План лекции 4
«Меры изменчивости »

1. Лимиты
2. Размах
3. Квантили
4. Размах от 90-го до 10-го процентиля
5. Полу-междуквартильный размах
6. Дисперсия
7. Свойства дисперсии
8. Стандартное отклонение
9. Среднее отклонение
10. Коэффициент вариации
11. Стандартизированные данные
12. Асимметрия
13. Эксцесс

Вариабельность 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

1. Лимиты
2. Размах
3. Квантили
4. Дисперсия
5. Стандартная ошибка
6. Среднее отклонение
7. 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=Xmax-Xmin.

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.

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).

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.

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

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

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

1. Прибавление константы *c* к каждому значению не влияет на дисперсию (а на среднее?)
2. Умножение каждого значения на константу *c* увеличивает дисперсию в c^2 раз.
3. 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)

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

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

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

$$v = s \Big/ \bar{x}.$$

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. Вычислить дисперсию тестового балла

№ п.п.	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.