ДЗ 1 задача 4

Юдина Екатерина БПИ198

Вариант 1

У с л о в и е :

Файл «youtube_1.csv» содержит следующие сведения о видеороликах на YouTube (сто роликов):

n — номер наблюдения,

id — идентификатор ролика,

framerate — число кадров в секунду,

frames — общее число кадров в видео,

bitrate — битрейт, Кбит/сек.

duration — продолжительность, сек.

size — размер видеофайла, байт.

---

1)Для признаков framerate, frames, bitrate, duration и size рассчитайте две корреляционные матрицы — на основании коэффициентов Пирсона и Спирмена.

2)Оцените значимость каждого коэффициента (проверьте гипотезу об отсутствии корреляции)

3)Представьте полученные результаты в виде таблицы

Коэффициенты корреляции Пирсона.

	framerate	**frames**	**bitrate**	**duration**	**size**
framerate	1.00	0.08	–0.02	0.04	0.02
frames	0.08	1.00	0.12	0.45**	0.29*
bitrate	–0.02	0.12	1.00	–0.03	0.72***
duration	0.04	0.45**	–0.03	1.00	0.36**
size	0.02	0.29*	0.72***	0.36**	1.00

* — коэффициент значим на уровне 5%,
** — коэффициент значим на уровне 1%,
*** — коэффициент значим на уровне 0.1%.
Коэффициенты, не отмеченные звёздочками, незначимы (нет оснований отвергнуть гипотезу об отсутствии корреляции на уровне 5%).

P.S. Сравните коэффициенты Пирсона и Спирмена, обратите внимание на случаи, когда два этих коэффициента существенно расходятся, если такие есть. Что такое «существенно», решайте сами. В случае существенного расхождения постройте диаграммы разброса для тех пар признаков, тесноту связи между которыми коэффициенты измеряют по-разному, и попытайтесь объяснить причину расхождения. Если вы не видите никаких существенных расхождений между двумя

матрицами, просто постройте диаграмму рассеяния для случая, где разность коэффициентов Пирсона и Спирмена наибольшая.

Р е ш е н и е

_

►Пирсон

◆ Рассчитаем коэффициенты Пирсона по формуле:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 * \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Будем называть $r_{X,Y}$(выборочным) коэффициентом корреляции Пирсона или выборочной корреляцией

Все вычисления производились в google.colab

◆ Коэффициенты Пирсона:

$$r_{framerate,framerate} = 1.0000$$

$$r_{framerate,frames} = 0.2676$$

$$r_{framerate,bitrate} = 0.2360$$

$$r_{framerate,duration} = 0.0908$$

$$r_{framerate,size} = 0.1919$$

$$r_{frames,frames} = 1.0000$$

$$r_{frames,bitrate} = 0.1752$$

$$r_{frames,duration} = 0.9464$$

$$r_{frames,size} = 0.8747$$

$$r_{bitrate,bitrate} = 1.0000$$

$$r_{bitrate,duration} = 0.1160$$

$$r_{bitrate,size} = 0.4582$$

$$r_{duration,duration} = 1.0000$$

$$r_{duration,size} = 0.7800$$

$$r_{size,size} = 1.0000$$

◆ Проверим гипотезы о независимости признаков (об отсутствии корреляции) с помощью коэффициента корреляции Пирсона.

Пусть $\rho = Corr(X_i, Y_i)$ — истинный (т.е. теоретический, не выборочный) коэффициент корреляции между (X_i, Y_i)

Гипотезы:

H_0 - $\rho = 0$ (X_i и Y_i независимы)

H_A - $\rho \neq 0$ (X_i и Y_i зависимы)

Условие:

Пусть $X_i \sim N(\mu_X, \sigma_X^2)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$, все случайные величины независимы. Тогда воспользуемся следующей статистикой..

Статистика:

$$t = \frac{r_{X,Y}\sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}}$$

Критическое правило:

Основная гипотеза, отвергается, если $|t| > t_{n-2,\frac{\alpha}{2}}$, где — α уровень значимости

[$t_{n-2,\frac{\alpha}{2}}$ -- распределение Стьюдента :)].

◆ Рассчитаем статистики для всевозможных пар признаков видеороликов с youtube

$$t_{12(framerate,frames)} = 2.75$$

$$t_{13(framerate,bitrate)} = 2.40$$

$$t_{14(framerate,duration)} = 0.90$$

$$t_{15(framerate,size)} = 1.94$$

$$t_{23(frames,bitrate)} = 1.76$$

$$t_{24(frames,duration)} = 28.99$$

$$t_{25(frames,size)} = 17.86$$

$$t_{34(bitrate,duration)} = 1.16$$

$$t_{35(bitrate,size)} = 5.10$$

$$t_{45(duration,size)} = 12.34$$

◆Теперь для каждой статистики сделаем вывод (примем или отвергнем основуню гипотезу) на уровне значимости 5%,1%,0.1%.

Табличные значения:

$t_{n-2,\frac{\alpha}{2}} = t_{98,\frac{0.05}{2}} = 1.98$

$t_{n-2,\frac{\alpha}{2}} = t_{98,\frac{0.01}{2}} = 2.63$

$t_{n-2,\frac{\alpha}{2}} = t_{98,\frac{0.001}{2}} = 3.39$

◆ Составим таблицу согласно полученным значениям

На диагонали будут стоять единицы (так как диаганальные значения отражают связь между одним и тем же признаком ⟶ связь прямая и строго линейная)

Коэффициент корреляции Пирсона					
-_-	framerate	frames	bitrate	duration	size
framerate	1	0.2676 **	0.236 *	0.0908	0.1919
frames	0.2676 **	1	0.1752	0.9464 ***	0.8747 ***
bitrate	0.236 *	0.1752	1	0.116	0.4582 ***
duration	0.0908	0.9464 ***	0.116	1	0.78 ***
size	0.1919	0.8747 ***	0.4582 ***	0.78 ***	1

*	значим на 5 %
**	значим на 1 %
***	значим на 0.1 %
	незначим (нет оснований отвергнуть гипотезу об отсутствии корреляции на уровне 5 %)

◆ *Пример вывода для таблицы Пирсона*

1. Для параметров duration (продолжительность) и frames (общее число кадров)

Есть достаточные основания отвергнуть основную гипотезу и считать что велечины зависимы на уровне значимости 0.1%. (Корреляция значима на уровне 0.1%). Связь между продолжительностью видео и общим числом кадров почти линейна.

1. Для параметров size и framerate (число кадров в секунду)

Нет оснований отвергнуть основную гипотезу и считать, что размер видеофайла зависит от его числа кадров в секунду в этом ролике. Корреляция между размером и числом кадров в секунду не отличается значимо от нуля на уровне 5%. (Корреляция не значима на уровне 5%)

Мы не можем сделать вывод, что связи между размером и числом кадров в секунду нет. Скорее, у нас просто недостаточно наблюдений, чтобы с уверенностью говорит о наличии связи, или могли произойти выбросы .. так как например, критерий Спирмена все-таки дает нам основание отвергнуть основную гипотезу для данной пары.. но об этом во второй части нашего выпуска.

►Спирмен

Проделаем аналогичные шаги для построения таблицы Спирмена

◆ Коэффициент ранговой корреляции Спирмена между признаками X и Y — это коэффициент Пирсона между рангами наблюдений по X и по Y:

$$r^{S}_{X,Y} = r_{rank(X),rank(Y)} = \frac{\sum_{i=1}^{n}(rank(X_i) - \overline{rank(X)})(rank(Y_i) - \overline{rank(Y)})}{\sqrt{\sum_{i=1}^{n}(rank(X_i) - \overline{rank(X)})^2 * \sum_{i=1}^{n}(rank(Y_i) - \overline{rank(Y)})^2}}$$

Здесь $rank(X_i)$ и $rank(Y_i)$— ранги наблюдения по X и по Y соответственно.

Все вычисления производились в google.colab

◆ Коэффициенты ранговой корреляции Спирмена:

$$r^{s}_{11(framerate,framerate)} = 1.0000$$

$$r^{s}_{12(framerate,frames)} = 0.3655$$

$$r^{s}_{13(framerate,bitrate)} = 0.4343$$

$$r^{s}_{14(framerate,duration)} = 0.1063$$

$$r^{s}_{15(framerate,size)} = 0.3823$$

$$r^{s}_{23(frames,bitrate)} = 0.1745$$

$$r^{s}_{24(frames,duration)} = 0.9391$$

$$r^{s}_{25(frames,size)} = 0.6728$$

$$r^{s}_{34(bitrate,duration)} = 0.0215$$

$$r^{s}_{35(bitrate,size)} = 0.7849$$

$$r^{s}_{45(duration,size)} = 0.5793$$

◆ Проверим гипотезы о независимости признаков с помощью коэффициента корреляции Спирмена.

Пусть $\rho = Corr(X_i, Y_i)$ — истинный (т.е. теоретический, не выборочный) коэффициент корреляции между (X_i, Y_i)

Гипотезы:

H_0 - $\rho = 0$ (X_i и Y_i независимы)

H_A - $\rho \neq 0$ (X_i и Y_i зависимы)

Условие:

Пусть в выборке $(X_1, Y_1), . . (X_N, Y_N)$ пары (X_i, Y_i) независимы и одинаково распределены. Тогда воспользуемся следующей статистикой..

Статистика:

$$t = \frac{r^S_{X,Y}\sqrt{n-2}}{\sqrt{1-(r^S_{X,Y})^2}}$$

Критическое правило:

Основная гипотеза, отвергается, если $|t| > t_{n-2,\frac{\alpha}{2}}$, где — α уровень значимости

[$t_{n-2,\frac{\alpha}{2}}$ -- распределение Стьюдента :)].

◆ Рассчитаем статистики (с использованием коэффициента Спирмена) для всевозможных пар признаков видеороликов с youtube

$$t_{12(framerate,frames)} = 3.89$$

$$t_{13(framerate,bitrate)} = 4.77$$

$$t_{14(framerate,duration)} = 1.06$$

$$t_{15(framerate,size)} = 4.10$$

$$t_{23(frames,bitrate)} = 1.75$$

$$t_{24(frames,duration)} = 27.06$$

$$t_{25(frames,size)} = 9.00$$

$$t_{34(bitrate,duration)} = 0.21$$

$$t_{35(bitrate,size)} = 12.54$$

$$t_{45(duration,size)} = 7.04$$

Коэффициент корреляции Спирмена					
_	framerate	frames	bitrate	duration	size
framerate	1	0.3655 ***	0.4343 ***	0.1063	0.3823 ***
frames	0.3655 ***	1	0.1745	0.9391 ***	0.6728 ***
bitrate	0.4343 ***	0.1745	1	0.0215	0.7849 ***
duration	0.1063	0.9391 ***	0.0215	1	0.5793 ***
size	0.3823 ***	0.6728 ***	0.7849 ***	0.5793 ***	1

*	значим на 5 %
**	значим на 1 %
***	значим на 0.1 %

	незначим (нет оснований отвергнуть гипотезу об отсутствии корреляции на уровне 5 %)

◆ *Пример вывода для таблицы Спирмена*

1. Для параметров duration (продолжительность) и frames (общее число кадров)

Есть достаточные основания отвергнуть основную гипотезу и считать что велечины зависимы на уровне значимости 0.1%. (Корреляция значима на уровне 0.1%). Связь между продолжительностью видео и общим числом кадров монотонна (почти строго монотонна) и значения сосредоточены тесно () .

1. Для параметров size (размер) и framerate (число кадров в секунду)

Есть достаточные основания отвергнуть основную гипотезу и считать что размер и число кадров в секунду зависимы на уровне значимости 0.1%.

В данном случае мы сделали противоположный вывод, тому что получислось при исследовании с помощью коэффициента Пирсона. Коэффициент Спирмена нечувствителен к выбросам в отличие от коэффициента Пирсона.

Для наглядности построим график рассеяности в третьей части нашей программы..

► Сравним таблицы:

Коэффициент корреляции Пирсона (r)					
-_-	framerate	frames	bitrate	duration	size
framerate	1	0,2676	0,236	0,0908	0,1919
frames	0,2676	1	0,1752	0,9464	0,8747
bitrate	0,236	0,1752	1	0,116	0,4582
duration	0,0908	0,9464	0,116	1	0,78
size	0,1919	0,8747	0,4582	0,78	1

Коэффициент корреляции Спирмена (rs)					
_	framerate	frames	bitrate	duration	size
framerate	1	0,3655	0,4343	0,1063	0,3823
frames	0,3655	1	0,1745	0,9391	0,6728
bitrate	0,4343	0,1745	1	0,0215	0,7849
duration	0,1063	0,9391	0,0215	1	0,5793
size	0,3823	0,6728	0,7849	0,5793	1

Разница между коэффициентами |r-rs|					
0_o	framerate	frames	bitrate	duration	size
framerate	0	0,0979	0,1983	0,0155	0,1904
frames	0,0979	0	0,0007	0,0073	0,2019
bitrate	0,1983	0,0007	0	0,0945	0,3267
duration	0,0155	0,0073	0,0945	0	0,2007
size	0,1904	0,2019	0,3267	0,2007	0

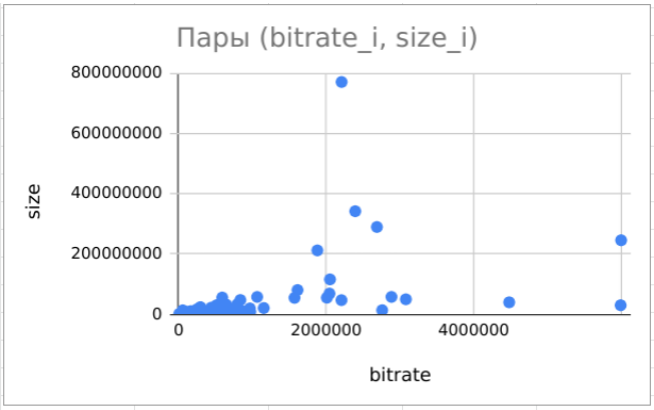
	значим на 5 %
	значим на 1 %
	значим на 0.1 %
	незначим на уровне 5 %

Можем заметить, что в принципе значения доволно близкие и иногда даже различаются меньше чем на тысячную. Но также присутствуют и расхождения, они чаще всего отмечаются там, где существуют так называемые 'выбросы' -значения которые выбиваются из общей массы.

Самые большие разницы между значениями коэффициентов присутствуют в парах, в которые входит параметр 'size', также доволно-таки сильно изменилось значение коэффициента 'bitrates+frames'

Поробуем выяснить причину, посмотрев на диаграммы рассеивания

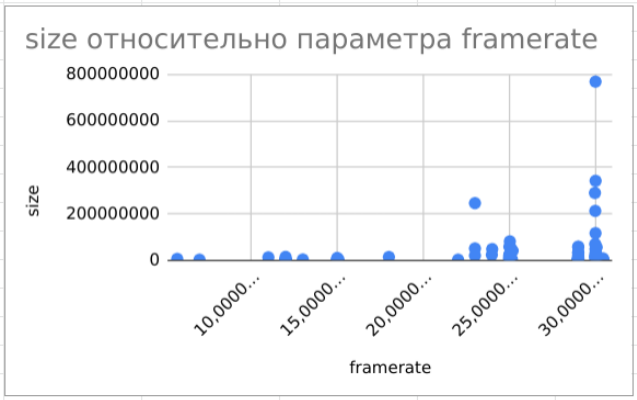
Рассмотрим одну из них с ('size,bitrate', так как судя по вычисленным разницам, здесь самое большое расхождение)

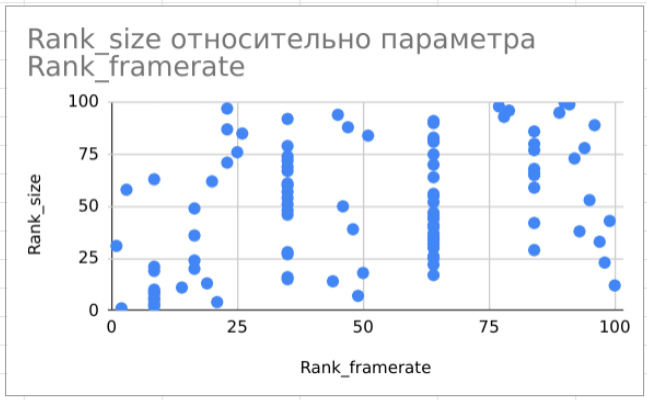


Действительно существуют выбросы, которые лежат поодаль от общей массы значений, что повлияло на значение коэффициента Пирсона.

Все $_$.

PS Ну, и еще один график для параметров size (размер) и framerate (число кадров в секунду) (те параметры для которых получились разные выводы при использовании разных коффицентов:





Можем увидеть небольшую тенденцию показывающую зависмость этих двух параметров на первом графике рассеивания, и относительно монотоноое распределение величин на втором и именно это показывает коэффициент Пирсона.