1 Гипотеза согласия

Рассмотрим простую гипотезу $H_0 : F = F_0$, где F_0 – заданное распределение. Альтернативу будем предполагать общей $H_1 : F \neq F_0$.

Критерий Колмогорова для простой гипотезы $H_0 : F = F_0$ с общей альтернативой имеет статистику

$$T_K = \sup_x |\widehat{F}_n(x) - F_0(x)|,$$

где $\widehat{F}_n(x)$ – ЭФР. Можно представить статистику в форме

$$T_K = \max\left(\frac{i}{n} - F_0(x_{(i)}), F_0(x_{(i)} - 0) - \frac{i-1}{n}\right).$$

При непрерывном распределении F_0 и верной гипотезе T_K имеет некоторое фиксированное распределение, не зависящее от F_0. При $n \to \infty$ величина $\sqrt{n}T_K$ сходится к некоторому распределению, называемому распределением Колмогорова. В Python его квантили есть в kolmogi.

Критерий Крамера-фон Мизеса для простой гипотезы $H_0 : F = F_0$ с общей альтернативой имеет статистику

$$T_{CvM} = n\int_{\mathbb{R}} (\widehat{F}_n(x) - F_0(x))^2 dF_0(x),$$

где $\widehat{F}_n(x)$ – ЭФР. Это выражение неудобно и используют явную формулу

$$T_{CvM} = \sum_{i=1}^{n}\left(\frac{2i-1}{2n} - F_0(X_{(i)})\right)^2 + \frac{1}{12n}.$$

Критерий Андерсона-Дарлинга имеет статистику

$$T_{AD} = n\int_{\mathbb{R}} \frac{(\widehat{F}_n(x) - F_0(x))^2}{F_0(x)(1-F_0(x))} dF_0(x),$$

где $\widehat{F}_n(x)$ – ЭФР. Это выражение неудобно и используют явную формулу

$$T_{AD} = -n - \sum_{i=1}^{n}\frac{2i-1}{n}(\ln F_0(X_{(i)}) + \ln(1 - F_0(X_{(n+1-i)}))).$$

В случае непрерывного распределения F_0 обе статистики имеют некоторые распределения, не зависящия от F_0, причем при $n \to \infty$ верны соотношения

$$T_{CvM} \xrightarrow{d} U, \quad T_{AD} \xrightarrow{d} V,$$

где

$$U = \sum_{j=1}^{\infty}\frac{Z_j^2}{j^2\pi^2}, \quad V = \sum_{j=1}^{\infty}\frac{Z_j^2}{j(j+1)},$$

где Z_j н.о.р. величины со стандартным нормальным распределением.

Еще один знакомый вам критерий – хи-квадрат, использующий статистику

$$T_\chi = \sum_{i=1}^{k}\frac{(\nu_i - nP_0(\Delta_i))^2}{nP_0(\Delta_i)},$$

где P_0 – мера, соответствующая F_0, ν_i – число наблюдений, попавших в Δ_i, Δ_i – разбиение прямой. Здесь ключевой вопрос в выборе Δ_i и k. Обычно предлагают $k = [log_2 n]$ или $k = [n^{1/5}]$. Что касается Δ_i, то их

стараются выбрать так, чтобы $P_0(\Delta_i)$ были близки, например, равными.

При верной гипотезе статистика сходится к распределению χ^2_{k-1}, откуда получаем критерий.

Фамилии, начинающиеся с буквы до К включительно решают пункт а), а после К – пункт б).

1. Реализовать критерии а) Крамера-фон Мизеса и б) Андерсона-Дарлинга, определяя p-value с помощью метода Монте-Карло. Построить график ЭФР p-value при верной гипотезе и проверить, что критерий работает верно. Учтите, что anderson из scipy.stats – это другое!

2. Построить графики p-value критериев Колмогорова, хи-квадрат, а) Крамера-фон Мизеса, б) Андерсона-Дарлинга для проверки гипотезы $H_0 : X_i \sim \mathcal{N}(0, 1)$ для а) $X_i \sim p\mathcal{N}(0, 1) + (1 - p)\mathcal{N}(0, 3)$ (под суммой имеется ввиду смесь) б) $X_i \sim 0.5\mathcal{N}(\mu, 1)+0.5\mathcal{N}(-\mu, 1)$, $p = 0.9$, $\mu = 0.1$. Подобрать n так, чтобы все критерии были чувствительны к гипотезе (то есть график p-value существенно отличался от биссектрисы, но не становился вертикальным). Какой критерий лучше справляется с задачей?

3. * Для предыдущей задачи построить график мощности всех четырех критериев уровня 95% как функции от параметра p или μ соответственно.

2 Проверка принадлежности параметрическому семейству

Мы будем рассматривать сложную гипотезу

$$H_0 : F \in \{F_\theta, \theta \in \Theta\}$$

с общей альтернативой, где $\{F_\theta\}$ – некоторое семейство распределений (например, нормальные распределения с неизвестными параметрами).

Сперва предложим общие подходы к такой задаче, а потом приведем частные случаи для двух популярных семейств.

2.1 Сложные критерии Крамера-фон Мизеса, Андерсона-Дарлинга и Колмогорова

Все три рассмотренных в прошлый критерии естественно модифицировать для проверки нашей гипотезы, заменив F_0 (которого мы теперь точно не знаем) на $F_{\widehat{\theta}}$, где $\widehat{\theta}$ – ОМП для θ в нашем семействе:

$$T_{KS} = \sup \left| \widehat{F}_n(x) - F_{\widehat{\theta}}(x) \right|,$$
$$T_{CvM} = \int_{\mathbb{R}} (\widehat{F}_n(x) - F_{\widehat{\theta}}(x))^2 dF_{\widehat{\theta}}(x),$$
$$T_{AD} = \int_{\mathbb{R}} \frac{(\widehat{F}_n(x) - F_{\widehat{\theta}}(x))^2}{F_{\widehat{\theta}}(x)(1 - F_{\widehat{\theta}}(x))} dF_{\widehat{\theta}}(x),$$

Все три статистики с теми же нормировками, что и раньше, в сильно регулярных абсолютно-непрерывных параметрических семействах будут сходиться к некоторому распределению. Однако теперь уже не к тем распределениям, что раньше, да и вообще к разным распределениям для разных семейств F_θ, но что хуже всего – может быть даже к разным распределениям при разных θ в рамках одного параметрического семейства. Это практически лишает нас возможности использовать эти подходы.

Впрочем, в семействах сдвига-масштаба все значительно лучше – точные распределения статистик не зависят от θ (но зависят от самого семейства), существуют предельные распределения для статистик с теми же нормировками (в случае сильно регулярных семейств), но предельные распределения станут другими и вновь будут зависеть от семейства.

Напомним, что семейства сдвига-масштаба имеют вид

$$F_\theta(x) = F_0\left(\frac{x - \theta_1}{\theta_2}\right),$$

где параметры необязательно независимы (а могут быть функциями одного параметра), а F_0 – известная ф.р.

Таким образом, мы можем строить критерии такого типа для отдельных семейств сдвига-масштаба, а для определения фактического уровня значимости использовать метод Монте-Карло, генерируя вспомогательные выборки из любого представителя нашего семейства.

2.2 Сложные критерии отношения правдоподобий и хи-квадрат

При использовании критерия хи-квадрат или к.о.п. для дискретных данных мы можем безболезненно подставлять ОМП в статистику критерия, используя теорему Уилкса или теорему о распределении статистики критерия множителей Лагранжа:

$$\sum_{i=1}^{k} \frac{(\nu_i - np_i(\widehat{\theta}))^2}{np_i(\widehat{\theta})^2} > y_{1-\alpha}, \quad \sum_{i=1}^{n} \nu_i \ln \frac{\nu_i}{np_i(\widehat{\theta})} > y_{1-\alpha},$$

где y – квантиль χ^2_{k-l}, где мы предполагаем, что поверхность $(p_1(\theta), \ldots, p_k(\theta))$ образует гладкое многообразие размерности l.

ОМП при этом определяется максимизацией правдоподобия

$$\prod_{i=1}^{k} p_i(\theta)^{\nu_i}. \tag{1}$$

Чтобы использовать этот критерий в непрерывном случае (впрочем, для дискретных с большим числом значений придется делать то же самое), мы дискретизируем данные и применяем к ним предыдущий критерий. Главная проблема в том, что ОМП $\widehat{\theta}$ необходимо рассчитывать максимизацией дискретного (!) правдоподобия (1). Таким образом, например, для нормального распределения мы вынуждены оценивать неизвестные параметры μ, σ, максимизируя

$$\prod_{i=0}^{k-1} \left(\Phi\left(\frac{a_{i+1}-\mu}{\sigma}\right) - \Phi\left(\frac{a_i-\mu}{\sigma}\right)\right)^{\nu_i},$$

где $a_0 = -\infty$, $a_k = +\infty$, a_i – точки разбиения прямой. Эту максимизацию сложно производить аналитически.

Второй проблемой является выбор деления на отрезки. Раньше мы могли выбирать отрезки так, что теоретическое распределение всех отрезков было одинаковым. Теперь мы вынуждены выбирать их глядя на данные, а ведь этот может нарушить работу критериев. Впрочем, существуют работы, в которых показано, что, например, деление выборки на равные части не изменяет предельного распределения статистики.

2.3 Три критерия проверки нормальности

Для часто возникающей проверки нормальности три, на наш взгляд, наиболее популярных критерия это критерий Андерсона-Дарлинга (с оговоренными выше поправками), критерий Шапиро-Уилка и К-критерий Д'Агостино.

1. Критерий Шапиро-Уилка реализован в scipy.stats.shapiro. Это один из лучших тестов для данной задачи с точки зрения многих исследователей. Он базируется на отношении

   $$T_{SW} = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

   где a_i – достаточно сложные коэффициенты:

   $$\vec{a} = \frac{m^T V^{-1}}{||m^T V^{-2} m||^{1/2}},$$

   где $(m_1, \ldots, m_n)$ – средние порядковых статистик из $\mathcal{N}(0,1)$ выборки, V – их ковариационная матрица. Распределение этой статистики ищется методом Монте-Карло.

2. Критерий Д'Агостино реализован в scipy.stats.normaltest. Идея этого и ряда других критериев (в том числе других версий критерий Д'Агостино) отличия нормального распределения от других ищутся на основе коэффициентов асимметрии

   $$\widehat{\mu}_3 = \frac{\overline{(X - \overline{X})^3}}{S^3}$$

   и эксцесса

   $$\widehat{\mu}_4 = \frac{\overline{(X - \overline{X})^4}}{S^4} - 3,$$

   где S – выборочная дисперсия. Критерий Д'Агостино преобразует эту пару статистик в некоторую

достаточно хитрую функцию вида

$$T_{DA} = C_1((C_2\widehat{\mu}_3))^2 + C_3\left(1 - \frac{1}{(C_4 + C_5\widehat{\mu}_4)^{1/3}}\right)^2,$$

которая имеет предельное χ^2_2 распределение. Здесь C_i – некоторые константы. Эти хитрые преобразования каждой из величин преобразованиями приближают их распределения к нормальному.

2.4 Четыре критерия проверки экспоненциальности

Для проверки другой частой гипотезы – гипотезы экспоненциальности, мы назовем четыре критерия. Первые два из них зачастую называются наиболее эффективными, а два других мы выбрали на основе обзора по ссылке). Итак, мы рассмотрим критерий Андерсона-Дарлинга (со своей поправкой), критерий Шапиро-Уилка для экспоненциальности, критерий Жанга и критерий Фрозини.

1. Критерий Андерсона Дарлинга использует привычную нам статистику, подставляя в качестве эталонного распределение экспоненциальное с оцененным параметром $1/\overline{X}$. При этом распределение статистики будет специфическим именно для экспоненциального распределения и квантили придется рассчитывать отдельно.

2. Критерий Шапиро-Уилка базируется на статистике

$$T_{SW,exp} = \frac{n(\overline{X} - X_{(1)})^2}{(n-1)\sum_{j=1}^n (X_j - \overline{X})^2}.$$

3. А-критерий Жанга имеет статистику

$$-\sum_{j=1}^n \left(\frac{\ln Z_{(j)}}{n - j + 1/2} + \frac{\ln(1 - Z_{(j)})}{j - 1/2}\right),$$

где $Z_j = F_{\widehat{\theta}}(X_j)$.

4. Критерий Фрозини использует статистику

$$T_{Fr} = \frac{1}{\sqrt{n}}\sum_{j=1}^n \left|1 - \exp\left(\frac{X_{(j)}}{\overline{X}}\right) - \frac{j - 0.5}{n}\right|.$$

Увы, ни одного из них в Python не реализовано.

2.5 Задачи

Пункт а, б или в определяется остатком по модулю 3 от длины фамилии (0, 1 или 2 соответственно), а i) или ii) – четностью длины имени (i четно, а ii – нечетно).

1. Реализовать сложные критерии а) Колмогорова, б) Крамера-фон Мизеса, в) Андерсона-Дарлинга для проверки принадлежности i) нормальному ii) экспоненциальному закону, используя метод Монте-Карло для вычисления предельного распределения p-value. Построить ЭФР p-value при гипотезе и при альтернативе, моделируя данные из распределения Стьюдента t_5 в первом случае и χ^2_2 во втором.

2. Реализовать критерий хи-квадрат для проверки i) нормальности, ii) экспоненциальности (ОМП по сгруппированным данным находить численно). Для разбиения на промежутки использовать следующие подходы: 1) делим так, чтобы в ячейки попадало поровну наблюдений, 2) берем квантили уровней i/k теоретического распределения, в качестве неизвестного параметра используем ОМП. Для подсчета p-value использовать предельное распределение.

3. Исследовать сгенерированные модельные данные X на нормальность, где

   - X_i имеют распределение Вейбулла с параметром формы λ;
   - X_i имеют χ^2_n распределение, $n = 5, 10, 25$
   - $X_i \sim Binom(n, 1/2)$.

   i) Использовать критерий из первого пункта, ii) использовать встроенный критерий Андерсона-Дарлинга (scipy.stats.anderson), а также критерии Шапиро-Уилко (scipy.stats.shapiro) и Д'Агостино (scipy.stats.normal Построить ЭФР p-value и график мощности. Подобрать размер выборки так, чтобы сравнение было осмысленным.

4. * Исследовать сгенерированные модельные данные X на экспоциальность, где

   (a) X_i имеют распределение $|\mathcal{N}(\mu, 1)|$;

   (b) X_i имеют χ^2_n распределение;

   (c) $X_i \sim Geom(p)$.

   Построить ЭФР p-value при фиксированных параметрах и график мощности в зависимости от параметра. Подобрать размер выборки и параметр так, чтобы сравнение было осмысленным. Использовать i) встроенный критерий Андерсона-Дарлинга ii) критерий из первого пункта, а также критерии а) Шапиро-Уилка, б) Жанга и в) Фрозини.

3 Проверка однородности

3.1 Теория

Задача проверки однородности двух выборок состоит в проверке гипотезе $H_0 : F = G$, где выборка $(X_1, \dots, X_n)$ имеет структуру $(Y_1, \dots, Y_{n_1}, Z_1, \dots, Z_{n_2})$, где $Y_i \sim F$, $Z_i \sim G$. Мы не касаемся так называемых парных повторных наблюдений, где Y_i и Z_i могут быть зависимыми.

Рассмотрим несколько подходов

1. Критерий хи-квадрат (для дискретных данных реализован в scipy). Данные дискретизируются, подсчитываются количества $\nu_{i,j}$ попадания i-й выборки в j-й интервал, вводится статистика

$$\sum_{i=1}^{2}\sum_{j=1}^{k}\frac{\left(\nu_{i,j}-\frac{n_i\nu_{\cdot,j}}{n}\right)^2}{\frac{n_i\nu_{\cdot,j}}{n}} > y_{1-\alpha}, \quad \nu_{\cdot,j} = \nu_{1,j} + \nu_{2,j},$$

   где y – квантиль χ^2_{k-1}.

2. Критерий Смирнова (реализован в scipy) имеет вид

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\sup|\widehat{F}_{n_1} - \widehat{G}_{n_2}| > k_{1-\alpha},$$

   где k – квантиль распределения Колмогорова.

3. Критерий Стивенса-Шольца (он же k-выборочный критерий Андерсона-Дарлинга) реализован в scipy) имеет вид

$$\frac{n_1 n_2}{n_1+n_2}\int\frac{(\widehat{F}_{n_1} - \widehat{G}_{n_2})^2}{\widehat{H}_{n_1,n_2}(x)(1-\widehat{H}_{n_1,n_2}(x))} > A_{1-\alpha},$$

   где A – квантиль распределения Андерсона-Дарлинга, $\widehat{H}$ – ЭФР объединенной выборки.

4. Критерий Баумгартнера-Вейсса-Шиндлера предлагает рассматривать статистику

$$\frac{1}{2n_1}\sum_{i=1}^{n_1}\frac{\left(R_i - \frac{(n_1+n_2)i}{n_1}\right)^2}{\frac{i}{n+1}\left(1-\frac{i}{n_1+1}\right)\frac{n_2(n_1+n_2)}{n_1}} + \frac{1}{2n_2}\sum_{i=1}^{n_2}\frac{\left(S_i - \frac{(n_1+n_2)i}{n_2}\right)^2}{\frac{i}{n_2+1}\left(1-\frac{i}{n_2+1}\right)\frac{n_1(n_1+n_2)}{n_2}},$$

   где R_i – ранги (упорядоченные по возрастанию) первой выборки, S_i – ранги (упорядоченные по возрастанию) второй выборки в общем вариационном ряду. Далее предлагают сравнивать ее с квантилями распределения Андерсона-Дарлинга, которые, опять же, можно определять методом Монте-Карло.

5. t-критерий основан на статистике

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\,\frac{\overline{Y}-\overline{Z}}{\sqrt{\frac{n_1 S_Y^2 + n_2 S_Z^2}{n_1+n_2-2}}}$$

   При верной гипотезе она сходится к $\mathcal{N}(0,1)$ распределению. Эту статистику используют для критерия однородности с альтернативой доминирования $F \le G$, то есть $F(x) \le G(x)$ при всех x, причем $F(x_0) < G(x_0)$ для некоторого x_0. В Python он есть здесь.

6. Критерий Манна-Уитни-Уилкоксона базируется на статистике

$$\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(I_{Y_i \ge Z_j} - \frac{1}{2}\right).$$

При выполнении гипотезы статистика при нормировке $\sqrt{(n_1+n_2)n_1n_2/12}$ сходится к величине с $\mathcal{N}(0,1)$ распределением, $n_1, n_2 \to \infty$. Отсюда получается соответствующий критерий для проверки гипотезы однородности с альтернативой доминирования.

Опишем ряд подходов, пригодных для m выборок.

1. Критерий хи-квадрат для m выборок имеет вид

$$\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(\nu_{i,j}-\frac{n_i\nu_{\cdot,j}}{n}\right)^2}{\frac{n_i\nu_{\cdot,j}}{n}} > y_{1-\alpha}, \quad \nu_{\cdot,j}=\nu_{1,j}+\nu_{2,j},$$

   где y – квантиль χ^2_{k-1}.

2. Критерий Стивенса-Шольца:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{n_i}{n}\int\frac{(\widehat{F}_{n_i}-\widehat{H})^2}{\widehat{H}_n(x)(1-\widehat{H}_n(x))}d\widehat{H}_n > A_{1-\alpha},$$

   где $\widehat{H}$ – ЭФР объединенной выборки. Критерий реализован в Python здесь.

3. Критерий Краскелла-Уоллиса (реализован в Python здесь):

$$\sum_{i=1}^{m} n_i\left(\overline{R}_i-\frac{n+1}{2}\right)^2,$$

   где $\overline{R}_i$ – среднее арифметическое рангов i-й выборки. При выполнении гипотезы статистика сходится к величине с распределением хи-квадрат с $m-1$ степенями. Отсюда получается соответствующий критерий для проверки гипотезы однородности с альтернативой, что хоть для одной пары выборок выполнена альтернатива доминирования.

3.2 Задачи

1. Рассмотрим t-критерий и критерий Манна-Уитни: применим их для сравнения однородности двух выборок из распределения из а) $\mathcal{N}(0,1)$ распределения и $\mathcal{N}(\mu,1)$ распределения, б) распределения $Laplace(0,1)$ и $Laplace(\mu,2)$ распределения, взяв размеры выборок равные а) 10, б) 50, в) 100. Используйте разные виды критериев, меняя настройки: для t-критерия equal_var и permutations, для Манна-Уитни – exact и asymptotic method. Построить график мощности всех версий критериев в зависимости от μ.

2. Реализовать критерий Баумгартнера-Вейсса-Шиндлера в перестановочной версии. Проверить его работу на а) $R[0,1]$ и $R[0,1]$ выборках б) $R[0,1]$ и $R[0.1,1.1]$ выборках. Сравнить с встроенным критерием Стивенса-Шольца (scipy.stats.anderson_ksamp).

3. Сравнить (построив ЭФР p-value) критерии Манна - Уитни, критерий Смирнова (scipy.stats.ks_2samp), критерий Стивенса-Шольца (scipy.stats.anderson_ksamp) и BWS на примере следующих модельных данных:

   (a) $X_i, Y_j \sim N(0,1)$,
   (b) $X_i \sim N(0,1), Y_j \sim N(0.3,1)$,
   (c) $X_i \sim N(0,1), Y_j \sim N(0,3/2)$,
   (d) $X_i \sim N(0,1), Y_j \sim t_k$, где t_k – распределение Стьюдента с k степенями свободы,
   (e) $X_i \sim N(0,1), Y_j$ – центрированная нормированная сумма k независимых с.в. из равномерного распределения $R[-1,1]$.

Размер выборки в каждом случае выбирать так, чтобы он был поменьше среди тех, когда часть критериев замечает разницу.

4. * Проверить на однородность k выборок, используя критерии Стивенса-Шольца, хи-квадрат и Краскелла-Уоллиса:

   (a) $X_{i,j} \sim N(0, 2+i)$, $i \leq 4$, $j \leq n$;

   (b) $X_{i,j} \sim N(i/4, 1)$, $i \leq 4$, $j \leq n$;

   (c) $X_{i,j} \sim t_{3+i}$, $i \leq 4$, $j \leq n$, где t – распределение Стьюдента.

4 Проверка независимости

4.1 Теория

Гипотеза независимости проверяется для двумерных выборок $(X_i, Y_i) \sim H$ и имеет вид

$$\mathrm{H}_0 : H(x, y) = F(x)G(y),$$

где $F(x)$, $G(y)$ – маргинальные распределения. Начнем с общей альтернативы $\mathrm{H}_1 : H(x, y) \neq F(x)G(y)$ для некоторых x, y.

4.1.1 Общая альтернатива

Прежде всего, заметим, что многие из наших прежних подходов остаются действенными.

1. Критерий хи-квадрат предлагает рассматривать статистику

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{1}{\frac{\nu_{i,\cdot}\nu_{\cdot,j}}{n}} \left(\nu_{i,j} - \frac{\nu_{i,\cdot}\nu_{\cdot,j}}{n}\right)^2, \quad \nu_{\cdot,j} = \sum_{i=1}^{m} \nu_{i,j}, \quad \nu_{i,\cdot} = \sum_{j=1}^{k} \nu_{\cdot,j}.$$

   Эту величину предлагается сравнивать с квантилью $\chi^2_{(k-1)(m-1)}$ распределения. Как мы видим, этот критерий тот же самый, что и критерий однородности для той же таблицы сопряженности. Аналогично дело обстоит с критерием отношения правдоподобия.

   Как обычно, мы дискретизируем значения переменных X, Y если они не дискретны, считаем количество попаданий в соответствующие ячейки по паре переменных и применяем к ним критерий хи-квадрат.

   Реализация используется все та же, что и ранее.

2. Критерий Смирнова можно адаптировать для гипотезы независимости со статистикой в форме

$$D_n = \sqrt{n} \sup_{x,y} |\widehat{H}_n(x, y) - \widehat{F}_n(x)\widehat{G}_n(y)|.$$

   При верной гипотезе $\mathrm{H}_0 : H(x, y) = F(x)G(y)$ данная статистика имеет некоторое распределение, которое не зависит от F и G. Таким образом, уровень значимости критерия можно определять методом Монте-Карло.

   В Python его, видимо, нет, реализуйте его методом Монте-Карло.

3. Более эффективным оказывается подход Секея и Риццо, предложенный уже в 21 веке. Их статистику можно представить в виде

$$D_n = n \int_{\mathbb{R}^2} |\widehat{\psi}_H(s, t) - \widehat{\psi}_F(s)\widehat{\psi}_G(t)|^2 \omega(s, t) ds dt,$$

   где $\widehat{\psi}_H$ – выборочная х.ф. вектора (X_i, Y_i), $\widehat{\psi}_F$, $\widehat{\psi}_G$ – выборочные х.ф. отдельных выборок, ω – некоторая весовая функция. Увы, распределение статистики зависит от H даже при верной гипотезе и определяется перестановочным методом. Данный критерий есть здесь.

4.1.2 Частная альтернатива

Рассмотрим также некоторые критерии, которые используют для более узкого спектра альтернатив. Описать конкретную альтернативу здесь не так просто, поэтому скажем условно – альтернатива "Если X большой, то и Y в среднем тоже".

1. Критерий Пирсона предлагает смотреть на коэффициент корреляции

$$\rho_P(X,Y) = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{S_X S_Y}.$$

При гипотезе независимости $\sqrt{n}\rho_P$ стремится к величине $Z \sim \mathcal{N}(0,1)$. Реализация есть здесь.

2. Критерий Спирмена предлагает считать тот же коэффициент для рангов (R_i, S_i), где R_i – ранг X_i среди $X_1, \ldots, X_n$, T_i – ранг Y_i среди $Y_1, \ldots, Y_n$:

$$\rho_S(X,Y) = \frac{\overline{RT} - \overline{R}\,\overline{T}}{n^{-1}\sqrt{\sum_{i=1}^n (R_i - \overline{R})^2 \sum_{i=1}^n (T_i - \overline{T})^2}}.$$

При гипотезе независимости $\sqrt{n-1}\rho_S$ стремится к величине $Z \sim \mathcal{N}(0,1)$. Критерий реализован здесь.

3. Коэффициент Кендалла предлагает рассматривать среди всех пар (X_i, Y_i), (X_j, Y_j) пары пар, для которых $X_i \le X_j$, $Y_i \le Y_j$ или $X_i > X_j, Y_i > Y_j$. Такие пары назовем *согласованными*. Пусть число согласованных пар N, а несогласованных – $M = C_n^2 - N$. Тогда

$$\rho_K = \frac{M - N}{M + N}.$$

При этом ρ_K при гипотезе имеет распределение, не зависящее от F, G, причем

$$\sqrt{\frac{9n(n-1)}{2(2n+5)}}\rho_K \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Конечно, можно упростить коэффициент до $9n/4$, но утверждается, что так точность аппроксимации выше. Критерий реализован здесь.

4.1.3 Коэффициенты корреляции

Отметим, что коэффициент корреляции зачастую используют не только для проверки гипотезы, но и для характеристики зависимости. Так для набора многомерных данных (записывая их по строкам) строят таблицу коэффициентов корреляции столбцов, из чего представляют, какие признаки связаны, а какие нет.

В некотором смысле, коэффициенты корреляции задают геометрию пространства случайных величин. Там обычный коэффициент корреляции можно описать следующим образом:

- Рассматриваем пространство $L^2(P)$ случайных величин со скалярным произведением EXY.
- Проецируем X и Y на ортогональное дополнение к пространству константа, получаем $\widetilde{X} = X - EX$, $\widetilde{Y} = Y - EY$.
- Считаем косинус угла между полученными величинами:

$$\frac{\mathbf{E}\widetilde{X}\widetilde{Y}}{\sqrt{\mathbf{E}\widetilde{X}^2}\sqrt{\mathbf{E}\widetilde{Y}^2}}.$$

Подобную интерпретацию можно дать и двум другим коэффициентам.

Зачастую возникает известная проблема зависимости через третье – может оказаться, что $corr(X,Y)$ большая, но просто потому, что обе переменные сильно коррелируют с некоторой Z. Для снижения этого фактора используют исключенные корреляции. Данная процедура работает так.

- Рассматриваем пространство $L^2(P)$ случайных величин со скалярным произведением EXY.

- Проецируем $\widetilde{X}$ и $\widetilde{Y}$ на ортогональное дополнение к пространству $\{cZ\}$, получаем $\widehat{X}$, $\widehat{Y}$.
- Считаем косинус угла между полученными величинами:

$$\frac{\mathbf{E}\widehat{X}\widehat{Y}}{\sqrt{\mathbf{E}\widehat{X}^2}\sqrt{\mathbf{E}\widehat{Y}^2}} = \frac{\rho_{X,Y} - \rho(X,Z)\rho(Y,Z)}{\sqrt{(1-\rho(X,Z)^2)(1-\rho(Y,Z)^2)}}.$$

Такой коэффициент называется частным коэффициентом корреляции X, Y при условии Z: $\rho(X,Y|Z)$. Если я хочу посчитать $\rho(X,Y|Z,W)$, то я провожу ту же процедуру и получаю

$$\rho(X,Y|Z,W) = \frac{\rho(X,Y|W) - \rho(X,Z|W)\rho(Y,Z|W)}{\sqrt{(1-\rho(X,Z|W)^2)(1-\rho(Y,Z|W)^2)}}.$$

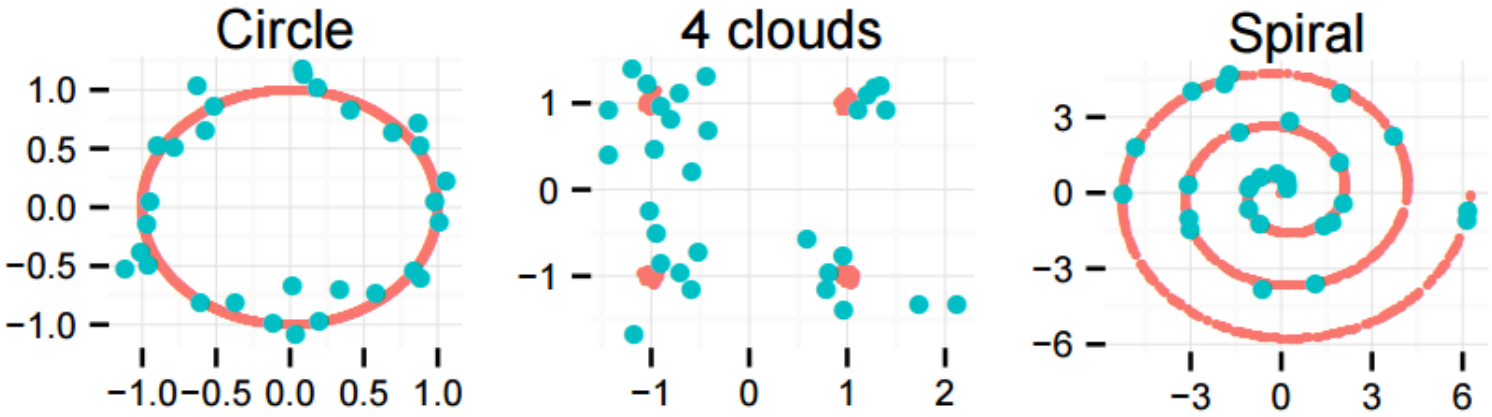
Эта процедура одна и та же для всех трех видов коэффициентов. Частные корреляции реализованы в пакете по ссылке.

4.2 Задачи

1. Найти коэффициенты корреляции баллов ЕГЭ и частные коэффициенты корреляции и сделать выводы о структуре их зависимости.

2. Для распределений, изображенных на рисунке 1 (сгенерируйте выборки самостоятельно, размер выборок 50 или 100), сравните критерии Секея-Риццо, Кендалла и хи-квадрат. Для распределения хи-квадрат ячейки предлагается выбирать, деля данные по каждой из строк на равные фрагменты.

Рис. 1: Три массива для задачи 2



3. Сравните критерии Смирнова, Пирсона, Кендалла, Спирмена и Секея-Риццо на выборках 1) $Y_i = X_i^2 + \varepsilon_i$, $X_i \sim R[-1,2]$, $\varepsilon_i \sim \mathcal{N}(0,0.5)$, 2) $Y_i = \sin X_i + \varepsilon_i$, $X_i \sim R[0,2\pi]$, $\varepsilon_i \sim \mathcal{N}(0,0.5)$.

5 Цепи Маркова

1. Дана м.в.п. однородной цепи Маркова

$$1)\begin{pmatrix} 1/4 & 3/4 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 3/4 & 1/4 \end{pmatrix},\ 2)\begin{pmatrix} 2/9 & 1/3 & 0 & 4/9 \\ 4/9 & 1/9 & 0 & 4/9 \\ 2/9 & 2/9 & 2/9 & 1/3 \\ 2/9 & 2/9 & 4/9 & 1/9 \end{pmatrix},\ 3)\begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/4 & 0 & 3/4 \\ 1/4 & 0 & 3/4 & 0 \end{pmatrix}.$$

   - Найти стационарное распределение.
   - Возвести матрицу в степень 5, 10, 20,30,50 и сравнить результаты.

2. Пусть ξ_n – ЦМ с м.в.п.

$$\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 3/4 & 0 & 1/4 \\ 1/2 & 1/6 & 1/3 \end{pmatrix}$$

   Найти стационарное распределение. Построить траекторию цепи. Сходится ли она п.н.? Найти долю посещений каждого из состояний за большое число n шагов. Сходится ли она п.н.? Моделируем 100 траекторий и для каждой вычтем из доли посещения первого состояния соответствующую стационарную вероятность, умножим результат на $\sqrt{n}$. Построить гистограмму полученного распределения.

3. В файле MarkovChain.txt находится м.в.п. цепи, не имеющей несущественных состояний. Написать программу, которая а) разделит состояния на неразложимые классы, б)* каждый из классов на подклассы состояний в соответствии с периодичностью цепи.

4. Пусть ξ_n – ЦМ с м.в.п.

$$\begin{pmatrix} 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 \\ 1/2 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}.$$

   Какую м.в.п. имеет обращенная цепь ξ_{N-n}? Исследовать к какому распределению сходится а) прямая цепь б) обращенная цепь?

5. Моделировать случайное блуждание а) на целочисленной решетке прямой, выбирающее равновероятно одну из двух соседних точек б) на целочисленной решетке плоскости, выбирающее равновероятно одну из четырех соседних точек в) в трехмерном пространстве, выбирающее равновероятно одну из шести соседних точек. Какие из них, исходя из моделирования, оказались возвратными?

6* *(Засчитывается как факультатив по теории случайных процессов)*. Построить кооп для проверки гипотезы H_0: X_i независимы с альтернативой H_1: X_i – однородная цепь Маркова. Использовать критерий (считая выполненной теорему Уилкса) для проверки гипотезы для данных с м.в.п.

$$\begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$$

   при а) $p = 1/2$, б) $p = 1/3$, в) $p = 0.52$.

6 Ветвящиеся процессы

Пусть Z_n – количество частиц в n-м поколении ветвящегося процесса (ВП) Гальтона-Ватсона ($Z_0 = 1$) с геометрическим распределением (с параметром p) числа потомков.

Прежде чем решать задачи, полезно вспомнить (или разобраться), как по параметру p определить, является ли процесс критическим, надкритическим или докритическим?

Задачи 1-4 посвящены моделированию ветвящихся процессов, задача 5 посвящена статистике.

1. Оценить вероятность вырождения ВП по множеству симуляций и сравнить с точным ответом.

2. Смоделировать критический процесс 1000 раз до момента вырождения или до 50-го хода, отобрать траектории, которые через 50 ходов еще жива, вывести их. Как их количество и численность соответствующих процессов согласуется с предельной теоремой для ВП?

   Сколько траекторий проживет 50 ходов в случае докритического процесса с геометрическ с $p > 1/2$? Моделировать процесс для различных p.

3. Смоделировать 20 шагов надкритического процесса с $p = 0.45$ построить график $\log Z_n$ от n. Построить несколько траекторий таких процессов.

4. Рассмотрим докритический ВП с иммиграцией в одну частицу. Найти численно вероятность того, что в процессе k частиц в момент n, где $k \leq 5$.

5*. Рассмотрим а) надкритический, б) критический ВП. Моделировать процесс в течении 20 поколений (дождавшись невырождения).

   Оценить с помощью ОМП среднее число потомков одной частицы и вероятность того, что частица дает ровно 0 потомков, а также вероятность того, что частица дает ровно одного потомка. Сравнить оценки с реальными значениями.
   Задача также приносит 0.5 баллов по факультативу за 5 семинар.

7 Спецкурс "Дополнительные главы теории вероятностей"

Вариант определяется по первой букве имени – А-К отправляется в первый вариант, Л-Я – во второй. Первый вариант получает пункты а), а второй – пункты б).
Решение всех задач со звездочкой добавляет +0.5 балла к сумме оценок за два семестра спецкурса "Дополнительные главы".

7.1 Первый семестр

1. Теорема Севастьянова.
   Ниже описан ряд величин. Построить гистограмму распределения соответствующей величины. Соответствует ли это теореме Севастьянова? Чтобы сопоставить модель теореме Севастьянова – рассмотрите индикаторы того, что в данном месте заканчивается некий удовлетворяющий условию блок.
   Симулировать многократно
   (a) серии из $c2^n$ бросков симметричной монеты и подсчитать число блоков из орлов длины не менее чем n бросков в каждой из серий, $c = 2$, $n = 14$;

   (b) серии розыгрышей из $cn!n!(2n+1)$ н.о.р. случайных величин из $R[0,1]$ и подсчитать в каждой серии число лесенок длины $2n+1$ ($a_1 < a_2 < \ldots < a_{n+1} > a_{n+2} > \ldots > a_{2n+1}$), $n = 5$, $c = 2$;

   (c) * проверить в каждом из случаев условия теоремы Севастьянова.

2. Пустые ячейки и высоковероятные слова.
   (a) Многократно симулировать распределение n частиц по n ячейкам, где $n = 50, 100, 200$, в каждом случае подсчитать число пустых ячеек. Произвести нормировку соответственно ЦПТ о размещении частиц по ячейкам и визуально продемонстрировать сходимость к $\mathcal{N}(0,1)$ распределению.

   (b) Сгенерировать $N = 1000000$ последовательностей слов из случайного алфавита $\{a, b\}$ с вероятностями 0.4 и 0.6 длины $T = 18$ и подсчитать частоты встречаемости различных слов. Какие частоты получили самые вероятные слова? Какая частота у большинства слов?

   (c) * Сравнить результаты с предельными теоремами, описанными в курсе.

3. * Расстояние по вариации.
   Пусть X, Y – случайные величины с распределениями F, G. Задать генератор двух зависимых случайных величин X', Y' с ф.р. $F_{1,1}$, $F_{1,2}$, для которых $\rho(X,Y) = P(X' \neq Y')$, где ρ – расстояние по вариации.
   (a) $F = Bern(1/2)$, $G = Bern(1/3)$.

   (b) $F = R[0,1]$, $G = x^2 I_{[0,1)} + I_{x>1}$

4. Теорема Линдеберга.
   Моделировать данные $X_i \sim P_i$, $i \leq n$. Исследовать предельное поведение $(S_n - \mathbf{E}S_n)/\sqrt{DS_n}$.
   (a) а) $P_n(n) = P_n(-n) = 1/(2n^2)$, $P_n(0) = 1 - 1/n^2$; б) $P_n(n) = P_n(-n) = 1/4$, $P_n(0) = 1/2$.

   (b) а) $P_n(\sqrt{n}) = P_n(-\sqrt{n}) = 1/2$, если n – полный квадрат, $P_n(1) = P_n(-1) = 1/2$ иначе, б) $P_n(n) = P_n(-n) = 1/(2\sqrt{n})$, $P_n(0) = 1 - 1/\sqrt{n}$.

   (c) * Проверить во всех случаях выполнение теоремы Линдеберга и сравнить полученные результаты с моделированием.

5. Безгранично делимые распределения.
   Построить предложенные ниже схемы серий и в каждой из них эмпирически определить предельное безграничное делимое распределение:
   (a) $X_{n,i} \sim Geom(1 - 1/n)$.

(b)

$$X_{n,i} = \begin{cases} \frac{(1-1/i)}{\sqrt{\ln n}}, & \frac{1}{i} \\ -\frac{1}{i\sqrt{\ln n}}, & 1 - \frac{1}{i} \end{cases}$$

(c) $X_{n,i} \sim NegBinom(1/n, 1/n)/n$.

8 Марковские процессы с непрерывным временем

1. Пусть X_t – марковский процесс с непрерывным временем, м.п.и. которого имеет вид

$$Q = \begin{pmatrix} -4 & 1 & 3 \\ 2 & -3 & 1 \\ 3 & 2 & -5 \end{pmatrix}.$$

   Найти м.в.п. вложенной цепи, моделировать X_t с помощью вложенной цепи.

2. Пусть X_t – марковский процесс с интенсивностью перехода i^2 из состояния i в $i+1$. Моделировать процесс и построить его траекторию.

3. В парикмахерскую с тремя парикмахерами приходят клиенты с интенсивностью λ, парикмахеры обслуживают клиентов с интенсивностью μ. Если все парикмахеры заняты, то клиент садится в очередь. Моделировать процесс при а) $\lambda = 1$, $\mu = 1$, б) $\lambda = 4, \mu = 1$.

4. Моделировать марковский процесс X_t

$$Q = \begin{pmatrix} -5 & 1 & 3 & 1 \\ 1 & -3 & 1 & 1 \\ 3 & 1 & -5 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}.$$

   Найдем стационарное распределение Q и сравним его со стационарными распределениями двух марковских цепей а) вложенной цепи б*) цепи, полученной измерением ЦМ через независимые exp(1) времена. Для оценки стационарного распределения марковских цепей будем считать долю времени, проведенного цепью в каждом из состояний за длительное время.

5. * В файле nerve.txt содержатся промежутки между нервными импульсами, проходящими по нейронам. Основная гипотеза заключается в том, что они независимые экспоненциальные. Построим процесс, который подпрыгивает на 1 через наши промежутки. Проверим, что длительность пребывания в состоянии с начала и до конца такая же как "осташаяся". Иначе говоря, возьмем исходную последовательность промежутков и построим ее гистограмму, а также выберем случайные точки на прямой (равномерно на накрытой промежутками части) и для каждой рассмотрим оставшееся время до ближайшего скачка нашего процесса, у полученного набора оставшихся времен построим гистограмму. Будут ли две наших гистограммы похожи (то есть выполнено ли свойство "новое такое же как старое"? Проверить эти две последовательности на однородность по критерию Манна-Уитни-Уилкоксона.

9 Процессы восстановления

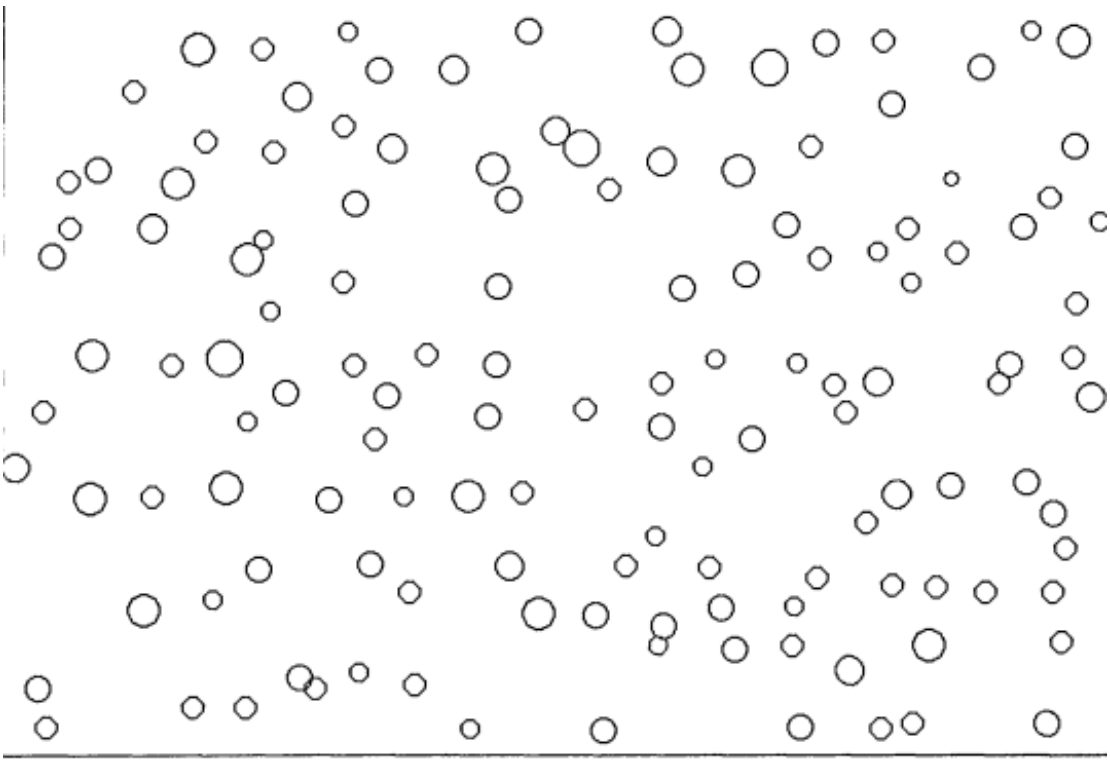
1. Моделировать процесс восстановления (то есть точечный процесс с расстояниями между скачками с указанным распределением) с а) $\exp(\lambda)$ шагами, б) $R[0,1]$ распределением. Исследовать численно предельное распределение эксцесса и сравнить с распределением шага.

2. Исследовать парадокс времени ожидания: моделировать а) пуассоновский поток с $\lambda = 2$ б) процесс с $Gamma(2, 1/4)$ распределением между скачками в) процесс с $R[0,1]$ распределением между скачками. Рассмотреть среднюю длину промежутка, накрывающего момент 100 и соотнести с математическим ожиданием между промежутками.

3. Исследуем условное свойство пуассоновского потока. Сгенерируйте точки пуассоновского потока а) на прямой б*) на плоскости и проверьте точки, попавшие на а) отрезок б*) квадрат на равномерность.

4. На картинке приведено расположений ели на участке хвойного леса. Считая, что они образуют двумерный одномерный пуассоновский поток, оценить число деревьев на небольшом участке и отсюда а) оценить число деревьев на картинке б) построить доверительный интервал (асимптотический) для этого количества. Повторить процедуру повторно и сравнить результаты.

Рис. 1: Ели в хвойном лесу



5. Проверить однородность потока на основе сравнения нескольких участков леса.

10 Гауссовские процессы

Вариант определяется остатком по модулю 3 + 1.

1. Построить частичные суммы в представлении Винера (в виде тригонометрических сумм) для броуновского движения. Проверить, что полученный процесс имеет независимые нормальные приращения – измерить значения процесса в моменты i/n, найти приращения, построить ЭФР и проверить ее на нормальность.

2. Построить траекторию броуновского движения, i) броуновского моста, ii) броуновской извилины и iii) броуновской экскурсии, используя то, что:

   - фрагмент броуновского движения от 0 до последнего нуля τ_0 совпадает по распределению (после растяжения) с броуновским мостом;
   - фрагмент броуновского движения от τ_0 до 1 совпадает по распределению (после растяжения) с броуновской извилиной.
   - фрагмент броуновского движения от τ_0 до первого нуля после точки 1 совпадает по распределению (после растяжения) с броуновской экскурсией

   Под растяжением подразумевается, что траектория растягивается по горизонтали в c раз до длины 1 и по вертикали в $\sqrt{c}$ раз.

3. Исследовать закон арксинуса для i) последнего пересечения нуля ii) момента достижения максимума iii) время выше оси для случайного блуждания с шагами а) $\mathcal{N}(0,1)$, б) $(1)-1$, в) $X = \delta\varepsilon$, $\mathbf{P}(\delta = 1) = \mathbf{P}(\delta = -1) = 1/2$, $\mathbf{P}(\varepsilon > x) = 1/x^{7/6}$, $x \geq 1$, где δ, ε предполагаются независимыми.

4. * Исследовать распределение времени, проведенного выше оси броуновским мостом, используя случайное блуждание, возвращающееся в ноль (как эффективно моделировать такое блуждание?).