Практикум 1. Описательная статистика

Нулевки

1. Сгенерировать реализацию выборки из распределения 1)$R[2,4]$, 2) $exp(2)$, 3) $\mathcal{N}(1,4)$ размера 50, вывести на экран первые 5 значений.

2. Найти выборочное среднее, стандартное отклонение, медиану.

3. Построить гистограмму.

4. Построить график плотности и функции распредления.

Первый вариант

1.1. Проделать следующие операции с массивом данных "flights":
1) вывести количество столбцов и строк;
2) вывести название третьего столбца;
3) вывести, сколько было пассажиров в июне 1952;
4) создать DataFrame из строк, в которых число перевезенных пассажиров не меньше 165;
5) создать DataFrame для 1951–1953 годов;
6*) найти суммарное число пассажиров за каждый год.
7*) создать DataFrame, в котором столбцами будут названия месяцев (данные – число пассажиров, строки – год);

1.2. Для массива данных "flights"
1) построить столбцовую диаграмму числа перевезенных пассажиров за 1952 год;
2) построить столбцовые диаграммы для разных лет, сравнить;
3) построить столбцовую диаграмму числа перевезенных пассажиров в марте;
4) сравнить 1951–1955 годы с помощью параллельных координат.

1.3. Независимые с.в. $X, Y, Z \sim \mathcal{N}(0,1)$. Построить диаграммы рассеяния для реализаций $(X-Y, Y)$ и $(X-Y,\ Z)$, сравнить. Построить одну диаграмму рассеяния для реализаций этих векторов, раскрасив точки в два цвета. Построить также ядерные оценки двумерных плотностей (kdplot()) для реализаций этих векторов и сравнить.

Второй вариант

2.1. Проделать следующие операции с массивом данных "tips":
1) вывести количество столбцов и строк;
2) вывести название второго столбца;
4) создать DataFrame из строк, в которых total_bill больше 10;
5) создать DataFrame для чаевых по воскресеньям;
6*) найти суммарные чаевые по дням недели.
7*) создать DataFrame, в котором столбцами будут названия дней недели, данные – суммарный размер чаевых, строки – курящие мужчины, некурящие мужчины, курящие женщины, некурящие женщины;

2.2. Для массива данных "tips"
1) построить диаграмму рассеяния для total_bill и tip, цветом задать время, формой точек – пол;
2) построить boxplot() для размера чаевых по дням недели.
3) сравнить курящих женщин, некурящих женщин, курящих мужчин и некурящих мужчин с помощью параллельных координат.

2.3. 1) Моделировать 1000 реализаций с.в. $X \sim Bin(100,\ 0.04)$ и с.в. $Y \sim Poisss(4)$. Построить столбцовые диаграммы отдельно и на одном графике, сравнить. То же для $X \sim Bin(1000,\ 0.004)$ и $X \sim Bin(10,\ 0.4)$.
2)С.в. $X \sim Bin(100,\ 0.4)$. Моделировать 1000 реализация с.в.$(X - \mathbf{E}X)/\sqrt{\mathbf{D}X}$, построить distplot().

Третий вариант

3.1. Проделать следующие операции с массивом данных "titanic":
1) вывести количество столбцов и строк;
2) вывести название пятого столбца;
4) создать DataFrame для людей, старших 19;
5) создать DataFrame для выживших пассажиров второго класса;
6*) найти средний fare по классам;
7*) создать DataFrame для числа выживших пассажиров: столбцы – pclass, строки – men, women, child.

3.2. Для массива данных "titanic"
1) построить диаграмму рассеяния для age и fare, цветом задать класс, формой точек – пол;
2) построить boxplot() для fare по классам.
3) сравнить выживших и невыживших пассажиров с помощью параллельных координат, используя столбцы survived, pclass, who, age, fare.

3.3. 1) Моделировать 1000 реализация с.в. $X \sim Cauchy$, и с.в. $Y \sim \mathcal{N}(0,1)$. Построить гистограммы отдельно и на одном графике, сравнить.
2) Независимые с.в. $X, Y \sim\sim \mathcal{N}(0,1)$. Моделировать 1000 реализация с.в.$(X+Y,\ X-Y)$, построить ядерную оценку двумерной плотности (kdplot()).