Сегодня мы попрактикуемся в визуализации данных. Простые графики удобно строить с помощью бибилотеки seaborn (также будет полезна matplotlib), для более "красивой"будем использовать plotly.

(Примеры можно посмотреть в файле Basic Examples)

Датасет с данными абитуриентов содержится в файле Priem.csv, остальные датасеты – в seaborn.

1. Построить на одном графике графики плотностей $\mathcal{N}(0, k), k = 1, ..., n$ (например, для $n = 7$).
2. Моделировать 1000 реализаций с.в. $X \sim Bin(100,\ 0.04)$ и с.в. $Y \sim Poiss(4)$. Построить столбцовые диаграммы отдельно и на одном графике, сравнить. То же для $X \sim Bin(1000,\ 0.004)$ и $X \sim Bin(10,\ 0.4)$.
3. Построить гистограммы баллов по математике и по русскому: а) по отдельности, б) на одном графике, сравнить их.
4. Построить диаграмму рассеяния баллов ЕГЭ по математике и по русскому.
5. Построить диаграмму рассеяния баллов ЕГЭ по математике и по русскому, разными цветами показав пол абитуриента.
6. Построить диаграмму рассеяния баллов ЕГЭ по математике и по русскому, размерами точек показав суммарный балл, цветом – пол, формой тип школы.
7. Для массива данных "tips" построить boxplot() для размера чаевых по дням недели.
8. Для массива данных "titanic" сравнить выживших и невыживших пассажиров с помощью параллельных координат, используя столбцы survived, pclass, who (нужно будет присвоить числовые значения), age. (Не стоит использовать весь массив, лучше выбрать случайным образом 30 строк).
9. * Решить задачу 1 с возможностью изменения n ползунком.
10. * Построить столбцовую диаграмму для биномиального распределения с возможностью менять параметры ползунками.
11. * Для массива данных "titanic" построить диаграмму "солнечные лучи расположим по слоям survived, pclass, who.
12. ** Построить гистограмму баллов по математике с изменяемым количеством бинов.

Тема 4. Центральная предельная теорема

Вспомогательная теория

Напомним формулировку центральной предельной теоремы.

Теорема 1. *Пусть X_i - н.о.р. случайные величины, $\mathbf{E}X_i = \mu$, $0 < \mathbf{D}X_i = \sigma^2 < \infty$. Тогда*

$$\frac{X_1 + \ldots + X_n - \mu n}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

То есть

$$\mathbf{P}\left(\frac{X_1 + \ldots + X_n - \mu n}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x),\ n \to \infty,\ \forall x,$$

где $\Phi(x)$ – функция распределения $\mathcal{N}(0,1)$.

Обратите внимание, что, вообще говоря, сходимость по распределению не гарантирует сходимости плотностей, однако, справедлив следующий замечательный результат с которым вы познакомитесь в курсе дополнительных глав теории вероятностей.

Теорема 2. *Пусть X_i - н.о.р. случайные величины, причем $\int_{\mathbb{R}} |\psi_X(t)|^a dt < +\infty$ при некотором $a > 0$, где ψ – характеристическая функция. При этом $\mathbf{E}X_i = \mu$, $0 < \mathbf{D}X_i = \sigma^2 < \infty$. Тогда*

$$f_{(S_n - \mu n)/(\sigma\sqrt{n})}(x) \xrightarrow{d} \phi(x),$$

где ϕ – плотность стандартной нормальной величины.

Помимо самого факта сходимости функций распределений центрированных нормированных сумм известны также следующие результаты.

Теорема 3 (Неравенство Берри–Эссеена.)**.** *Пусть выполнены условия ЦПТ и дополнительно $\mathbf{E}|X|^3 < \infty$. Тогда*

$$\left|\mathbf{P}\left(\frac{X_1 + \ldots + X_n - \mu n}{\sigma\sqrt{n}} \leq x\right) - \Phi(x)\right| \leq C\frac{\mathrm{E}|X_1 - \mathrm{E}X_1|^3}{(\mathrm{D}X_1)^{3/2}\sqrt{n}},$$

где C – некоторая константа, не зависящая от распределения X_i. По последним данным $C \leqslant 0.4784$.

Теорема 4. *Пусть $X_1, \cdots, X_n$ н.о.р. $\mathbf{E}X = \mu$, $\mathbf{D}X = \sigma^2$, $\mathbf{E}(X-\mu)^3 = \rho_3$.*

Справедливо также следующее асимптотическое разложение.

Пусть $\mathbf{E}(X-\mu)^3 = \rho_3$, $a_3 := \rho_3/\sigma^3$ – коэффициент асимметрии. Тогда

$$\mathbf{P}\left(\frac{X_1 + \ldots + X_n - \mu n}{\sigma\sqrt{n}} \leq x\right) - \Phi(x) = \frac{a_3}{6\sqrt{2\pi n}}(1-x^2)\exp\left(-\frac{x^2}{2}\right) + o\left(\frac{1}{\sqrt{n}}\right)$$

при $n \to \infty$.

Теорема 4 дает более точные приближения чем теорема 3, зато теорема 3 не предельная, а верна при всех n.

Задачи

1. Моделировать выборки $X_{i,j}$, $i \leq 1000$, $j \leq n$, где i) n=20 ii) n=100 величин из распределений а) Bern(1/2), б) $R[0,1]$, в) exp(1), г) Коши. Найти $S_{i,n} = \sum_{j=1}^{n} X_{i,j}$ и построить на одном графике ЭФР $S_{n,i}$ и ф.р. $\mathcal{N}(n\overline{X}, nS^2)$, где $\overline{X}$, S^2 – выборочное среднее и выборочная дисперсия всех имеющихся наблюдений. Похожи ли визуально полученные графики?

2. Пусть $X \sim Gamma(n, 4)$ Построить на одном графике графики плотности распределения с.в. $(X - \mathbf{E}X)/\sqrt{\mathbf{D}X}$ и плотности $\mathcal{N}(0, 1)$ для различных n.

3. Построить гистограмму по набору значений с.в. $S_n = X_1 + \ldots + X_n - \mu n$ (генерируем k выборок $X_1, \ldots, X_n$, по каждой находим одно значение суммы). На том же графике построить плотность распределения с.в. S_n (для дискретных – дискретное распределение) и плотность $\mathcal{N}(0, \sigma^2 n)$.

   Здесь распределения X_i рассматриваются следующие:

   - 1 вариант: $Poiss(\lambda)$, 2 вариант: $Geom(p)$,
   - 1 вариант: $exp(\lambda)$, 2 вариант: $Gamma(a, b)$.
   - * Для всех вариантов $R[0, 1]$ (для поиска плотности распределения суммы можно использовать sympy.stats.UniformSum() или написать формулу самостоятельно, см. распределение Ирвина–Холла),

4. Обозначим $Y = (S_n - n\mu)/(\sigma\sqrt{n})$. Построить на одном графике: $F_Y(x) - \Phi(x)$, правую часть неравенства Берри–Эссеена, ее же, умноженную на -1, правую часть асимптотического разложения. Рассмотреть $n = 5, 10, 20, 50, 100, 500$. Соотнести полученные результаты с теоремами 3 и 4. Рассмотрите следующие распределения X_i: $Bern(p)$, $exp(\lambda)$.

Тема 5. Генерация случайных величин.

1. Моделировать выборку из а) распределения $exp(1)$ б) распределения Коши с помощью метода обратной функции. Построить гистограмму, сравнить со встроенным методом.

2. Моделировать выборку из равномерного распределения на единичном круге: а) методом выбора с отклонениями, б)* методом условных распределений, в) с помощью полярных координат.

3. Моделировать выборку с плотностью а) Лапласа б) $\exp(-x)/2$, $x > 0$ и $\exp(2x)$, $x < 0$. Построить гистограмму. (Использовать смеси распределений.)

4. * Моделировать выборку с плотностью

$$\frac{1}{3}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) + \exp(-|x|)\right).$$

   Построить гистограмму

5. Смоделировать с помощью алгоритма Acceptance-Rejection выборку из а) треугольного распределения (с плотностью $(1 - |x|)\,\mathrm{I}(x \in [-1, 1])$); б)* нормального распределения с помощью распределения Лапласа.

6 Состоятельность. Асимптотическая нормальность

На семинарах вы уже освоили такие свойства оценок, как состоятельность и асимптотическая нормальность, и научились их доказывать. Сформулируем еще некоторые полезные результаты в этой области (для более глубокого ознакомления с темой рекомендруем книгу М.Б. Лагутина "Наглядная математическая статистика").

Для начала напомним (или даже введем) некоторые понятия:

Определение 1. Выборочной медианой MED называют оценку

$$\text{MED} = \begin{cases} X_{(k+1)}, & n = 2k+1, \\ \dfrac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k, \end{cases},$$

усеченным среднем $\overline{X}_\alpha$

$$\overline{X}_\alpha = \frac{1}{n-2k}(X_{(k+1)} + \cdots + X_{(n-k)}), \quad k = [\alpha n].$$

Выборочная медиана оценивает теоретическую медиану $x_{1/2} = F^{-1}(1/2)$. Усеченное среднее как правило используют для оценки центра симметрии у симметричных распределений.

Теорема 1. *Пусть распределение F таково, что $F(x_{1/2}+\varepsilon) > 1/2$ при всех $\varepsilon > 0$. Тогда* MED *будет состоятельной оценкой $x_{1/2}$.*

Теорема 2. *Пусть $X_1, \ldots, X_n$ выборка из распределения с плотностью f, причем $f(x) > 0$ в некоторой окрестности $x_{1/2}$; здесь $x_{1/2}$ – медиана распеделения с.в. X_1. Тогда выборочная медиана* MED *является асимптотически нормальной оценкой $x_{1/2}$:*

$$\sqrt{n}(\text{MED} - x_{1/2}) \xrightarrow{d} Z \sim \mathcal{N}\left(0, \frac{1}{4f^2(x_{1/2})}\right), \; n \to \infty.$$

Теорема 3. *Пусть $X_1, \ldots, X_n \sim F(x-\theta)$, где F обладает следующими свойствами: найдется такое $0 < c \leq +\infty$, что $F(-c) = 0$, $F(c) = 1$ и на $(-c, c)$ $F(x)$ имеет четную, непрерывную и положительную плотность $f(x)$.*

Тогда усеченное среднее $\overline{X}_\alpha$ при $0 < \alpha < 1/2$ является асимптотически нормальной оценкой θ:

$$\sqrt{n}(\overline{X}_\alpha - \theta) \xrightarrow{d} Z \sim \mathcal{N}\left(0, \sigma_\alpha^2\right), \; n \to \infty, \quad \sigma_\alpha^2 = \frac{2}{(1-2\alpha)^2}\left[\int_0^{x_{1-\alpha}} t^2 f(t)\,dt + \alpha x_{1-\alpha}^2\right],$$

где x_γ – решение уравнения $F(x_\gamma) = \gamma$.

Задачи

1. $X_1, ..., X_n \sim \text{R}[0, \theta]$.

   (a) Построить гистограммы для $X_{(n)}$ при разных n и сравнить с нормальной плотностью (с такими же математическим ожиданием и дисперсией, как у $X_{(n)}$).

   (b) Построить гистограммы для $n(\theta - X_{(n)})$ при разных n и сравнить с плотностью распределения $\exp(1)$.

   (c) * Изобразить гистограммы $\sqrt{n}(2\overline{X} - \theta)$ и $\sqrt{n}(2\overline{X}_\alpha - \theta)$ на одном графике, сравнить разбросы (обе ли оценки асимптотически нормальны, чья асимптотическая дисперсия меньше?).

   (d) Сравнить, какая из оценок $((n+1)/n)X_{(n)}$ и $2\overline{X}$ чаще оказывается ближе к θ при разных n. Для этого смоделировать по 1000 реализаций (для каждого n) и найти, в какой доле из этих 1000 ближе оказалась $((n+1)/n)X_{(n)}$.

2. $X_1, \ldots, X_n$ имеет распределение Коши $f_\theta(x) = (\pi(1 + (x - \theta)^2))^{-1}$.

   (a) Построить гистограммы для $\overline{X}$ при разных n. Является ли эта оценка состоятельной?
   *Сравнить гистограмму/оценку плотности с нормальной плотностью.

   (b) Построить гистограммы $\sqrt{n}(\text{MED} - \theta)$, сравнить с соответсвующей нормальной плотностью (см. теорему 1).

3. * $X_1, ..., X_n \sim Bern(p)$, где а) $p = 1/3$ б) $p = 1/2$. Будет ли MED состоятельна? Асимптотически нормальна?
   Постройте гистограммы $\sqrt{n}(\text{MED} - 1/2)$, похоже ли распределение на нормальное?

4. $X_1, ..., X_n \sim R([\theta - 2, \theta - 1] \cup [\theta + 1, \theta + 2])$. Будут ли выборочная медиана и усеченные средние а) состоятельны, б) асимптотически нормальны?

5. Смоделировать выборку из распределения Лапласа и численно сравнить асимптотическую дисперсию медианы, выборочного среднего, $\overline{X}_\alpha$ с $\alpha = 0.1$ и $\alpha = 0.3$. Для этого построить гистограммы или плотности каждого из распределений.

7 ОММ, ОМП, ОМС

Мальчики получают первый вариант, а девочки – второй.

Не забудьте, что правдоподобие удобнее логарифмировать перед максимизацией.

1. Построить график функции логарифмического правдоподобия (в пункте б) – обычного) для следующих моделей при $n = 1, 5, 20, 100, 1000$:
   (a) 1 вариант – $X_1, \ldots, X_n \sim \mathcal{N}(0, \theta)$, 2 вариант $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$.
   (b) $X_1, \ldots, X_n \sim R[0, \theta]$.
   (c) * $X_1, \ldots, X_n$ – выборка из распределения, являющегося смесью $\mathcal{N}(\theta_1, \theta_2)$ и $\mathcal{N}(0, 1)$ с весами $1/2$, $1/2$; здесь требуется построить графики $L(\theta_2)$ при фиксированном значении первого параметра: а) θ_1 – настоящее значение (с которым генерировалась выборка), б) $\theta_1 = X_1$, в) любое число, не равное настоящему значению и не совпадающее ни с одним из элементов выборки.

2. X_i имеют распределение Коши, где в варианте 1 неизвестный параметр — сдвиг θ, а в варианте 2 – масштаб. Построить ОМП по выборке размера $n = 5, 10, 20, 50, 100$. Для каждого n генерировать $k = 500$ выборок $X_1, \ldots, X_n$, для каждой найти значение ОМП и $\widehat{\theta}$, найти выборочное среднее и выборочную дисперсию ОМП и $\widehat{\theta}$ и сравнить их. Здесь $\widehat{\theta}$ для первого варианта – это выборочная медиана, а для второго – половина интерквартильного размаха, то есть полуразность верхнего и нижнего выборочного квартиля.

3. 1 вариант – $X_1, ..., X_n \sim \mathcal{N}(\theta, 1)$, 2 вариант – $X_1, \ldots, X_n \sim \mathcal{N}(0, \theta)$, $\widehat{\theta}_1$ – ОММ, $\widehat{\theta}_2$ – ОМС, $\widehat{\theta}_3$ – ОМП.
   (a) Построить по выборке $\widehat{\theta}_i$, $i = 1, 2, 3$ (найти численно или аналитически, как удобнее).
   (b) Сравнить, какая из оценок чаще оказывается ближе к θ при разных n, смоделировав для этого по 1000 реализаций (для каждого n).
   (c) Построить гистограммы для $\sqrt{n}(\widehat{\theta}_i - \theta)$ на одном графике, сравнить разбросы.
   (d) * Посчитать асимтотическую дисперсию этих оценок, построить график (по θ, сравнить с выборочной дисперсией (нормированной).

4. * 1-й вариант: $X_1, ..., X_n \sim Beta(a, b)$, 2-й вариант: $X_1, ..., X_n \sim Gamma(a, b)$,
   $\theta = (a, b)$, $\widehat{\theta}_1 = (\widehat{a}_1, \widehat{b}_1)$ – ОМП, $\widehat{\theta}_2 = (\widehat{a}_2, \widehat{b}_2)$ – ОМС.
   (a) Построить по выборке оценки $\widehat{\theta}_1$ и $\widehat{\theta}_2$ (найти численно).
   (b) Сравнить, какая из оценок чаще оказывается ближе к θ при разных n (отдельно по каждой координате и в смысле расстояния на плоскости), смоделировав для этого по 1000 реализаций (для каждого n).
   (c) Построить гистограммы (одномерные) для $\sqrt{n}(\widehat{\theta}_i - \theta)$ на одном графике, сравнить разбросы.

9 Доверительное оценивание

1. Построить график функции $y_{1-\alpha+\beta} - y_\beta$ для $\beta \in (0, \alpha)$, где y_t – квантиль распределения

   (a) $N(0, 1)$,

   (b) $Gamma(n, 1)$, $n = 1, 2, 5, 10, 100$,

   (c) $R[0, 1]$,

   (d) $Beta(a, b)$, $a = b = 5$, $a = 10$, $b = 2$, $a = 20$, $b = 1$.

   Рассмотреть одно любое значение α, например, $\alpha = 0.001$, 0.05, 0.1. Сделать вывод о выборе оптимального β для построения доверительного интервала на основе статистики с нашим распределением.

2. $X_1, ..., X_n \sim \mathrm{R}[0, \theta]$.

   (a) Построить асимптотический доверительный интервал, используя $\overline{X}$. Найти эмпирически доверительную вероятность этого интервала (построить 1000 выборок, подсчитать долю тех, для которых интервал накрыл истинное значение параметра, для $n = 20, 50, 100$.

   (b) Построить точный доверительный интервал, используя достаточную статистику. Сравнить средние длины точного и асимптотического интервалов при $n = 20, 50, 100$.

3. $X_1, ..., X_n \sim Bern(\theta)$. Построить асимптотический доверительный интервал двумя способами с помощью $\overline{X}$, сравнить средние длины полученных интервалов (генерировать 1000 выборок, по каждой строить оба интервала, посчитать и показать средние длины) для $\theta = 0.1$, 0.4, 0.5, 0.9 и $n = 20$, 50, 100.

4. * $X_1, ..., X_n \sim Gamma(\theta, 1)$. Построить асимптотический доверительный интервал для θ на основе ОМП.

5. ** Построить доверительный эллипс для параметра (μ_1, μ_2) по выборке из $\mathcal{N}(\vec{\mu}, \Sigma)$ распределения, где а) $\Sigma = E$ б) Σ имеет 1 и 2 на диагонали и 0.5 вне. Для построения можно использовать confidence_ellipse из matplotlib. Как меняется эллипс при измении размера выборки: взять $n = 10, 100, 500$.

6. ** $X_1, ..., X_n \sim \mathrm{R}[\theta_1, \theta_2]$. Построить доверительное множество для (θ_1, θ_2) с помощью $X_{(1)}$, $X_{(n)}$, изобразить для разных (θ_1, θ_2).

11 Критерии: основные понятия

1. Пусть $X_i \sim Bern(\theta)$, для $H_0 : \theta = 1/2$ и $H_1 : \theta = \theta_1$, где а) $\theta_1 = 1/3$, б) $\theta_1 = 2/3$. Рассмотрим критерий $\{\sum_{i=1}^n X_i > C\}$, $n = 10$. Построить графики вероятностей ошибки I рода, ошибки II рода и мощности критерия в зависимости от C. Для какой альтернативы осмысленно использовать этот критерий?

2. График ЭФР p-value.

   (a) Генерируем выборку, находим значение статистики критерия T ($T = X_{(1)}$ или $T = X_{(n)}$). Находим функцию распределения $F_T(x)$ нашей статистики.Вычисляем $p - value = 1 - F_T(T)$, для критических множеств вида $\{T > C\}$, $p - value = F_T(T)$ для критических множеств вида $\{T < C\}$. Повторяем $m \geq 100$ раз. Получился массив $p_1, \ldots, p_m$, упорядочиваем его по возрастанию.

   (b) Строим график: по оси Ox – значения $p_1, \ldots, p_m$, по оси Oy – числа $1/m, 2/m, \ldots, 1$. Иными словами, мы строим график эмпирической функции распределения p-value.

   Мы знаем, что если $F(x)$ непрерывна, то $F_T(T) \sim R[0, 1]$. Значит, при гипотезе точки должны быть близки к прямой $y = x$. При альтернативе мы ожидаем увидеть отклонение от этой прямой.

   Посмотрим как это работает на синтетическом наборе данных: пусть $X_i \sim \mathcal{N}(\theta, 1)$, $H_0 : \theta = 0$, $H_1 : \theta = \theta_1$. Постройте а) критерий Неймана-Пирсона для $\theta_1 > 0$ б) для $\theta_1 < 0$ в) асимптотический критерий $|MED - 1/2| > C$. Рассчитайте для них ф.р. статистик критерия, найдите p-value критериев.

   (a) Построить графики p-value всех трех критериев, выбирая данные а) при верной нулевой гипотезе б) при каждой из альтернатив ($\theta_1 = -1$ и $\theta_1 = 1$). Построить их на одном графике.

   В какую сторону отклоняются графики от прямой $y = x$?

   (b) Как влияет размер выборки на отклонение от $y = x$?

   (c) Какой критерий самый лучший?

3. $X_1, \ldots, X_n \sim Bern(\theta)$, $H_0 : \theta = 1/2$, $H_1 : \theta = 1/3$. При каких n можно построить критическое множество вида $\{\sum_{i=1}^n X_i < C\}$ так, чтобы вероятности ошибок первого и второго рода не превышали 0.05? Построить графики вероятностей ошибок первого и второго рода (как функции от C) для разных n.

4. * Построить в предыдущей задаче рандомизированный критерий Неймана-Пирсона уровня значимости 0.05. Эмпирически исследовать вероятность ошибки I рода критерия и убедиться, что она действительно 0.05.

10 Байесовский анализ

1. $X_1, ..., X_n \sim Bern(\theta)$, $\theta \sim \mathrm{R}[0,1]$. Посчитать апостериорную плотность и построить ее график.

   а) для выборок размера $n = 5, 10, 20, 50, 100$ для $\theta = 1/2$, $\theta = 1/3$;

   б) вместо генерации выборки положите $\sum_{i=1}^{n} X_i$ равной $9n/10$ или $99n/100$.

2. Пусть $X_1, ..., X_n \sim \mathcal{N}(\theta, 1)$, $\theta \sim \mathcal{N}(\mu, \sigma^2)$. Апостериорную плотность можно не считать, а взять из таблицы. Посмотрим, как влияют параметры априорного распредления на итоговый результат. Для этого будем генерировать выборки для какого-то одного фиксированного θ (возьмите любое число из $[0,1]$, а также посмотрите на какое-нибудь θ, близкое к 0 или 1).

   (a) Сравнить (визуально на графике) апостериорные плотности для нескольких разных значений (μ, σ^2).

   (b) Построить пример, когда при $n = 10000$ оценка θ достаточно сильно отличается от настоящего значения несмотря на размер выборки.

   (c) Сравнить байесовские оценки для квадратичного риска при разных (μ, σ^2).

3. $X_1, ..., X_n \sim \mathcal{N}(\theta, 1)$. Построить (на листочке) байесовский криретерий для проверки $H_0 : \theta = 0$ против $H_1 : \theta = 1$, если априорная вероятность $\mathbf{P}(\theta = 0) = p$. Построить графики зависимости ошибок 1-го и 2-го рода от p.

4. * Пусть $X_i \sim \exp(\theta)$, а $\theta \sim Gamma(a, b)$. а) Построить байесовские оценки для абсолютной и квадратичной функций потерь и сравнить у таких оценок среднюю а) квадратичную б) абсолютную ошибку. б) Построить байесовский доверительный интервал уровня 95% и эмпирически исследовать ее уровень доверия.

12 Критерии хи-квадрат и отношения правдоподобий для дискретных данных

Статистика критерия хи-квадрат для простой гипотезы выглядит как

$$\sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}.$$

При верной гипотезе она сходится к χ^2_{k-1}. Статистика критерия отношения правдоподобий предлагает взамен брать

$$\sum_{i=1}^{k} \nu_i \ln \frac{\nu_i}{np_i}.$$

Наконец, более общий подход Кресси-Рида предлагает рассматривать

$$\frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} \nu_i \left(\left(\frac{\nu_i}{np_i} \right)^{\lambda} - 1 \right),$$

где при $\lambda \in \{-1, 0\}$ выражение доопределяется из соображений по непрерывности.

1. Начнем с проверки простой гипотезы.

   Найдите первые 1000 цифр числа π после запятой. С помощью критерия хи-квадрат проверьте, можно ли при уровне значимости 0.05 считать эти цифры случайными равномерными?

2. Проверим однородность и независимость. Использовать данные из файла Priem.csv и встроенный критерий.

   (a) Ответить на вопрос - отличаются ли мальчики и девочки в плане успешности сдачи ЕГЭ? Для этого попарно проверьте на однородность суммарные баллы, баллы по русскому, баллы по математике.

   (b) Правда ли, что оценки по математике и русскому независимы?

3. Построим критерий Кресси-Рида для проверки простой гипотезы о полиномиальном распределении. Давайте сравним наши критерии для различных λ. Рассмотрите λ, равные $-1, 0, 0.5, 1, 2$. Постройте график p-value для каждого из них и выберите наиболее удачный критерий. Используйте исходное равномерное распределение (все p_i равны) и неравномерное на свой вкус.

4. * Переходим к параметрической гипотезе.

   Среди 2020 семей, имеющих 2 детей, 527 семей, в которых 2 мальчика, и 476 - две девочки. Можно ли при уровне значимости 0.05 считать, что количество мальчиков – биномиальная случайная величина?

   В этой задаче нужно сначала найти ОМП для параметрической гипотезы (формулу для нее) на листочке, затем вычислить ее значение для данных из условия задачи и воспользоваться встроенным критерием хи-квадрат.

   Теперь исследуем работу получившегося критерия на модельных данных (нужно использовать ту же формулу для ОМП, что и раньше). Рассмотреть следующие распределения:

   1) $Binom(2, 1/2)$, $Binom(2, 1/8)$,

   2) равномерное распределение $\mathcal{R}\{0, 1, 2\}$,

   3) $\mathbf{P}(X = 0) = \mathbf{P}(X = 2) = 3/8, \mathbf{P}(X = 1) = 1/4$ (X – число мальчиков).

   Для каждого распределения сгенерировать по 100 выборок. К каждой выборке применить построенный критерий, получить p-value. Построить графики p-value для каждого распределения, сравнить их.

12 ЭФР, критерий Колмогорова

Определение 1. Эмпирической функцией распределения называют функцию

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}.$$

Теорема 1 (Колмогорова)**.** *Пусть $X_i \sim F$, где F непрерывна, тогда*

$$\mathbf{P}(\sqrt{n}\sup_x |\widehat{F}_n(x) - F(x)| \le y) \to K(y),$$

где

$$K(y) = \sum_{k=-\infty}^{\infty} e^{-2k^2y^2}$$

называют функцией распределения Колмогорова.

Отметим, что если распределение не непрерывно, то

$$\mathbf{P}(\sqrt{n}\sup_x |\widehat{F}_n(x) - F(x)| \le y) \to K'(y),$$

где $K'(y)$ зависит от F, но $K'(y) \ge K(y)$ при всех y.

Теорема 2 (Неравенство Дворецкого-Кифера-Вольфовица)**.** *Пусть $X_i \sim F$, тогда*

$$\mathbf{P}(\sup_x |\widehat{F}_n(x) - F(x)| \ge y) \le 2e^{-2ny^2}.$$

Отсюда мы можем узнать погрешность оценки $F(x)$ функцией $\widehat{F}_n(x)$, построить доверительную полосу для $F(x)$. Отметим, что непосредственно критерий не слишком мощный и на практике, как правило, используют более мощный критерий Андерсона-Дарлинга (об этом в следующем семестре).

Критерий Колмогорова реализован в пакете scipy.

1. Пусть $X_i \sim \mathcal{N}(0,1)$.
   1) Построить эмпирическую функцию распределения (ЭФР) на одном графике с теоретической функцией распределения при разных n (проиллюстрировать сходимость ЭФР к ф.р.). То же для $X_i \sim R[0,1]$, $X_i \sim Bin(m,p)$, где $m = 3$, $m = 10$.
   2) Построить 95% доверительную полосу для ф.р., используя а) критерий Колмогорова. б) неравенство Дворецкого-Кифера-Вольфовица.

2. С помощью критерия Колмогорова проверить гипотезу $H_0 : X_i \sim \mathcal{N}(0,1)$, если
   $X_i = (Y_i - \mathbf{E}Y_i)/\sqrt{\mathbf{D}Y_i}$, где
   (a) $Y_i \sim \mathcal{N}(-3,8)$,
   (b) $Y_i \sim Bin(m, 1/2)$,
   (c) Y_i имеют распределение Ирвина–Холла (сумма m независимых равномерных с.в.).

   Построить графики p-value для $m = 1, 2, 5, 20$. Рассмотреть выборки длины $n = 50, 100, 500$.

3. Если выборка имеет нормальное распределение с неизвестными параметрами, то можно попробовать перейти от выборки X_i к $(X_i - \overline{X})/S$ и применить к ним критерий Колмогорова нормальности $\mathcal{N}(0,1)$. Построить график p-value при верной гипотезе и посмотреть корректно ли работает "модифицированный" критерий.