

Многомерный анализ биологических данных на компьютере

БСИ ДВО РАН

Кислов Д.Е.
16 октября 2019 г.

Структура лекционной программы

- общие представления о теории вероятностей;
- введение в среду статистического анализа Python
- многомерный анализ данных;

Общие представления о теории вероятностей

- Понятие вероятности, случайной величины, дискретно и непрерывно распределенные случайные величины, функции и плотности распределения вероятности, условные и безусловные вероятности, примеры;
- серия независимых испытаний (модель Бернулли), центральная предельная теорема;
- основные распределения используемые в ТВ;

- задача теории вероятностей и математической статистики (примеры), понятие статистической оценки, оценки вероятностей событий в серии независимых испытаний, точечные и интервальные оценки параметров распределений;
- проверка статистических гипотез, понятия ошибок первого и второго рода, критерии согласия (тесты Колмогорова-Смирнова, χ^2 , Шапиро-Уилка и др.), анализ таблиц сопряженности (критерий χ^2 , точный тест Фишера);
- корреляционный анализ, метод наименьших квадратов, линейная и нелинейная регрессии (примеры), классический дисперсионный анализ и его обобщения;

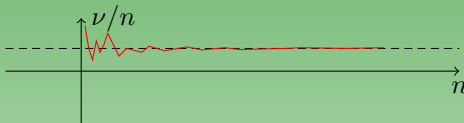
- задачи классификации многомерных данных, метод k -средних, иерархическая кластеризация, оценка качества кластеризации и сравнение кластерных структур;
- метод главных компонент (PCA), дискриминантный анализ (DA), многомерное шкалирование (MDS, NMDS), анализ соответствий (CA);
- классификация по прецедентам и основные методы решения задачи (k -NN, линейная классификация, машина опорных векторов, деревья решений и случайный лес), оценки качества решения задачи классификации по прецедентам (точность, матрица ошибок, ROC/AUC характеристика);

Основания теории вероятностей

Если ν – число осуществлений некоторого события, то $\frac{\nu}{n}$ – его частота реализации (появления).

ОСНОВНОЕ МОДЕЛЬНОЕ ПОЛОЖЕНИЕ

Частота появления события при многократном повторении эксперимента должна проявлять устойчивость: осуществляя колебания, она должна стремиться к определенному значению.



Что такое вероятность?

Под термином "вероятность" события будем понимать некоторое число, характеризующее частоту его реализации при многократном повторении эксперимента.

Теория вероятностей (ТВ) и математическая статистика (МС)

Задача ТВ

Построение математических моделей случайных явлений, *проявляющих свойство статистической устойчивости.*

Задача МС

Формирование выводов на основе данных опыта и представлений теории вероятностей.

Задача

Симметричную монету подбросили 100 раз, из которых 42 раза выпала «решка» и 58 — «орел». Построены 2 модели этого явления: 1) $P(\text{«орел»})=1/2$, $P(\text{«решка»})=1/2$; 2) $P(\text{«орел»})=2/3$, $P(\text{«решка»})=1/3$. Какую модель следует выбрать?

Основания теории вероятностей

Таблица: Соответствие вероятностных и теоретико-множественных представлений

Множественное понятие	Понятие теории вероятностей
1. Множество A пусто ($A = \emptyset$).	1. Событие A невозможно.
2. $A \cap B = \emptyset$.	2. Два события несовместны.
3. $A_1 \cap A_2 \cap \dots \cap A_k = X$.	3. Событие X состоит в одновременном наступлении событий A_1, A_2, \dots, A_k .
4. $A_1 \cup A_2 \cup \dots \cup A_k = X$.	4. Событие X состоит в наступлении одного из событий A_1, A_2, \dots, A_k .
5. Дополнительное множество к A (\overline{A}).	5. Событие состоит в ненаступлении события A (в этом случае говорят, что наступило противоположное A событие).
6. $B \subseteq A$.	6. Наступление события B влечет наступление события A .
7. $A = \Omega$.	7. Событие A достоверно.

Задача 1

Эксперимент состоит в n -кратном повторении опыта с двумя исходами. Вероятность «успеха» в опыте равна p , вероятность «неудачи» — q ($p + q = 1$). Определить вероятность k успехов при выполнении эксперимента.

Решение

Рассмотрим событие, состоящее в том, что первые k испытаний окончились «успехом», а остальные $n - k$ — «неудачей». Вероятность такого события — $p^k \cdot q^{n-k}$. Общее число подобных событий в эксперименте, отличающихся порядком «успехов» и «неудач» равно количеству k -элементных подмножеств n -элементного множества, т. е. C_n^k . Следовательно, искомая вероятность определяется выражением: $C_n^k p^k q^{n-k}$. $C_n^k = n! / (k!(n - k)!)$

Задача 2

При посадке тиса приживаемость составляет 10%. Какова вероятность, что из 10 посаженных образцов приживется хотя бы один.

Задача 3

При посадке тиса приживаемость составляет 10%. Какова вероятность, что из 10 посаженных образцов приживется хотя бы один.

Решение

- $(9/10)^{10}$ – не приживется ни один, $1 - (9/10)^{10}$ – приживется хотябы один.

Задача 4

При посадке тиса приживаемость составляет 10%. Какова вероятность, что из 10 посаженных образцов приживется хотя бы один.

Решение

- $(9/10)^{10}$ – не приживется ни один, $1 - (9/10)^{10}$ – приживется хотя бы один.
- $\sum_{k=1}^{10} C_{10}^k (1/10)^k (9/10)^{n-k}$

Задача 5

При посадке тиса приживаемость составляет 10%. Какова вероятность, что из 10 посаженных образцов приживется хотя бы один.

Решение

- $(9/10)^{10}$ – не приживется ни один, $1 - (9/10)^{10}$ – приживется хотя бы один.
- $\sum_{k=1}^{10} C_{10}^k (1/10)^k (9/10)^{n-k}$

Задача 6

При высаживании неприкированной рассады помидоров только 80% растений приживаются. Найдите вероятность того, что из десяти посаженных кустов приживется не менее 7.

Задача 7

Приживаемость саженцев составляет в среднем 30%. Каков должен быть минимальный объём посадок, чтобы можно было гарантировать выживаемость 50 экземпляров с доверительной вероятностью не меньшей 90%?

Задача 8

Приживаемость саженцев составляет в среднем 30%. Каков должен быть минимальный объём посадок, чтобы можно было гарантировать выживаемость 50 экземпляров с доверительной вероятностью не меньшей 90%?

Решение

$$\min_N \sum_{k=50}^N C_n^k 0.3^k 0.7^{n-k} \geq 0.9$$

Задача 9

Приживаемость саженцев составляет в среднем 30%. Каков должен быть минимальный объём посадок, чтобы можно было гарантировать выживаемость 50 экземпляров с доверительной вероятностью не меньшей 90%?

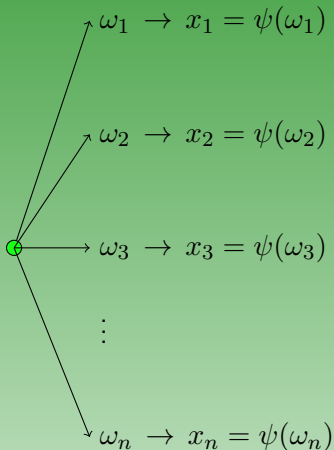
Решение

$$\min_N \sum_{k=50}^N C_n^k 0.3^k 0.7^{n-k} \geq 0.9$$

Решить такую задачу можно численно, используя язык программирования.

Случайные величины и их распределения

Эксперимент $\omega_i \in \Omega$:



Определение

Если $P(\omega_i) = p_i$ определены, то преобразование ψ совместно с p_i и Ω определяют дискретную случайную величину ξ : $P(\xi = x_i) = p_i$.
Множество $\{(x_i, p_i)\}$ образуют закон распределения случайной величины ξ .

Пример

Количество очков, выпавшее при подбрасывании игральной кости, — дискретная случайная величина.

Функция распределения случайной величины

Определение

Функция $F_\xi(x) = P(\xi < x)$, где $x \in \mathbb{R}$, называется функцией распределения случайной величины ξ .

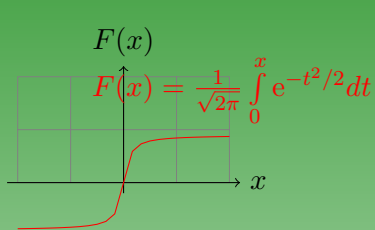
Замечание

Случайная величина называется *распределенной непрерывно*, если соответствующая функция распределения является *непрерывной*.

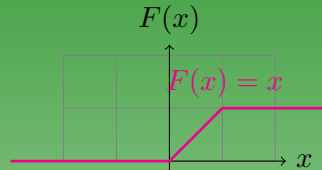
Замечание

Для случайных величин, имеющих дискретное распределение, функция распределения терпит разрывы. Аксиомы теории вероятностей определяют основные свойства функции распределения.

Примеры непрерывных и разрывной функций распределения



Стандартная нормальная функция распределения.



Функция распределения случайной величины равномерно распределенной на интервале $[0,1]$.

Разрывная функция распределения



Функция распределения случайной величины — числа очков при подбрасывании кубика.

Плотность распределения случайной величины

Определение

Плотностью распределения ($f_\xi(x)$) случайной величины ξ называется производная по x (если таковая существует) от функции распределения $F_\xi(x)$:

$$f_\xi(x) = \frac{dF_\xi(x)}{dx}.$$

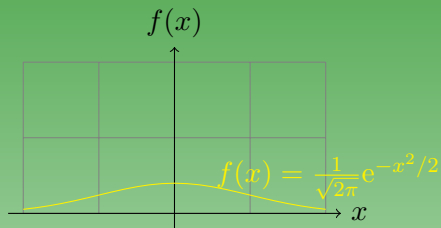
Замечание

Плотность распределения — скорость изменения вероятности события $\{\xi < x\}$, зависящая от x .

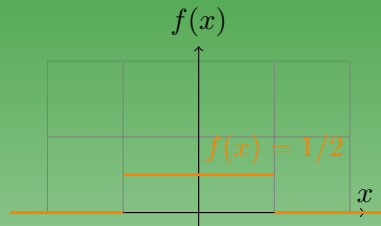
Замечание

Плотность распределения в точке x характеризует вероятность принадлежности случайной величины достаточно малой окрестности, содержащей точку x .

Примеры плотностей распределения случайных величин



Плотность стандартного нормального распределения.



Плотность случайной величины ξ , распределенной равномерно на интервале $[-1, 1]$.

Утверждение

Если $f_\xi(x)$ плотность распределения случайной величины ξ , то

$$\blacktriangleright F_\xi(x) = \int_{-\infty}^x f_\xi(\tau) d\tau;$$

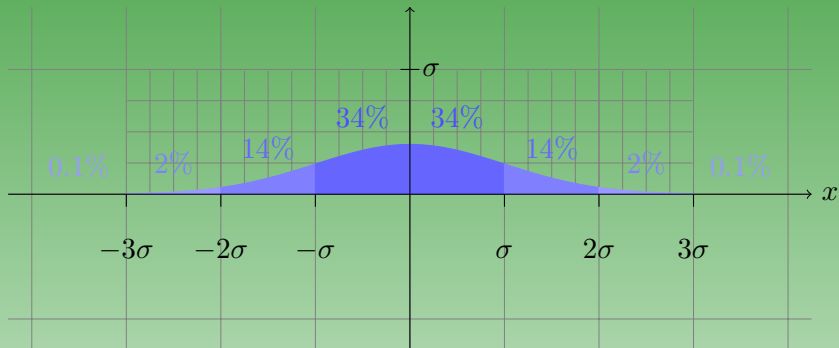
$$\blacktriangleright \lim_{x \rightarrow \pm\infty} f_\xi(x) = 0;$$

$$\blacktriangleright \int_{-\infty}^{\infty} f_\xi(x) = 1;$$

$$\blacktriangleright P(a \leq \xi < b) = \int_a^b f_\xi(x) dx, a, b \in \mathbb{R}.$$

Доверительные интервалы нормального распределения

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right)$$



Несмещенное нормальное распределение с дисперсией σ^2 .

Многомерное нормальное распределение

Определение

Пусть $\Upsilon = (\eta_1, \dots, \eta_n) \in \mathcal{N}(0, 1)$ – независимы в совокупности случайные величины; тогда $\Xi = A\Upsilon + b$, где $A = A_{n \times n}$, $\dim b = n$, имеет многомерное нормальное распределение.

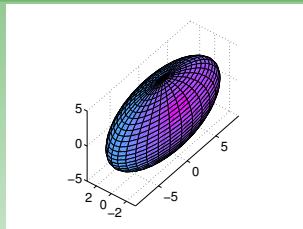
Утверждение

Если матрица A невырождена, то Ξ имеет плотность

$$f_{\Xi}(x) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(A^T A)}} e^{-(x-b)^T (A^T A)^{-1} (x-b)/2}.$$

Геометрическая интерпретация

Пример многомерного нормального распределения с диагональной матрицей A ;



Вычислительные среды для анализа данных

- R, <http://r-project.org> + Packages (<https://cran.r-project.org/>)
- Python, <http://python.org>
 - Pandas, <http://pandas.pydata.org>
 - SciPy+NumPy, <http://scipy.org>
 - Matplotlib, <http://matplotlib.org/>
 - Scikit-learn, <http://scikit-learn.org/>
 - ... Packages
- Statistica, <http://statsoft.com>
- MatLab, <http://www.mathworks.org>

Интерактивные среды

RStudio Server

www.rstudio.com

Jupyter Notebook

ipython.org/notebook.html