Компьютерная обработка результатов измерений

Лекция 2. Статистика и вероятность. Случайные величины и распределения

Емельянов Эдуард Владимирович

Специальная астрофизическая обсерватория РАН
Лаборатория физики оптических транзиентов

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.

Вероятностью наступления события называют предел относительной частоты наступления данного события n_k/N:

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

Если событие **невозможно** ($\emptyset$), его вероятность равна нулю. Однако, обратное в общем случае неверно (например, вероятность попасть в конкретную точку мишени равна нулю, но это событие не является невозможным).
Вероятность **достоверного** события равна 1.

Условная вероятность

Условной вероятностью двух событий A и B (вероятность появления A при условии B) называют отношение числа опытов, в которых A и B появились вместе, к полному числу опытов, в которых появилось B:

$$P(A|B) = \frac{n_{AB}}{n_B} = \frac{n_{AB}/N}{n_B/N} = \frac{P(AB)}{P(B)}.$$

У **независимых** событий $P(A|B) = P(A)$, $P(B|A) = P(B)$. А т.к. $P(A|B) = \frac{P(AB)}{P(B)}$, получим для независимых событий:

$$P(AB) = P(A) \cdot P(B).$$

Умножение вероятностей

$$P(AB) = P(B) \cdot P(A|B) = P(A) \cdot P(B|A).$$

Сложение вероятностей

Несовместные события

$A_iA_j = \emptyset \quad \forall i \neq j$, $P(A_i + A_j) = P(A_i) + P(A_j)$.

Совместные события

$P(A_i + A_j) = P(A_i) + P(A_j) - P(A_iA_j)$.

Независимые совместные события

$$P(\overline{A}\,\overline{B}) = P(\overline{A})\cdot P(\overline{B}) = (1-P(A))\cdot(1-P(B)) = 1-P(A)-P(B)+P(A)\cdot P(B)$$

$$P(A+B) = P(A) + P(B) - P(AB) \quad \Rightarrow$$

$$1 - P(A+B) = P(\overline{A}\,\overline{B}) \quad \text{или} \quad P(A+B) = 1 - P(\overline{A}\,\overline{B}).$$

Полная вероятность

Полная вероятность (вероятность события, зависящего от условий опыта) является следствием правил сложения и умножения вероятностей.

N условий опыта должны быть взаимоисключающими, т.е. несовместными: $P(H_i H_j) = 0$ для $j \neq j$. И они должны формировать **полную группу**, т.е. $\sum P(H_i) = 1$. Тогда $P(A) = \sum P(AH_i)$. А т.к. $P(AH_i) = P(H_i) \cdot P(A|H_i)$, получим:

$$P(A) = \sum_{i=1}^{N} P(H_i) \cdot P(A|H_i).$$

Здесь $P(H_i)$ — **априорная вероятность** (известна до проведения опыта). Вероятность $P(A|H_i)$ мы узнаем из опыта, ее называют **апостериорной**.

Полная вероятность

Пример

Среди наблюдаемых спиральных галактик 23% имеют тип Sa, 31% – тип Sb и 45% – тип Sc. Вероятность вспышки сверхновой в течение года в галактике Sa составляет 0.20%, в Sb – 0.35%, в Sc – 0.55%. Найти вероятность вспышки сверхновой в спиральной галактике, тип которой не удается определить. $P(S_a) = 0.23$, $P(S_b) = 0.31$, $P(S_c) = 0.46$. Вероятность вспышки в галактике типа X есть $P(F|X)$. Тогда полная вероятность вспышки равна $P(F) = \sum P(X)P(F|X)$. То есть:

$$P(F) = 0.23 \cdot 0.002 + 0.31 \cdot 0.0035 + 0.46 \cdot 0.0055 = 0.0041 = 41\%.$$

Формула (теорема) Байеса

Как и для полной вероятности, гипотезы H_i считаем несовместными, образующими полную группу. Событие A считаем уже произошедшим. В этом случае можно пересчитать априорные вероятности $P(H_i)$ с учетом этого. Найдем $P(H_i|A)$. Известно, что $P(H_iA) = P(H_i) \cdot P(A|H_i)$ или $P(H_iA) = P(A) \cdot P(H_i|A)$.

$$P(A) \cdot P(H_i|A) = P(H_i) \cdot P(A|H_i), \quad \Rightarrow$$

Формула Байеса

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{P(A)},$$

где $P(A) = \sum P(H_i)P(A|H_i)$.

Формула (теорема) Байеса

Пример

В течение часа наблюдений была обнаружена вспышка сверхновой в спиральной галактике неизвестного типа. Определить вероятность того, что галактика принадлежит каждому из подтипов Sa, Sb или Sc.

По формуле Байеса, $P(X|F) = \dfrac{P(X)P(F|X)}{P(F)}$. В предыдущем примере мы уже нашли: $P(F) = 0.0041$, следовательно

$$P(S_a|F) = \frac{0.23 \cdot 0.0020}{0.0041} = 0.11,$$

$$P(S_b|F) = \frac{0.31 \cdot 0.0035}{0.0041} = 0.27,$$

$$P(S_c|F) = \frac{0.46 \cdot 0.0055}{0.0041} = 0.62.$$

Итог: свойства вероятности

$P(\emptyset) = 0$	
$\forall A \subset B \quad P(A) \leqslant P(B)$	B включает в себя A
$0 \leqslant P(A) \leqslant 1$	
$\forall A \subset B \quad P(B \setminus A) = P(B) - P(A)$	B наступит без A
$P(\overline{A}) = 1 - P(A)$	
$P(A+B) = P(A) + P(B) - P(AB)$	вероятность одного из событий
$P(A\vert B) = {}^{P(AB)}/_{P(B)}$	условная вероятность (A при B) $\Longrightarrow$
$P(AB) = P(B) \cdot P(A\vert B)$	или $P(AB) = P(A) \cdot P(B\vert A) \Longrightarrow$
$P(A\vert B) = \dfrac{P(A) \cdot P(B\vert A)}{P(B)}$	(теорема Байеса)
$P(AB) = P(A) \cdot P(B)$	для независимых событий

Комбинаторика

Размещение

Количество способов, которыми можно разместить n элементов по k ячейкам.

Без повторений: $A_n^k = n(n-1)\cdots(n-k+1) = \dfrac{n!}{(n-k)!} = \binom{n}{k}k!$.

С повторениями (каждый предмет можно взять до k раз): $\overline{A}_n^k = n^k$.

Размещение без повторений встречается в задачах на составление k-значных чисел из n цифр, причем, каждая цифра может использоваться лишь однократно. Размещение с повторениями показывает все возможные комбинации n цифр в k разрядах (например, количество чисел до k-го разряда по основанию n).

Комбинаторика

Перестановка

Без повторений: $P_n = A_n^n = n!$.

С повторениями (n элементов m типов). n_i – количество элементов каждого типа (т.е. $\sum n_i = n$). $P(n_1, \ldots, n_m) = \dfrac{n!}{\prod n_i!}$.

Задача на перестановки без повторений является частным случаем задачи размещения без повторений, когда $k = n$.

Пример задачи на перестановки с повторениями — формирование разных слов (даже лишенных смысла) из букв заданного слова. Например, из слова «собака» можно составить $6!/(1!1!1!2!1!) = 720/2 = 36$.

Комбинаторика

Сочетание

Неупорядоченный набор из k элементов n-элементного множества. Т.о. сочетание — это такое размещение n по k, где не учитывается порядок следования членов (напр., размещения 123, 213, 321 и т.д. считаются одним сочетанием).

Без повторений: $C_n^k = \binom{n}{k} = \dfrac{n!}{k!(n-k)!}$.

С повторениями: $\overline{C}_n^m = \binom{n+k-1}{n-1} = \binom{n+k-1}{k} = \dfrac{(n+k-1)!}{k!(n-1)!}$.

Схема испытаний Бернулли: $P_n^k = C_n^k p^k (1-p)^{n-k}$ (вероятность, что событие наступит k раз в n испытаниях).
$1 = (p + [1-p])^n = \sum C_n^k p^{n-k} (1-p)^k = \sum P_n^k$.

Для непрерывных случайных величин, X, вводят понятия **Функции распределения**, $F(x)$ и **плотности вероятности**, $\rho(x)$: $F(x) = P(X < x)$.

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

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

Генеральная совокупность — набор всех возможных значений случайной величины. **Выборка** — конечное число значений (подвыборка генеральной совокупности). **Энтропия** выборки:

$$E = -\sum_{k=1}^{n} p(x_k) \lg p(x_k).$$

Характеристики случайных величин

Среднее арифметическое и математическое ожидание

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

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

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);
- $\overline{f(x)} = \int_{-\infty}^{\infty} f(x)\varphi(x)\,dx$ (для непрерывных случайных величин).

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.
Центральный момент второго порядка называют **дисперсией**: $D(X) = \overline{(x - \overline{x})^2} \equiv \overline{x^2} - \overline{x}^2$. **СКО**: $\sigma = \sqrt{D}$.

Свойства дисперсии:

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

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

Неравенство Чебышёва: $P(|X - \overline{X}| \geqslant \varepsilon) \leqslant D(X)/_{\varepsilon^2} \quad \Rightarrow$
$P(|X - \overline{X}| < \varepsilon) = 1 - P(|X - \overline{X}| \geqslant \varepsilon) \geqslant 1 - D(X)/_{\varepsilon^2}$.

$$\lim_{n\to\infty} P\left(\left|\frac{\sum X_n}{n} - \frac{\sum \overline{X_n}}{n}\right| < \varepsilon\right) = 1, \quad \text{т.к. } D(\textstyle\sum X_n/_n) = D(X)/_n$$

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

Квантиль – значение, которое случайная величина не превышает с фиксированной вероятностью. α-квантиль, x_α: $P(X \leqslant x_\alpha) = \alpha$, $P(X \geqslant x_\alpha) = 1 - \alpha$.

$P(X \leqslant x_{\frac{1+\alpha}{2}}) = \frac{1+\alpha}{2}$, $P(X \leqslant x_{\frac{1-\alpha}{2}}) = \frac{1-\alpha}{2}$, следовательно, свойство:

$$P(x_{\frac{1-\alpha}{2}} \leqslant X \leqslant x_{\frac{1+\alpha}{2}}) = \frac{1+\alpha}{2} - \frac{1-\alpha}{2} = \alpha.$$

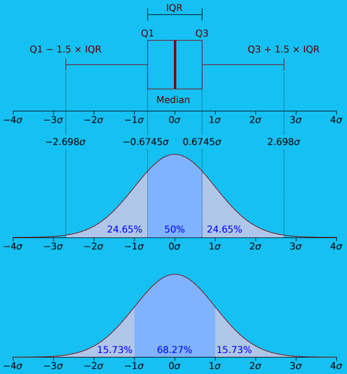
Процентиль (перцентиль) – квартиль, выраженная в процентах. Например, «70-й перцентиль» (величина с вероятностью 70% меньше этого значения). *Квартиль* – деление на четыре (первый, второй и третий квартили). **Медиана** – второй квартиль. $IQR = x_{0.75} - x_{0.25}$ – интерквартильный интервал.

Квантили нормального распределения

P – вероятность, x_P – квантиль (в RMS от мат. ожидания), $P_c = P(-x_P \leqslant X - \overline{X} \leqslant x_P)$.

P	99.99	99.90	99.00	97.72	97.50
x_P	3.719	3.090	2.326	1.999	1.960
P_c	99.98	99.80	98.00	95.44	95.00

P	95.00	90.00	84.13	50.00
x_P	1.645	1.282	1.000	0.000
P_c	90.00000	80.00	68.27	0.00



Octave: пакет `statistics`, функция `norminv`. Например:

```
norminv([0.9 0.95 0.99 0.999 0.9999])
ans =
1.2816   1.6449   2.3263   3.0902   3.7190
```

Можно также задать $\overline{X}$ и σ_X (скажем, квантиль 90% при $\overline{X} = 25$ и $\sigma_X = 3$):

```
norminv(0.9, 25, 3)
ans =  28.845
```

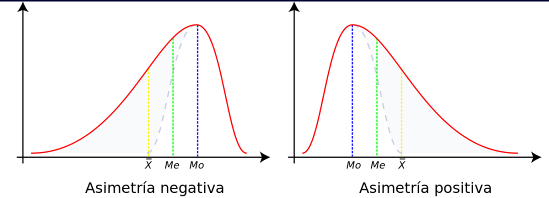
Для расчета вероятности $P(X \leqslant x_0)$ функция `normcdf` (интегральное распределение). Например, посчитаем вероятности нахождения в интервале $\overline{X} \pm k\sigma$:

```
k=[1:0.5:5];
normcdf(k)-normcdf(-k)
ans =
0.68269   0.86639   0.95450   0.98758   0.99730   0.99953
0.99994   0.99999   1.00000
```

Характеристические значения распределений

Медиана и мода

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



Поиск медианы

Самый медленный — сортировкой ряда данных, $O(n \ln n)$. Quick Select, $O(n)$. Гистограмма (в т.ч. дерево гистограмм), $O(n)$. Для фиксированных n — opt_med („Numerical Recipes in C“), $O(n)$.

Законы распределения

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

Функция распределения

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

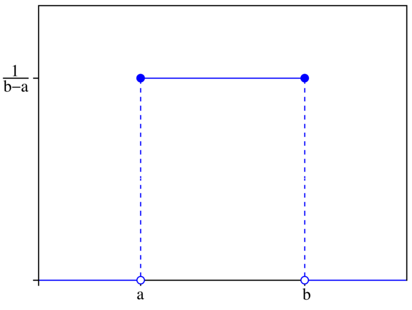
$$P(a \leqslant X \leqslant b) = F(b) - F(a).$$

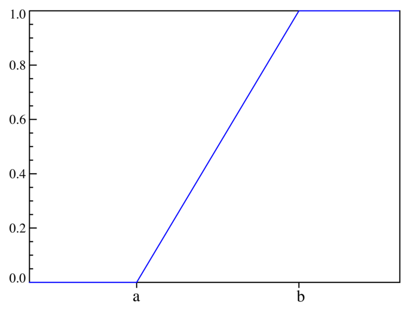
Равномерное распределение

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

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

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



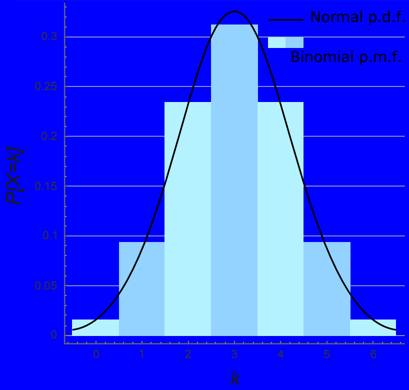


Биномиальное распределение

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



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

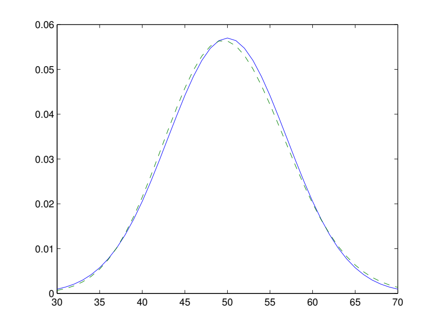
$\overline{X} = np$, $\mathrm{Mo}(X) = \lfloor (n+1)p \rfloor$,
$\lfloor np \rfloor \leqslant \mathrm{med}(X) \leqslant \lceil np \rceil$, $\sigma_X^2 = npq$.

Распределение Пуассона

Распределение вероятности *редких событий*. 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.

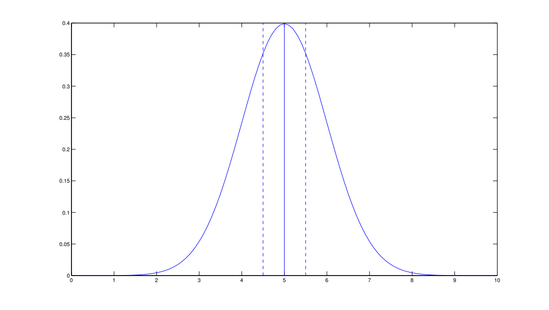


Распределение Гаусса

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

$\mathrm{Mo}(X)=\mathrm{med}\,X=\overline{X}$. $P(\alpha<X<\beta)=\Phi\left(\frac{\beta-\overline{x}}{\sigma}\right)-\Phi\left(\frac{\alpha-\overline{x}}{\sigma}\right)$,

функция Лапласа $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x\exp\left(-t^2/_2\right)$.

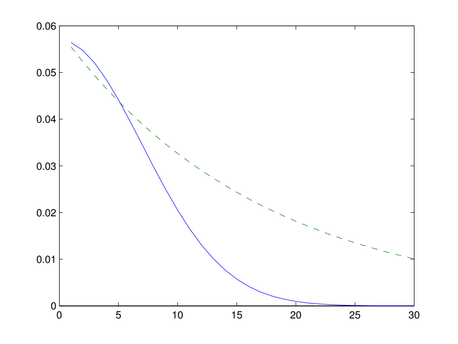


Показательное (экспоненциальное) распределение

Время между двумя последовательными свершениями события

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

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



Корреляция и ковариация

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

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

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

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

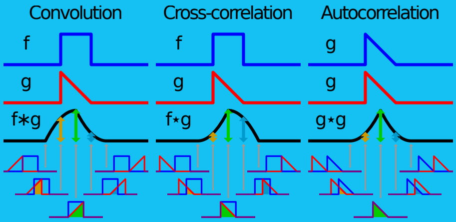
Взаимно корреляционная функция

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

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

свертка:

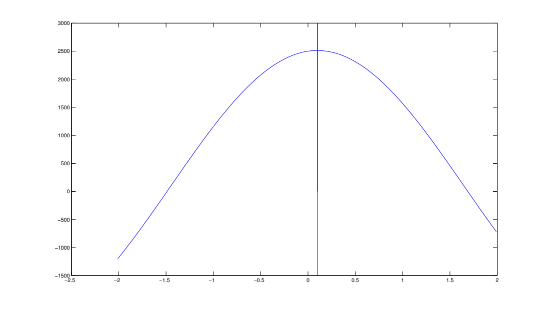
$$(f * g)(x) \stackrel{\text{def}}{=} \int\limits_{-\infty}^{\infty} f(y)\, g(x-y)\, dy = \int\limits_{-\infty}^{\infty} f(x-y)\, g(y)\, dy.$$



Если X и Y — две независимых случайных величины 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: $\mathcal{F}(f \star g) = \mathcal{F}(f)^* \cdot \mathcal{F}(g)$.



Применение корреляции

- Расчет спектральной плотности энергии и энергетического содержимого сигнала. $\mathcal{F}(\Psi(\tau)) = G_E(f)$ – образ Фурье автокорреляционной функции есть спектральная плотность энергии; $\Psi(0) = E$ – полная энергия сигнала.
- Детектирование и оценка периодических сигналов в шуме.
- Корреляционное детектирование.

Шум

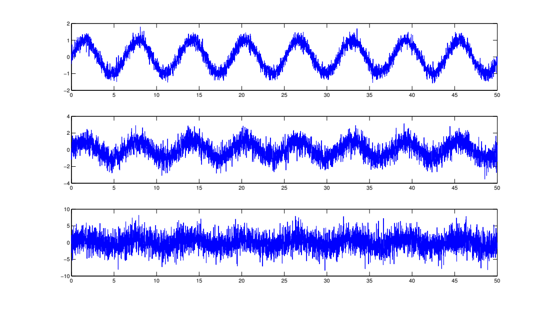
Шум — беспорядочные колебания различной физической природы, отличающиеся сложной временной и спектральной структурой.

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