Международная конференция "Современные проблемы прикладной математики и механики: теория, эксперимент и практика", посвященная 90-летию со дня рождения академика Н.Н. Яненко

ЧИСЛЕННЫЕ ОПЕРАЦИИ НАД ПЛОТНОСТЯМИ ВЕРОЯТНОСТЕЙ СЛУЧАЙНЫХ ВЕЛИЧИН

Борис Добронец[1] Ольга Попова[2]

[1]Институт Космических и Информационных Технологий
Сибирский Федеральный Университет, Красноярск
e-mail: BDobronets@sfu-kras.ru

[2]СИБАДИ, Омск
e-mail: olgaarc@yandex.ru

Новосибирск, Россия, 30 мая — 4 июня 2011 г.

Distributions are the numbers of the future

Schweizer, B. (1984). Distributions are the numbers of the future. In Proceedings of the mathematics of fuzzy systems meeting (pp. 137–149). Naples: University of Naples.

Неопределенность

Случайная	Значения показателей есть случайные числа, которые задаются некоторыми вероятностными распределениями их возможных значений
Интервальная	Состояние неполного (частичного знания) об интересующей нас величине, когда нам известна лишь ее принадлежность некоторому интервалу, т. е. мы можем указать границы возможных значений этой величины или пределы их изменения
Нечеткая	Задаются лингвистически сформулированными распределениями их значений

aleatory & epistemic uncertainty

Неопределённость, которая является неотъемлемым атрибутом случайных событий, называют элитерной (aleatory) неопределённостью. Теория вероятностей предназначена для моделирования, оценки и оперирования именно элитерными неопределённостями. В свою очередь, неопределённость самих вероятностных оценок называют эпистемической неопределённостью (epistemic uncertainty). Эпистемическая неопределённость прямо связана с объёмом и достоверностью информации, на основании которой получаются эти оценки.

Арифметика неопределенных значений

Цель: осуществление расчетов таким образом, чтобы получить реальные результаты, с тем чтобы не получить дополнительные неопределенности!!!!.

Hans Schjaer-Jacobsen
Representation and calculation of economic uncertainties: Intervals, fuzzy numbers, and probabilities // Int. J. Production Economics 78 (2002) 91–98
Department of Production Engineering, Copenhagen Engineering College, 15 Lautrupvang, DK-2750 Ballerup, Denmark

Гистограммная арифметика

$$\{+, -, \cdot, /, \uparrow\}$$

Гистограммные числа

Гистограмма P — кусочно-постоянная функция определяется сеткой $\{x_i | i = 0, ..., n\}$, на отрезке $[x_{i-1}, x_i]$ гистограмма принимает постоянное значение p_i.

Первые работы

Герасимов В.А., Добронец Б.С., Шустров М.Ю. Численные операции гистограммной арифметики и их применения // АиТ. 1991, №2. С. 83–88.

Berleant D. Automatically verified reasoning with both intervals and probability density functions // Interval Computations. 1993 No. 2, P. 48–70

Аналитические формулы

$$p_{x_1+x_2}(x) = \int_{-\infty}^{\infty} p(x-v, v)dv = \int_{-\infty}^{\infty} p(v, x-v)dv, \qquad (1)$$

$$p_{x_1/x_2} = \int_0^{\infty} vp(xv, v)dv - \int_{-\infty}^{0} vp(v, xv)dv. \qquad (2)$$

$$p_{x_1 x_2} = \int_0^{\infty} (1/v)p(x/v, v)dv - \int_{-\infty}^{0} (1/v)p(v, x/v)dv. \qquad (3)$$

Построение гистограммы суммы двух случайных величин

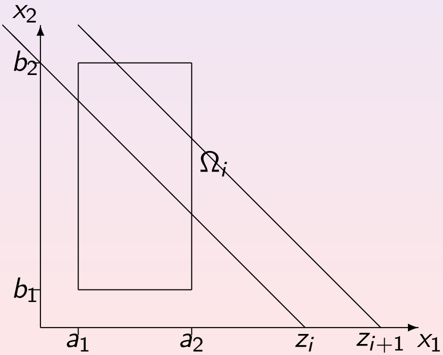
Пусть $\mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2$, и носители $\mathbf{x}_1$ — $[a_1, a_2]$, $\mathbf{x}_2$ — $[b_1, b_2]$, $p(x_1, x_2)$ — плотность распределения вероятностей случайного вектора $(\mathbf{x}_1, \mathbf{x}_2)$. Прямоугольник $[a_1, a_2] \times [b_1, b_2]$ — носитель плотности распределения вероятностей $p(x_1, x_2)$ и плотность вероятности $\mathbf{z}$ отлична от нуля на интервале $[a_1 + b_1, a_2 + b_2]$. Обозначим $z_i, \ i = 0, 1, \ldots, n$ — точки деления этого интервала на n отрезков.

Тогда вероятность попадания величины $\mathbf{z}$ в интервал $[z_i, z_{i+1}]$ определяется по формуле

$$P(z_i < \mathbf{z} < z_{i+1}) = (\int\int_{\Omega_i} p(x_1, x_2) dx_1 dx_2,$$

Построение гистограммы суммы двух случайных величин

$$p_{zi} = (\int\int_{\Omega_i} p(x_1, x_2) dx_1 dx_2 / (z_{i+1} - z_i),$$
где $\Omega_i = \{(x_1, x_2) | z_i \leq x_1 + x_2 \leq z_{i+1}\}$.



Сумма нескольких случайных величин

$$\mathbf{z} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + ... + a_n\mathbf{x}_n$$

$p(x_1, x_2, ..., x_n)$ — плотность распределения вероятностей случайного вектора $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$. Тогда вероятность попадания $\mathbf{z}$ в интервал (z_i, z_{i+1}) соответственно равна

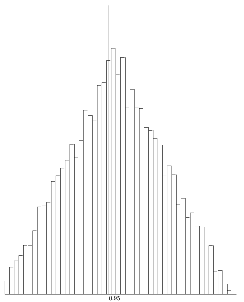
$$P(z_i < \mathbf{z} < z_{i+1}) = \int ... \int_{\Omega_i} p(x_1, x_2, ..., x_n) dx_1 dx_2 .... dx_n,$$

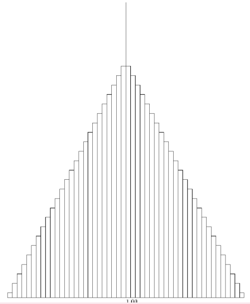
где $\Omega_i = \{(x_1, x_2, ..., x_n) | z_i < a_1x_1 + a_2x_2 + ... + a_nx_n < z_{i+1}\}$, p_{zi} имеет вид

$$p_{zi} = \int ... \int_{\Omega_i} p(x_1, x_2, ..., x_n) dx_1 dx_2 .... dx_n / (z_{i+1} - z_i).$$

Сравнение с Монте–Карло

Пример вычисления суммы двух равномерно распределенных случайных величин на интервале [0,1]





Метод Монте–Карло 10000 бросаний M=0.95
Гистограммная арифметика M=1.0

max & min

Рассмотрим операцию $\max(\mathbf{x}, \mathbf{y})$. Вероятность того, что $\max(\mathbf{x}, \mathbf{y}) < z$ — определяется через функцию распределения F

$$F(z) = \int_{-\infty}^{z} P_x(\xi) d\xi \int_{\infty}^{z} P_y(\xi) d\xi.$$

Используя функцию распределения F, можно построить гистограмму для $\max(\mathbf{x}, \mathbf{y})$ на сетке $\{z_i\}$. Тогда $p_i = F(z_{i+1}) - F(z_i)$.

Законы распределения функций случайных аргументов

Имеется непрерывная случайная величина **x** с плотностью распределения p_x. Другая случайная величина **y** связана с нею функциональной зависимостью:

$$\mathbf{y} = f(\mathbf{x}).$$

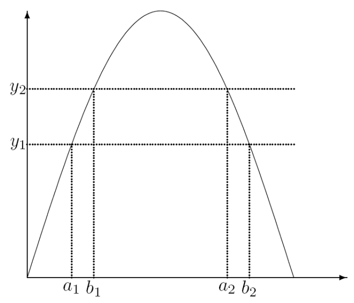
$$p_y(y) = f(f^{-1}(y))|(f^{-1})'(y)|,$$

Плотность распределения величины **y** в случае монотонной функции f определяется следующим образом: где f^{-1} — функция обратная к f. Однако, пользоваться таким представлением в ряде случаев довольно затруднительно.

Законы распределения функций случайных аргументов

Пусть $[a_j, b_j]$, $j = 1, ..., m$ такие, что $[y_{i-1}, y_i] = f([a_j, b_j])$.

$$p_{yi} = \sum_{j=1}^{m} \int_{a_j}^{b_j} p_x(\xi) d\xi.$$



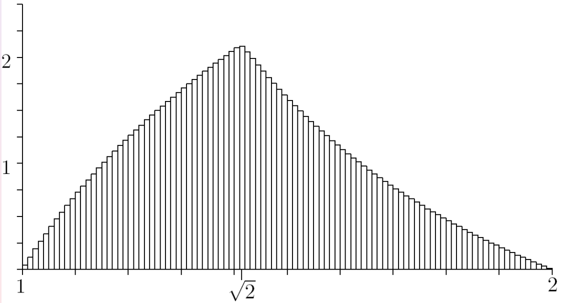
Решение нелинейных уравнений

$f(x,k) = 0$, где k — вектор случайных параметров. Пусть корень локализован на отрезке $[a,b]$. Необходимо найти плотность распределения $\mathbf{x}$. Будем предполагать, что можно с достаточной точностью для любого $z \in [a,b]$ вычислять плотность распределения ϕ_z случайной величины $f(z,\mathbf{k})$. Тогда, $P(z)$ есть вероятность, что корень лежит левее (правее) точки z:

$$P(z) = \int_{-\infty}^{0} \phi_z(\xi) d\xi.$$

Решение простейшего квадратного уравнения

$\mathbf{a}x^2 - \mathbf{b} = 0$, где **a**, **b** — случайные величины с равномерным законом распределения заданные соответственно на отрезках $[1, 2]$, $[2, 4]$.



Решение систем линейных алгебраических уравнений

Рассмотрим решение систем линейных алгебраических уравнений

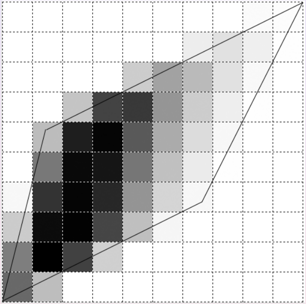
$$Ax = b.$$

Для простоты изложения предположим, что случайный вектор $\mathbf{b}$ состоит из независимых компонент $\mathbf{b}_1$, $\mathbf{b}_2$, каждая из которых, случайная величина равномерно распределенная на отрезке $[0, 1]$.
матрица матрица A имеет вид

$$A = \begin{pmatrix} a_{11} & -1 \\ -1 & 2 \end{pmatrix}.$$

Для простоты изложения предположим, что $\mathbf{a}_{11}$ — случайная величина равномерно распределенная на отрезке $[2, 4]$.

Решение систем линейных алгебраических уравнений



На рисунке приведено кусочно-постоянное с шагом 0.1 приближение совместной плотности вероятности вектора **x**. Сплошной линией приведена граница множества решений исходной системы.

Решение систем нелинейных уравнений

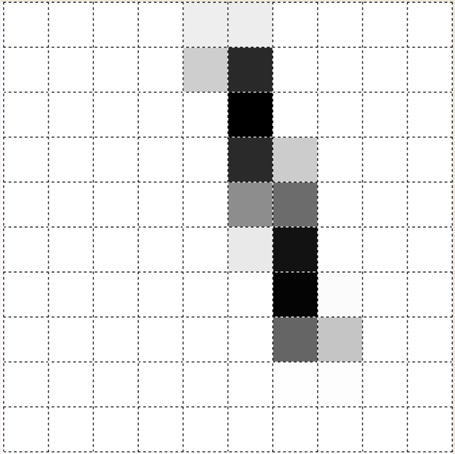
Рассмотрим пример системы нелинейных уравнений

$$\mathbf{a}x^2 + \mathbf{b}y^2 - 4 = 0,$$

$$xy - \mathbf{c} = 0,$$

где $\mathbf{a}, \mathbf{b}, \mathbf{c}$ — равномерные случайные величины, плотности вероятности которых имеют носители $[1, 1.1]$, $[2, 2.1]$, $[0.505, 0.51]$.

Решение систем нелинейных уравнений



Задачи интерполяции и экстраполяции

Рассмотрим случай линейной интерполяции. Пусть относительно некоторой функции f в точках x_1, x_2 значения f_1, f_2 представлены гистограммами F_1 и F_2.
Необходимо построить линейную гистограммную функцию $L(x)$ такую, что выполнены условия интерполяции $L(x_1) = F_1$ и $L(x_2) = F_2$.
Тогда линейная интерполяция L функции f в некоторой точке x может быть представлена в следующем виде

$$L(x) = F_1 \frac{x_2 - x}{x_2 - x_1} + F_2 \frac{x - x_1}{x_2 - x_1}.$$

Заметим, что условия интерполяции выполнены и $L(x)$ принимает соответствующие значения в узлах интерполяции.

Задачи интерполяции и экстраполяции

Далее, если необходимо восстановить оценку плотности вероятности f в точке x, то необходимо знать априорную информацию $\mathbf{f}''$ о плотностях вероятности f'' на отрезке $[x_1, x_2]$, тогда

$$\mathbf{f}(x) = L(x) + \frac{(x - x_1)(x - x_2)}{2}\mathbf{f}''(\xi).$$

Поскольку точка ξ неизвестна, то для оценки воспользуемся интервальной статистикой. Для этих целей построим интервальное расширение функции $\mathbf{f}''(\xi) \in \mathbf{F}'' = [\underline{f}'', \overline{f}'']$. Таким образом, следуя подходу интервальной статистики, для плотности вероятности $\mathbf{f}(x)$ получаем включение

$$\mathbf{f}(x) \in L(x) + \frac{(x - x_1)(x - x_2)}{2}[\underline{f}'', \overline{f}''].$$

Задачи интерполяции и экстраполяции

Таким образом, построение уже простейших интерполяционных формул приводит к необходимости расширить стандартные действия над гистограммными переменными элементами интервальных гистограмм или гистограмм второго порядка. О последних будет сказано более подробно ниже.

Гистограммы второго порядка

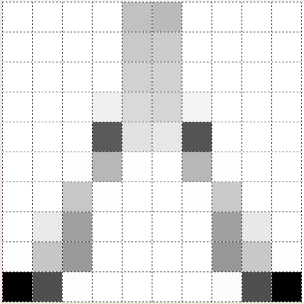
Гистограмма второго порядка — гистограмма, каждый столбец которой — гистограмма.

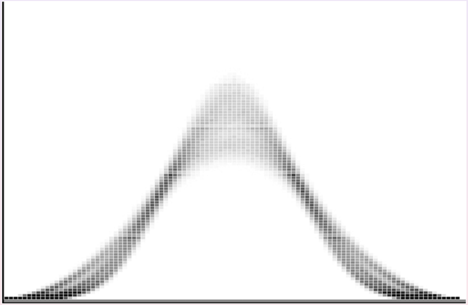
Пример гистограммы второго порядка, приближающая случайную сумму

$$\mathbf{S} = \frac{1}{\mathbf{n}} \sum_{i=1}^{\mathbf{n}} \mathbf{y}_i,$$

где $\mathbf{n}$ — дискретная случайная величина равномерно распределенная на отрезке $[2, 6]$, $\mathbf{y}_i$ — равномерно распределенные случайные величины с носителем $[0, 1]$.

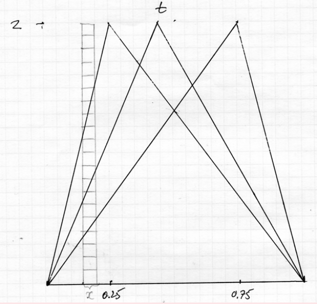
Гистограммы второго порядка



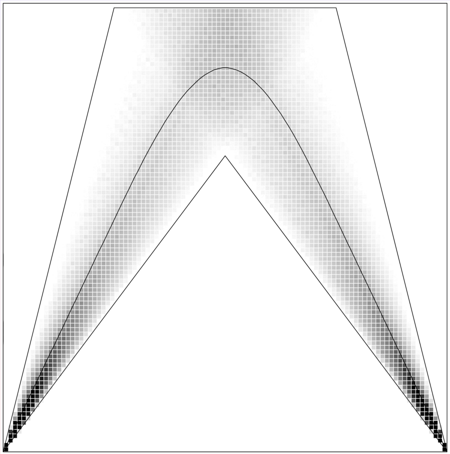


Гистограммы второго порядка. Пример

Пусть случайная величина S_t имеет треугольное распределение P_t на отрезке $[0, 1]$, высота $h = 2$ и вершина в некоторой точке $(t, 2)$, t — случайная величина с треугольным распределением на отрезке $[0.25, 0.75]$ с вершиной $(0.5, 4)$.



Гистограммы второго порядка

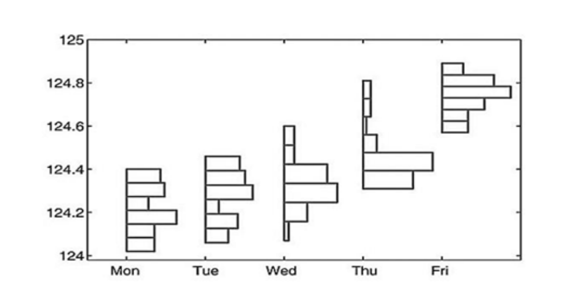


Гистограммы второго порядка

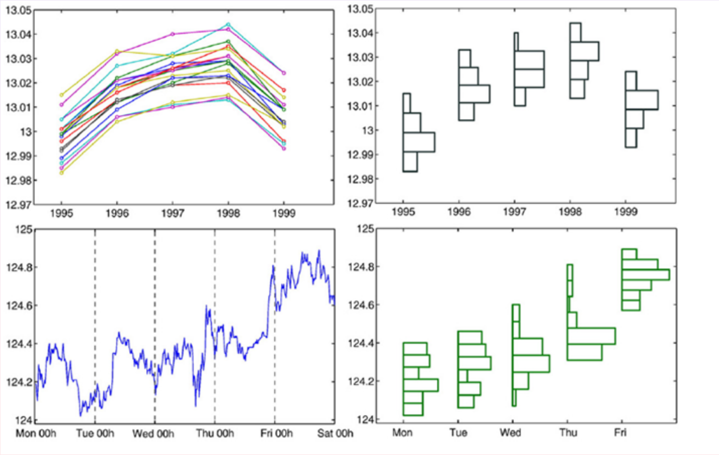
Интервальное распределение (максимальное и минимальное $P_t(x)$ для всех t) изображено на рис. граничными линиями. Внутренняя линия определяет *"эффективную" плотность вероятности* гистограммы второго порядка — математическое ожидание плотностей вероятности $P_t(x)$ в точке x.
endtabular

Гистограммный ряд

Гистограммный ряд:
последовательность (ряд) плотностей вероятности, представленных в виде гистограмм.



Гистограммный подход как метод исследования временных рядов



Временные ряды

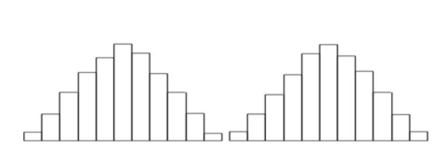
В самом общем подходе аддитивная модель временного ряда может быть представлена в виде

$$x(t_i) = f(t_i) + \eta(t_i),$$

где $f(t)$ — некоторая функция, η — случайная величина, $x(t_0), x(t_1), \dots, x(t_n)$ — случайные величины, определяющие ряд наблюдений, которые произведены в определенные моменты времени t_i.

Модельный пример

$f = 0.1t + 1$, случайная величина η имеет треугольное распределение, имеет нулевое математическое ожидание и ее носитель менялся по закону $[-0.2 + 0.001t, 0.2 - 0.001t]$, шаг дискретизации ряда 0.01.



Гистограмма η (слева) и гистограмма треугольного распределения полученная методом Монте–Карло при 10000 бросаний (справа)

$f(x) \sim 0.098x + 1.019$

Модель производство — продажи

Пусть завод производит некоторую продукцию и осуществляет ее реализацию. По наблюдениям известна гистограмма Y_i покупок i-го товара, например, за неделю. Это означает, что известна плотность вероятности продажи товара p_i, $Y_i \in [\underline{y}_i, \overline{y}_i]$.
Пусть доход D_i линейно зависит от продажи товара $D_i = \alpha Y_i$. Будем считать, что производство i-го товара не зависит от производства других товаров. Тогда общий доход будет просто сумма всех доходов по продажам всех товаров. Поэтому мы можем рассмотреть производство только одного вида продукции и для простоты индекс опустим.
Гипотеза. *Завод несет убытки если товар произведен в объеме R, но продано $Y < R$. Величина убытков U линейно зависит от разницы $\beta(Y - R) \leq 0$.*

Модель производство — продажи

Пусть завод произвел товар объемом R. Оценим возможный доход

$$D = \alpha \int_{\underline{y}}^{R} p(y) y dy,$$

и убыток

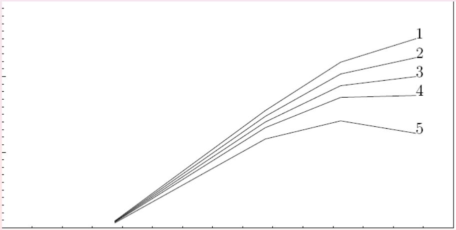
$$U = \beta \int_{\underline{y}}^{R} p(y)(y - R) dy.$$

Заметим, что в этой постановке нам надо максимизировать сумму $D+U\to\max$

$$f(R)=(D+U)(R)=\alpha\int_{\underline{y}}^{R}p(y)ydy+\beta\int_{\underline{y}}^{R}p(y)(y-R)dy=$$

$$=(\alpha+\beta)\int_{\underline{y}}^{R}p(y)ydy-\beta R\int_{\underline{y}}^{R}p(y)dy\to\max\,.$$

На рис. приведены графики зависимости производство — прибыль при различных значениях параметров α и β. В расчетах величина $\alpha = 1$, кривые 1–5 соответствуют значениям $\beta = 0.3$ $\beta = 0.4$, $\beta = 0.5$, $\beta = 0.7$, $\beta = 1$. Как видим при небольших значениях β увеличение производства покрывает возможные убытки перепроизводства, но при сравнительно больших значениях β величину производства R следует ограничивать исходя из максимального значения $D + U$.



Принятие решения об инвестировании

Компания рассматривает вопрос о приобретении для последующего производства патента нового лекарственного препарата. Стоимость патента составляет $3,4 млн. Обычно решение принимается на основе анализа дисконтированных денежных потоков NPV и IRR. Горизонт расчетов составляет три года.

Стандартная финансовая модель

	Год 0	Год 1	Год 2	Год 3
Цена упаковки		$6,00	$6,05	$6,10
Количество проданных, шт.		802000	967000	1132000
Выручка		$4 812 000	$5 850 350	$6 905 200
Себестоимость		$2 646 600	$3 217 693	$3 797 860
Валовая прибыль		$2 165 400	$2 632 658	$3 107 340
Операционные издержки		$324 810	$394 899	$466 101
Чистый доход до налогов		$1 840 590	$2 237 759	$2 641 239
Налоги		$588 989	$716 083	$845 196
Стартовые инвестиции	–$3 400 000			
Чистый доход	–$3 400 000	$1 251 601	$1 521 676	$1 796 043

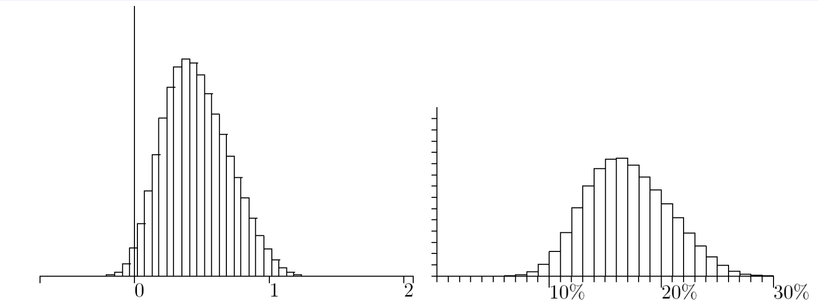
NPV & IRR

Внутренняя норма доходности определяет максимально приемлемую ставку дисконта, при которой можно инвестировать средства без каких-либо потерь для собственника: $IRR = r$, при котором $NPV(r) = 0$

$$NPV(r) = \sum_{t=0}^{n} \frac{CF_t}{(1+r)^t} - \sum_{t=0}^{n} \frac{I_t}{(1+r)^t} = 0.$$

CF_t — приток денежных средств в период t; I_t — сумма инвестиций (затраты) в t-ом периоде; n - суммарное число периодов (интервалов, шагов) t = 0, 1, ..., n.
NPV — чистая приведенная стоимость.

NPV & IRR



Из анализа гистограмм NPV и IRR видно, что вероятны как крайне негативные исходы, так и возможна значительная прибыль в сравнении со стандартным анализом.

Сравнение подходов интервальной математики и численных операций над плотностями вероятности случайных величин, показывает что практически значимые плотности вероятности занимают лишь небольшую часть полученных интервалов. Поэтому, в таких случаях, границы решений мало информативны. Часто при работе со случайными величинами ограничиваются вычислением лишь небольшого числа характеристик: математическое ожидание и дисперсия.

Гистограммный подход

- Относится к сфере символического анализа данных (BillardDiday, 2003

Гистограммный подход

- Относится к сфере символического анализа данных (BillardDiday, 2003
- Относится к новой, развивающейся парадигме статистики, которая «воздействует на наборы данных (Noirhomme, 2008

Гистограммный подход

- Относится к сфере символического анализа данных (BillardDiday, 2003
- Относится к новой, развивающейся парадигме статистики, которая «воздействует на наборы данных (Noirhomme, 2008
- Представляет инструмент ВИМ-технологии для поддержки принятия решений

Symbolic Data Analysis

В символическом анализе данных и Data Mining гистограммы используются для исследования множества различных процессов и применяются для описания изменчивости количественных признаков.

Billard, L., & Diday, E. (2006). Symbolic data analysis: conceptual statistics and data mining. Chichester: Wiley & Sons.

Почему гистограммы?

- Позволяют достаточно точно представлять произвольные распределения случайных величин

Почему гистограммы?

- Позволяют достаточно точно представлять произвольные распределения случайных величин
- Позволяют описать изменчивость количественных признаков

Почему гистограммы?

- Позволяют достаточно точно представлять произвольные распределения случайных величин
- Позволяют описать изменчивость количественных признаков
- Развитая арифметика для работы с гистограммными переменными.

Выводы:

Все формы неопределенности являются прекрасным средством коммуникации между финансовыми и другими специалистами

...

Hans Schjaer-Jacobsen
Representation and calculation of economic uncertainties: Intervals, fuzzy numbers, and probabilities // Int. J. Production Economics 78 (2002) 91–98

Публикации

В. А. Герасимов, Б. С. Добронец, М. Ю. Шустров. Численные операции гистограммной арифметики и их применения. АиТ. 1991, №2. С. 83–88.

Б. С. Добронец *Интервальная математика*. Красноярск: КГУ, 2004

Б. С. Добронец, О. А. Попова. Применение гистограммной математики в задачах принятия экономических решений. — *Труды IX международной ФАМЭТ 2010 конференции.*— Красноярск: КГТЭН, СФУ, 2010. С.127-130

Публикации

Б. С. Добронец, О. А. Попова. Гистограммая арифметика для визуально-интерактивного моделирования в задачах принятия экономических решений. — Актуальные проблемы анализа ипостроения информационных систем и процессов: *сб. статей Международной научно-технической конференции*. Таганрог: Издательство Технологичсекого института ЮФУ, 2010. С. 53–57

Б. С. Добронец, О. А. Попова. Гистограммный подход к согласованию интересов в условиях неопределенности в задачах принятия экономических решений // XIV Международная ЭМ конференция по эвентологической математике и смежным вопросам, Красноярск, 2010

Публикации

Б. С. Добронец, О. А. Попова. Численные операции над случайными величинами и их приложения // Journal of Siberian Federal University. Mathematics Physics 2011, 4(2), 229–239

Б. С. Добронец, О. А. Попова. Гистограммные временные ряды // Труды международной ФАМЭТ'2011 конференции, Красноярск, 2011.

Б. С. Добронец, О. А. Попова. Применение гистограммной математики в экономических задачах исследования операций. — VI Московская международная конференция по исследованию операций (ORM2010): Москва, 19–23 октября 2010 г.: Труды / Отв. ред. П.С. Краснощеков, А.А. Васин. — М.: МАКС Пресс, 2010. С. 90–92.

Спасибо за внимание!