Описание характеристик задания №1

Дискретный вариационный ряд – это структура содержащая в себе значение колонки иследуемой выборки и частоту данного значение в отсорторованном по возрастанию виде.

Среднее значение выборки (Mean) – среднее арифметическое среди всех значений выборки.

Мода (Mode) – наиболее часто встречаемое значение(я) в выборке.

Медиана (Median) – значение что стоит на центральной позиции в выборке.

Дисперсия (Dispersion) – значение, что характеризует меру отклонения значений от среднего арифметического (Не применимо для непосредственного сравнения с значениями).

Среднее квадратическое (Standard deviation) - значение, что характеризует меру отклонения значений от среднего арифметического (Применимо для непосредственного сравнения с значениями).

Размах (Range) – значение, что является разностью максимального и минимального значений выборки. Оценивает размер рассматриваемого интервала.

Коефициент ассимметрии (Coefficient of skew) – значение, что характеризует ассиметрию распределения ряда относительно среднего значения. (Положительное значение близкое к нулю означает, что ряд почти симметрично расположен вокруг его среднего значения, но его правая часть немного длиннее)

Коефициент эксцесса (Coefficient of kurtosis) – значение, что характеризует остроту пика распределения ряда. (Отрицательное значение означает, что пик довольно гладкий, то есть нет частоты, что сильно выбивалась бы из общей картины)

Полигон частот – график что отображает зависимость частот значений от самих значений.

Кумулятивная кривая - график что отображает зависимость накопленных частот значений от самих значений.

Мои выводы: частоты случайной величины распределены довольно равномерно вокруг среднего значения. Значений, что имели бы значительно превосходящую частоту от остальных нет (острых пиков). Значений, что могли бы считаться аномальными нет. (Речь идет о значений что не входили бы в интервал среднее значение +- 2 средних квадратических). Скорее всего эта величина имеет нормальное распределение.

Описание результатов задания №2

Корреляция между колонками первой таблицы равна 0.960247157312

Корреляция между колонками второй таблицы равна 0.0310293561543

Для расчетов я использовал корреляцию Пирсона, т.к. корреляция Кендалла и Спирмена применяется для значений, которые можно разбить по рангам(большему значению – больший ранг). А данные во второй таблице в первой колонке никак нельзя разбить по рангам для получения адекватного представления выборки. Так что пришлось уповать на корреляцию Пирсона, что работает с величинами, измеренными в одной и той же шкале. (Хотя видя данные я не сильно рассчитывал на адекватные результат).  
Вывод: колонки в первой таблице имеют сильную зависимость одна от другой. Во второй же таблице данные в колонках никак не связаны.

Описание результатов задания №3

Для выполнения этого задания первоначально я использовал стандартную линейную(т.к. данные имеют довольно линейную зависимость) регрессию из пакета sklearn и получил коэффициенты а=15.905656572 b=37.0734370687. Однако провизуализировав предоставленные данные и его регрессионную прямую я решил, что точность недостаточно велика. Потому я решил построить собственную регрессию. Коэффициенты вышли a = 16.0000102585 b = 5.48501031981. Однако на сей раз регрессионная прямая совпадала с датасетом куда больше. Потому моими ответами будут a = 16.0000102585 b = 5.48501031981.

Описание результатов задание №4

Первое с чего бы стоило начать это определить характеристики по которым я буду характеризовать пассажиропоток и движение транспорта. Все данные я беру раздельно по маршрутам, т.к. показатели от маршрута к маршруту сильно отличаются и то, что для одного маршрута считается нормальным для другого может быть аномальным и соответственно будет мешать при анализе.   
Характеристикой движения транспорта будет интервал между прибытиями на станцию. Других параметров из приведенной выборки данных просто не может быть.   
Характеристикой же пассажиропотока будет кол-во пассажиров что зашли в текущий троллейбус за интервал между прибытием пред идущего троллейбуса и текущего.

Маршрут на который поступил жалоба определенно №51. Причин у жалобы сразу несколько:  
1) Начиная приблизительно с 20.40 и до 20.53 значительно повышенная концентрация новоприбывших людей, что ожидают троллейбуса №51. В то время как у всех маршрутов кол-во ожидающих троллейбуса снижается или остается той же, что и была (очень редко повышается на 1-2 за минут 5-6), у 51 маршрута с 20.42 до 20.53 кол-во ожидающих выросло на 8 (пусть даже и за 11 минут это все равно вдвое выше максимального повышения в другое время). Это уже сулит проблемы, т.к. в такой период, где повышенная концентрация людей стоит запускать дополнительный троллейбус.  
2) Следующий после 20.42 троллейбус маршрута №51 прибыл аж в 20.53. Интервал между ними составил 11 минут, что на 4 минуты выше среднего значения между интервалами (такое отклонение не обычно. Обычное составляет 1 минуту в верхнюю или нижнюю сторону). Следующий же троллейбус после 20.53 приехал в 20.56 и интервал между ними меньше среднего на 3 минуты, что тоже довольно необычно. Из выше сказанного могу сделать вывод, что троллейбус приехавший в 20.53 действительно опоздал на 3-4 минуты. Или же расписание было составлено нелогично.