0.1 Getting started

Полезные ссылки - Python

https://habr.com/ru/sandbox/80371/

https://pythonworld.ru/samouchitel-python

https://pythonworld.ru/uploads/pythonworldru.pdf

https://jenyay.net/Matplotlib/Matplotlib

Полезные ссылки - R

https://www.statmethods.net/

http://www.inp.nsk.su/ baldin/DataAnalysis/R/R-01-intro.pdf

https://stepik.org/course/497/promo

https://stepik.org/course/129/promo

0.2 Теория вероятностей

0.2.1 Центральная предельная теорема

1.1. Независимые с.в. $\xi_i \sim exp(1)$, $S_n = \sum_{i=1}^n \xi_i$. Посторить графики плотностей S_n, $n = 1, 2, \ldots, 10$. Изобразить на одном графике плотность центрированной и нормированной с.в. S_n и плотность $N(0, 1)$ для $n = 1, 2, \ldots, 20$. Также изобразить графики ф.р.

1.2. С. в. ξ_i независимы и имеют плотность $(e^{-(x-1)}\mathbb{I}\{x > 1\} + e^{-x}\mathbb{I}\{x > 0\})/2$, $S_n = \sum_{i=1}^n \xi_i$. Найти формулу для плотности распределения с.в. S_n. 1) Построить графики плотности для $n = 1, 2, \ldots, 10$ на одном графике.
2) Построить графики плотности центрированной и нормированной суммы и стандартного нормального распределения (2 плотности на одном графике) для $n = 1, 2, \ldots, 10$.

1.3. Распределение случайной величины ξ является смесью двух нормальных распределений: $\mathcal{N}(a, b)$ с весом 1/3 и $\mathcal{N}(c, d)$ с весом 2/3. Найти формулу для плотности распределения суммы n независимых реализаций случайной величины ξ.
1) Построить графики плотности для $n = 1, 2, \ldots, 10$. По графикам найти числа локальных максимумов плотности для каждого из 10 значений n. Параметры $a,\ b,\ c,\ d$ выбрать так, чтобы плотность ξ имела два выраженных локальных максимума.
2) Построить графики плотности центрированной и нормированной суммы и стандартного нормального распределения (2 плотности на одном графике) для $n = 1, 2, \ldots, 10$.

0.2.2 Байесовские оценки

2.1. $X_1, ..., X_n \sim Bern(\theta)$, $\theta \sim \mathrm{R}[0, 1]$. Построить график апостреиорной плотности θ а) для выборок размера $n = 5, 10, 20, 50, 100$ (можно $n = 10, 20, \ldots, 100,$) б) для значения $\sum_{i=1}^n X_i = 9n/10,\ 99n/100$.

2.2. Сделать то же, что в 2.1 для $X_1, ..., X_n \sim Bern(\theta)$, $\theta \sim Beta(a, b)$.

2.3. $X_1, ..., X_n \sim Bern(\theta)$. Сравнить отклонения оценок $\widehat{\theta}_1 = \overline{X}$ и $\widehat{\theta}_2 = \overline{X} + \left(\frac{1}{2} - \overline{X}\right)/(1+\sqrt{n})$ от θ : генерировать 100 выборок и оценить вероятность того, что $\widehat{\theta}_2$ ближе к θ, чем $\widehat{\theta}_1$ (при разных значениях θ.)

0.2.3 Оценки метода моментов, оценки максимального правдоподобия.

3.1. В партии пушек с некоторой вероятностью появляется пушка с браком. В файле Sample1.txt записаны точки падения снаряда при стрельбе одиночной качественной пушки по точке 0. В файле Sample2.txt — точки падения снаряда при стрельбе одиночной бракованной пушки. В Sample3.txt — итоги стрельбы батареи из большого числа различных пушек. Оценить вероятность брака.
Ссылка на файлы
https://yadi.sk/d/bMIsG3mLrTJlKA
https://yadi.sk/d/Sdtf7w8AguCBIw
https://yadi.sk/d/zkhA5XfCbxAWRw

3.2. $X_i \sim p(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$. Построить ОМП по выборке размера $n = 5, 10, 20, 50, 100$. Для каждого n генерировать $k = 500$ выборок $X_1, \dots, n$, для каждой найти значение ОМП и MED, найти выборочное среднее и выборочную дисперсию ОМП и MED, сравнить.

3.3. Распределение случайной величины ξ является смесью двух нормальных распределений: $\mathcal{N}(a, b)$ с весом p и $\mathcal{N}(c, d)$ с весом $1 - p$. Построить график функции правдоподобия как функции параметра d при фиксированных p, b, c и $a = X_1$. (Сначала сгенерировать выборку для фиксированных a, b, c, d, p, затем подставлять в функцию правдоподобия $a = X_1$ и разные значения b, c, p.) Построить также график $L(b, d,$ $L(b, d,$ $L(p, d)$, $L(c, d)$.

0.2.4 Доверительные интервалы

4.1. $X_1, ..., X_n \sim Bern(\theta)$. Построить 2 асимптотических доверительных интервала с помощью $\overline{X}$, сравнить их средние длины (генерировать 1000 выборок, по каждой строить оба интервала, посчитать и показать средние длины) для $\theta = 0.1,\ 0.4,\ 0.5,\ 0.9$ и $n = 20,\ 50,\ 100$.

4.2. $X_1, ..., X_n \sim \mathcal{N}(\theta, \theta_2)$, $\theta > 0$. Построить 3 точных доверительных интервала (центральная функция, S^2, $(\overline{X}, S^2)$. Сравнить их средние длины для $\theta = 1,\ 10,\ 100$ и $n = 5,\ 20,\ 50,\ 100$.

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

0.2.5 Критерии

5.1. Критерий Колмогорова.
https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.kstest.html
https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.kstest.html $Y_{ij} \sim \mathrm{R}[0, 1]$, i $= 1, \dots,$ k, j $= 1, \dots$ независимы.
$X_j = \sum_{i=1}^{k} Y_{ij}, j = 1, \dots, n$, $F(x)$ – функция распределения с.в. $(X_1 - MX_1)/\sqrt{DX_1}$.
Гипотеза $H_0 : F(x) = \Phi(x) \sim \mathcal{N}(0, 1)$. Проверить H_0 с помощью критерия Колмогорова, нарисовать график p-value для $k = 1, 2, 10; n = 10, 50, 200$.
Алгоритм действий.
1) Генерировать $Y_1, \dots, Y_k \sim R[0, 1]$, вычислить X_1. Повторить n раз.

2) К выборке $X_1, \dots, X_n$ применить scipy.stats.kstest и вычислить $p - value$.
3) Повторить 1)-2) m раз (m=100,500), получится массив $p_1, \dots, p_m$. Упорядочить их и построить график (эмпирическую функцию распределения) этих p-value. (по x – упорядоченные p_j, по y – makelist(j/m,j,1,m)).
4) Проанализировать графики при разных k, n, m.

5.2. Критерий Андерсона–Дарлинга.
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html
$Y_{ij} \sim$ R[0, 1], i = 1, . . . , k, j = 1, . . . независимы. $X_j = \sum_{i=1}^{k} Y_{ij}, j = 1, \dots, n$, $F(x)$ – функция распределения с.в. $(X_1 - MX_1)/\sqrt{DX_1}$ Гипотеза $H_0 : F(x) = \Phi(x) \sim \mathcal{N}(0, 1)$. Проверить H_0 с помощью критерия Андерсона-Дарлинга и сравнить с его с критерием Колмогорова, нарисовать график p-value для $k = 1, 2, 10; n = 10, 50, 200$.
Алгоритм действий.
1) Генерировать $Y_1, \dots, Y_k \sim R[0, 1]$, вычислить X_1. Повторить n раз.
2) К выборке $X_1, \dots, X_n$ применить scipy.stats.anderson и вычислить $p - value$.
3) Повторить 1)-2) m раз (m=100,500), получится массив $p_1, \dots, p_m$. Упорядочить их и построить график (эмпирическую функцию распределения) этих p-value.
(x – упорядоченные p_j, y : makelist(j/m,j,1,m).
4) Проанализировать графики при разных k, n, m.
5) Построить при разных k, n, m на одном графике эмпирическую функцию распределения p-value для критериев Колмогорова и Андерсона–Дарлинга, сравнить.

5.3. $X_1, ..., X_n \sim Poiss(\theta)$. Построить РНМ критерий для проверки $H_0 : \theta = \theta_0$ против $H_1 : \theta > \theta_0$. Найти мощность критерия. Построить графики $c_\alpha(\theta)$, $\gamma(\theta)$, $\beta(\theta)$ для $\alpha = 0.01$, 0.05. Построить графики $p(\alpha)$, $\beta(\alpha)$.

5.4. С помощью критерия хи-квадрат проверить гипотезу о равномерном распределении цифр числа π. Построить гистограмму частот появления цифр.