Компьютерная обработка результатов измерений

Лекция 2. Статистика и вероятность. Случайные величины и распределения

Емельянов Эдуард Владимирович

Специальная астрофизическая обсерватория РАН
Лаборатория обеспечения наблюдений

12 июля 2016 года

Случайные величины, вероятность

Случайной величиной называется величина X, если все ее возможные значения образуют конечную или бесконечную последовательность чисел $x_1, \ldots, x_N$, и если принятие ею каждого из этих значений есть случайное событие.

Вероятность наступления данного события — это предел относительной частоты наступления данного события n_k/N:

$$P(x_k) = \lim_{N\to\infty} \frac{n_k}{N}.$$

Совместные и несовместные события, полная группа, свойства вероятности. Для непрерывных случайных величин вводят понятие **плотности вероятности**:

$$\rho(x) = \lim_{\Delta x\to 0} \frac{P(x < X < x + \Delta x)}{\Delta x} = \frac{dP}{dx}.$$

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

Характеристики случайных величин

Независимые случайные величины

$P(x_n y_n) = P(x_n)P(y_n)$.

Среднее арифметическое и математическое ожидание

$$<X> = {}^1/_N \sum_{n=1}^{N} x_n,$$

$$M(X) \equiv \overline{X} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x_n \quad \text{и} \quad M(X) = \int_{-\infty}^{\infty} x\phi(x)\,dx.$$

Свойства математического ожидания

- $\overline{\mathfrak{const}} = \mathfrak{const}$;
- $\overline{\sum \mathfrak{C}_n \cdot X_n} = \sum \mathfrak{C}_n \cdot \overline{X_n}$, где $\mathfrak{C}_n$ – постоянная величина;
- $\overline{\prod X_n} = \prod \overline{X_n}$ (для независимых случайных величин);
- $\overline{f(x)} = \int\limits_{-\infty}^{\infty} f(x)\phi(x)\,dx$ (для непрерывных случайных величин).

$\overline{X} \Leftrightarrow <X>$? Закон больших чисел

Неравенство Чебышёва: $P(|X - \overline{X}| < \epsilon) \geq 1 - D(X)/\epsilon^2 \quad \Rightarrow$

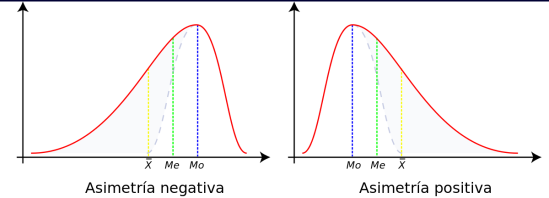
$$\lim_{n\to\infty} P\left(\left|\frac{\sum X_n}{n} - \frac{\sum \overline{X_n}}{n}\right| < \epsilon\right) = 1.$$

Теорема Бернулли: $\lim\limits_{n\to\infty} P(|m/n - p| < \epsilon) = 1$.

Характеристические значения распределений

Медиана и мода

Мода — наиболее часто встречающееся значение (но вполне могут быть мультимодальные распределения). **Медиана** делит площадь распределения пополам.



Характеристические значения распределений

Моменты

Если $f(x) = (x - x_0)^n$, то $\overline{f(x)}$ — момент порядка n. Если $x_0 = 0$ — начальный момент, если $x_0 = \overline{X}$— центральный момент.
Моменты нулевого порядка равны 1, начальный момент первого порядка равен математическому ожиданию случайной величины; центральный момент первого порядка равен нулю.
Центральный момент второго порядка называют **дисперсией**:
$D(X) = \overline{(x - \overline{x})^2} \equiv \overline{x^2} - \overline{x}^2$, $\sigma = \sqrt{D}$.

Свойства дисперсии:

- $D(\mathfrak{const}) = 0$;
- $D(\mathfrak{const}X) = C^2 D(X)$, где $\mathfrak{C}$ – постоянная величина;
- $D(\sum X_n) = \sum D(X_n)$.

Законы распределения

Закон распределения *дискретной* случайной величины — соответствие между возможными значениями и их вероятностями.

Функция распределения

$$F(x) \equiv P(X \le x) = \int\limits_{-\infty}^{x} \phi(x)\,dx, \qquad \int\limits_{-\infty}^{\infty} \phi(x)\,dx = 1.$$

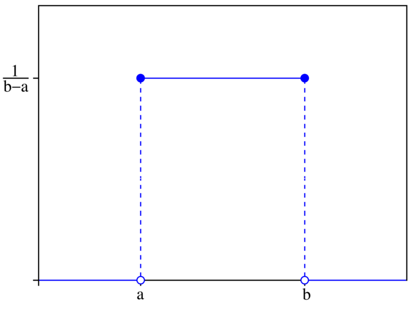
$$P(a \le X \le b) = F(b) - F(a).$$

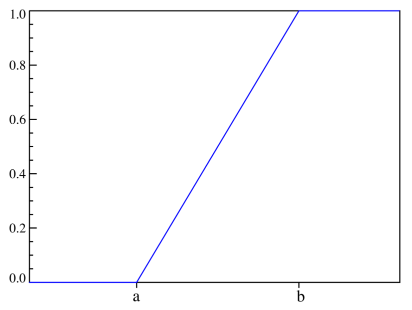
Равномерное распределение

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

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x < b \\ 1, & x \ge b \end{cases}.$$

$$\overline{X} = \mathrm{med}(X) = (a+b)/2,$$
$$\mathrm{Mo}(X) = \forall x \in [a,b],$$
$$\sigma_X^2 = \frac{(b-a)^2}{12}.$$



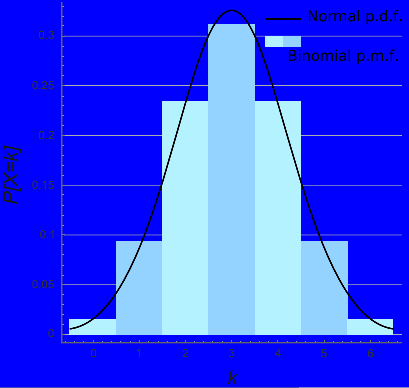


Биномиальное распределение

Формула Бернулли: $P_n(k) = C_n^k p^k q^{n-k}, \quad C_n^k = \dfrac{n!}{k!(n-k)!}, \quad q = 1 - p.$

$$(p+q)^n = C_n^n p^n + \cdots + C_n^k p^k q^{n-k} + \cdots + C_n^0 q^k.$$

Описывает вероятность наступления события k раз в n независимых испытаниях



$$F(k; n, p) = P(X \le k) = \sum_{i=0}^{\lfloor k \rfloor} C_n^i p^i (1-p)^{n-i}.$$

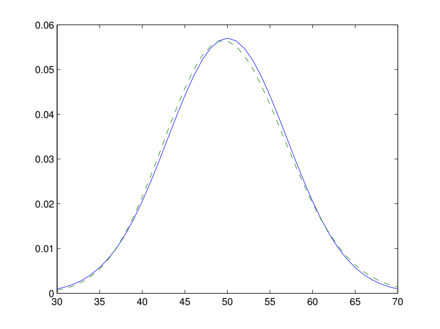
$\overline{X} = np$, $\mathrm{Mo}(X) = \lfloor (n+1)p \rfloor$,
$\lfloor np \rfloor \le \mathrm{med}(X) \le \lceil np \rceil$, $\sigma_X^2 = npq$.

Распределение Пуассона

При $n \to \infty$ распределение Бернулли преобразуется в распределение Пуассона ($\lambda = np$):

$$P_n(k) = \frac{\lambda^k}{k!}\exp(-\lambda).$$

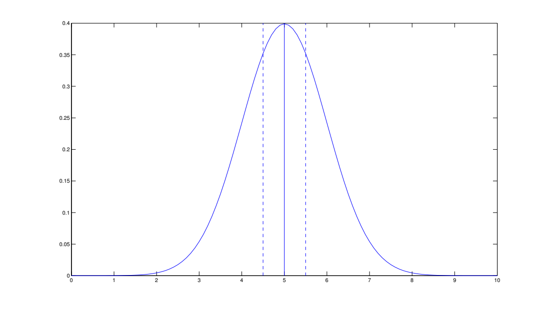
$F(k,\lambda) = \dfrac{\Gamma(k+1,\lambda)}{k!}$, $\overline{X} = \lambda$,
$\mathrm{Mo}(X) = \lfloor\lambda\rfloor$,
$\mathrm{med}\,X \approx \lfloor\lambda + 1/3 - 0.02/\lambda\rfloor$,
$\sigma_X^2 = \lambda$.
С ростом λ распределение Пуассона стремится к распределению Гаусса.



Распределение Гаусса

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

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\left(-\frac{(t-\overline{x})^2}{2\sigma^2}\right) dt, \ \mathrm{Mo}(X) = \mathrm{med}\, X = \overline{X}.$$

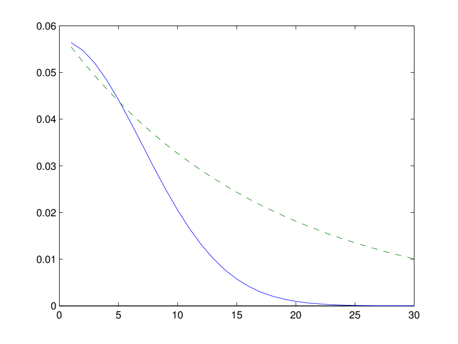


Показательное (экспоненциальное) распределение

Время между двумя последовательными свершениями события

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

$\overline{X}=\lambda^{-1}$, $\mathrm{Mo}(X)=0$, $\mathrm{med}\,X=\ln(2)/\lambda$, $\sigma_X^2=\lambda^{-2}$.



Корреляция и ковариация

Ковариация является мерой линейной зависимости случайных величин и определяется формулой: $\mathrm{cov}(X,Y) = \overline{(X-\overline{X})(Y-\overline{Y})} \Longrightarrow \mathrm{cov}(X,X) = \sigma_X^2$.
Ковариация независимых случайных величин равна нулю, обратное неверно.

Если ковариация положительна, то с ростом значений одной случайной величины, значения второй имеют тенденцию возрастать, а если знак отрицательный — убывать.
Масштаб зависимости величин пропорционален их дисперсиям $\Longrightarrow$ масштаб можно отнормировать (**коэффициент корреляции** Пирсона):

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma X \sigma Y}, \qquad \mathbf{r} \in [-1,1].$$

Коэффициент корреляции равен ± 1 тогда и только тогда, когда X и Y линейно зависимы. Если они независимы, $\rho_{X,Y} = 0$ (**обратное неверно!**). Промежуточные значения коэффициента корреляции не позволяют однозначно судить о зависимости случайных величин, но позволяет предполагать степень их зависимости.

Корреляционная функция

Одна из разновидностей — **автокорреляционная функция**:

$$\Psi(\tau) = \int f(t) f(t-\tau)\,dt \equiv \int f(t+\tau) f(t)\,dt.$$

Для дискретных случайных величин автокорреляционная функция имеет вид

$$\Psi(\tau) = <X(t) X(t-\tau)> \equiv <X(t+\tau) X(t)>.$$

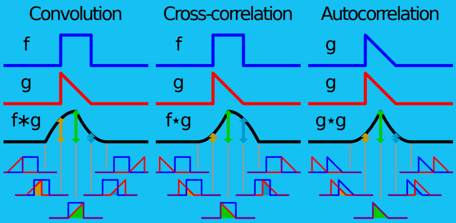
Взаимно корреляционная функция

Другая разновидность — **кросс–корреляционная функция**:

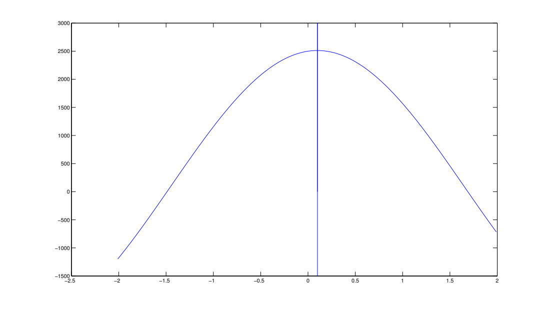
$$(f \star g)(t) \stackrel{\text{def}}{=} \int\limits_{-\infty}^{\infty} f^*(\tau) g(t+\tau)\, d\tau$$

свертка:

$$(f * g)(x) \stackrel{\text{def}}{=} \int\limits_{-\infty}^{\infty} f(y)\, g(x-y)\, dy = \int\limits_{-\infty}^{\infty} f(x-y)\, g(y)\, dy.$$



Если X и Y — два независимых случайных числа с функциями распределения вероятностей f и g, то $f \star g$ соответствует распределению вероятностей выражения $-X + Y$, а $f * g$ — распределению вероятностей суммы $X + Y$.
ВКФ часто используется для поиска в длинной последовательности более короткой заранее известной, определения сдвига (см. рис).
Связь со сверткой: $f(t) \star g(t) = f^*(-t) * g(t)$, если f и g четны, то $f(t) \star g(t) = f(t) * g(t)$. Через преобразование Фурье: $\mathcal{F}(f \star g) = \mathcal{F}(f)^* \cdot \mathcal{F}(g)$.



Шум

Шум — беспорядочные колебания различной физической природы, отличающиеся сложной временной и спектральной структурой.

Белый шум, $\xi(t)$, имеет время корреляции много меньше всех характерных времен физической системы; $\overline{\xi(t)} = 0$, $\Psi(t, \tau) = <\xi(t + \tau)\xi(t)> = \sigma^2(t)\delta(\tau)$. Разновидность — AWGN.

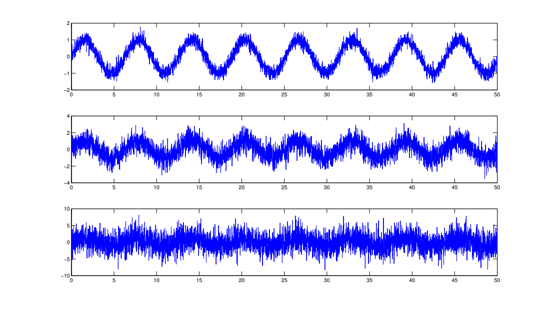
Дробовой шум имеет пуассонову статистику $\Longrightarrow \sigma_X \propto \sqrt{x}$ и $\mathrm{SNR}(N) \propto \sqrt{N}$. Суточные и вековые корреляции.

Шум вида **«соль–перец»** обычно характерен для изображений, считываемых с ПЗС.

SNR

SNR — безразмерная величина, равная отношению мощности полезного сигнала к мощности шума.

$$\mathrm{SNR} = \frac{P_\mathrm{signal}}{P_\mathrm{noise}} = \left(\frac{A_\mathrm{signal}}{A_\mathrm{noise}}\right)^2, \quad \mathrm{SNR}(dB) = 10\lg\left(\frac{P_\mathrm{signal}}{P_\mathrm{noise}}\right) = 20\lg\left(\frac{A_\mathrm{signal}}{A_\mathrm{noise}}\right).$$



(10, 0, -10 дБ.)

Спасибо за внимание!

mailto

eddy@sao.ru
edward.emelianoff@gmail.com