Математика для Data Science. Теория вероятностей.
Решения задач

Содержание

Замечание. Вот этим цветом отмечены ссылки на страницы внутри этого файла.

Арифметика случайных величин и нормальное распределение

Задача 2

Даны совместно независимые случайные величины X_1, X_2, X_3, такие что

- $E[X_1] = 0, Var(X_1) = 1$,
- $E[X_2] = 11, Var(X_2) = 3$,
- $E[X_3] = 8, Var(X_3) = 4$.

Найдите математическое ожидание и дисперсию случайной величины $\frac{2X_1+4X_2-X_3}{6} - 4$.

Подсказка. Воспользуйтесь линейностью математического ожидания.

Вспомните, что для любой случайной величины X и любого числа $c \in \mathbb{R}$ выполнено $Var(X+c) = Var(X)$ и $Var(cX) = c^2 Var(X)$. А ещё для независимых случайных величин дисперсия суммы равна сумме дисперсий.

Решение.

$$E\left[\frac{2X_1+4X_2-X_3}{6} - 4\right] = E\left[\frac{2X_1+4X_2-X_3}{6}\right] - 4 = \frac{E[2X_1+4X_2-X_3]}{6} - 4 =$$
$$= \frac{2E[X_1]+4E[X_2]-E[X_3]}{6} - 4 = \frac{2\cdot 0+4\cdot 11-8}{6} - 4 = 2$$

$$Var\left(\frac{2X_1+4X_2-X_3}{6} - 4\right) = Var\left(\frac{2X_1+4X_2-X_3}{6}\right) = \left(\frac{1}{6}\right)^2 Var(2X_1+4X_2-X_3) =$$
$$= \frac{1}{36}\cdot(Var(2X_1)+Var(4X_2)+Var(-X_3)) = \frac{1}{36}\cdot(2^2\cdot Var(X_1)+4^2\cdot Var(X_2)+(-1)^2\cdot Var(X_3)) =$$
$$= \frac{1}{36}\cdot(2^2\cdot 1+4^2\cdot 3+(-1)^2\cdot 4) = \frac{56}{36} = \frac{14}{9} \approx 1.55555556$$

Нормальное распределение

Задача 1

Независимые случайные величины X и Y имеют распределения $N(4,5)$ и $N(3,9)$ соответственно. Найдите распределение случайной величины $X + \frac{Y}{3} - 4$.

Решение. Случайная величина $\frac{Y}{3}$ нормально распределена и имеет математическое ожидание

$$E\left[\frac{Y}{3}\right] = \frac{1}{3}\cdot E[Y] = \frac{1}{3}\cdot 3 = 1$$

и дисперсию

$$Var\left(\frac{Y}{3}\right) = \left(\frac{1}{3}\right)^2 \cdot Var(Y) = \frac{1}{9}\cdot 9 = 1.$$

Значит, случайная величина $X + \frac{Y}{3} - 4$ имеет нормальное распределение с математическим ожиданием равным

$$E[X] + E\left[\frac{Y}{3}\right] + E[-4] = 4 + 1 - 4 = 1$$

и дисперсией

$$Var(X) + Var\left(\frac{Y}{3}\right) + Var(-4) = 5 + 1 + 0 = 6$$

(мы воспользовались независимостью X и Y.) Итак, мы получили распределение $N(1,6)$.

Задача 2

Независимые случайные величины X, Y и Z имеют распределения $N(1,2), N(3,4)$ и $N(5,3)$ соответственно. Найдите распределение случайной величины $\frac{X+Y+Z}{3}+2$.

Решение. Случайная величина $\frac{X+Y+Z}{3}+2$ имеет нормальное распределение. Найдём его математическое ожидание:

$$E\left[\frac{X+Y+Z}{3}+2\right]=E\left[\frac{X+Y+Z}{3}\right]+2=\frac{E[X+Y+Z]}{3}+2=\frac{E[X]+E[Y]+E[Z]}{3}+2=\frac{1+3+5}{3}+2=5$$

И теперь найдём его дисперсию, пользуясь тем, что дисперсия суммы независимых случайных величин равна сумме их дисперсий:

$$Var\left(\frac{X+Y+Z}{3}+2\right)=Var\left(\frac{X+Y+Z}{3}\right)=\left(\frac{1}{3}\right)^2\cdot Var(X+Y+Z)=$$
$$=\frac{1}{9}\cdot(Var(X)+Var(Y)+Var(Z))=\frac{1}{9}\cdot(2+4+3)=\frac{1}{9}\cdot 9=1$$

Итого мы получили распределение $N(5,1)$.

Задача 3

Все совместно независимые случайные величины $X_1,\dots,X_n$ имеют одинаковое распределение $N(\mu,\sigma^2)$. Найдите распределение случайной величины $X_1+\dots+X_n$.

Решение. Случайная величина $X_1+\dots+X_n$ имеет нормальное распределение. Найдём его математическое ожидание:
$E[X_1+\dots+X_n]=E[X_1]+\dots+E[X_n]=n\cdot\mu$
И теперь найдём его дисперсию, пользуясь совместной независимостью величин $X_1,\dots,X_n$:
$Var(X_1+\dots+X_n)=Var(X_1)+\dots+Var(X_n)=n\cdot\sigma^2$
Значит, искомое в задаче распределение — это $N(n\mu,n\sigma^2)$.

Задача 4

Все совместно независимые случайные величины $X_1,\dots,X_n$ имеют одинаковое распределение $N(\mu,\sigma^2)$. Найдите распределение случайной величины $\frac{X_1+\dots+X_n}{n}$.

Решение. По предыдущей задаче случайная величина $X_1+\dots+X_n$ имеет распределение $N(n\mu,n\sigma^2)$. А тогда
$E\left[\frac{X_1+\dots+X_n}{n}\right]=\frac{1}{n}\cdot(E[X_1+\dots+X_n])=\frac{1}{n}\cdot n\mu=\mu$
$Var\left(\frac{X_1+\dots+X_n}{n}\right)=\frac{1}{n^2}\cdot(Var(X_1+\dots+X_n))=\frac{1}{n^2}\cdot n\sigma^2=\frac{\sigma^2}{n}$
Значит, искомое в задаче распределение — это $N(\mu,\frac{\sigma^2}{n})$.

Статистический тест

Задача 1

Докажите, что на прошлом шаге мы правильно нашли распределение статистики T при условии, что верна гипотеза H_0. А именно, докажите такую последовательность утверждений:

- $x_1+\dots+x_n$ имеет распределение $N(n\mu_0,n\sigma^2)$
- $\frac{x_1+\dots+x_n}{n}$ имеет распределение $N\left(\mu_0,\frac{\sigma^2}{n}\right)$
- $\frac{x_1+\dots+x_n}{n}-\mu_0$ имеет распределение $N\left(0,\frac{\sigma^2}{n}\right)$

- $T(x_1, \dots, x_n) := \dfrac{\frac{x_1+\dots+x_n}{n} - \mu_0}{\sigma/\sqrt{n}}$ имеет распределение $N(0,1)$

Подсказка. Здесь пригодится утверждение из предыдущего урока про сумму независимых случайных величин с нормальным распределением.

Решение.

- То, что $x_1 + \dots + x_n$ имеет распределение $N(n\mu_0, n\sigma^2)$, мы доказали в пятой задаче
- Пользуясь предыдущим пунктом и свойствам математического ожидания и дисперсии, получаем:

$$E\left[\frac{x_1 + \dots + x_n}{n}\right] = \frac{E\left[x_1 + \dots + x_n\right]}{n} = \frac{n\mu_0}{n} = \mu_0$$

$$Var\left(\frac{x_1 + \dots + x_n}{n}\right) = \frac{Var\left(x_1 + \dots + x_n\right)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

  Теперь, чтобы доказать, что $\frac{x_1+\dots+x_n}{n}$ имеет распределение $N\left(\mu_0, \frac{\sigma^2}{n}\right)$, осталось вспомнить, что сумма независимых нормально распределённых случайных величин тоже имеет нормальное распределение.

- $\frac{x_1+\dots+x_n}{n} - \mu_0$ имеет распределение $N\left(0, \frac{\sigma^2}{n}\right)$, поскольку $E\left[\frac{x_1+\dots+x_n}{n} - \mu_0\right] = E\left[\frac{x_1+\dots+x_n}{n}\right] - E[\mu_0] = \mu_0 - \mu_0 = 0$ и $Var\left(\frac{x_1+\dots+x_n}{n} - \mu_0\right) = Var\left(\frac{x_1+\dots+x_n}{n}\right) = \frac{\sigma^2}{n}$
- Наконец, разберёмся с распределением величины $T(x_1, \dots, x_n) := \dfrac{\frac{x_1+\dots+x_n}{n} - \mu_0}{\sigma/\sqrt{n}}$. Его математическое ожидание равно

$$E\left[\frac{\frac{x_1+\dots+x_n}{n} - \mu_0}{\sigma/\sqrt{n}}\right] = \frac{E\left[\frac{x_1+\dots+x_n}{n} - \mu_0\right]}{\sigma/\sqrt{n}} = \frac{0}{\sigma/\sqrt{n}} = 0.$$

  А его дисперсия равна

$$Var\left(\frac{\frac{x_1+\dots+x_n}{n} - \mu_0}{\sigma/\sqrt{n}}\right) = \frac{Var\left(\frac{x_1+\dots+x_n}{n} - \mu_0\right)}{\sigma^2/n} = \frac{\sigma^2/n}{\sigma^2/n} = 1.$$

Задача 2

Найдите распределение T при условии, что выполнена H_1. Это нужно нам для нахождения вероятности ошибки второго рода

Ясно, что из-за того, что μ_1 и μ_0 не указаны, нельзя указать вероятность ошибки второго рода. Обсудите с преподавателем, как будет вести себя β (вероятность ошибки второго рода) с увеличением $\mu_1 - \mu_0$.

Комментарий. Обратите внимание, что вероятность ошибки первого рода не зависит от μ_1, ведь вероятность ошибки первого рода всегда равна $\alpha = 0.05$. А вот вероятность ошибки второго рода зависит от того, насколько μ_1 далеко от μ_0, то есть от $\mu_1 - \mu_0$. Это логично: чем дальше μ_1 от μ_0, тем легче должно быть отличить H_1 от H_0. Можно сказать, что чем больше $\mu_1 - \mu_0$, тем мощнее наш критерий (как мы помним, мощность критерия это $(1 - \beta)$, где β это вероятность ошибки второго рода).

Подсказка. На шаге ранее мы уже проводили нужные вычисления для конкретного случая, остаётся их повторить для общего случая.

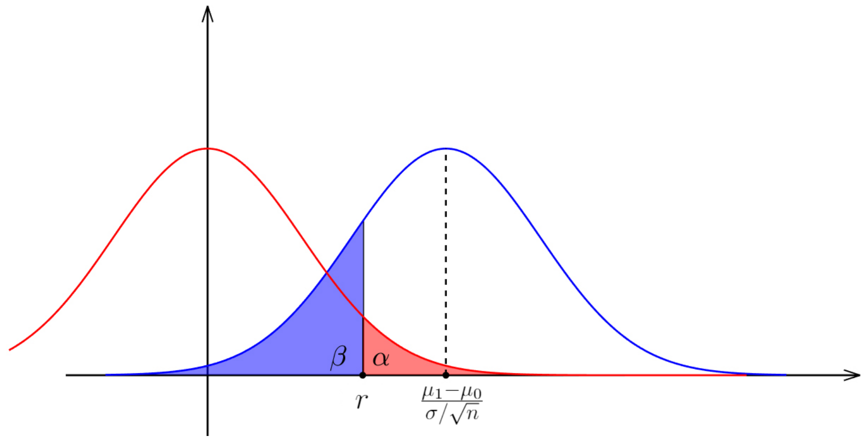
Решение.

Напомним, что гипотеза H_1 — добавка увеличивает среднюю продолжительность жизни. Согласно этой гипотезе продолжительность жизни мышей, употребляющих добавку, имеет распределение $N(\mu_1, \sigma^2)$. Аналогично рассуждениям в предыдущей задаче мы получаем, что при выполнении H_1

- $x_1 + \dots + x_n$ имеет распределение $N(n\mu_1, n\sigma^2)$

- $\frac{x_1+\dots+x_n}{n}$ имеет распределение $N\left(\mu_1, \frac{\sigma^2}{n}\right)$
- $\frac{x_1+\dots+x_n}{n} - \mu_0$ имеет распределение $N\left(\mu_1 - \mu_0, \frac{\sigma^2}{n}\right)$
- $$T(x_1,\dots,x_n) := \frac{\frac{x_1+\dots+x_n}{n} - \mu_0}{\sigma/\sqrt{n}} \text{ имеет распределение } N\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}, 1\right)$$

Посмотрим на графики плотностей полученных нормальных распределений статистики T. Красный график показывает функцию плотности распределения T, если выполнено H_0, а синий – функцию плотности распределения T если выполнено H_1.



Здесь $[r, +\infty)$ — наше критическое множество. При верной H_1 ошибка второго рода происходит тогда, когда T принимает значение меньше r – ведь именно в этих случаях мы принимаем гипотезу H_0. Вероятность этого равна площади подграфика функции плотности распределения $N\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}, 1\right)$ над лучом $(-\infty, r]$. Итак, на картинке выше площадь, заштрихованная синим – это вероятность ошибки второго рода.

С увеличением $\mu_1 - \mu_0$ синяя кривая будет сдвигаться вправо и соответствующая площадь подграфика будет уменьшаться. То есть будет уменьшаться и β (вероятность ошибки второго рода).

ЗБЧ и ЦПТ

Задача 1

Рассмотрим несимметричную монетку — она имеет распределение Бернулли с параметром $p = 0.25$. То есть $P(\xi_i' = 0) = 0.75$ и $P(\xi_i' = 1) = 0.25$. Воспользуйтесь ЦПТ и найдите параметры нормального распределения, к которому близко распределение величины $\eta_{100}' := \sum_{i=1}^{100} \xi_i'$.

Подсказка. Решение аналогично рассуждению с предыдущего шага. Только надо будет вспомнить (или вывести) математическое ожидание и дисперсию распределения Бернулли.

Решение. По ЦПТ мы можем считать, что η_{100}' распределена нормально (достаточно близка к нормальному, чтобы разницей можно было пренебречь). Чтобы найти параметры этого нормального распределения, нам потребуется мат.ожидание и дисперсия ξ_i'. Выведем ещё раз эти формулы:

$$E[\xi_i'] = 0 \cdot 0.75 + 1 \cdot 0.25 = 0.25$$

$$Var(\xi_i') = E[(\xi_i')^2] - (E[\xi_i'])^2 = 0^2 \cdot 0.75 + 1^2 \cdot 0.25 - 0.25^2 = 0.25 \cdot (1 - 0.25) = 0.25 \cdot 0.75 = 0.1875$$

А тогда

$$\mu' = E[\eta_{100}'] = 100 \cdot E[\xi_i'] = 100 \cdot 0.25 = 25$$

$$\sigma'^2 = Var(\eta'_{100}) = 100 \cdot Var(\xi'_i) = 100 \cdot\ 0.1875 = 18.75 \approx 4.3301^2$$

А значит, искомая в задаче сумма $\mu' + \sigma' \approx 25 + 4.3301 = 29.3301$

Задача 2

Докажите, что стандартное отклонение случайной величины, имеющей распределение Бернулли, не превосходит 0.5.

Замечание. Это задача скорее на матан, чем тервер.

Подсказка. Пусть p — параметр в распределении Бернулли. Тогда стандартное отклонение — функция от p. Мы умеем находить её максимум.

Ещё одно замечание: стандартное отклонение максимально тогда, когда максимальна дисперсия. А максимум дисперсии считать чуть прятнее.

Решение. Пусть случайная величина ξ имеет распределение Бернулли с параметром p. Вспомним, как находится $Var(\xi)$:

Поскольку $P(\xi\ = 0) = 1 - p$ и $P(\xi = 1) = p$, то

$$E[\xi] = 0 \cdot (1 - p) + 1 \cdot p = p$$

При этом

$$Var(\xi) = E[(\xi)^2] - (E[\xi])^2 = 0^2 \cdot (1 - p) + 1^2 \cdot p - p^2 = p - p^2$$

Итак, мы хотим доказать, что $\sqrt{Var(\xi)} \leq \frac{1}{2}$. Возведя это равенство в квадрат, получим $Var(\xi) \leq \frac{1}{4}$. Итак, нужно доказать, что $p - p^2 \leq \frac{1}{4}$. А это равносильно $p^2 - p + \frac{1}{4} \geq 0$, что в свою очередь равносильно верному неравенству $\left(p - \frac{1}{2}\right)^2 \geq 0$.