ГУАП

КАФЕДРА № 41

ОТЧЕТ
ЗАЩИЩЕН С ОЦЕНКОЙ
ПРЕПОДАВАТЕЛЬ

канд. техн. наук, доцент		Г.В. Титова
должность, уч. степень, звание	подпись, дата	инициалы, фамилия

ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ № 2

СТАТИСТИЧЕСКАЯ ПРОВЕРКА КРИТЕРИЕВ

по курсу:

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ИНФОРМАЦИИ

РАБОТУ ВЫПОЛНИЛ

СТУДЕНТ гр. № 4318		А.Д. Борисоглебский
	подпись, дата	инициалы, фамилия

Санкт-Петербург 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Цель работы: провести статистическую проверку критериев.

Данные взяты с сайта статистики Испании, из раздела демографии. Данные содержат информацию о рождениях, естественному приросту населения и случаях внутриутробной смерти в абсолютный величинах (кол-во случаев).

Ссылка: https://www.ine.es/jaxiT3/Datos.htm?t=6567#_tabs-tabla

Также для проведения более полного исследования были взяты данные о случаях домашнего насилия в регионах, также в абсолютных величинах.

Ссылка: https://www.ine.es/dynInfo/Infografia/Territoriales/capitulo.html#!tabla

Данные по рождаемости относятся к 2023 году. Данные по насилию к 2024-му. Однако разница не будет ощутима, так как глобальные изменения в поведении населения не возникают за такие короткие промежутки времени.

1 Описание данных

Данные взяты по 19-ти регионам Испании.

Из основного источника взяты признаки: рождаемость, естественный прирост, случаи смерти плода. Из дополнительных данных взята статистика по жертвам насилия, домашнего и в общем.

Итоговые датафреймы представлены на рисунках 1–2.

	Autonomous Communities and Cities	Nacimiento	Natural growth	Late foetal deaths
0	01 Andalucia	61397.0	-13299.0	153.0
1	02 Aragon	8669.0	-5188.0	27.0
2	03 Asturias, Principado de	4545.0	-8467.0	29.0
3	04 Balears, Illes	8738.0	99.0	37.0
4	05 Canarias	11998.0	-5832.0	37.0
5	06 Cantabria	2976.0	-3056.0	2.0
6	07 Castilla y Leon	12496.0	-15606.0	29.0
7	08 Castilla - La Mancha	14075.0	-5372.0	30.0
8	09 Catalunna	54174.0	-13488.0	131.0
9	10 Comunitat Valenciana	35378.0	-11305.0	121.0
10	11 Extremadura	6810.0	-4551.0	26.0
11	12 Galicia	14004.0	-18778.0	22.0
12	13 Madrid, Comunidad de	50299.0	1653.0	94.0
13	14 Murcia, Region de	12860.0	660.0	41.0
14	15 Navarra, Comunidad Foral de	4496.0	-1351.0	19.0
15	16 Pa�s Vasco	13462.0	-8866.0	54.0
16	17 Rioja, La	1999.0	-1237.0	9.0
17	18 Ceuta	683.0	123.0	3.0
18	19 Melilla	771.0	271.0	6.0
19	Foreign	826.0	-1878.0	6.0
20	Total	320656.0	-115468.0	876.0

Рисунок 1 — Датафрейм с данными о рождаемости

	Víctimas de violencia de género	Víctimas de violencia doméstica
0	8.020	1.851
1	994	237
2	704	189
3	1.307	415
4	2.441	500
5	681	132
6	1.879	538
7	1.889	394
8	2.768	940
9	4.858	1.106
10	938	255
11	1.350	384
12	3.432	1.025
13	1.588	342
14	426	123
15	766	249
16	358	117
17	115	32
18	170	31

Рисунок 2 — Датафрейм с данными о жертвах насилия

2 Наглядное представление

Распределение данных о рождении детей и естественном приросте по регионам представлено на рисунке 3.

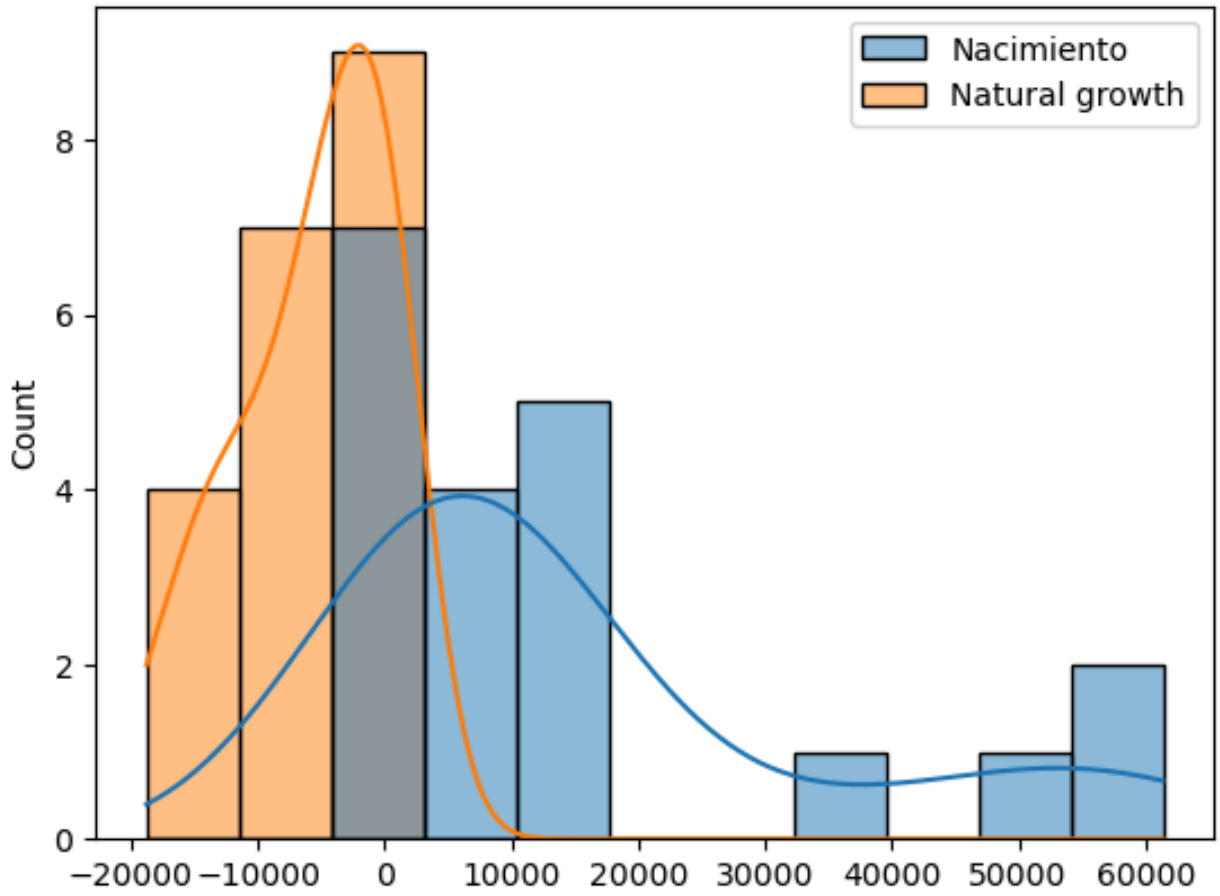

Рисунок 3 — График распределения признаков

Распределение данных о случаях смерти по регионам представлено на рисунке 4.

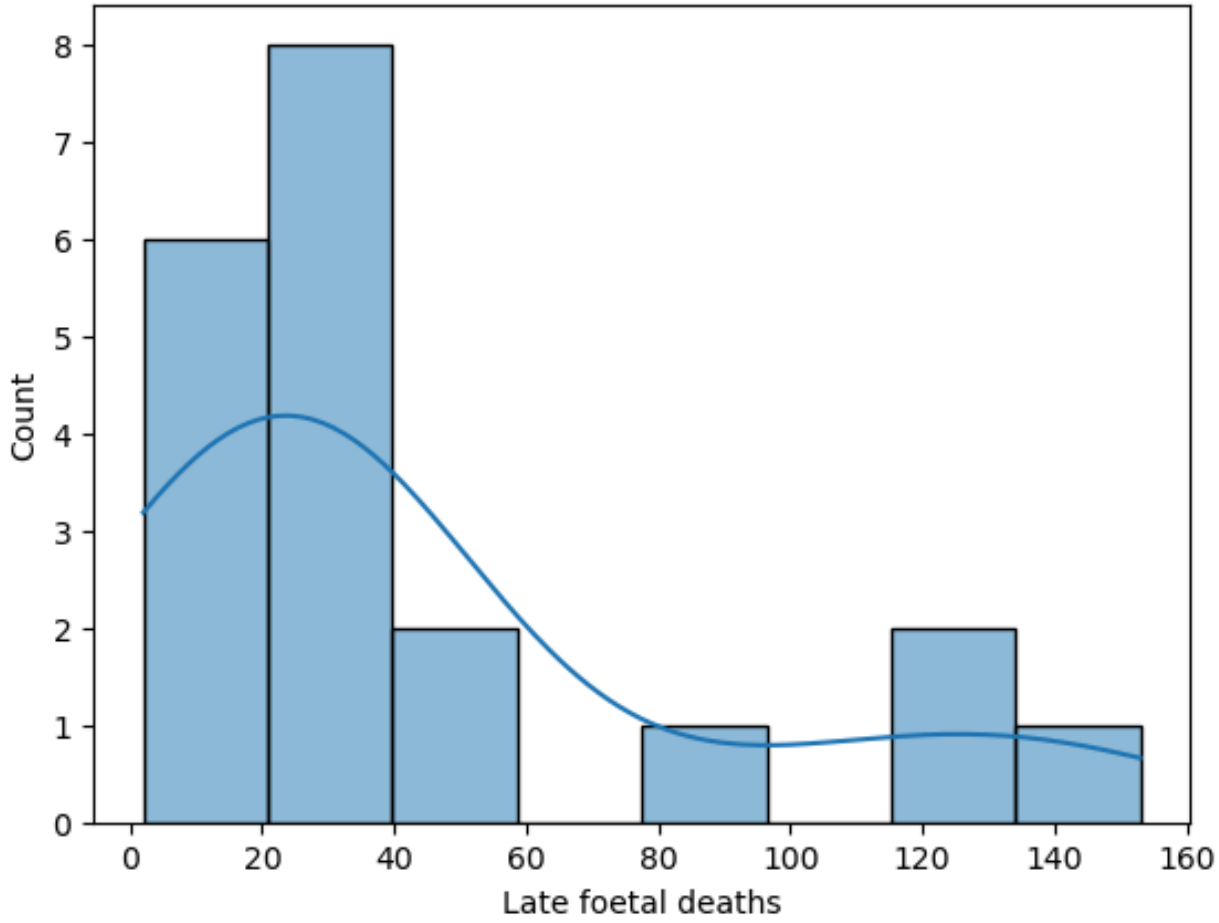

Рисунок 4 — График распределения признаков

Распределение данных о случаях домашнего и общего насилия по регионам представлено на рисунке 5.

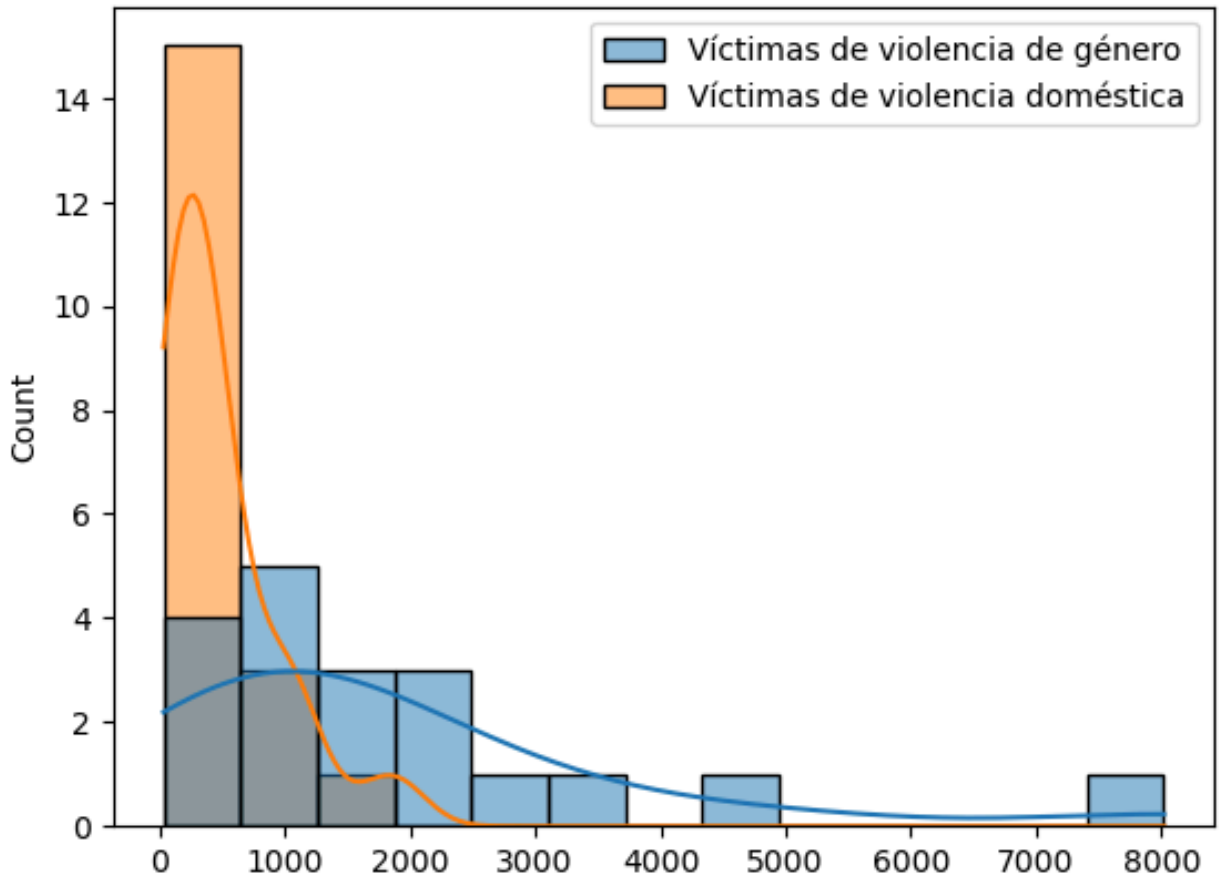

Рисунок 5 — График распределения признаков по насилию

3 Критерий Шапиро-Уилка

Для чего используется:

Критерий проверяет, является ли распределение данных нормальным. Если доказана нормальноть, то можно использовать параметричечские методы (корреляция Пирсона), а если нет — нужно использовать непараметрические (Спирмен, Манн-Уитни). По результатам этого теста решается, какие критерии применять в дальнейшем исследовании.

Условия применения:

– $3 \leq$ n ≤ 50 (при большем количестве снижается точность)

– Проверка нормальности

Для признаков: Рождения, естественный прирост, случаи смерти.

Нулевая гипотеза:

$H_0 = \{f(x) = N(a, \sigma^2)\}$

Данные распределены нормально

Альтернативная гипотеза:

$H_1 = \{f(x) \neq N(a, \sigma^2)\}$

Данные не распределены нормально

Формула:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{1}$$

где:

– $x_{(i)}$ — упорядоченные значения выборки

– a_i — коэффициенты, рассчитываемые на основе ожидаемых значений порядковых статистик нормального распределения

– $\overline{x}$ — среднее значение выборки

– n — объём выборки

Команда в Python:

```
from scipy.stats import shapiro
stat, p = shapiro(x)
```

Тест Шапиро-Уилка показал, что данные не распределены нормально (табл. 1).

Таблица 1 — Результаты критерия Шапиро-Уилка

Признак	**W**	**p-value**	**Сравнение**	**Вывод**
Рождения	0.7446	0.0002	$p \leq 0.05$	Отвергаем H_0, значит данные НЕ нормальны
Естественный прирост	0.9234	0.1306	$p \geq 0.05$	Не может отвергнуть H_0, однако выборка из 18 значений и статистика 0.92 свидетельствует о недостаточной нормальности распределения
Сметри плода	0.7983	0.0011	$p \leq 0.05$	Отвергаем H_0, значит данные НЕ нормальны
Случаи домашнего насилия	0.8029	0.0013	$p \leq 0.05$	Отвергаем H_0, значит данные НЕ нормальны

В результате все данные не подчиняются нормальному закону распределения (в полной мере). Для дальнейшего исследования используем непараметрические критерии.

4 Критерий корреляции Спирмена

Для чего используется:

Критерий показывает, есть ли связь между двумя признаками. Если один признак растёт, растёт ли вместе с ним другой (положительная связь), или наоборот — один растёт, а другой падает (отрицательная связь).

Используем для проверки связи между домашним насилием и случаями смерти детей.

Условия применения:

– n $\geq$ 5

– Непараметрический

– Не требует нормальности

Для признаков:

Домашнее насилие и случаи смерти детей. Попытаемся определить, есть ли зависимость.

Нулевая гипотеза:

$H_0 = \{\rho_s = 0\}$

Корреляция отсутствует

Альтернативная гипотеза:

$H_1 = \{\rho_s > 0\}$ — положительная корреляция

$H_2 = \{\rho_s < 0\}$ — отрицательная корреляция

$H_3 = \{\rho_s \neq 0\}$ — корреляция существует (выборано через alternative=„two-sided“)

Корреляция существует

Формула:

Коэффициент корреляции Спирмена:

$$\rho_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2}$$

где:

– d_i — разность между рангами соответствующих значений X и Y

– n — объём выборки

– $\rho_s \in [-1, 1]$ (от -1 до 1)

Команда в Python:

```
from scipy.stats import spearmanr
corr, p = spearmanr(violencia_domestica, deaths)
```

Паргументы по умолчанию:

– `axis=0` — вычисление корреляции по столбцам

– `nan_policy='propagate'` — обработка пропущенных значений (NaN тоже учитывает)

– `alternative='two-sided'` — двусторонняя альтернатива ($\rho \neq 0$)

Результат: p-value ≤ 0.05

Вывод: Отвергаем H_0. Существует высокая коррелция ($\rho = 0.8516$).

Интерпретация: Положительная корреляция. Данные изменяются вместе, однако нельзя утверждать о наличии причинно-следственной связи. Вероятнее всего есть общий показатель, влияющий на исследуемые значения (например, численность населения регионов).

5 Критерий корреляции Кендалла

Для чего используется:

Критерий Кендалла проверяет связь между двумя признаками. Основное отличие — Кендалл более устойчив к выбросам и лучше подходит для небольших выборок. Используется как дополнительная проверка корреляции. В работе применяется для анализа связи между

Условия применения:

– n ≥ 5

– Устойчив к выбросам

– Не требует нормальности

Для признаков:

Домашнее насилие и случаи смерти детей. Берём те же данные, чтобы проверить ту же гипотезу другим способом.

Нулевая гипотеза:

$H_0 = \{\tau = 0\}$

Корреляция отсутствует

Альтернативная гипотеза:

$H_1 = \{\tau \neq 0\}$

Корреляция существует

Формула:

Коэффициент корреляции Кендалла:

$$\tau = \frac{C - D}{\frac{n(n-1)}{2}} \tag{3}$$

где:

– C — количество согласованных (конкордантных) пар

– D — количество несогласованных (дискордантных) пар

– n — объём выборки

– $\tau \in [-1, 1]$ (от −1 до 1)

Команда в Python:

```
from scipy.stats import kendalltau
corr, p = kendalltau(a, b)
```

Аргументы по умолчанию:

– `method='auto'` — автоматический выбор метода вычисления (может быть asymptotic для больших выборок либо exact для точности и маленьких выборок)

– `variant='b'` — вариант статистики τ Кендалла («b» - для пар наблюдений, а «c» - для прямоугольных таблиц сопряженности)

– `alternative='two-sided'` — двусторонняя альтернатива

– `nan_policy='propagate'` — обработка пропущенных значений

Результат: p-value ≤ 0.05

Вывод: Отвергаем H_0. Корреляция существует ($\tau = 0.6647$)

Интерпретация: Положительная связь — данные изменяются согласованно. Двумя способами получен одинаковый вывод.

6 Критерий Манна-Уитни

Для чего используется:

Критерий сравнивает две группы и проверяет, различаются ли они статистически. В работе используется для сравнения естественного прироста в регионах — проверяем, действительно ли две группы отличаются друг от друга.

Условия применения:

– 2 независимые группы

– $n_1, n_2 \geq 3$

– Не требует нормальности

Для признаков:

Естественный прирост (2 группы, $n_1 = 9, n_2 = 10$).

Выборка разделена на две группы по медиане:

– Группа 1: регионы с рождаемостью < 8738

– Группа 2: регионы с рождаемостью $\geq$ 8738

Нулевая гипотеза:

$H_0 = \left\{F_{X_1}(x) = F_{X_2}(x)\right\}$

Распределения идентичны (две независимые выборки принадлежат однородным генеральным совокупностям)

Альтернативная гипотеза:

$H_1 = \left\{F_{X_1}(x) \neq F_{X_2}(x)\right\}$

Распределения различаются

Формула:

U-статистика Манна-Уитни:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{4}$$

где:

– n_1, n_2 — объёмы первой и второй выборок

– R_1 — сумма рангов элементов первой выборки в общей упорядоченной выборке

– Вычисляется также $U' = n_1 n_2 - U$, используется $\min(U, U')$

Команда в Python:

```
from scipy.stats import mannwhitneyu
stat, p = mannwhitneyu(low,high, alternative='two-sided')
```

Аргументы по умолчанию:

– `use_continuity=True` — применение поправки на непрерывность (True рекомендуется для малых выборок)

– `alternative='two-sided'` — двусторонняя альтернатива

– `method='auto'` — автоматический выбор метода вычисления

– `nan_policy='propagate'` — обработка пропущенных значений

– `keepdims=False` — не сохранять размерность массива (False - скалярный результат, True - массив)

Результат:

U = 73.0, p = 0.0247

p-value ≤ 0.05 — отвергаем H_0

Вывод: Регионы с высокой рождаемостью ($\geq$8738) имеют статистически значимо отличающийся естественный прирост по сравнению с регионами с низкой рождаемостью.

7 Медианный критерий

Для чего используется:

Критерий сравнивает медианы двух или более групп. Работает проще, чем Манн-Уитни — делит все данные на две категории (выше и ниже общей медианы) и сравнивает пропорции. Более устойчив к сильным выбросам. Используется как проверка различий между группами регионов по смертности.

Условия применения:

– 2+ группы

– Устойчив к выбросам

– Не требует нормальности

Для признаков:

Смертность (2 группы, $n_1 = 9, n_2 = 10$).

Выборка разделена на две группы по медиане:

– Группа 1: регионы с рождаемостью < 8738

– Группа 2: регионы с рождаемостью $\geq$ 8738

Нулевая гипотеза:

$H_0 = \left\{ x_1 = x_2 = ... = x_n \right\}$

Медианы всех выборок равны (несколько независимых выборок принадлежат однородным генеральным совокупностям)

Альтернативная гипотеза:

$H_1 = \{\text{хотя бы одно равенство не выполняется}\}$

Хотя бы одна медиана отличается от других

Формула:

Медианный критерий использует статистику хи-квадрат:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{5}$$

где:

– O_{ij} — наблюдаемые частоты (количество значений выше/ниже общей медианы в каждой группе)

– E_{ij} — ожидаемые частоты при условии равенства медиан

– k — количество групп

Команда в Python:

```
from scipy.stats import median_test
stat, p, med, contingency = median_test(low, high)
```

Аргументы по умолчанию:

– `ties='below'` — обработка значений, равных медиане (below - считать их ниже медианы)

– `correction=True` — применение поправки Йейтса (True - улучшает точность на малых выборках)

– `lambda_=1` — параметр для вычисления статистики (1 - стандартная статистика Пирсона X^2)

– `nan_policy='propagate'` — обработка пропущенных значений

Результат:

$\chi^2 = 6.4653$

p-value = 0.0110

Вывод: Существует статистически значимая связь между уровнем рождаемости и уровнем смертности. Регионы с низкой и высокой рождаемостью по-разному распределяются относительно медианного уровня смертности.

Интерпретация:

Группа с низкой рождаемостью (< 8738):

– 1 регион имеет смертность выше медианы;

– 8 регионов имеют смертность ниже медианы;

– 89% регионов с низкой рождаемостью имеют низкую смертность.

Группа с высокой рождаемостью (≥ 8738):

– 8 регионов имеют смертность выше медианы;

– 2 региона имеют смертность ниже медианы;

– 80% регионов с высокой рождаемостью имеют высокую смертность.

7.1 Критерий Краскела-Уоллиса

Для чего используется:

Критерий сравнивает три или больше групп одновременно. Проверяет, различаются ли группы по уровню признака. В работе используется для проверки того, отличаются ли регионы с разным уровнем домашнего насилия (низкий, ниже среднего, выше среднего, высокий) по рождаемости.

Условия применения:

– 3+ группы

– Не требует нормальности

Для признаков:

Рождаемость (4 группы). Разделение уровня насилия по квантилям.

Нулевая гипотеза:

$H_0 = \left\{F_{X_1}(x) = F_{X_2}(x) = ... = F_{X_k}(x)\right\}$

Все выборки принадлежат однородным генеральным совокупностям (медианы равны)

Альтернативная гипотеза:

$H_1 = \{\text{хотя бы одно равенство не выполняется}\}$

Хотя бы одна медиана отличается

Формула:

Н-статистика Краскела-Уоллиса:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \tag{6}$$

где:

– k — количество групп

– n — общий объём выборки ($n = \sum_{i=1}^{k} n_i$)

– n_i — объём i-й группы

– R_i — сумма рангов i-й группы в общей упорядоченной выборке

Команда в Python:

```
from scipy.stats import kruskal
stat, p = kruskal(low, lower, higher, high)
```

Аргументы по умолчанию:

– `nan_policy='propagate'` — обработка пропущенных значений

Результат:

Статистика H: 11.8121

p-value: 0.0081 $\leq$ 0.05

Вывод: Отвергаем H_0, хотя бы одна медиана отличается от других

Интерпретация: Существует статистически значимая связь между уровнем домашнего насилия и рождаемостью в регионах. Регионы с разной интенсивностью домашнего насилия систематически различаются по показателям рождаемости.

8 Вывод

В работе была проведена статистическая проверка шести критериев на данных по рождаемости 19 регионов Испании в 2023 году в дополнении с информацией о случаях домашнего насилия в каждом регионе за 2024 год.

Код реализации всех критериев на языке Python доступен по ссылке: