**В.Ф.Писаренко**

# Стохастическое моделирование, бутстреп

# Генерирование случайных чисел с заданным распределением.

Функция распределения $F(x)$ случайной величины $\xi$ :

$$F(x) = Pr\{\xi \le x\}.$$

**Если $F(x)$ *непрерывна*, то случайная величина $F(\xi)$ имеет равномерное на отрезке [0, 1] распределение.**

$Rn = (r_1, \ldots, r_n)$ – вектор [0, 1]-случайных величин. Пусть существует такой случайный вектор $Zn$, что выполняется

$$F(Zn) = Rn . \qquad (*)$$

Применим к уравнению (*) функцию $Finv(.)$ обратную по отношению к функции $F(x)$: $Zn = Finv(Rn)$. Тогда все компоненты вектора $Zn$ имеют функцию распределения $F(x)$.

Пример: $F(x) = 1 - exp(-b \cdot x);\ x \ge 0.$ $\qquad Zn = -(1/b) \cdot log(1- Rn).$

Обратную функцию *Finv* в явном виде можно найти не для всех функций распределения.

Что делать в таких случаях? Ищем ***приближенную*** обратную функцию (с очень высокой степенью приближения). Определим интервал заведомо включающий весь возможный для реальных наблюдений диапазон [*m*, *M*]. Задаем на отрезке [*m, M*] очень мелкую, равномерную решетку

$$MN = m : 1/N : M,$$

где *N* – очень большое число (оно ограничено только возможностями компьютера). Вектор *MN* состоит из *N* компонент: $MN = (m_1, \ldots, m_N)$. Вычисляем прямую функцию распределения от аргументов $(m_1, \ldots, m_N)$:

$$z_j = F(m_j);\ j = 1,\ldots,N; \qquad Z_N = (z_1, \ldots, z_N).$$

Пусть задано множество ***n*** случайных точек $R_n$ из [0, 1]: $R_n = (r_1, \ldots, r_n)$.

На множестве $R_n$ получаем значения магнитуд $M_n$ с помощью интерполяции с решетки $(M_N, Z_N)$.

В пакете MATLAB эти операции можно компактно выполнить с помощью стандартной операции линейной интерполяции ***interp1:***

***$M_n$ = interp1($Z_N$, $M_N$, $R_n$, 'linear').***

Если шаг решетки *1/N* достаточен для детального описания функции *F(x)*, то можно считать, что компоненты вектора $M_n$ имеют функцию распределения *F(x)* с приемлемой точностью.

# Многомерные гауссовские случайные векторы с заданной ковариационной матрицей.

Плотность многомерного гауссовского вектора $(\xi_1, \dots \xi_n)$ имеет вид:

$$f(x_1,\dots x_n / A, B) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(B)}} \cdot \exp\{-\frac{1}{2}\Sigma\beta_{kj}(x_k - a_k)(x_j - a_j)\}; \qquad (1)$$

$E(\xi_k) = a_k$ ; $A = (a_1, \dots a_n)$; $B = \{ E(\xi_k - a_k)(\xi_j - a_j) \}$; $\{\beta_{kj}\} = B^{-1}$;

$E$ – символ математического ожидания; вектор $A$ – вектор средних значений; матрица B – матрица ковариаций. Они заданы.

Требуется генерировать вектор случайных гауссовских величин со средними значениями $A$ и матрицей ковариаций $B$. Обозначим через $G$ корень из матрицы $B$: $G = B^{1/2} = \{ \gamma_{kj} \}$.

Берем вектор стандартных независимых гауссовских значений ($\eta_1, \ldots \eta_n$ ); (их генерирует компьютер стандартной командой **randn** ) и получаем из него нужные нам гауссовские величины ($\xi_1, \ldots \xi_n$ ) с помощью линейных комбинаций:

$$\xi_k = a_k + \sum_{j=1}^{n} \gamma_{kj} (\eta_j - a_j); \quad k = 1, \ldots, n.$$

Вектор ($\xi_1, \ldots \xi_n$ ) имеет плотность вероятности (1).

**Статистическое тестирование критериев значимости, основанных на эмпирической функции распределения, при наличии неизвестных параметров.**

Пусть $(x_1 < x_2 \ldots < x_n)$ – упорядоченная выборка независимых наблюдений случайной величины с функцией распределения $F(x / \boldsymbol{\alpha})$, где $\boldsymbol{\alpha}$ - некоторый параметр (возможно, векторный). Нужно проверить гипотезу о том, что выборка соответствует функции $F(x / \boldsymbol{\alpha})$ при каком-то неизвестном значении параметра $\boldsymbol{\alpha}$. Оценим $\boldsymbol{\alpha}$ методом максимального правдоподобия, оценку обозначим $\hat{\alpha}$. Подставим эту оценку в $F(x / \hat{\alpha})$ и рассмотрим расстояние Колмогорова $KD$ между $F(x / \hat{\alpha})$ и выборочной функцией распределения $\hat{F}_n(x)$:

$$KD = n^{1/2} max| F(x / \hat{\alpha}) - \hat{F}_n(x)|.$$ ( MATLAB: **kstest**)

Если бы функция распределения была известна точно (были бы точно известны параметры), то распределение расстояния $KD$ можно было бы определить точно (оно называется распределением Колмогорова, есть таблицы). Далее нужно выбрать порог допустимых расстояний $H$, который будет превышаться лишь с малой вероятностью $p$ (скажем, $p = 0.05$), и отвергать нашу гипотезу, если наблюдается $KD > H$. Для случая с подгонкой неизвестного параметра расстояние Колмогорова будет меньше, чем без подгонки, и поэтому оно не будет соответствовать распределению Колмогорова. Какой выход их этого положения?

*Предлагается смоделировать процедуру оценки параметра, повторить её N раз (скажем 10000 раз) и получить большую выборку расстояний Колмогорова ($KD_1,\ldots,KD_N$).*

*Затем взять выборочный квантиль уровня (1-p) ,т.е. то значение из выборки ($KD_1,\ldots,KD_N$), которое превзойдут лишь $p \cdot N$ значений этой выборки. Его-то и нужно взять в качестве порога для расстояния Колмогорова в случае подгонки параметров. Оно будет несколько больше порога H, а вероятность его превзойти с большой точностью равна p.*

# Бутстреп-метод

Понятие бутстреп (bootstrap) введено в 1977 году Брэдли Эфроном (Bradly Efron). Суть метода состоит в том, чтобы по имеющейся выборке построить эмпирическое распределение. Используя это распределение как теоретическое распределение вероятностей, можно с помощью датчика псевдослучайных чисел сгенерировать практически неограниченное количество псевдовыборок произвольного размера. С помощью этих псевдовыборок можно оценить анализируемые статистические характеристики и изучить их вероятностные распределения. Таким способом, например, можно оценить дисперсию любой статистики независимо от её сложности.

Пусть $(x_1, \ldots, x_n)$ выборка из неизвестного распределения $F(x)$. Производим с помощью датчика случайных чисел $\boldsymbol{n}$-кратную выборку (**выбор с возвращением**) из величин $(x_1, \ldots, x_n)$; полученный вектор обозначим $\boldsymbol{y}^{(1)} = (y_1^{(1)}, \ldots, y_n^{(1)})$.

Генерируем $N$ таких векторов: $\boldsymbol{y}^{(1)} \ldots \boldsymbol{y}^{(N)}$.

Считая, что эти векторы достаточно представительно характеризуют распределение $F(x)$ (что справедливо лишь с некоторым приближением, которое улучшается при $n \to \infty$), можно статистически оценить любую интересующую нас характеристику распределения. При этом надо помнить, что любые такие оценки построены фактически на одной исходной выборке $(x_1, \ldots, x_n)$ и поэтому, строго говоря, являются **условными** (при условии реализации данной выборки). Они стремятся к безусловным оценкам при $n \to \infty$. Степень близости надо оценивать в каждом конкретном случае, имитируя реальную ситуацию, на примерах с известным ответом. По этой причине бутстреп-оценка дисперсии какой-то статистики, как правило, несколько ниже её истинной дисперсии.

# Бутстреп с непрерывной функцией распределения

Классический бутстреп предполагает использование дискретного (полиномиального) распределения с одинаковыми вероятностями *1/n* для каждого значения выборки $x_k$ . Это соответствует процедуре «выборка с возвращением». Но если мы знаем, что исследуемая случайная величина непрерывная, то естественно как-то выправить это несоответствие. Для этого можно дискретное полиномиальное распределение приблизить непрерывной функцией распределения. Способы приближения могут быть различны. Например, стандартная команда вычисления выборочных квантилей (процентилей) **prctile(X, q%)** выполняет такое приближение с помощью линейной интерполяции исходного полиномиального распределения в интервале [min(X); max(X)]. Пусть $R_n$ – вектор случайных величин из [0, 1]: $R_n = (r_1, \ldots, r_n)$. Тогда вектор $\boldsymbol{Z}$ = ***prctile(X,*** $R_n$ ***)*** имеет (**приближенно**) функцию распределения исходного вектора ***X*.** Можно генерировать любое количество независимых векторов ***Z*** и использовать их как независимые наблюдения для оценивания любых интересующих исследователя статистик с необходимой точностью.

Приведем альтернативный способ приближения исходной функции распределения – кернел-метод (kernel – ядро, функция типа колокола). Этот метод дает оценку $\hat{F}(z)$ в точке $z$ для функции распределения:

$$\hat{F}(z) = mean( normcdf( (X - z), s );$$

$s$ – параметр сглаживания; ***normcdf*** – (кумулятивная) Гауссовская функция распределения (можно брать и другие ядра) со средним значением $(X - z)$ и стандартным отклонением $s$ ; усреднение проводится по $n$ значениям выборки; параметр $s$ – подбирается так, чтобы сгладить выбросы от отдельных наблюдений и не «загладить» распределение (возможная рекомендация $s \approx 0.3 \cdot std(X)/n^{1/5}$

Имея функцию распределения, можно генерировать сколь угодно большие выборки и статистически оценивать любые статистики со сколь угодно большой точностью (условной, поскольку все это делается в рамках бутстрепа, на основе единственной исходной выборки).

## Аппроксимация выборочной функции равномерного распределения F(x) на [0; 2],

n = 4, показанного черными стрелками с помощью:

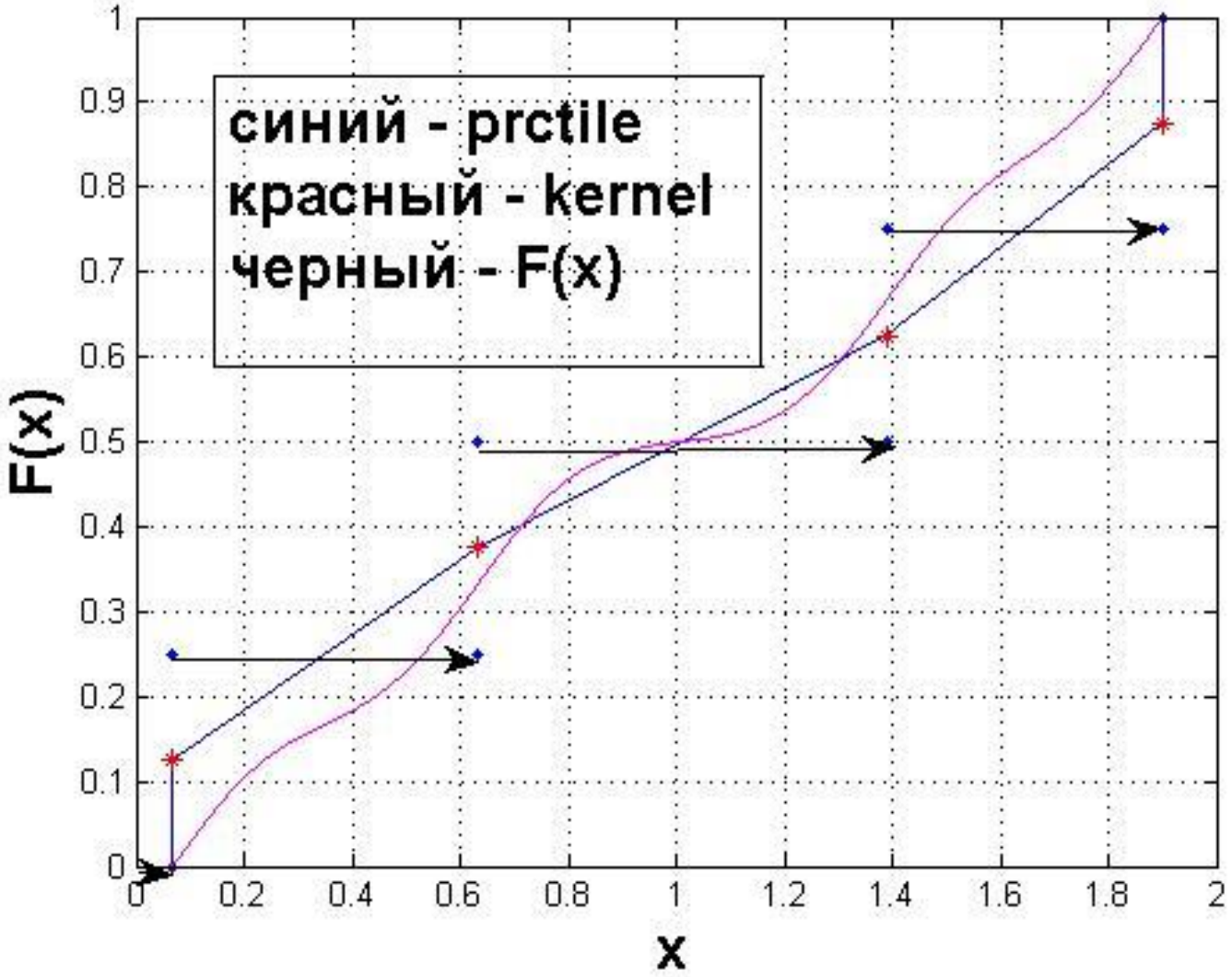
красный – кернел-метод;

синий – процентили;

черные стрелки – выборочная F(x).

Метод процентилей линейно аппроксимирует **середины скачков** выборочной F(x), они отмечены звездочками.

Кернел-метод дает более плавную аппроксимацию к истинной F(x) – диагонали прямоугольника.



# Использования статистического моделирования и бутстреп-метода для оценки параметров функции распределения.

Рассмотрим предложенную недавно в работах Писаренко и Родкина новую составную модель для закона повторяемости землетрясений. Закон повторяемости землетрясений задаётся моделью, состоящей их двух ветвей. Повторяемость в диапазоне средних и слабых землетрясений описывается линейным законом Гутенберга-Рихтера (Г-Р), а в диапазоне сильных - Обобщенным распределением Парето (Generalized Pareto Distribution, GPD).

Составная модель содержит три параметра, подлежащих оценке: наклон графика повторяемости в диапазоне средних и слабых магнитуд – $b$, параметр $h$, задающий точку гладкого сочленения двух ветвей модели и параметр формы GPD-распределения $\xi$, характеризующий быстроту убывания плотности при приближении к крайней предельной точке.

$$C_1\{1 - exp[-b\cdot(m - m_0)]\};\qquad m_0 \le m \le h;$$

$$F(m/b,h,\xi) =$$

$$C_3+C_2\{1-[1+\frac{b\xi}{1+\xi}(m-h)]^{-1/\xi}\};\quad h \le m \le M_{max} = h - \frac{1+\xi}{b\xi},\quad -1 < \xi < 0;$$

$$C_1 = 1/(1+\xi\, exp[-b\cdot(h - m_0)]);$$

$$C_2 = (1 + \xi)\, exp[-b\cdot(h - m_0)]\ /(1+\xi\, exp[-b\cdot(h - m_0)]);$$

$$C_3 = \{1 - exp[-b\cdot(h - m_0)]\}/(1+\xi\, exp[-b\cdot(h - m_0)]).$$

Здесь верхняя ветвь соответствует распределению Г-Р; нижняя ветвь – распределению GPD; $m_0$ – нижняя граница диапазона магнитуд; $b$ - наклон графика повторяемости ветви Г-Р; $h$ – параметр, задающий точку гладкого сочленения двух ветвей модели; $\xi$ - параметр формы GPD-распределения; $C_1$, $C_2$, $C_3$ – нормирующие константы, зависящие от параметров и обеспечивающие непрерывность плотности распределения и её первой производной в точке сочленения $m = h$.

# Оценка параметров составной модели

Дана выборка магнитуд $X = (m_1, \ldots, m_n)$. Нужно оценить $h$ , $\xi$ , $b$.

В качестве оценки параметра $h$ берется выборочный квантиль $X$ уровня $q_h\%$ :

$$\hat{h} = \boldsymbol{prctile(X, q_h).}$$

Уровень $q_h$ подбирается с учетом выборки X таким образом, чтобы число наблюдений выше $\hat{h}$ было не менее 20. Обычно $q_h = 50 \div 80\%$. Параметры $(\xi , b)$ оцениваются методом максимального правдоподобия с помощью стандартной программы поиска максимума функции от нескольких переменных (двух в данном случае).

# Подгонка Модели-2

Каталог SMT 1976-2016

Континентальные землетрясения Евразии:

$27 < \lambda < 47;\ 10 < \varphi < 95;\ h \le 70\ km;$

$\boldsymbol{m \ge 5.5;\ q_h = 80\%.}$

$\boldsymbol{n = 436;\ n_1\ (m<h)=349;\ n_2(m \ge h)=87.}$

**MaxLikEstimates:** $\hat{\xi} = -0.104;$

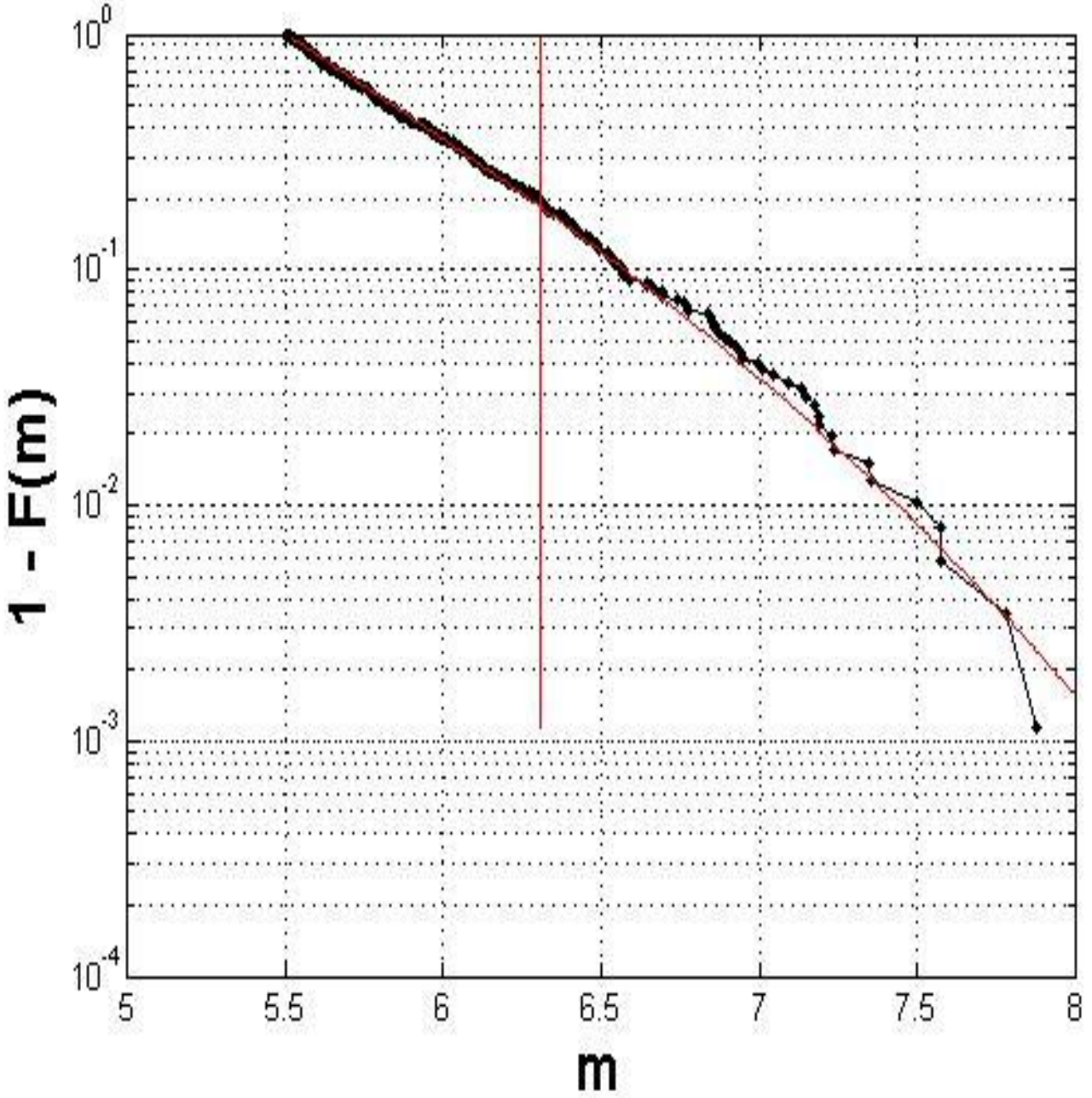
$\hat{b} = 0.863;$

$\hat{h} = prctile(X,80) = 6.31$

$KD = 0.803;\quad pvKD = 0.49.$

Вероятность $pvKD = P\{KD > 0.803\}$ оценивалась методом случайного моделирования, описанным выше.

Погрешность оценок $\hat{\xi}$ , $\hat{b}$ оценивалась двумя способами



# Погрешность оценок $\hat{\xi}$ , $\hat{b}$

Для оценки погрешности $\hat{\xi}$ , $\hat{b}$ нужно иметь достаточно много (псевдо)-выборок. Классический бутстреп-метод генерирует их с помощью выборки с возвращением. Альтернативный метод состоит в том, чтобы подставить в модельную функцию распределения $F$ выборочные оценки $\hat{\xi}$ , $\hat{b}$ и рассматривать $F(m/\hat{\xi},\hat{b})$ как теоретическую функцию распределения, которая, как мы видели ранее, позволяет генерировать сколь угодно большие выборки того же объема $n$ , что и исходная выборка. Конечно, $F(m/\hat{\xi},\hat{b})$ отличается от истинной функции распределения (поскольку в ней стоят не истинные значения параметров, а их оценки), но это будут уже *отличия второго порядка* и ими, как правило, можно пренебречь при оценке разброса $(\hat{\xi},\hat{b})$. При увеличении объема исходной выборки $n$ эти отличия исчезают.

Применяя эти два метода, мы получили следующие оценки погрешности:

$std(\hat{\xi}) = 0.0501;\ std(\hat{b}) = 0.134;$ бутстреп-метод;

$std(\hat{\xi}) = 0.0542;\ std(\hat{b}) = 0.130;$ метод генерирования выборок с помощью $F(m/\hat{\xi},\hat{b})$.

# Метод случайного моделирования

## Резюме.

Итак, метод случайного моделирования состоит из двух этапов:

1. Из исходной выборки объема $n\ (x_1,\ldots,x_n)$ получаем тем или иным способом большое число $N$ псевдо-выборок. Бутстреп делает это с помощью выборки с возвращением, а способ иммитации аппроксимирует функцию распределения и с её помощью генерирует псевдо-выборки.
2. Имея большое число псевдо-выборок можно по ним взять выборочное среднее значение любой статистики (теоретически вычислять её распределение иногда очень трудно). Псевдо-выборки зависимы, но среднее значение (даже от зависимых одинаково распределенных слагаемых) всегда равно среднему по ансамблю.

**СПАСИБО ЗА ВНИМАНИЕ!**