ОСНОВЫ МАТЕМАТИЧЕСКОЙ ОБРАБОТКИ ИНФОРМАЦИИ

ЛЕКЦИИ 5,6

Лектор:
Поздняков Станислав Александрович,
кандидат технических наук, доцент

Зачем нужны меры центральной тенденции?

- Это наиболее важная статистика больших массивов информации (статистика – это любая функция данных).
- Средние значения обладают большей устойчивостью.
- Средние значения – это наиболее репрезентативные значения.
- Если нужно заменить весь массив одним числом – то нужно использовать среднее значение.
- Разные виды средних обладают разными свойствами. Выбор вида среднего выбирается в каждой конкретной ситуации.

Меры центральной тенденции

- Среднее арифметическое
- Среднее гармоническое
- Среднее квадратическое
- Среднее кубическое
- Среднее геометрическое
- Мода
- Медиана

Виды средних

- Автомобиль движется из пункта А в пункт Б с постоянной скоростью 80 км/час, а из пункта Б в пункт А с постоянной скоростью 40 км/час.
- Определить среднюю скорость движения автомобиля.

$$V = \frac{s_{общ}}{t_{общ}} = \frac{2s}{\frac{s}{v_1} + \frac{s}{v_2}} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}} = \frac{2}{\frac{1}{80} + \frac{1}{40}} = 55,3$$

Виды средних

- Диаметр одной корзины подсолнуха равен 10 см, диаметр другой корзины подсолнуха равен 30 см.
- Определить средний диаметр корзин подсолнуха.

$$s_{\hat{i}\acute{a}\grave{u}} = 2\pi\left(\frac{\bar{d}_{\cdot}}{2}\right)^2 = \pi\left(\frac{\bar{d}_1}{2}\right)^2 + \pi\left(\frac{\bar{d}_2}{2}\right)^2$$

$$\bar{d}_{\cdot} = \sqrt{\frac{d_1^2 + d_2^2}{2}} = \sqrt{\frac{100 + 900}{2}} \approx 22,4$$

Виды средних

- Диаметр одного яйца равен 5 см, диаметр другого яйца равен 3 см.
- Определить средний диаметр яиц.

$$v_{общ} = \frac{4}{3}\pi\left(\frac{\bar{d}_{\cdot}}{2}\right)^3 = \frac{4}{3}\pi\left(\frac{\bar{d}_1}{2}\right)^3 + \frac{4}{3}\pi\left(\frac{\bar{d}_2}{2}\right)^3$$

$$\bar{d}_{\cdot} = \sqrt[3]{\frac{d_1^3 + d_2^3}{2}} = \sqrt[3]{\frac{125+27}{2}} \approx 4,24$$

Используемые обозначения

Точка (.) вместо индекса обозначает суммирование по этому индексу

$$x_{\cdot} = \sum_{i=1}^{6} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6$$

Черточка над переменной $(\overline{x}_{\cdot})$ обозначает усреднение по индексам, по которым производится суммирование

$$\overline{x}_{\cdot} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6}{N}$$

Среднее арифметическое и его свойства

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

- Если каждое значение совокупности уменьшить или увеличить на одно и то же число, то среднее ?
- Если каждое значение совокупности умножить или разделить на одно и то же число, то среднее ?

Среднее арифметическое и его свойства

- Среднее двух совокупностей является взвешенным средним этих совокупностей ?
- Сумма отклонений значений совокупности от ее среднего равно ?
- Сумма квадратов отклонений от их средней меньше суммы квадратов отклонений тех же значений от любой другой величины.

Среднее арифметическое и его свойства

$$\sum_{i=1}^{n}\left[x_i-\left(\bar{x}_{\cdot}+c\right)\right]^2=\sum_{i=1}^{n}\left[\left(x_i-\bar{x}_{\cdot}\right)-c\right]^2=$$

$$\sum_{i=1}^{n}\left(x_i-\bar{x}_{\cdot}\right)^2-2c\sum_{i=1}^{n}\left(x_i-\bar{x}_{\cdot}\right)+nc^2$$

Откуда

$$\sum_{i=1}^{n}\left(x_i-\bar{x}_{\cdot}\right)^2\le\sum_{i=1}^{n}\left[x_i-\left(\bar{x}_{\cdot}+c\right)\right]^2$$

Среднее, мода и медиана объединенных групп

$$\bar{x}_{.} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3}{n_1 + n_2 + n_3}$$

- Для того, чтобы найти объединенное среднее, необходимо знать число элементов в подгруппах.
- Для того, чтобы найти объединенную моду, необходимо знать какие элементы встречаются наиболее часто во всех подгруппах.
- Для того, чтобы найти объединенную медиану, необходимо знать распределение во всех подгруппах.

Структурные средние

Мода – это то значение, которое в выборке встречается наиболее часто.

Медиана – это то значение, относительно которого упорядоченная по возрастанию или по убыванию выборка делится пополам.

Как считать доход на душу населения? (как среднее или как медиану?)

Мода

Мода – это наиболее частое значение, а не частота этого значения.

1. Если все значения встречаются в массиве одинаково часто, то массив не имеет моды.

2. Если два соседних значения имеют одинаковую частоту и они больше частоты любого другого значения, то мода есть среднее этих двух значений

3. Если два несмежных значения в массиве имеют равные частоты и они больше частоты любого значения, то массив является бимодальным

Свойства моды

- Мода вычисляется наиболее просто – ее можно определить на глаз.
- Для очень больших массивов данных это достаточно стабильная мера центра распределения.
- Во многих задачах мода близка к двум другим мерам – медиане и среднему.

Вычислить меры центральной тенденции

Диаметры корзинок подсолнухов:

15, 13, 11, 16, 8, 13, 15, 16, 17, 15

Вычислить

Mo =

Me =

$\bar{x}. =$

Интерпретация моды, медианы и среднего

Интерпретация осуществляется в терминах ошибок, возникающих из-за того, что все значения в выборке заменяются одним значением (наиболее репрезентативным)

Мода – наиболее репрезентативное значение в том смысле, что совпадает с наибольшим числом значений в выборке.

Интерпретация моды, медианы и среднего

Медиана – это такая точка на числовой оси, для которой сумма абсолютных разностей всех значений меньше суммы разностей для любой другой точки.

Среднее – обеспечивает минимальное значение суммы квадратов отклонений значений от среднего.

Критерии выбора меры центральной тенденции

1. В малых группах мода очень нестабильна (1,1,1,3,5,7,7,8) Mo =1. Но если 1□0 и 1□2, то Mo =7.

2. На медиану не влияют большие и малые (экстремальные) значения

3. На величину среднего влияет каждое значение. (Как?)

Для каких массивов среднее, мода и медиана совпадают?

Задача 1. Где строить дом?

	п.1		п.2			п.3	п.4	п.5
0	1	2	3	4	5	6	7	8

Задача 2. Какую меру центральной тенденции выбрать?

Доходы 5 мужчин:

1. 25 центов
2. 25 центов
3. 2 000 долларов
4. 15 000 долларов
5. 5 000 000 долларов

Как охарактеризовать их средний доход?

В США средний доход – это медиана, а не среднее

Рекомендуемая литература

1. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 2004, 479 с.

2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 2004, 400 с.

3. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. Пер. с англ. – М.: Издательство «Прогресс», 1976. -496 с.

4. Маслак А.А. Основы планирования и анализа сравнительного эксперимента в педагогике и психологии. – Курск: РОСИ, 1998. – 167 с.

Меры вариабельности данных

- Меры центральной тенденции говорят нам о концентрации данных на числовой оси. Каждая такая мера в каком-то смысле наилучшим образом «представляет» данные.
- Меры центральной тенденции игнорируют различия между данными.
- Для измерения вариабельности данных требуются другие описательные статистики.

Зачем нужны меры вариабельности данных?

- Научная работа связана с понятием вариабельности данных. Если есть много необъяснимых причин вариабельности, прогнозы будут неточными.
- Задача науки найти причины вариабельности данных и тем самым увеличить точность прогноза.
- Например установлено, что наследственность и окружающая среда влияют на IQ ребенка. Поэтому информация о родителях ребенка и его воспитании позволяет более точно прогнозировать его умственное развитие в зрелости. Без такой информации прогноз будет менее точным.

Наиболее часто используемые меры вариабельности данных

- Лимиты
- Размах
- Квантили
- Дисперсия
- Стандартная ошибка
- Среднее отклонение
- Коэффициент вариации

Лимиты

- Это самая простая мера изменчивости.
- Определяется минимальное (Xmin) и максимальное значение (Xmax) массива данных. Между этими статистиками находятся все данные массива.
- Несмотря на свою простоту эта мера используется редко, потому что экстремальные значения сильно подвержены ошибкам.
- Поэтому трудно определить влияние факторов на вариабельность данных.

Размах

Определяет расстояние на числовой оси, в пределах которого варьируются данные.

$$R=X_{max}-X_{min}.$$

Исключающий размах – это разность максимального и минимального значений.

Включающий размах – это разность между естественной верхней границей интервала, содержащего максимальное значение и естественной нижней границей интервала, содержащего минимальное значение.

Размах

- Например рост 5 мальчиков равен:

150, 155, 157, 165 и 168

- Исключающий размах равен:

168-150=18

- Включающий размах равен:

168,5 – 149,5=19

Квантили

Это характеристики вариационного ряда, которые отсекают определенную его часть. Наиболее часто используются квартили, децили и процентили.

Квартиль – это статистика, отсекающая ¼ часть ряда. Три квартиля Q1, Q2 и Q3 делят ряд на четыре, равные по объемы части (кварты).

Квантили

Дециль (Di) – это статистика, отсекающая 1/10 часть ряда. Девять децилей делят ряд на 10 равных частей.

Процентиль (Pi) - это статистика, отсекающая 1/100 часть ряда. Девяносто девять процентилей делят ряд на 100 равных частей.

Зачем нужны квантили?

Квантили, как и медиана, - это важные характеристики вариационного ряда, особенно для асимметричных распределений.

Часто квантили используются для установления границ тех или иных нормативов.

Зачем нужны квантили?

Размах от 90-ого до 10-ого процентиля является более стабильной мерой, чем размах.

Полу-междуквартильный размах Q3-Q1 содержит 50% наблюдений вариационного ряда.

Дисперсия

- При вычислении всех предыдущих мер вариабельности не учитывалось каждое отдельное значение массива данных.
- Отклонения наблюдений от мер центральной тенденции несут информацию о вариабельности данных. Чем больше отклонения, тем больше вариабельность.

Однако: $$\sum_{i=1}^{n}\left(y_i - \bar{y}_.\right) \equiv 0$$

Формула для вычисления дисперсии

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Свойства дисперсии

- Прибавление константы *c* к каждому значению не влияет на дисперсию (а на среднее?)
- Умножение каждого значения на константу *c* увеличивает дисперсию в c^2 раз.
- Дисперсия объединенной совокупности зависит как от дисперсий, так и от средних объединяемых групп

$$s^2 = \frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2 + n_a(\bar{x}_{.a} - \bar{x}_{..})^2 + n_b(\bar{x}_{.b} - \bar{x}_{..})^2}{n_a + n_b - 1}$$

Задача 3. Вычислить средние и дисперсии совокупностей:

A (3, 3, 3, 3) и B (7,7,7,7)

$$\bar{x}_a = \quad \bar{x}_a = \quad \bar{x}_{a+b} =$$

$$s_a^2 = \quad s_a^2 = \quad s_{a+b}^2 =$$

Стандартное отклонение

Эта мера тесно связана с дисперсией. Стандартное отклонение – это положительный корень из дисперсии.

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

Стандартное отклонение измеряется в тех же единицах, что и исходные данные. Например, как интерпретировать $кг^2$ или $л^2$?

Полезность этой меры еще и в том, что для многих распределений мы знаем, какая доля наблюдений находится внутри одного, двух, трех и более стандартных отклонений. Поэтому эта мера используется наиболее часто.

Среднее отклонение

Формула имеет вид

$$\sum_{i=1}^{N} \left| x_i - \bar{x}_{\cdot} \right| / N$$

- Несмотря на легкость вычисления и простоту интерпретации эта мера используется редко.
- Это объясняется тем, что эта мера неудобна для аналитический преобразований (например необходимо брать производную для поиска минимума функции).
- Эта формула неудобна также для вычисления стандартизированных отклонений.

Коэффициент вариации

Формула для вычисления имеет вид:

$$\nu = s / \bar{x}.$$

Эта мера позволяет сравнивать вариабельность признаков имеющих разные единицы измерения.

Эта мера часто используется в биологии и других науках, где измеряемые признаки отличны от нуля.

Стандартизированные данные

Формула для вычисления имеет вид:

$$z_i = \frac{x_i - \bar{x}_.}{s_x}$$

- Таким образом любое множество данных на основе вычисленных среднего и стандартного отклонения можно преобразовать в стандартизированное множество с нулевым средним и единичной дисперсией.
- Это удобно для проверки различных статистических гипотез.

Задача 4. Вычислить средние и дисперсии двух массивов

x_1	10	15	20	25	30	35	40	45	50	x_1.
x_2	10	28	28	30	30	30	32	32	50	x_2.
$(x_1-x.)$										□
$(x_2-x.)$										□
$(x_1-x.)^2$										□
$(x_2-x.)^2$										□

Задача 5.
Вычислить дисперсию тестового балла

№	x_i	$(x_i - \bar{x}.)$	$(x_i - \bar{x}.)^2$
1	6	0	0
2	4	-2	4
3	7	1	1
4	10	4	16
5	7	1	1
6	2	-4	16
Сумма	36	0	38

$$S^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x}.)^2}{N-1} = \frac{38}{5} \qquad S = \sqrt{S^2} = \sqrt{7,6} = 2,76$$

Рекомендуемая литература

1. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 2004, 479 с.

2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 2004, 400 с.

3. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. Пер. с англ. – М.: Издательство «Прогресс», 1976. -496 с.

4. Маслак А.А. Основы планирования и анализа сравнительного эксперимента в педагогике и психологии. – Курск: РОСИ, 1998. – 167 с.