Непрерывные случайные величины

Непрерывные случайные величины. Плотность распределения вероятностей. Функции от случайных величин. Примеры распределений: равномерное, нормальное, экспоненциальное, Стьюдента. Характеристики и свойства распределений. Многомерные распределения. Совместное и маргинальное распределение. Энтропия.

Даниил Корбут
Специалист по Анализу Данных

Окончил бакалавриат ФИВТ МФТИ (Анализ данных) в 2018г
Учусь на 2-м курсе магистратуры ФИВТ МФТИ
Работал в Statsbot и Яндекс. Алиса.
Сейчас в Insilico Medicine, Inc, занимаюсь генерацией активных молекул и исследованиями старения с помощью DL.

Даниил Корбут
DL Researcher
Insilico Medicine, Inc

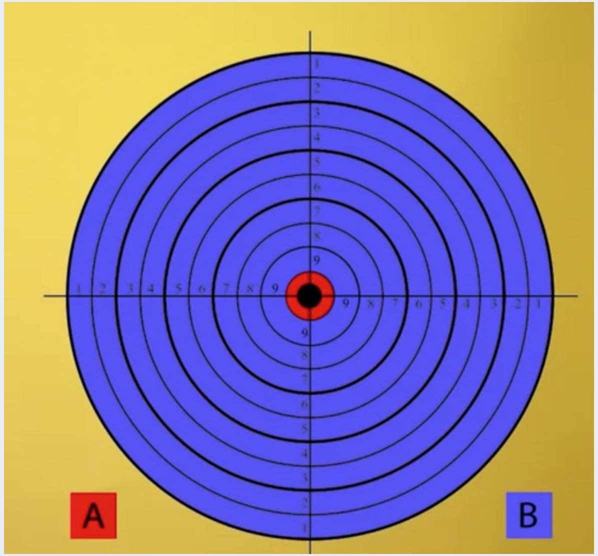
Условная вероятность (повторение)

$$A|B : P(A|B) = \frac{P(AB)}{P(B)}, \quad P(B) > 0$$

$$P(B) = 0.8, P(A) = 0.05$$
$$P(AB) = P(A) = 0.05 \Rightarrow$$
$$P(A|B) = 0.05/0.8 = 0.0625$$

Формула полной вероятности:

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$$



НЕТОЛОГИЯ

https://www.coursera.org/learn/mathematics-and-python

Непрерывные случайные величины

Вспомним, как мы задавали дискретные случайные величины на прошлой лекции:

- X принимает счётное множество значений $A = \{a_1, a_2, a_3, \ldots\}$ с вероятностями $p_1, p_2, p_3, \ldots$ где $p_i \geq 0 \; \forall i$ и $\sum_{i=1}^{\infty} p_i = 1$
- $P(X = a_i) = p_i$ – функция вероятности

! Ключевой момент: в силу счётности X мы можем определить функцию вероятности для каждого фиксированного ai из A.

Непрерывные случайные величины

В случае абсолютно непрерывных случайных величин так сделать нельзя, потому что вероятность каждого значения с.в. будет нулевой!

Поэтому непрерывные с.в. нельзя задавать с помощью функции вероятности.

Один из способов задания непрерывной случайной величины является **функция распределения**.

НЕТОЛОГИЯ

https://www.coursera.org/learn/mathematics-and-python

Непрерывные случайные величины

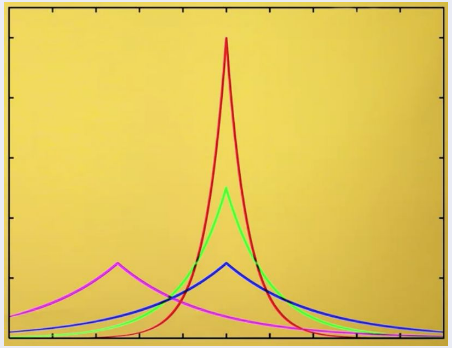
Другим способом задания непрерывной случайной величины является **плотность распределения** случайной величины f(x), которая тесно связана с функцией распределения непрерывной случайной величины.

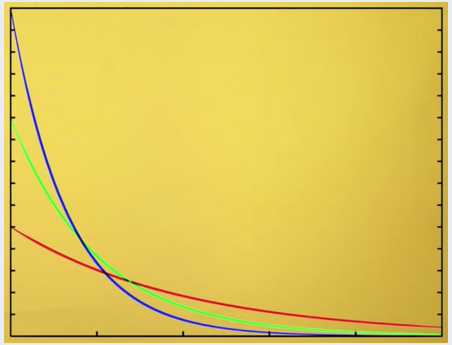
- $f(x): \int_a^b f(x)dx = P(a \leq X \leq b)$ – плотность распределения
- $F(x) = \int_{-\infty}^{x} f(u)du$
- $\int_{-\infty}^{+\infty} f(u)du = P(-\infty \leq X \leq +\infty) = 1$

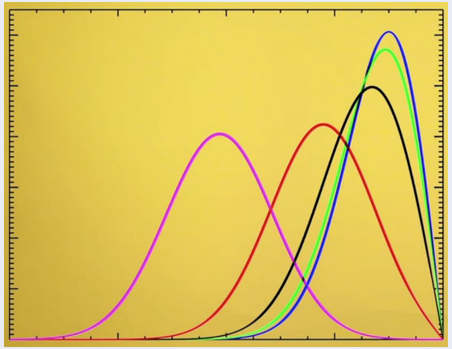
https://www.coursera.org/learn/mathematics-and-python

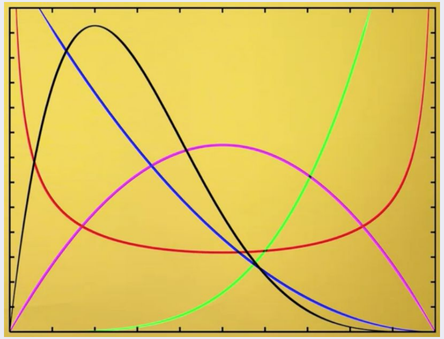
Непрерывные случайные величины

Функция плотности распределения, в отличие от неубывающей функции распределения, может вести себя совершенно по-разному. В этом их достоинство: проще отличать семейства распределений по плотностям.





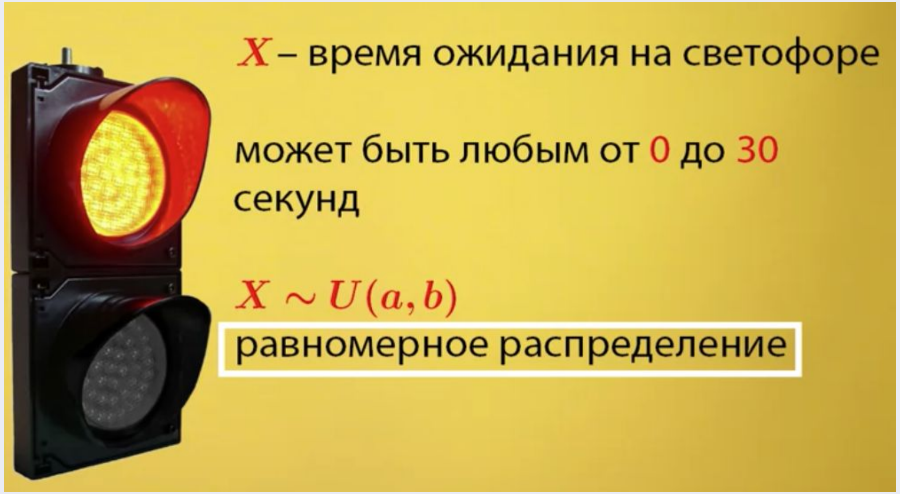


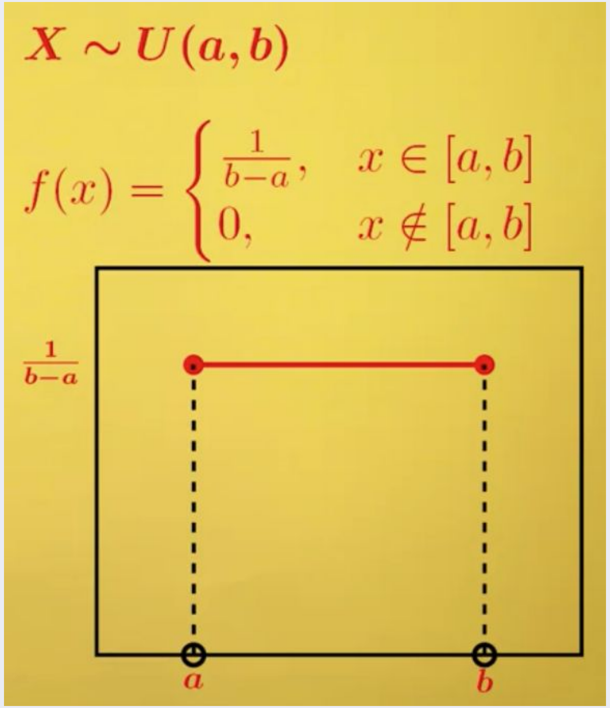


https://www.coursera.org/learn/mathematics-and-python

Примеры непрерывных случайных величин (равномерное распределение)

Ярким примером непрерывной случайной величины, распределённой **равномерно**, является время ожидания перехода дороги со светофором без секунд.





https://www.coursera.org/learn/mathematics-and-python

Примеры непрерывных случайных величин (нормальное распределение)

Ярким примером непрерывной случайной величины, распределённой **нормально**, является время прихода на работу, если вы всегда старайтесь приходить в офис, например, около 12:00.

- X – время прихода на работу
- $X \sim N(\mu, \sigma^2)$

нормальное (Гауссово) распределение

Сумма слабо зависимых случайных факторов

Как вы думаете, какие из приведённых ниже величин тоже можно моделировать нормальным распределением?

1. Погрешность барометра
2. Длина листьев одного дерева
3. Число опечаток на страницу текста в длинной рукописи
4. Размер выигрыша лотерейного билета

НЕТОЛОГИЯ

https://www.coursera.org/learn/mathematics-and-python

Примеры непрерывных случайных величин (нормальное распределение)

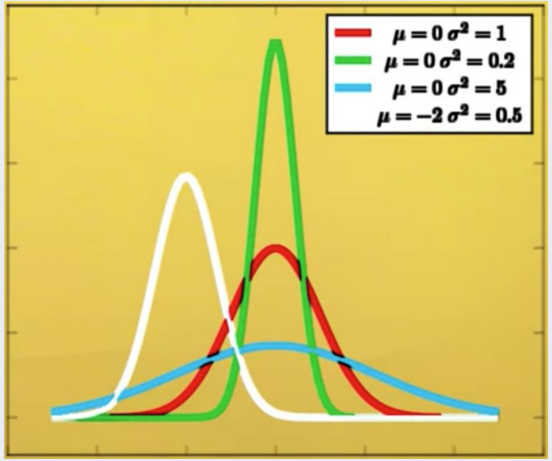
Верно! Погрешность барометра и длину листьев одного дерева.

X – время прихода на работу

$X \sim N(\mu, \sigma^2)$

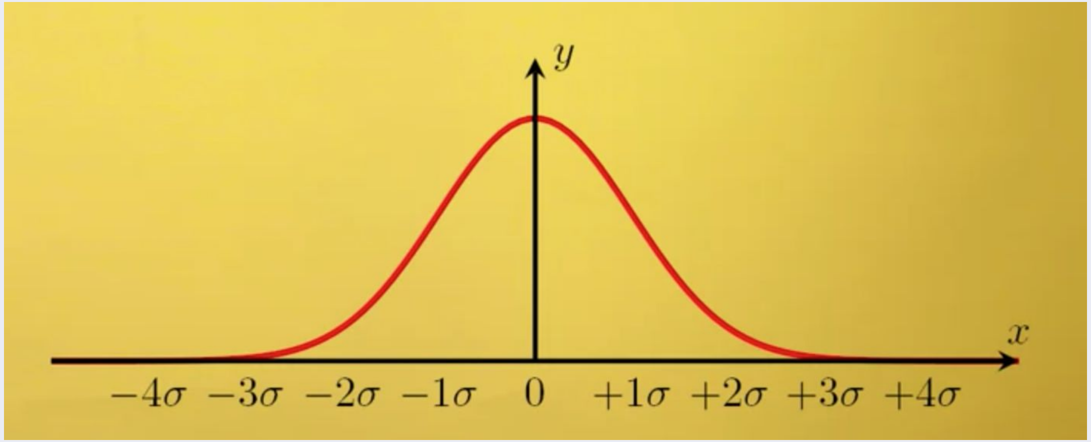
среднее время прихода

разброс вокруг среднего



$$X \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



https://www.coursera.org/learn/mathematics-and-python

Примеры непрерывных случайных величин (экспоненциальное распределение)

Ещё одним наиболее часто встречающимся непрерывным распределением является **экспоненциальное** распределение с.в.

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases} \qquad F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases}$$

Здесь **λ** - единственный параметр данного распределения, полностью определяющий его свойства. В частности, числовые характеристики выражаются через этот параметр: E(X)=1/λ, D(X)=1/λ^2.

Экспоненциальное распределение моделирует время между двумя последовательными свершениями события, а параметр **λ** описывает среднее число наступлений события в единицу времени. Обычно с помощью этого закона описывают:

- продолжительность обслуживания покупателя
- время жизни оборудования до отказа
- промежуток времени между поломками

Примеры непрерывных случайных величин (распределение Стьюдента)

Некоторые распределения связаны между собой (помните дискретные с.в.?). Одним из таких семейств для непрерывных с.в. является **распределение Стьюдента**.
Пусть Yi - независимые стандартные нормальные случайные величины, тогда

$$Y_i \sim \mathrm{N}(0,1),\ i = 1,\ldots,n.$$

$$f_t(y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}}$$

$$t = \frac{Y_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}}$$

имеет распределение Стьюдента с n степенями свободы.

Распределение Стьюдента симметрично. В частности если t имеет распределение Стьюдента с n степенями свободы, то -t имеет то же распределение.

НЕТОЛОГИЯ

Многомерные распределения

Зачастую наш эксперимент зависит далеко не от одного параметра, и хочется каким-то образом построить распределение над векторами параметров.
Случай дискретных переменных (таблица совместного распределения):

	c=0	c=1	c=2
g=0	0.1	0.1	0.1
g=1	0.2	0.4	0.1

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \mathsf{P}(\xi_1 = a_i,\ \xi_2 = b_j) = 1.$$

$$\{\xi_1 = a_i\} = \bigcup_{j=1}^{\infty} \{\xi_1 = a_i, \xi_2 = b_j\}.$$

Маргинальное распределение

$$\mathsf{P}(\xi_1 = a_i) = \sum_{j=1}^{\infty} \mathsf{P}(\xi_1 = a_i, \xi_2 = b_j), \quad \mathsf{P}(\xi_2 = b_j) = \sum_{i=1}^{\infty} \mathsf{P}(\xi_1 = a_i, \xi_2 = b_j).$$

https://nsu.ru/mmf/tvims/chernova/tv/
https://habr.com/ru/post/351400/

Многомерные распределения

Аналогично можно определить совместное распределение для случая абсолютно непрерывных случайных величин:

$$\mathsf{P}((\xi_1,\xi_2) \in B) = \iint\limits_B f_{\xi_1,\xi_2}(s_1,s_2)ds_1ds_2.$$

$$F_{\xi_1,\xi_2}(x_1,x_2) = \mathsf{P}(\xi_1 < x_1, \xi_2 < x_2) = \int\limits_{-\infty}^{x_1}\left(\int\limits_{-\infty}^{x_2} f_{\xi_1,\xi_2}(s_1,s_2)\,ds_2\right)ds_1.$$

Свойства плотности ничем не отличаются от случая одномерного распределения:

$$f_{\xi_1,\xi_2}(x_1,x_2) \geqslant 0 \text{ для любых } x_1, x_2 \in \mathbb{R};$$

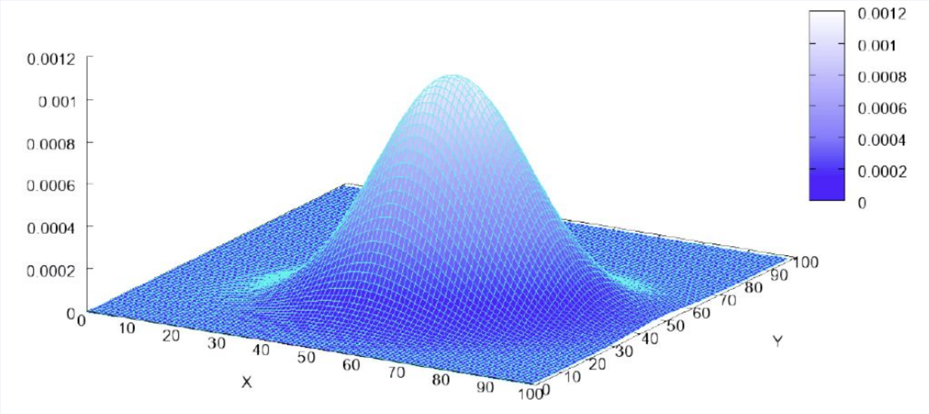
$$\iint\limits_{\mathbb{R}^2} f_{\xi_1,\xi_2}(x_1,x_2)\,dx_1\,dx_2 = 1.$$

НЕТОЛОГИЯ

Многомерные распределения (примеры)

Многомерное нормальное:

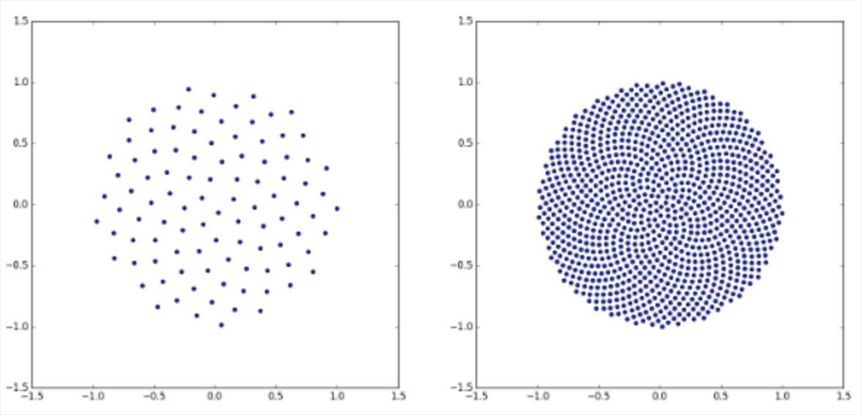
- $X \sim N(\mu, \Sigma),\ \mu \in \mathbb{R}^k$
  $\Sigma \in \mathbb{R}^{k \times k}$ положительно определена,
- $f(x) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T |\Sigma^{-1}| (x-\mu)}$



Многомерное равномерное:

$$f_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n) = \begin{cases} \dfrac{1}{\lambda(S)}, & \text{если } (x_1,\ldots,x_n) \in S, \\ 0, & \text{если } (x_1,\ldots,x_n) \notin S. \end{cases}$$

$$\int\limits_{\mathbb{R}^n} f_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n)\, dx_1 \ldots\ dx_n = \frac{1}{\lambda(S)} \int\limits_S dx_1 \ldots\ dx_n = \frac{1}{\lambda(S)} \lambda(S) = 1.$$



https://askdev.ru/q/ravnomerno-raspredelyaya-n-tochek-na-sfere-26399/
https://www.coursera.org/learn/mathematics-and-python

Энтропия

Одно из важнейших понятий теории информации, напрямую связанное с теорией вероятности.
Информационная энтропия — мера неопределённости некоторой системы, в частности непредсказуемость появления какого-либо символа первичного алфавита. Например, в последовательности букв, составляющих какое-либо предложение на русском языке, разные буквы появляются с разной частотой, поэтому неопределённость появления для некоторых букв меньше, чем для других.

Информационная двоичная энтропия для независимых случайных событий x с n возможными состояниями, распределённых с вероятностями pi, рассчитывается по формуле Шеннона:

!

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

https://ru.wikipedia.org/wiki

Энтропия (пример)

В случае равновероятных событий формула Шеннона упрощается до формулы Хартли:

$$I = -\log_2 p = \log_2 N,$$

где I – количество передаваемой информации, p – вероятность события, N – возможное количество различных (равновероятных) сообщений.

Пример: В колоде 36 карт. Какое количество информации содержится в сообщении, что из колоды взята карта с портретом “туз”; “туз пик”?

Вероятность p1 = 4/36 = 1/9, а p2 = 1/36. Используя формулу Хартли имеем:

$$I_1 = -\log_2 p_1 = \log_2 \frac{1}{\frac{1}{9}} = \log_2 9 \approx 3.17$$
$$I_2 = -\log_2 p_2 = \log_2 \frac{1}{\frac{1}{36}} = \log_2 36 \approx 5.17$$

Заметим (из второго результата), что для кодирования всех карт, необходимо 6 бит.

НЕТОЛОГИЯ

http://inf-w.ru/?page_id=597

Энтропия (пример)

Пример: В колоде 36 карт. Из них 12 карт с “портретами”. Поочередно из колоды достается и показывается одна из карт для определения изображен ли на ней портрет. Карта возвращается в колоду. Определить количество информации, передаваемой каждый раз, при показе одной карты.

$$I = -(p_{ic} \log_2 p_{ic} + p_{ot} \log_2 p_{ot}) = \frac{12}{36} \log_2 \frac{1}{\frac{12}{36}} + \frac{36-12}{36} \log_2 \frac{1}{\frac{36-12}{36}} = \frac{\ln 3}{3 \ln 2} + \frac{2 \ln \frac{3}{2}}{3 \ln 2} \approx 0.91$$

Пример: Документация некоторого учреждения размещена в 4-х комнатах. В каждой комнате находится 16 шкафов. Каждый шкаф имеет 8 полок. Определить количество информации, которое несет сообщение о том, что нужный документ находится в третьей комнате, в тринадцатом шкафу на пятой полке.

Для независимых $x_1, \ldots, x_n$ справедливо:

$$I(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} I(x_i)$$

$$I = \log_2 4 + \log_2 16 + \log_2 8 = 9$$

http://inf-w.ru/?page_id=597

Спасибо за внимание!

НЕТОЛОГИЯ