Лекции по ООРИ

К.С. Пилипенко

26 марта 2023 г.

Этапы статистического анализа

1. Что изучается? Какие статистические гипотезы?
2. Какие данные?
3. Выбор критерия.
4. Интерпретация результатов
5. Вывод

1 Стандартная ошибка средней арифметической

Стандартная ошибка средней арифметической Описание данных

Среднее арифметическое, как известно, используется для получения обобщающей характеристики некоторого набора данных. Если данные более-менее однородны и в них нет аномальных наблюдений (выбросов), то среднее хорошо обобщает данные, сведя к минимуму влияние случайных факторов (они взаимопогашаются при сложении).

Когда анализируемые данные представляют собой выборку (которая состоит из случайных значений), то среднее арифметическое часто (но не всегда) выступает в роли приближенной оценки математического ожидания. Почему приближенной? Потому что среднее арифметическое – это величина, которая зависит от набора случайных чисел, и, следовательно, сама является случайной величиной. При повторных экспериментах (даже в одних и тех же условиях) средние будут отличаться друг от друга.

Для того, чтобы на основе статистического анализа данных делать корректные выводы, необходимо оценить возможный разброс полученного результата. Для этого рассчитываются различные показатели вариации. Но то исходные данные. И как мы только что установили, среднее арифметическое также обладает разбросом, который необходимо оценить и учитывать в дальнейшем (в выводах, в выборе метода анализа и т.д.).

Интуитивно понятно, что разброс средней должен быть как-то связан с разбросом исходных данных. Основной характеристикой разброса средней выступает та же дисперсия.

2 Математическое ожидание

Свойства математического ожидания:

1. M[X+Y] = M[X] + M[Y], где P(X|Y) = P(X) и P(Y|X) = P(Y)
   Доказательство

$$M[X+Y] = \sum_i \sum_j (x_i + y_i) p_{xi} p_{yj} = \sum_i \sum_j x_i p_{xi} p_{yj} + \sum_i \sum_j y_j p_{xi} p_{yj} = \sum_i x_i p_{xi} \sum_j p_{yj} + \sum_j p_{yj} y_j \sum_i p_{xi}$$

2. $M[XY] = M[X]M[Y]$
   Доказательство

$$M[XY] = \sum_i \sum_j x_i y_i p_{xi} p_{yj} = \sum_i x_i p_{xi} \sum_j x_j p_{xj} = M[X]M[Y]$$

Свойства дисперсии:

1. $D[X] = M[X^2] - M^2[X]$;
2. $D[a] = M[a^2] - M^2[a] = a^2 - a^2 = 0$;
3. $D[X+Y] = D[X] + D[Y]$
   Доказательство

$$D[X+Y] = M[(X+Y)^2] - M^2[X+Y] =$$
$$= M[X^2] + 2M[XY] + M[Y^2] - (M^2[X] + 2M[X]M[Y] + M^2[Y]) =$$
$$= M[X^2] - M^2[X] + M[Y^2] - M^2[Y] = D[X] + D[Y]$$

Замечание $D[X-Y] = D[X] + D[Y]$!

4. $D[X+a] = D[X]$;

5. $D[aX] = a^2 D[X]$

Дисперсия выборочных данных – это средний квадрат отклонения от средней, и рассчитать ее по исходным данным не составляет труда, например, в Excel предусмотрены специальные функции. Однако, как же рассчитать дисперсию средней, если в распоряжении есть только одна выборка и одно среднее арифметическое? Расчет дисперсии и стандартной ошибки средней арифметической

Чтобы получить дисперсию средней арифметической нет необходимости проводить множество экспериментов, достаточно иметь только одну выборку. Это легко доказать. Для начала вспомним, что средняя арифметическая (простая) рассчитывается по формуле:

формула средней арифметической

где xi – значения переменной, n – количество значений.

Теперь учтем два свойства дисперсии, согласно которым, 1) — постоянный множитель можно вынести за знак дисперсии, возведя его в квадрат и 2) — дисперсия суммы независимых случайных величин равняется сумме соответствующих дисперсий. Предполагается, что каждое случайное значение xi обладает одинаковым разбросом, поэтому несложно вывести формулу дисперсии средней арифметической:

Формула дисперсии средней арифметической

Используя более привычные обозначения, формулу записывают как:

Дисперсия средней арифметической

где σ2 – это дисперсия, случайной величины, причем генеральная.

На практике же, генеральная дисперсия известна далеко не всегда, точнее совсем редко, поэтому в качестве оной используют выборочную дисперсию:

Дисперсия средней арифметической по выборке

Стандартное отклонение средней арифметической называется стандартной ошибкой средней и рассчитывается, как квадратный корень из дисперсии.

Формула стандартной ошибки средней при использовании генеральной дисперсии

Стандартная ошибка средней

Формула стандартной ошибки средней при использовании выборочной дисперсии

Стандартная ошибка средней по выборке

Последняя формула на практике используется чаще всего, т.к. генеральная дисперсия обычно не известна. Чтобы не вводить новые обозначения, стандартную ошибку средней обычно записывают в виде соотношения стандартного отклонения выборки и корня объема выборки. https://statanaliz.info/statistica dannyx/dispersiya-srednej-arifmetiskoj/

3 Статистическое определение вероятности. Метод Монте-Карло

4 Виды распределений

1. **Равномерное распределение** *Функция распределения*

$$f(x) = \begin{cases} 0, x < a \\ \frac{x-a}{b-a}, a \leqslant x < b \\ 1, x \geqslant b \end{cases} \tag{1}$$

Плотность распределения

$$\omega(x) = \begin{cases} \frac{1}{b-a}, x \in [a,b] \\ 0, x \notin [a,b] \end{cases} \tag{2}$$

2. **Распределение Пуассона**

Описывает дискретную случайную величину, представляющую собой чис-

ло событий, произошедших за фиксированное время, при условии, что данные события происходят независимо друг от друга с некоторой фиксированной средней интенсивностью.

Распределение Пуассона применимо, если:

(a) случайная величина принимает только положительные значения,

(b) если длина интервала (например, t – время наблюдения) стремится к нулю, то вероятность одного события также стремится к нулю,

(c) события, относящиеся к неперекрывающимся интервалам, являются статистически независимыми.

Вероятность наблюдения n событий, произошедших за время t определяется формулой (функция вероятности):

$$P(n) = \frac{\bar{n}^n}{n!} e^{-\bar{n}}, \tag{3}$$

$\bar{n}$ – математическое ожидание случайной величины(среднее количество событий за промежуток времени t) *Плотность распределения*

$$\omega(n) = e^{-\bar{n}} \sum_{j=0}^{n} \frac{\bar{n}^j}{j!}, \tag{4}$$

Пример. Устройство состоит из 1000 элементов, работающих независимо один от другого. Вероятность отказа любого элемента в течение времени T равна 0,002. Найти вероятность того, что за время T откажут ровно три элемента. Решение
Математическое ожидание в этом случае будет определяться как

$$N_0 = \sum_{i=1}^{n} P_i, \tag{5}$$

где n — это число элементов, а P_i — вероятность выхода из строя одного элемента и тогда $N_0 = 2$. И тогда вероятность выхода из строя трёх элементов будет определяться следующим образом:

$$P_3 = \frac{2^3}{3!} e^{-2} \approx 0{,}18 \tag{6}$$

Для расчёта дисперсия будем считать $\bar{n} = M[n]$, тогда:

$$D[n] = M[n^2] - \bar{n}^2 = \sum_0^\infty n^2 \frac{\bar{n}^n}{n!} e^{-\bar{n}} - \bar{n}^2 = \bar{n}e^{-\bar{n}} \sum_0^\infty n \frac{\bar{n}^{n-1}}{(n-1)!} - \bar{n}^2 =$$
$$= \bar{n}e^{-\bar{n}} \left(\sum_0^\infty (n-1) \frac{\bar{n}^{n-1}}{(n-1)!} + \sum_0^\infty \frac{\bar{n}^{n-1}}{(n-1)!} \right) - \bar{n}^2 =$$
$$= \bar{n}e^{-\bar{n}} \left(\bar{n} \sum_0^\infty \frac{\bar{n}^{n-2}}{(n-2)!} + \sum_0^\infty \frac{\bar{n}^{n-1}}{(n-1)!} \right) - \bar{n}^2 =$$
$$= \bar{n}e^{-\bar{n}}(\bar{n}e^{\bar{n}} + e^{\bar{n}}) - \bar{n}^2 = \bar{n}. \quad (7)$$

В выводе исполозовано разложение функции $e^{\bar{n}}$ в ряд Маклорена $\left(e^{\bar{n}} = \sum_0^\infty \frac{\bar{n}^n}{n!}\right)$

3. **Нормальное распределение. Расп-е Гаусса.**

Плотность распределения

$$\omega(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \quad (8)$$

Функция распределения

$$f(x) = \frac{1}{2} \left[1 + erf\left(\frac{x - \bar{x}}{\sigma\sqrt{2}} \right) \right], \quad (9)$$

где $erf(\frac{x-\bar{x}}{\sigma\sqrt{2}})$ — функция ошибок ($erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$)

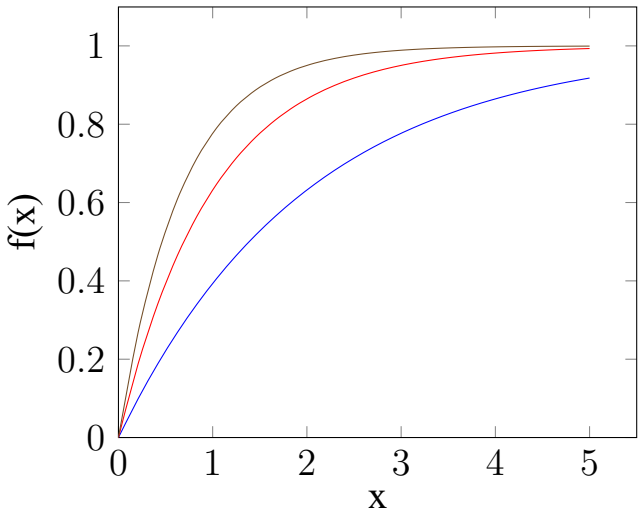
4. **Экспоненциальное распределение.**

Плотность распределения

$$\omega(x) = \begin{cases} \lambda e^{-\lambda x}, x \geqslant 0, \\ 0, x \leqslant 0 \end{cases} \quad (10)$$

Функция распределения

$$f(x) = \begin{cases} 1 - e^{-\lambda x}, x \geqslant 0, \\ 0, x \leqslant 0 \end{cases} \quad (11)$$



5. **Биномиальное распределение**

   Функция вероятности

$$p(x) = C_n^k p^k q^{n-k}, \tag{12}$$

   где $C_n^k = \frac{n!}{(n-k)!k!}$; p, q — вероятность успеха и неудачи, соотвественно, k — количество успехов, n — общее количество неудач.

6. Геометрическое распределение

5 Распределение χ^2. Критерий Пирсона

До конца XIX века нормальное распределение считалась всеобщим законом вариации данных. Однако К. Пирсон заметил, что эмпирические частоты могут сильно отличаться от нормального распределения. Встал вопрос, как это доказать. Требовалось не только графическое сопоставление, которое имеет субъективный характер, но и строгое количественное обоснование.

Так был изобретен критерий χ^2 (хи квадрат), который проверяет значимость расхождения эмпирических (наблюдаемых) и теоретических (ожидаемых) частот. Это произошло в далеком 1900 году, однако критерий и сегодня на ходу. Более того, его приспособили для решения широкого круга задач. Прежде всего, это анализ категориальных данных, т.е. таких, которые выражаются не количеством, а принадлежностью к какой-то категории. Например, класс автомобиля, пол участника эксперимента, вид растения и т.д. К таким данным нельзя применять математические операции вроде сложения и умножения, для них можно только подсчитать частоты.

Плотность распределения χ^2 [**?**]:

$$p_n(z) \begin{cases} 0, z < 0 \\ \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} z^{\frac{n-2}{2}} e^{-\frac{z}{2}}, z > 0, \end{cases} \tag{13}$$

где $\Gamma(x) = \int\limits_0^\infty x^{z-1}e^{-z}dz$ — гамма-функция Эйлера. Виды распределений

- Распределение Стьюдента (t-распределение). Необходимые условия для критерия Стьюдента
  - Нормальное распределение
  - Количественные данные
  - две независимые выборки
- Распределение Фишера