вариационного ряда (обозначается символом *Mo)* называется то из значений x_1, x_2, x_3, ... , x_n, ***которому соответствует наибольшая частота***.
- ***Медиана*** - это значение варианта, которое является ***серединой вариационного ряда***, то есть половина вариантов имеют значения большие, чем медиана, а половина вариантов имеют значения меньшие, чем медиана. Если вариантов четное количество, то медиана вычисляется как среднее двух вариантов, находящихся в середине множества.
- Доля единиц, обладающих тем или иным признаком в генеральной совокупности, называется ***генеральной долей*** и обозначается *p*.

- ***Статистическим распределением выборки*** называют перечень возможных значений признака x_i и соответствующих ему частот или относительных частот (частостей).
- ***К выборочным статистикам*** относятся:

  $\overline{X}_{выб}$ - выборочная средняя;

  $\sigma^2_{выб}$ - выборочная дисперсия;

  $\sigma_{выб}$ - выборочное среднее квадратическое отклонение;

  ω - выборочная доля - это доля в выборке элементов, которые обладают некоторым свойством

  $$\omega = m/n,$$

  где n - объем выборки,

  m - количество единиц выборочной совокупности, обладающих этим свойством.

- Статистики, получаемые по различным выборкам, как правило, отличаются друг от друга. Например, если использовать случайный отбор студентов из всего колледжа для определения среднего балла по математическим дисциплинам (из 600 студентов случайным образом отбирают 60 человек), то, проведя этот отбор несколько раз, можно получить разные значения выборочных статистик.
- В первой выборке средний балл может быть равным 3,82, во второй - 3,89, в третьей - 3,78. Поэтому ***статистика, полученная из выборки, отличается от соответствующего параметра в генеральной совокупности***, но ***является оценкой неизвестного параметра генеральной совокупности.***
- ***Оценкой параметра*** называется определенная ***числовая характеристика, полученная из выборки***. Когда оценка определяется одним числом, ее называют ***точечной оценкой.***

- ***В качестве точечных оценок параметров генеральной совокупности используются соответствующие выборочные характеристики***.
- Теоретическое обоснование возможности использования этих выборочных оценок для суждений о характеристиках и свойствах генеральной совокупности дают закон больших чисел и центральная предельная теорема Ляпунова.
- ***Выборочная средняя является точечной оценкой генеральной средней***, т.е. $\overline{X}_{выб} = \overline{X}$.
- Генеральная дисперсия имеет две точечные оценки:

  $\sigma^2_{выб}$ - ***выборочная дисперсия***, исчисляется при $n \geq 30$

$$\sigma^2_{выб} = \frac{\sum_{i=1}^{k} (X_i - \overline{X}_{выб})^2 m_i}{n},$$

  S^2 - ***исправленная выборочная дисперсия***, при $n < 30$

$$S^2 = \frac{n}{n-1} \sigma^2_{выб}.$$

- При больших объемах выборки $\sigma^2_{выб}$ и S^2 практически совпадают.
- Оценка генерального среднеквадратического отклонения производится с использованием формул дисперсий $\sigma^2_{выб}$ и S^2 .