Теория вероятностей. Лекция седьмая
Дисперсия и не только

Дмитрий Валерьевич Хлопин
glukanat@mail.ru

Институт математики и механики им. Н.Н.Красовского

12.10.2018

Напомним про математическое ожидание

Пусть в вероятностном пространстве $(\Omega, \mathcal{F}, \mathbb{P})$ задана дискретная случайная величина $\xi : \Omega \to \mathbb{R}$ с распределением

x_1	x_2	x_3	x_4	$\ldots$
p_1	p_2	p_3	p_4	$\ldots$

Для всякой дискретной случайной величины ξ значение выражения

$$\mathbb{E}\xi \triangleq \sum_{i \in \mathbb{N}} p_i x_i$$

назовем математическим ожиданием случайной величины ξ.
Будем говорить, что "математическое ожидание существует", если значение корректно определено и конечно.

Что дальше?

- распределение случайных величин
- медиана, математическое ожидание
- независимость случайных величин
- дисперсия
- производящие функции как матожидание
- ковариация и корреляция
- совместное распределение случайных величин, маргинальные распределения
- условное математическое ожидание

Дисперсия

Дисперсией называют центральный момент второго порядка

$$\mathbb{D}\xi = \mathbb{E}\Big((\xi - \mathbb{E}\xi)^2\Big).$$

Корень из нее называют **среднеквадратичным отклонением**:

$$\sigma = \sqrt{\mathbb{D}\xi}.$$

Грубо говоря, дисперсия охарактеризует разбросанность распределения, среднеквадратичное отклонение — типичный размах вокруг матожидания.

Терминологическое замечание 3. В зарубежных книжках обычно дисперсию обозначают $Var\,\xi$, встречалось и обозначение σ^2.

Две формулы для дисперсии

$$\mathbb{D}\xi \triangleq \mathbb{E}\Big((\xi - \mathbb{E}\xi)^2\Big) = \sum_{i\in\mathbb{N}} p_i\Big(x_i - \sum_{j\in\mathbb{N}} p_j x_j\Big)^2.$$

Еще один способ считать дисперсию:

$$\begin{aligned}
\mathbb{D}\xi &= \mathbb{E}\Big((\xi - \mathbb{E}\xi)^2\Big) = \mathbb{E}\Big(\xi^2\Big) - \mathbb{E}\Big(2\xi\mathbb{E}\xi\Big) + \mathbb{E}\Big((\mathbb{E}\xi)^2\Big) \\
&= \mathbb{E}\Big(\xi^2\Big) - \mathbb{E}\xi\mathbb{E}\Big(2\xi\Big) + (\mathbb{E}\xi)^2 \\
&= \mathbb{E}\Big(\xi^2\Big) - (\mathbb{E}\xi)^2.
\end{aligned}$$

$$\mathbb{D}\xi \triangleq \mathbb{E}\Big(\xi^2\Big) - (\mathbb{E}\xi)^2 = \sum_{i\in\mathbb{N}} p_i x_i^2 - \Big(\sum_{i\in\mathbb{N}} p_i x_i\Big)^2.$$

Свойства дисперсии [с-но]

1^0 $\mathbb{D}\xi$ неотрицательна;

2^0 $\mathbb{D}\xi = 0$ тогда и только тогда, когда ξ вырождена (равна константе);

3^0 $\mathbb{D}(\xi_1 + \dots + \xi_n) = \mathbb{D}\xi_1 + \dots + \mathbb{D}\xi_n$ в случае попарно независимых случайных величин $\xi_1, \dots, \xi_n$;

4^0 $\mathbb{D}(c\xi)$ равна $c^2\mathbb{D}\xi$ для всякой константы $c \in \mathbb{R}$;

5^0 $\mathbb{D}\xi$ существует и конечна тогда и только тогда, когда конечно $\mathbb{E}|\xi|^2$.

Про случайную величину ξ говорят, что она интегрируема с квадратом, если конечно $\mathbb{E}|\xi|^2$.

Весьма пользительное наблюдение. Пусть $\xi_1, \xi_2, \dots, \xi_n$ одинаково распределены и интегрируемы с квадратом. Посчитайте $\mathbb{D}\frac{\xi_1+\xi_2+\dots+\xi_n}{n}$ для случая, когда $\xi_1, \dots, \xi_n$ попарно независимы, и для случая, когда $\xi_1, \dots, \xi_n$ совпадают. Сделайте вывод самостоятельно.

Типичные задачки на подсчет матожидания и дисперсии

[суммирование по частям] Урна содержит 100 шаров с номерами от 1 до 100. Пусть K - наибольший номер, полученный при 10 их поштучных извлечениях с возвращением. Найдите $\mathbb{E}K$ и $\mathbb{D}K$.
[сумма матожиданий равна матожиданию суммы] Как с помощью листа тетрадки в линейку (с расстоянием между линиями 1 см) и иголки длиной $1,5$ см, согнутой в кочергу (буквой Γ), определить число π? То же с иголкой, согнутой в форме буквы Π. А для круглой иголки произвольной длины?
[производящая функция] В бар с целью застрелить Бессмертного Джо за час в среднем заходит 1 снайпер и 2 ламера. Предполагая, что каждый заходящий не уйдет, пока не пристрелит Джо, при этом ламер попадает с вероятностью $1/4$, снайпер — с вероятностью $3/4$, найдите среднее за час число выстрелов.

Мощный метод посчитать: производящие функции

Пусть случайная величина ξ принимает лишь целочисленные неотрицательные значения. Производящей функцией (pgf: probability generating function) случайной величины ξ называют ряд

$$\phi_\xi(t) \triangleq \sum_{k=0}^{\infty} \mathbb{P}(\xi = k) t^k = \mathbb{E} t^\xi \qquad \forall t.$$

Подумать: откуда, из какого множества, здесь t?

0^0 $\phi_\xi(1) = 1$;

1^0 $\mathbb{P}(\xi = k) = \frac{1}{k!}\phi_\xi^{(k)}(0)$.

Подумать: а производная точно существует?

Подумать: не только по дискретному распределению можно получить производящую функцию, но и по функции можно восстановить распределение.

Производящие функции: свойства [с-но]

Пусть случайные величины $\xi, \xi_1, \ldots, \xi_k, \ldots$ принимают лишь целые неотрицательные значения.

- 0^0 $\phi_\xi(1) = 1$;
- 1^0 $\mathbb{P}(\xi = k) = \frac{1}{k!}\phi_\xi^{(k)}(0)$;
- 2^0 $\phi'_\xi(1) = \mathbb{E}\xi$, $\phi_\xi^{(k)}(1) = \mathbb{E}\xi(\xi-1)\ldots(\xi-k+1)$, если соответствующие матожидания существуют;
- 3^0 $\phi_{a\xi+b}(t) = t^b\phi_\xi(t^a)$ для целых неотрицательных a, b;
- 4^0 $\phi_{\xi_1+\xi_2+\dots+\xi_k} = \phi_{\xi_1}\phi_{\xi_2}\ldots\phi_{\xi_k}$ для независимых (как?) случайных величин $\xi_1, \xi_2, \ldots, \xi_k$;
- 5^0 [0,5 баллов] для независимых (как?) случайных величин $\xi_1, \xi_2, \ldots, \xi_k$ и независящей от них случайной величины η, принимающей натуральные значения, выполнено $\phi_{\xi_1+\xi_2+\dots+\xi_\eta} = \phi_\eta(\phi_{\xi_1})$.

Посчитать матожидание через производящие функции: геометрическое распределение

Пример 1. Для $\xi \in Geom^1(p)$ получаем

$$
\begin{aligned}
\phi_\xi(t) &= \sum_{k=1}^{\infty} q^{k-1} p t^k = pt \sum_{k=0}^{\infty} (qt)^k = \frac{pt}{1-qt}, \\
\phi'_\xi(t) &= \frac{d}{dt}\left(\frac{pt}{1-qt}\right) = \frac{p(1-qt)+pqt}{(1-qt)^2} = \frac{p}{(1-qt)^2}, \\
\phi''_\xi(t) &= \frac{d}{dt}\left(\frac{p}{(1-qt)^2}\right) = \frac{2pq}{(1-qt)^3}, \\
\mathbb{E}\xi = \phi'_\xi(1) &= \frac{1}{p}, \quad \mathbb{E}(\xi^2) = \mathbb{E}\xi(\xi-1) + \mathbb{E}\xi = \phi''_\xi(1) + \frac{1}{p} = \frac{2q}{p^2} + \frac{1}{p}, \\
\mathbb{D}\xi &= \frac{2q}{p^2} + \frac{p}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}.
\end{aligned}
$$

Посчитать матожидание через производящие функции: биномиальное распределение

Пример 2. Для распределенной по Бернулли случайной величины $\xi \in B(1,p)$ выполнено

$$\phi_\xi(t) = q + pt,\ \mathbb{E}\xi = \phi'_\xi(1) = p,\ \mathbb{D}\xi = \phi''_\xi(1) + \phi'_\xi(1) - (\phi'_\xi(1))^2 = pq.$$

Пример 3. Для случайной величины $\eta \in B(n,p)$, как суммы n независимых распределенных по Бернулли случайных величин, имеем

$$\phi_\eta(t) = (q + pt)^n,\ \mathbb{E}\eta = np,\ \mathbb{D}\eta = npq.$$

Ковариация

Здесь и далее рассматриваем лишь дискретные случайные величины, интегрируемые с квадратом ($\mathbb{E}|\xi|^2 < +\infty$).

Ковариацией двух случайных величин назовем значение выражения

$$cov(\xi_1, \xi_2) = \mathbb{E}(\xi_1 - \mathbb{E}\xi_1)(\xi_2 - \mathbb{E}\xi_2).$$

Как и дисперсию, ковариацию можно выразить по-другому:

$$cov(\xi_1, \xi_2) = \mathbb{E}(\xi_1 - \mathbb{E}\xi_1)(\xi_2 - \mathbb{E}\xi_2) = \mathbb{E}(\xi_1\xi_2) - \mathbb{E}\xi_1\mathbb{E}\xi_2.$$

Подумать: ковариация инвариантна относительно сдвига на число любого из своих аргументов.

Свойства ковариации [с-но]

Для любых дискретных случайных величин ξ_1, ξ_2, η, для любых $c_1, c_2 \in \mathbb{R}$:

1^0 $\mathbb{D}(\xi_1 + \xi_2) = \mathbb{D}\xi_1 + \mathbb{D}\xi_2 + 2cov(\xi_1, \xi_2)$;

2^0 $cov(\xi_1, \xi_2) = cov(\xi_2, \xi_1)$;

3^0 $cov(\xi_1 + \xi_2, \eta) = cov(\xi_1, \eta) + cov(\xi_2, \eta)$;

4^0 $cov(\xi_1, \xi_2) = 0$ в случае независимых ξ_1, ξ_2;

5^0 $cov(\xi_1 + c_1, \xi_2 + c_2) = cov(\xi_1, \xi_2)$;

6^0 $cov(\xi_1, \xi_1) = \mathbb{D}\xi_1$;

7^0 $cov(c_1\xi_1, c_2\xi_2) = c_1c_2cov(\xi_1, \xi_2)$.

Коррелированные случайные величины

Корреляцией двух случайных величин ξ_1, ξ_2 называют выражение

$$\rho(\xi_1, \xi_2) = \frac{cov(\xi_1, \xi_2)}{\sqrt{\mathbb{D}\xi_1 \cdot \mathbb{D}\xi_2}}.$$

Если $cov(\xi_1, \xi_2) = 0$, то случайные величины называют *некоррелированными*, в частности отличные от констант независимые случайные величины всегда некоррелированны (функционально зависимы).

Подумать: корреляция инвариантна относительно сдвига и растяжения (на число) каждого аргумента.

Некоррелированность, независимость и линейная зависимость

$\rho(\xi_1, \xi_2) = 0$ выполнено в точности для некоррелированных случайных величин (отличных от константы), в частности это так для независимых случайных величин.

[С-но] Приведите пример некоррелированных, но тем не менее зависимых случайных величин.

Теорема. Если коэффициент корреляции существует, то
1) $\rho(\xi_1, \xi_2) \in [-1, 1]$;
2) в случае $|\rho(\xi_1, \xi_2)| = 1$ для некоторых не равных нулю одновременно чисел b, c выполнено $\xi_2 = c + b\xi_1$ почти для всех ω, и b имеет знак $\rho(\xi_1, \xi_2)$.

Сильная коррелированность

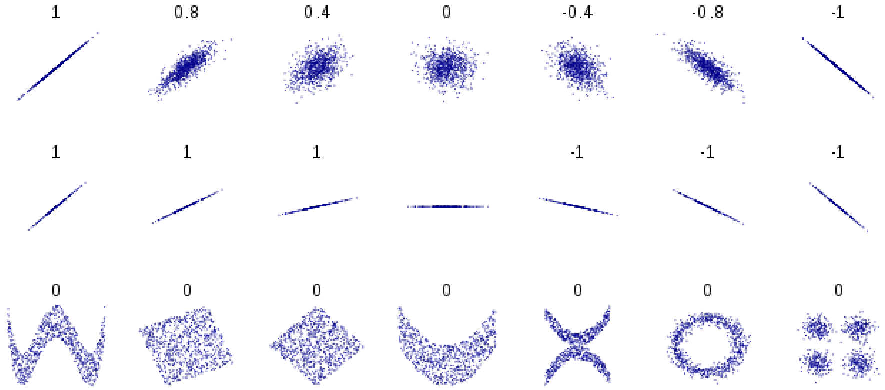
Теорема. Если коэффициент корреляции существует, то
1) $\rho(\xi_1, \xi_2) \in [-1, 1]$;
2) в случае $|\rho(\xi_1, \xi_2)| = 1$ для некоторых не равных нулю одновременно чисел b, c выполнено $\xi_2 = c + b\xi_1$ почти для всех ω, и b имеет знак $\rho(\xi_1, \xi_2)$.

Доказательство. Можно считать, что $\mathbb{E}\xi_1 = \mathbb{E}\xi_2 = 0, \mathbb{D}\xi_1 = \mathbb{D}\xi_2 = 1$. Теперь $2|cov(\xi_1, \xi_2)| = |\mathbb{E}(2\xi_1\xi_2)| \le \mathbb{E}\xi_1^2 + \mathbb{E}\xi_2^2 = 2$, и пункт 1 доказан. В случае равенства имеем $\mathbb{E}(\xi_1 \pm \xi_2)^2 = 0$, то есть $\xi_1 \pm \xi_2 = 0$, и пункт 2 также доказан.

Подумать: верно ли, что в формулировке теоремы можно написать $\xi_1 = c + b\xi_2$?

Подумать: верно ли, что в формулировке теоремы можно гарантировать, что $b \neq 0$?

Корреляция наглядно



Линейная регрессия

Даны две случайные величины $\xi, \eta : \Omega \to \mathbb{R}$. Найти такие константы $\bar{a}, \bar{b}$, что

$$\inf_{a,b\in\mathbb{R}} \mathbb{E}(\xi - a\eta - b)^2 = \mathbb{E}(\xi - \bar{a}\eta - \bar{b})^2.$$

Прямой подстановкой имеем:

$$\begin{aligned} f(a,b) &\triangleq \mathbb{E}(\xi - a\eta - b)^2 \\ &= \mathbb{E}\xi^2 + a^2\mathbb{E}\eta^2 + b^2 - 2a\mathbb{E}(\xi\eta) - 2b\mathbb{E}\xi + 2ab\mathbb{E}\eta. \end{aligned}$$

Поскольку функция гладкая, то для $(\bar{a}, \bar{b})$ получаем систему уравнений

$$\begin{aligned} 0 = \frac{\partial f(\bar{a},\bar{b})}{\partial a} &= 2\bar{a}\mathbb{E}\eta^2 - 2\mathbb{E}(\xi\eta) + 2\bar{b}\mathbb{E}\eta, \\ 0 = \frac{\partial f(\bar{a},\bar{b})}{\partial b} &= 2\bar{b} - 2\mathbb{E}\xi + 2\bar{a}\mathbb{E}\eta. \end{aligned}$$

Линейная регрессия

Отсюда, подставляя $\bar{b} = \mathbb{E}\xi - \bar{a}\mathbb{E}\eta$, получаем

$$\begin{aligned} 0 &= \bar{a}\mathbb{E}\eta^2 - \mathbb{E}(\xi\eta) + (\mathbb{E}\xi - \bar{a}\mathbb{E}\eta)\mathbb{E}\eta \\ &= \bar{a}\big(\mathbb{E}\eta^2 - (\mathbb{E}\eta)^2\big) - \mathbb{E}(\xi\eta) + \mathbb{E}\xi\mathbb{E}\eta \\ &= \bar{a}\mathbb{D}\eta - cov(\xi,\eta), \end{aligned}$$

то есть $\bar{a} = \frac{cov(\xi,\eta)}{\mathbb{D}\eta} = \frac{cov(\xi,\eta)}{cov(\eta,\eta)}$, $\bar{b} = \mathbb{E}\xi - \frac{cov(\xi,\eta)}{\mathbb{D}\eta}\mathbb{E}\eta$.

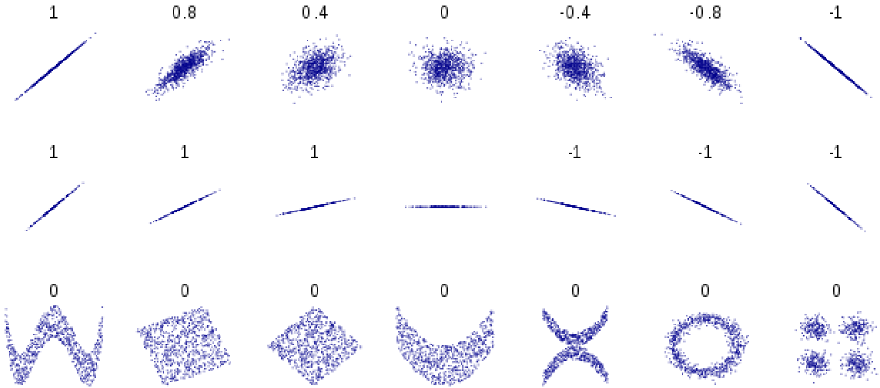
Таким образом, наилучшая линейная оценка $\hat{\xi}$ величины ξ с помощью η равна

$$\hat{\xi} \triangleq \mathbb{E}\xi + \frac{cov(\xi,\eta)}{cov(\eta,\eta)}(\eta - \mathbb{E}\eta).$$

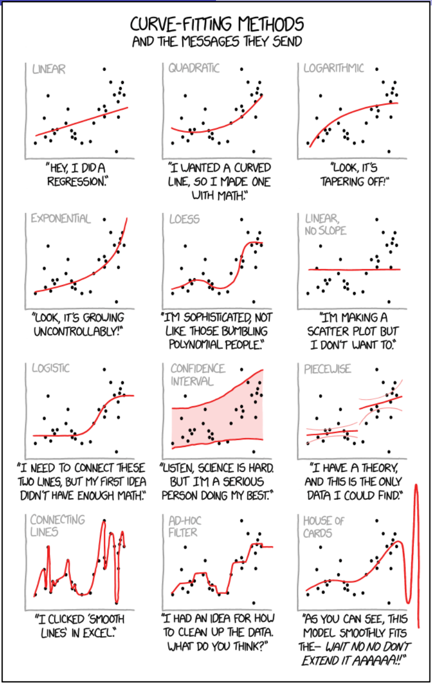
[С-но] Проверьте, что

$$\mathbb{E}\hat{\xi} = \mathbb{E}\xi,\ \ \mathbb{D}\hat{\xi} = \rho^2(\xi,\eta)\mathbb{D}\xi,\ \ \mathbb{D}(\hat{\xi} - \xi) = \big(1 - \rho^2(\xi,\eta)\big)\mathbb{D}\xi.$$

Корреляция наглядно снова: $\mathbb{D}(\hat{\xi} - \xi) = \big(1 - \rho^2(\xi, \eta)\big)\mathbb{D}\xi$



Но можно использовать не только линейную регрессию!


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Что будет дальше?

- распределение случайных величин
- медиана, математическое ожидание
- независимость случайных величин
- производящие функции
- дисперсия, ковариация и корреляция
- совместное распределение случайных величин, маргинальные распределения
- условное математическое ожидание