Лекция 2.

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

ПРЕДМЕТ:	***ЗАДАЧА:***
Анализ экспериментальных данных – значений количествен-ного признака (артериальное давление, пульс). **Такой признак – <u>случайная величина</u>.**	**изучить законы распределения иссле-дуемых случайных величин, их характеристики, проверить ряд гипотез, установить, есть ли между величинами связь.**

Часть I.

БАЗОВЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1. ПОНЯТИЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ И ВЫБОРКИ

- **ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ** – ВСЕ МНОЖЕСТВО ОБЪЕКТОВ, ОБЛАДАЮЩИХ ДАННЫМ ПРИЗНАКОМ.
- **ВЫБОРКА** – ЧАСТЬ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ.

- **ЭЛЕМЕНТЫ ВЫБОРКИ** – значения изучаемого признака у входящих в выборку объектов.
- ***ОБЪЕМ*** **ВЫБОРКИ N** – число элементов в ней.
- ***ВАРИАНТЫ*** – отличающиеся друг от друга, различные элементы выборки.

РЕПРЕЗЕНТАТИВНАЯ ВЫБОРКА

Чтобы по выборке можно было судить о генеральной совокупности, выборка должна быть РЕПРЕЗЕНТАТИВНОЙ.

РЕПРЕЗЕНТАТИВНОЙ называется выборка, верно отражающая основные законо-мерности генеральной совокупности.

<u>Условия репрезентативности</u>:

- случайный отбор
- достаточно большой объем

2. СПОСОБЫ ПРЕДСТАВЛЕНИЯ ВЫБОРКИ

- ***ПРОСТОЙ СТАТИСТИЧЕСКИЙ РЯД***
- ***РАНЖИРОВАННЫЙ*** **РЯД**
- ***ВАРИАЦИОННЫЙ*** **РЯД**
- ***ИНТЕРВАЛЬНЫЙ РЯД***

ПРОСТОЙ СТАТИСТИЧЕСКИЙ РЯД –

ПЕРЕЧИСЛЕНИЕ ЭЛЕМЕНТОВ ВЫБОРКИ

В ПОРЯДКЕ ИХ ПОЛУЧЕНИЯ.

ПОСТРОЕНИЕ РАНЖИРОВАННОГО И ВАРИАЦИОННОГО РЯДОВ

РАНЖИРОВАННЫЙ РЯД – ПЕРЕЧИСЛЕНИЕ ЭЛЕМЕНТОВ ВЫБОРКИ В ПОРЯДКЕ ИХ ВОЗРАСТАНИЯ (ИЛИ УБЫВАНИЯ).

При этом каждое значение повторяется столько раз, сколько оно встречается в выборке.

Число появлений данного значения, т.е. варианты, в выборке называется частотой этой варианты, n.

Отношение частоты к объему выборки называется относительной частотой варианты, $W = n / N$.

ВАРИАЦИОННЫЙ РЯД

ВАРИАЦИОННЫЙ РЯД –

ПЕРЕЧИСЛЕНИЕ ВАРИАНТ

В ПОРЯДКЕ ИХ ВОЗРАСТАНИЯ (ИЛИ УБЫВАНИЯ)

С УКАЗАНИЕМ СООТВЕТСТВУЮЩИХ ЧАСТОТ

ИЛИ ОТНОСИТЕЛЬНЫХ ЧАСТОТ.

Таблица вариационного ряда

напоминает ряд распределения ДСВ.

Графическим изображением

вариационного ряда является ***полигон***.

ТАБЛИЦА ВАРИАЦИОННОГО РЯДА

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k
W_i	W_1	W_2	…	W_k

$x_1 < x_2 <\ldots < x_k$

$n_1 + n_2 + \ldots + n_k = N$

$W_1 + W_2 + \ldots + W_k = 1$,
проявление УСЛОВИЯ НОРМИРОВКИ
в статистике.

ПОЛИГОН ЧАСТОТ
или ОТНОСИТЕЛЬНЫХ ЧАСТОТ

- На оси абсцисс - значения x_i ,
  на оси ординат - частоты n_i или
  относительные частоты W_i.

- Точки с координатами (x_i, n_i) соединяются отрезками прямых.

Полученная ломаная – **полигон**.

ПОСТРОЕНИЕ ИНТЕРВАЛЬНОГО РЯДА

ЕСЛИ ОБЪЕМ ВЫБОРКИ ВЕЛИК,
ВАРИАЦИОННЫЙ РЯД ПРЕОБРАЗУЮТ
В ИНТЕРВАЛЬНЫЙ.
В этом случае не перечисляют все варианты, а разбивают вариационный ряд на несколько интервалов и указывают число значений в каждом из них.

№ интервала, k	Границы интервала	Частота, n
1		
2		
…		
m		

Алгоритм построения интервального ряда

1. Определение *разумного* ***числа интервалов***:

$$m = \log_2 N,$$

округляем до целого числа.

2. ***Размах распределения***:

$$L = x_{max} - x_{min}.$$

3. ***Шаг разбиения***, или ***ширина интервала***:

$$h = \Delta x = L / m = \frac{x_{max} - x_{min}}{m}$$

4. ***Границы интервалов***: получаются добавлением шага к предыдущей границе.

Граница может входить только в один интервал, предыдущий или последующий.

- [- граница включается в данный интервал;
- (- граница не включается в интервал.

5. Подсчет ***частоты*** **n** - числа значений, попавших в данный интервал,

и ***относительной частоты***

W = n / N.

ГИСТОГРАММА

Графическое изображение
интервального ряда – ***ГИСТОГРАММА***:
фигура, состоящая из прямоугольников.
Основание каждого
прямоугольника - соответствующий интервал,
высота равна частоте или относительной частоте.

Пример.

У 12 больных гриппом,
прошедших предварительно
вакцинацию,
замерили температуру
в первые сутки болезни.

Получены значения – простой статистический ряд:

37,5; 39,0; 38,1; 38,4; 37,9; 38,4; 38,4; 38,1; 38,6; 38,4; 38,6; 38,4.

Ранжированный ряд:

37,5; 37,9; 38,1; 38,1; 38,4; 38,4; 38,4; 38,4; 38,4; 38,6; 38,6; 39,0.

Вариационный ряд:

x_i	37,5	37,9	38,1	38,4	38,6	39,0
n_i	1	1	2	5	2	1
W_i	1/12	1/12	2/12	5/12	2/12	1/12

ИНТЕРВАЛЬНЫЙ РЯД:

$$m = \log_2 12 \approx 3;$$

$$L = 39{,}0 - 37{,}5 = 1{,}5;$$

$$\Delta x = 1{,}5 / 3 = 0{,}5.$$

Определяем границы первого интервала:

левая граница – $x_{min} = 37{,}5$,

правая граница - $x_{min} + 0{,}5 = 38{,}0$.

Левую границу включаем в первый интервал, правую – нет.

С нее начнется второй интервал.

Таблица интервального ряда

№ интер-вала, k	Границы интервала	Частота, n_k	Относит. частота, W_k
1	[37,5; 38,0)	2	2/12 = 1/6
2	[38,0; 38,5)	7	7/12
3	[38,5; 39,0]	3	3/12

3. ХАРАКТЕРИСТИКИ ВЫБОРКИ

- Средняя выборочная $\bar{x}$
- Выборочная дисперсия $D_в = \sigma^2_в$
- Выборочное средне-квадратическое отклонение $\sigma_в$
- *Мода* Мо
- *Медиана* Ме

СРЕДНЯЯ ВЫБОРОЧНАЯ

вариационного ряда:

$$\bar{x} = \frac{\Sigma x_i n_i}{N}$$

Если все $n_i = 1$, то

$$\bar{x} = \frac{\Sigma x_i}{N}$$

<u>интервального ряда</u>:

$$\overline{x}_и = \frac{\sum c_k n_k}{N}$$

Здесь c_k – середины интервалов:

$c_k = (a + b) / 2 = a + \Delta x / 2$

(a - левая граница интервала,
b - правая граница интервала).

Иными словами,
при вычислении характеристик интервального ряда его заменяют (приближенно)
на вариационный вида:

c_k	c_1	c_2	…	c_m
n_k	n_1	n_2	…	n_m

ВЫБОРОЧНАЯ ДИСПЕРСИЯ

вариационного ряда:

$$\sigma^2_в = \frac{\Sigma (x_i - \bar{x})^2 n_i}{N}$$

Если все $n_i = 1$, то

$$\sigma^2_в = \frac{\Sigma (x_i - \bar{x})^2}{N}$$

интервального ряда:

$$\sigma^2_в = \frac{\Sigma (c_k - \bar{x}_и)^2 n_k}{N}$$

ВЫБОРОЧНОЕ СРЕДНЕКВАДРАТИЧНОЕ ОТКЛОНЕНИЕ

$$\sigma_в = \sqrt{\sigma^2_в}$$

МОДА, МЕДИАНА

- **МОДА – варианта с <u>наибольшей частотой</u>.**

- **МЕДИАНА делит вариационный ряд пополам:**

слева от нее столько же элементов,

сколько справа.

В случае четного числа элементов медиана равна среднему арифметическому двух центральных.

<u>Определяется легко по ранжированному ряду.</u>

В нашем примере

Мо = Ме = 38,4.

4. ТОЧЕЧНЫЕ ОЦЕНКИ ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ ПО ПАРАМЕТРАМ ВЫБОРКИ

ПАРАМЕТРЫ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ – числовые характеристики исследуемой СВ:

- **математическое ожидание (средняя генеральная, средняя теоретическая) μ**
- **дисперсия σ^2**
- **среднеквадратическое отклонение σ**

ИХ *ТОЧЕЧНЫЕ ОЦЕНКИ* - НАИБОЛЕЕ БЛИЗКИЕ К НИМ (согласно теории) ПАРАМЕТРЫ ВЫБОРКИ.

А именно:

точечная оценка

- **средней теоретической – средняя выборочная,**

$$\mu \approx \bar{x}$$

Точечные оценки

- генеральной дисперсии – ***исправленная дисперсия***, s^2:

$$\sigma^2 \approx s^2$$

- среднеквадратичного отклонения – ***стандартное отклонение***, s:

$$\sigma \approx s$$

Чтобы «исправить» выборочную дисперсию, нужно ввести <u>поправочный коэффициент</u>:

$$s^2 = \sigma^2_{в} \cdot \frac{N}{N-1}$$

Таким образом,

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2 n_i}{N - 1}$$

$$s^2_и = \frac{\Sigma (c_k - \bar{x}_и)^2 n_k}{N - 1}$$

Далее $s = \sqrt{s^2}$

Обратите внимание: точечные оценки – **приблизительные** и **случайные** (так как выборка сделана из генеральной совокупности случайным образом, то ее элементы и параметры можно считать случайными величинами)

5. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Дать *ИНТЕРВАЛЬНУЮ ОЦЕНКУ* того или иного параметра генеральной совокупности – значит указать случайный интервал, который с заданной вероятностью γ (гамма) содержит данный параметр.

Этот интервал называется *ДОВЕРИТЕЛЬНЫМ*, а γ – *ДОВЕРИТЕЛЬНОЙ ВЕРОЯТНОСТЬЮ*, или *НАДЕЖНОСТЬЮ.*

Наряду с доверительной вероятностью

используют также понятие

УРОВЕНЬ ЗНАЧИМОСТИ

$$\beta = 1 - \gamma,$$

т.е. вероятность того,

что доверительный интервал *НЕ* содержит в себе оцениваемый параметр.

Доверительный интервал для средней теоретической нормально распределенной величины

Имеет вид

$(\bar{x} - \Delta\ ,\ \bar{x} + \Delta)$.

Здесь Δ – абсолютная погрешность

интервальной оценки μ

по средней выборочной $\bar{x}$.

Но называть ее принято ТОЧНОСТЬЮ оценки.

В данном случае надежность

$\gamma = P(\bar{x} - \Delta < \mu < \bar{x} + \Delta)$

- вероятность того, что

доверительный интервал будет содержать в себе

среднюю теоретическую.

Доверительную вероятность задаем сами,

обычно в медицине это 95%,

то есть $\gamma = 0{,}95$.

Точность Δ рассчитывается по формуле:

$$\Delta = \frac{ts}{\sqrt{N}}$$

Среднюю выборочную и стандартное отклонение находим по выборке.

t определяется

по надежности с помощью известной формулы теории вероятности:

$$\gamma = 2\Phi(t) - 1.$$

Отсюда

$$2\Phi(t) = 1 + \gamma,$$

$$\Phi(t) = \frac{1 + \gamma}{2}$$

Зная Ф (t),

по таблицам нормального распределения

находим **t**.

Так,

если $\gamma = 0{,}95$, то

$$\Phi(t) = 0{,}975$$

и $t \approx 2$.

Если объем выборки невелик, то вместо таблицы нормального распределения нужно воспользоваться таблицей *РАСПРЕДЕЛЕНИЯ СТЬЮДЕНТА*.

Значение t в таблице этого распределения находят по заданным N и γ.

Запишем

АЛГОРИТМ

построения доверительного интервала для средней теоретической нормально распределенной величины.

1. Вычислить $\overline{x}$ и s.
2. По заданной γ рассчитать Ф (t).
3. По значению Ф (t) в таблице найти значение t.
4. Рассчитать точность Δ оценки μ по $\overline{x}$.

5. Записать ответ в виде:

$$\overline{x} - \Delta < \mu < \overline{x} + \Delta.$$

Возможна краткая запись

$$\mu = \overline{x} \pm \Delta$$

ОПРЕДЕЛЕНИЕ МИНИМАЛЬНОГО ОБЪЕМА ВЫБОРКИ,
необходимого для достижения заданной точности с заданной надежностью

Итак, известны γ (и t) и Δ, а найти надо N.

Пользуемся формулой:

$$\Delta = \frac{ts}{\sqrt{N}}$$

Отсюда:

$$\sqrt{N} = \frac{ts}{\Delta}$$

и

$$N = \frac{t^2 s^2}{\Delta^2}$$

Округлить до ближайшего большего целого!