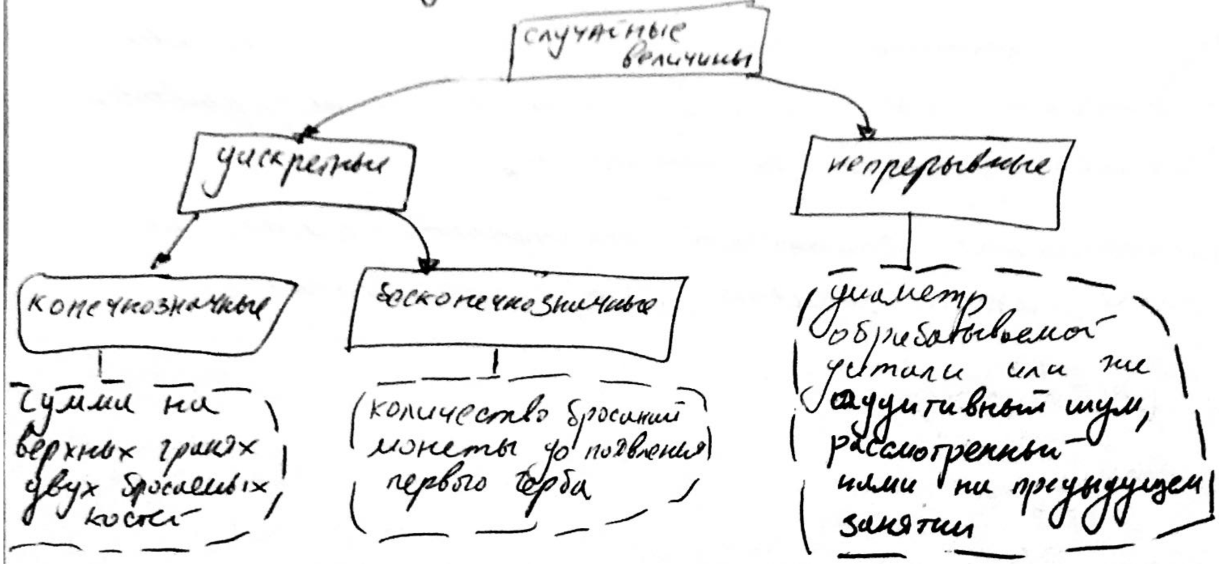
Введение в математическую статистику и теорию вероятностей

Как правило, на практике в различных задачах обработки экспериментальных данных мы имеем дело со случайными величинами. Если говорить строго математически, то случайной величиной на вероятностном пространстве $(\Omega, \mathcal{F}, P)$ называется числовая функция $X: \Omega \to R$, которая ставит в соответствие каждому элементарному исходу $\omega \in \Omega$ число $X(\omega)$ — значение случайной величины на этом исходе. Простым языком это есть ни что иное как величина, которая в результате опыта принимает то или иное значение. Выделяют 3 различных вида случайных величин, а именно:



Также случайные величины могут быть как скалярными так и многомерными. Последние представляют собой совокупность нескольких рассматриваемых совместно случайных величин.

Рассмотрим пока скалярные случайные величины.

I. Дискретные случайные величины.

Пусть X — дискретная случайная величина с возможными значениями $x_1, x_2 \dots x_n \dots$. В результате опыта X принимает одно из своих значений x_i. При этом действует правило нормировки, говорящее о том, что $\sum_{i=1}^{n} p_i = 1$, где

$p_k = P(X = x_k)$. Последнее выражение носит название "закон распределения" и является, по сути, функциональной зависимостью p_k от x_k.

Пример — закон распределения для дискретной бесконечнозначной величины — количества бросаний монеты до появления первого герба $\quad p_k = \frac{1}{2^k} \quad k \in \mathbb{N}$

Такое распределение — геометрическое т.к. условие нормировки представляет собой сумму бесконечно убывающей геометрической прогрессии.

В случае же конечных величин распределение удобно задавать в виде таблицы значений. Ещё одной хар-кой случайных величин, применимой как для дискретных так и непрерывных величин является функция распределения.

Ф-ция распределения случайной величины — вероятность принятия его значений меньших конкретного числа x

$$F(x) = P(X < x)$$

<u>Св-ва:</u> $0 \le F(x) \le 1$

$$F(-\infty) = P(X < -\infty) = 0$$

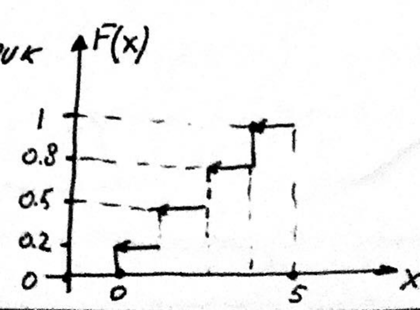
$$F(\infty) = P(X < +\infty) = 1$$

$$\forall x_2 > x_1 \quad F(x_2) \ge F(x_1)$$

$$P(x_1 \le X < x_2) = F(x_2) - F(x_1)$$

Для дискретной величины $F(x) = \sum\limits_{\forall x_k < x} p_k$

Из-за чего мы имеем ступенчатый график



II Непрерывные случайные величины

1 Случайная величина называется непрерывной в случае непрерывности соответствующей ей функции распределения.

Также для непрерывной случайной величины вводят такое понятие как плотность распределения как предел отношения вероятности её попадания в малый интервал шириной Δx вблизи точки x к ширине интервала Δx при $\Delta x \to 0$

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x \le X < x + \Delta x)}{\Delta x} = \frac{dF(x)}{dx} \geq 0$$

ввиду того, что $F(x)$ неубывающая

Кроме того $F_X(x) = \int\limits_{-\infty}^{x} f_X(t)\,dt$

Также, используя факт что $F(-\infty) = 0$ и $F(+\infty) = 1$ можно получить условие нормировки плотности распределения

$$\int\limits_{-\infty}^{+\infty} f(x)\,dx = 1$$

Кроме того имеет место и следующее свойство

$$P(x_1 < X < x_2) = \int\limits_{x_1}^{x_2} f(x)\,dx$$

Как же мы можем характеризовать случайную величину?

~~[illegible]~~

<u>Математическое ожидание</u> – сумма произведений всех возможных значений величины X на вероятности ~~их~~ появления.

$$m_X = \int\limits_{-\infty}^{+\infty} x f(x)\,dx$$ ← непрерывная случайная величина.

$$m_X = \sum_{k=1}^{\infty} x_k p_k$$ ← дискретная случайная величина

Мода случайной величины – наиболее вероятное её значение. Для дискретной случайной величины это такое x_k для которого максимальна p_k, а для непрерывной такое x при котором максимальна $f(x)$. Также есть распределения у которых всевозможные значения – моды. Такие распределения – равномерные.

Медиана случайной величины – такое значение x_m при котором $F(x_m)=0{,}5$

Моментом (начальным моментом) m-го порядка называется сумма произведений всех возможных значений величины X в m-ой степени на вероятности их появления.

$$\alpha_m=\sum_{k=1}^{\infty} x_k^m p_k \leftarrow \text{дискретная случайная величина}$$

$$\alpha_m=\int_{-\infty}^{+\infty} x^m f(x)dx \leftarrow \text{непрерывная случайная величина}$$

Центральным моментом m-го порядка называется сумма произведений всех всевозможных значений величины X из которых вычтено её математическое ожидание в m-й степени на вероятности их появления

$$\mu_m=\sum_{k=1}^{\infty}(x_k-m_x)^m p_k \leftarrow \text{дискретная случайная величина}$$

$$\mu_m=\int_{-\infty}^{+\infty}(x-m_x)^m f(x)dx \leftarrow \text{непрерывная случайная величина}$$

Первый центральный момент всегда равен нулю.

Центральные моменты выражаются через начальные

$$\mu_2=\alpha_2-\alpha_1^2$$

$$\mu_3=\alpha_3-3\alpha_2\alpha_1+2\alpha_1^3$$

$$\mu_4=\alpha_4-4\alpha_3\alpha_1+6\alpha_1^2\alpha_2-3\alpha_1^4$$

Дисперсией случайной величины называется её второй случайный момент $D=\mu_2=\int_{-\infty}^{+\infty}(x-m_x)^2 f(x)dx \geq 0$

Среднеквадратичное отклонение случайной величины — квадратный корень из её дисперсии.

Как правило случайные величины подчинены какому-то закону и реализованы в среде MATLAB. Благодаря чему легко можно узнать их параметры. Рассмотрим ниже основные распределения и соответствующие им функции языка MATLAB.

Биномиальное распределение

Проводится n одинаковых независимых опытов в каждом из которых возможен один из двух взаимоисключающих исходов (успех/неуспех) $P(У) = p$

$P(Н) = 1 - p = q$ ← схема Бернулли

Рассмотрим случайную величину X — количество успехов в n испытаниях по схеме Бернулли. Ее распределение называют биномиальным.

$$p_k = \binom{n}{k} p^k q^{n-k}$$

В MATLAB вычисление закона биномиального распределения реализовано с помощью ф-ции binopdf, и ф-ции распределения binocdf.

$m_x = np$ (если проводится n независимых опытов, и вероятность успеха в каждом из них равна p, то в среднем ожидается np успехов)

$D_x = npq$

Вычисление МО и дисперсии биномиального распределения реализовано в MATLAB в виде функции binostat.
Также с помощью ф-ции binornd можно сгенерировать случайные числа с биномиальным распределением, а с помощью функции binofit оценить параметр p биномиального распределения по данным наблюдения.

Геометрическое распределение - распределение случайной величины X-количества испытаний по схеме Бернулли до первого появления события У.

$$p_k = pq^{k-1}$$

$$m_x = p\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{p}$$

$$D_x = p\sum_{k=1}^{\infty} k^2 q^{k-1} - \frac{1}{p^2} = \frac{q}{p^2}$$

geopdf - закон распределения;
geocdf - функция распределения;
geornd - генератор случайных величин с геом. р-ем;
geostat - МО и дисперсия.

Распределение Пуассона

Поток событий - одноименные события, происходящие в случайные моменты времени. Мы хотим определять вероятность наступления того или иного количества событий потока на заданном отрезке $[t_1, t_2]$. X - количество событий на $[t_1, t_2]$

Поток без последствий - вероятность наступления того или иного количества событий в $[t_1, t_2]$ не зависит от того, сколько событий произошло до момента t_1.

Ординарный поток - вероятность наступления одного события за бесконечно малое время dt является бесконечно малой величиной одного порядка малости с dt, а вероятность наступления более чем одного события за бесконечно малое время dt является бесконечно малой величиной высшего порядка малости по сравнению с dt. В таком потоке за время dt может произойти или одно событие с вероятностью λdt или 0 событий с вероятностью $1-\lambda dt$

Пуассоновский поток - ординарный поток в котором отсутствуют последствия

k - число событий

$$p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad k = 0, 1, 2 \ldots$$

величины

Пуассоновским называется распределение случайной величины X X-количество событий в пуассоновском потоке за заданное фиксированное время

$$p_k = \frac{a^k}{k!} e^{-a} \quad k = 0, 1, 2 \ldots$$

$$\lambda t = a = const$$

$$m_x = a$$

$$D_x = a$$

poisspdf – закон распределения;

poisscdf – ф-ция распределения;

poissrnd – случ. числа по пуассоновскому закону;

poisstat – МО и дисперсия

poissfit – оценка пар-ров

Равномерное распределение

Непрерывное распределение величины X называется равномерным на отрезке [a, b] если все её возможные значения сосредоточены на этом отрезке, а плотность распределения на этом отрезке постоянна.

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

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & x \in [a, b] \\ 1, & x > b \end{cases}$$

$$m_x = \frac{b+a}{2}$$

$$D_x = \frac{(b-a)^2}{12}$$

unifpdf – плотность распределения;

unifcdf – ф-ция распределения;

unifrnd – генератор случайных чисел с равномерным непр.
~rand распределением

unifstat – МО и дисперсия

unifit – оценка параметров равномерного р-я

Нормальное (Гауссово) распределение

Распределение непрерывной величины называется нормальным если его плотность имеет вид $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-m}{\sigma}} e^{-u^2/2} du$$

$m_x = m$

$D_x = \sigma^2$

normpdf – закон распределения;
normcdf – функция распределения;
normrnd – генератор случайных чисел с нормальным распр-ем;
normstat – МО и дисперсия нормального распределения;
normfit – оценки параметров нормального распределения

<u>Математическая статистика</u> – приложение теории вероятностей к практическим задачам обработки результатов эксперимента. Как правило, на практике, необходимо обобщить результаты исследования т.е. по части данных дать результат по всем.
Исследуемая величина – генеральная совокупность, полученный из неё набор экспериментальных данных – выборочной совокупностью или выборкой.
Одна из основных характеристик выборки – <u>выборочная функция распределения</u>. Рассмотрим случайную величину X с функцией распределения $F(x)$ и возьмем из неё выборку объёма n:
$x_1, x_2 \ldots x_n$ и упорядочим эти числа в порядке возрастания
$x_1 \le x_2 \le \ldots x_n$ и по этим числам построим функцию:

$$F^*(x)=\begin{cases} 0 & x\le x_1 \\ \frac{1}{n} & x_1<x\le x_2 \\ \frac{2}{n} & x_2<x\le x_3 \\ \dots & \\ \frac{n-1}{n} & x_{n-1}<x\le x_n \\ 1 & x>x_n \end{cases}$$

В разных реализациях выборки эта функция разная, но если $F^*(x)$ хорошо приближает $F(x)$, то и остальные выборочные параметры хорошо описывают соответствующие генеральные параметры. Соответственно возникает задача оценки параметров распределения по выборке из него. Соответственно мы строим какие-то оценки параметров распределения желая, чтобы оценка была состоятельной, несмещенной и эффективной. (последнее при наличии нескольких оценок выбирает ту, у которой дисперсия меньше при каждом n).

Выборочное математическое ожидание

$$M_x^*=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad \boxed{mean}$$

Выборочная дисперсия

$$\Delta_x^*=\frac{1}{n}\sum_{i=1}^{n}(X_i-M_x^*)^2$$

$$D_x^*=\frac{n}{n-1}\Delta_x^*=\boxed{\frac{1}{n-1}}\sum_{i=1}^{n}(X_i-m_x^*)$$

(из-за требования несмещенности
M(оценки) = зн-е оцен. пар-ра)

$\delta=n-1$ — число степеней свободы выборки из n элементов.

$\boxed{var}$

Медиана выборки — median

Размах — разность м/у X_{max} и X_{min} range

5