ФГАОУ ВО «УрФУ им. Первого Президента России Б. Н. Ельцина»

**Статистическое исследование зависимости бюджета, выделяемого на фильмы, от года их производства**

Студент(ка): Кочергин Михаил Александрович

Преподаватель: Анашкина Екатерина Дмитриевна

Группа: РИ-220934

2023

Содержание

Список обозначений

Введение

I. Работа с одномерной выборкой

II. Работа с двумерной выборкой

Список обозначений

N – кол-во измерений

Xmin, Ymin – минимальные значения выборочных совокупностей

Xmax, Ymax – максимальные значения выборочных совокупностей

k – кол-во интервалов

h – размах интервала

nix, niy – частота появления

Xср, Yср – среднее значение интервала выборки

wi – плотность частоты

F’(x) – эмпирическая функция распределения

Xв, Yв – среднее выборочное

Dв – выборочная дисперсия

S2 – исправленная дисперсия

S – исправленное среднее квадратичное отклонение

А - асимметрия

Е - эксцесс

Χк2 – критический критерий Пирсона

Χн2 – наблюдаемый критерий Пирсона

pi – теоретическая вероятность

α – уровень значимости

r – число степеней свободы

m – кол-во зависимых критериев

Kxy – коэффициент ковариации

rxy – коэффициент корреляции

Введение

В данной расчётно-графической работе исследуется зависимость бюджета, выделенного на съёмки фильмов, выпущенных в период с 1970 по 2023 год от года выпуска.

Случайная величина X – год выпуска фильма.

Случайная величина Y – бюджет в миллионах долларов США.

Цель работы: установить вид распределения СВ X, проследить зависимость случайных величин.

Данные для выборки были взяты с сайта “Кинопоиск” при помощи специального сервиса для подбора случайных фильмов: <https://www.kinopoisk.ru/chance/>

I. Работа с одномерной выборкой

Выборка из собранных данных представлена в таблице:

|  |  |  |  |
| --- | --- | --- | --- |
| X | Y | X | Y |
| 1972 | 3,5 | 1995 | 16,5 |
| 1975 | 1,2 | 1997 | 70 |
| 1978 | 2,3 | 1998 | 60 |
| 1978 | 1,5 | 1998 | 95 |
| 1980 | 4,5 | 1999 | 23 |
| 1981 | 10 | 2000 | 12 |
| 1981 | 12 | 2000 | 90 |
| 1982 | 12 | 2000 | 41 |
| 1982 | 12 | 2000 | 25 |
| 1984 | 6,5 | 2000 | 14 |
| 1984 | 22 | 2001 | 58 |
| 1985 | 9 | 2001 | 27 |
| 1985 | 15 | 2001 | 53 |
| 1986 | 6 | 2002 | 21 |
| 1987 | 20 | 2002 | 4 |
| 1988 | 17 | 2002 | 52 |
| 1989 | 22,5 | 2002 | 102 |
| 1989 | 26 | 2003 | 79 |
| 1989 | 16 | 2005 | 15 |
| 1990 | 31 | 2005 | 126 |
| 1990 | 22 | 2005 | 113 |
| 1991 | 28 | 2005 | 22 |
| 1991 | 17 | 2006 | 8,3 |
| 1991 | 26 | 2006 | 1 |
| 1992 | 32 | 2006 | 11 |
| 1992 | 4,5 | 2006 | 16,5 |
| 1993 | 30 | 2007 | 110 |
| 1994 | 5 | 2007 | 12 |
| 1994 | 30 | 2007 | 25 |
| 1995 | 65 | 2007 | 15 |
| 1995 | 17 | 2008 | 32 |
| 1995 | 90 | 2008 | 12,5 |
| 1995 | 18 | 2008 | 150 |
| 2008 | 2,9 | 2015 | 40 |
| 2009 | 130 | 2016 | 22,5 |
| 2009 | 5 | 2017 | 22 |
| 2010 | 38 | 2017 | 7 |
| 2010 | 6 | 2017 | 30 |
| 2010 | 18 | 2018 | 4 |
| 2011 | 8 | 2018 | 52 |
| 2011 | 15 | 2019 | 6,2 |
| 2011 | 10 | 2019 | 55 |
| 2012 | 100 | 2019 | 90 |
| 2012 | 16,6 | 2019 | 18 |
| 2012 | 26 | 2019 | 14 |
| 2013 | 3 | 2020 | 250 |
| 2013 | 130 | 2021 | 165 |
| 2014 | 50 | 2022 | 1,7 |
| 2015 | 75 | 2022 | 350 |

1. Обработаем данные для случайной величины X.

N = 100; Xmin = 1972; Xmax = 2022

По формуле Стерджесса: k = 1 + 3,32 \* lg(N) = 7

h = (Xmax – Xmin) / k = (2022 – 1972) / 7 = 7

2. Разобьём выборку на интервалы и занесём данные в таблицу, где Xср – среднее значение, nix – частота появления, wi –плотность частоты.

|  |  |  |  |
| --- | --- | --- | --- |
| Интервал | Xср | ni | wi |
| [1972-1979] | 1976 | 4 | 0,5 |
| (1979-1986] | 1983 | 11 | 1,57 |
| (1986-1993] | 1990 | 14 | 2 |
| (1993-2000] | 1997 | 16 | 2,28 |
| (2000-2007] | 2004 | 20 | 2,86 |
| (2007-2014] | 2011 | 18 | 2,57 |
| (2014-2022] | 2018 | 17 | 2,13 |

3. По полученным данным построим гистограмму и полигон частот.

Изображение выглядит как График, линия, текст, диаграмма

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, Шрифт

Автоматически созданное описание

4. Найдём эмпирическую функцию распределения и построим её график.

Изображение выглядит как текст, снимок экрана, число, линия

Автоматически созданное описание

5. Посчитаем числовые характеристики СВ Х.

Среднее выборочное:

Xв =

Выборочная дисперсия:

Dв =

Исправленная дисперсия:

S2 =

Исправленное среднее квадратичное отклонение:

S =

Асимметрия:

А =

Эксцесса:

E=

Вывод: исходя из построенного полигона и гистограммы частот, предполагаем, что распределение для выборки согласуется с нормальным распределением. График имеет левостороннюю симметрию, т.к A<0, а также более плосковершинную форму, т.к E<0.

6. Прежде чем приступать к постановке гипотез, проведём аналогичные вычисления для СВ Y.

N = 100; Ymin = 1; Ymax = 350

По формуле Стерджесса: k = 1 + 3,32 \* lg(N) = 7

h = (Xmax – Xmin) / k = (2022 – 1972) / 7 = 7

|  |  |  |  |
| --- | --- | --- | --- |
| Интервал | Yср | ni | wi |
| [1-50] | 26 | 75 | 1,47 |
| (50-100] | 75 | 15 | 0,3 |
| (100-150] | 125 | 7 | 0,14 |
| (150-200] | 175 | 1 | 0,02 |
| (200-250] | 225 | 1 | 0,02 |
| (250-300] | 275 | 0 | 0 |
| (300-350] | 325 | 1 | 0,02 |

Среднее выборочное:

Yв =

Выборочная дисперсия:

Dв =

Исправленная дисперсия:

S2 =

Исправленное среднее квадратичное отклонение:

S =

7. Перейдём к постановке гипотез.

Выдвинем нулевую и обратную гипотезы:

H0 – распределение СВ X согласуется с нормальным распределением

H1 – распределение СВ Х не является нормальным

8. Для того, чтобы проверить наши гипотезы найдём наблюдаемый критерий согласия Пирсона и сравним его с критическим.

Вычислим число степеней свободы и возьмём некоторые значения уровня значимости чтобы найти критический критерий Пирсона.

r = k – (m + 1) = 7 – (2 + 1) = 4

Для α = 0,01; 0,025; 0,05 и r=4 Xк2 соответственно равен:

Xк2 = 13,3; 11,1; 9,5

Найдём теоретическую вероятность pi и вычислим Xн2.

Для нормального распределения:

pi(xi-1 < Xcp < xi) =

p1(1972 < Xcp < 1979) =

p2(1979 < Xcp < 1986) =

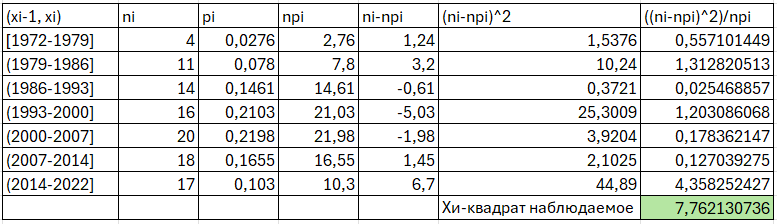
p3(1986 < Xcp < 1993) =

p4(1993 < Xcp < 2000) =

p5(2000 < Xcp < 2007) =

p6(2007 < Xcp < 2014) =

p7(2014 < Xcp < 2022) =



Xн2 =

Xн2 < Xк2(0,01; 4) следовательно гипотеза H0 верна при уровне значимости α = 0,01

9. Выдвинем новые гипотезы.

H0 – СВ Х согласуется с распределением Коши

H1 – СВ Х не согласуется с распределением Коши

Аналогично, находим наблюдаемый критерий Пирсона и сравниваем его с критическим при α = 0,01.

r = k – (m + 1) = 7 – (3 + 1) = 3

Χк2(0,01; 3) = 11,3

Для распределения Коши:

pi(xi-1 < Xcp < xi) =

p1(1972 < Xcp < 1979) =

p2(1979 < Xcp < 1986) =

p3(1986 < Xcp < 1993) =

p4(1993 < Xcp < 2000) =

p5(2000 < Xcp < 2007) =

p6(2007 < Xcp < 2014) =

p7(2014 < Xcp < 2022) =

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Xн2 =

Xн2 > Xк2(0,01; 3) следовательно гипотезу H0 мы отвергаем.

Вывод: в ходе вычислений гипотеза H0 о том, что СВ Х согласуется с нормальным законом, подтвердилась. Вместе с тем случайная величина не согласуется с распределением Коши, что также было доказано в процессе исследования.

II. Работа с двумерной выборкой

10. Для того, чтобы проверить зависимость величины Х от величины Y найдём их корреляционную зависимость, построим теоретические и эмпирические графики регрессии.

Для начала определим первые начальные и вторые центральные моменты;

Первыми начальным моментами для СВ является среднее выборочное:

Xв = 2001; Yв = 46,75

Вторыми центральными моментами для СВ является выборочная дисперсия:

Dвx=149,053; Yвх=2149,179

Построим корреляционную таблицу.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| y\x | [1972-1979] | (1979-1986] | (1986-1993] | (1993-2000] | (2000-2007] | (2007-2014] | (2014-2022] |
| [1-50] | 4 | 11 | 14 | 10 | 12 | 14 | 10 |
| (50-100] | 0 | 0 | 0 | 6 | 4 | 1 | 4 |
| (100-150] | 0 | 0 | 0 | 0 | 4 | 3 | 0 |
| (150-200] | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| (200-250] | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| (250-300] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (300-350] | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

*Kxy = XY – Xв\*Yв =*

*rxy =Kxy / (Sx\*Sy) = 181,2125 / (12,27 \* 46,35) = 0,3185*

*Хотя, зависимость величин вероятна, значение корреляционного момента 0,3185 указывает на несильную зависимость между случайными величинами.*

*Чтобы построить графики регрессии найдём теоретическую и эмпирическую линию регрессии. Допустим, что теоретическая линия является прямой.*

*Формула прямой:*

*(Y-Yв) = rxy \**

*Найдём функцию регрессии X по Y и Y по X.*

*(Y-46,75)*

*Y по X: Y = 1,203 \* X – 2360,463*

*(X-2001) =*

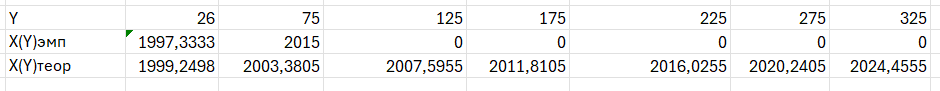
*X по Y: X = 0,0843 \* Y + 1997,058*

*Также найдём точки графика опытным путём и составим таблицу, по которой построим линии регрессии.*

**

*Изображение выглядит как линия, График, диаграмма, скат

Автоматически созданное описание*

**

*Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание*

*Общий вывод: в ходе исследований был установлен нормальный вид распределения величины X и опровергнута гипотеза о распределении Коши. Был выявлен коэффициент корреляции, который говорит о довольно слабой зависимости величин. Вероятно, это связано с тем, что величины зависят в том числе от других факторов. Графики регрессии так же говорят о слабой зависимости, однако достоверность эмпирической регрессии малодостоверна из-за отсутствия подходящих данных.*