Студент: Давыдов Михаил
Группа: 0303
Вариант: 7
Дата: 30 мая 2022 г.

Теория вероятностей и математическая статистика

Индивидуальное домашнее задание №5

Задание 1. *Из файла population.csv для столбца ν сформировать выборку объема n согласно правилу:*

$$n = 100 + 7 mod 21 = 107$$

Выбрать программное обеспечение/язык программирования для обработки результатов. Обосновать выбор.

Решение. Для обработки результатов был выбран язык Python, потому что он лучше всего подходит для быстрого написания программ, особенно когда количество обрабатываемых данных не шибко велико (как в нашем случае). Кроме того в нем есть библиотеки для работы с выборками: например, была использована библиотека random. Получение выборки происходило путем комбинирования нескольких способ создание выборок по следующему алгоритму: 1) Все данные были разбиты на группы с шагом в 74 (чтобы группы получились равные). 2) В каждой группе случайным образом было выбрано количество элементов прямо пропорционально размеру получившейся группы таким образом, чтобы в сумме выборка была из 107 элементов.

Получившаяся выборка: 378, 386, 330, 391, 379, 390, 382, 362, 377, 392, 362, 320, 383, 372, 442, 426, 422, 446, 400, 436, 426, 437, 470, 428, 429, 405, 407, 397, 461, 463, 463, 423, 452, 459, 418, 452, 448, 458, 429, 414, 442, 470, 427, 440, 463, 463, 404, 438, 462, 428, 411, 406, 406, 443, 452, 419, 434, 470, 453, 449, 421, 411, 443, 458, 421, 457, 434, 438, 418, 493, 548, 496, 544, 512, 495, 525, 536, 500, 475, 517, 498, 525, 521, 474, 480, 502, 493, 541, 525, 501, 475, 518, 506, 483, 483, 532, 493, 510, 523, 544, 481, 553, 623, 593, 576, 585, 591 □

Задание 2. *Последовательно преобразовать выборку в ранжированный, вариационный и интервальный ряды. Результаты содержрательно проинтерпретировать и сделать выводы.*

Решение. Преобразование выборки в ранжированный ряд: 320, 330, 362, 362, 372, 377, 378, 379, 382, 383, 386, 390, 391, 392, 397, 400, 404, 405, 406, 406, 407, 411, 411, 414, 418, 418, 419, 421, 421, 422, 423, 426, 426, 427, 428, 428, 429, 429, 434, 434, 436, 437, 438, 438, 440, 442, 442, 443, 443, 446, 448, 449, 452, 452, 452, 453, 457, 458, 458, 459, 461, 462, 463, 463, 463, 463, 470, 470, 470, 474, 475, 475, 480, 481, 483, 483, 493, 493, 493, 495, 496, 498, 500, 501, 502, 506, 510, 512, 517, 518, 521, 523, 525, 525, 525, 532, 536, 541, 544, 544, 548, 553, 576, 585, 591, 593, 623.

Преобразование выборки в вариационный ряд:

x_i	n_i	P_i^*
320	1	1/107
330	1	1/107
362	2	2/107
372	1	1/107
377	1	1/107
378	1	1/107
379	1	1/107
382	1	1/107
383	1	1/107
386	1	1/107
390	1	1/107
391	1	1/107
392	1	1/107
397	1	1/107
400	1	1/107
404	1	1/107
405	1	1/107
406	2	2/107
407	1	1/107
411	2	2/107
414	1	1/107
418	2	2/107
419	1	1/107
421	2	2/107
422	1	1/107
423	1	1/107
426	2	2/107
427	1	1/107
428	2	2/107
429	2	2/107
434	2	2/107
436	1	1/107
437	1	1/107
438	2	2/107
440	1	1/107
442	2	2/107
443	2	2/107
446	1	1/107
448	1	1/107
449	1	1/107

x_i	n_i	P_i^*
452	3	3/107
453	1	1/107
457	1	1/107
458	2	2/107
459	1	1/107
461	1	1/107
462	1	1/107
463	4	4/107
470	3	3/107
474	1	1/107
475	2	2/107
480	1	1/107
481	1	1/107
483	2	2/107
493	3	3/107
495	1	1/107
496	1	1/107
498	1	1/107
500	1	1/107
501	1	1/107
502	1	1/107
506	1	1/107
510	1	1/107
512	1	1/107
517	1	1/107
518	1	1/107
521	1	1/107
523	1	1/107
525	3	3/107
532	1	1/107
536	1	1/107
541	1	1/107
544	2	2/107
548	1	1/107
553	1	1/107
576	1	1/107
585	1	1/107
591	1	1/107
593	1	1/107
623	1	1/107

Преобразование выборки в интервальный ряд:

$[x_i; x_{i+1}]$	n_i	P_i^*
320;363.2857	4	4/107
363.2857;406.571	16	16/107
406.5714;449.8571	32	32/107
449.8571;493.1429	27	27/107
493.1429;536.4286	18	18/107
536.4286;579.7143	6	6/107
579.7143;623	4	4/107

При преобразовании в вариационный ряд, выборка приобрела некоторую читаемую структуру и немного сократилась в представлении. Дальнейшее преобразование в интервальный ряд и вовсе сделало выборку компактной и удобной для представляения и анализа данных. □

Задание 3. *Для интервального ряда абсолютных частот построить и отобразить графически полигон, гистрограмму и эмпирическую функцию. Сделать выводы.*

Решение. Были простроены графики:

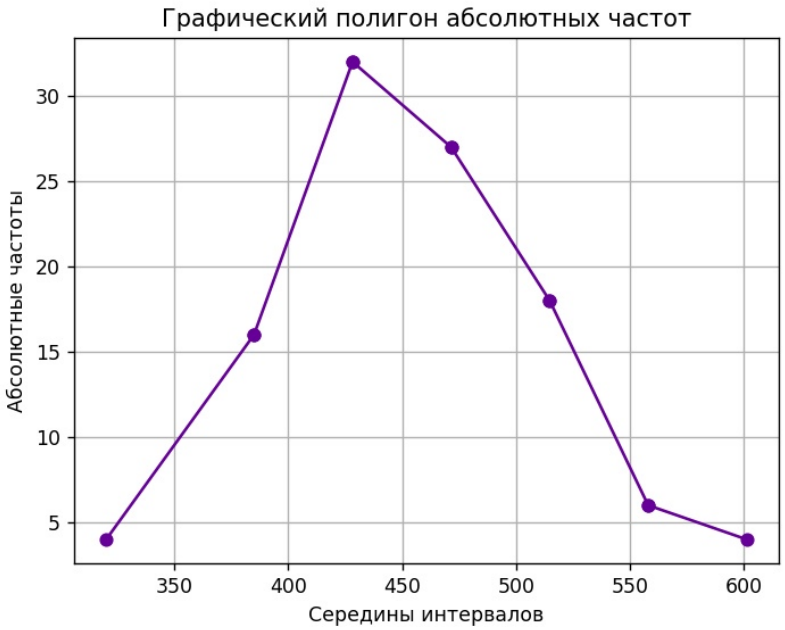

Рис. 1 – Графический полигон для абсолютных частот

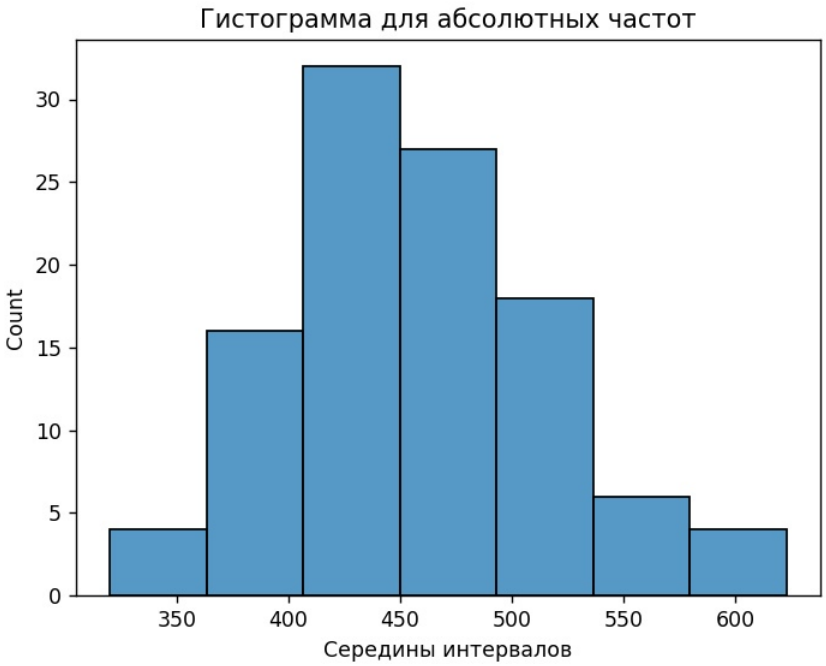

Рис. 2 – Гистрограмма для абсолютных частот

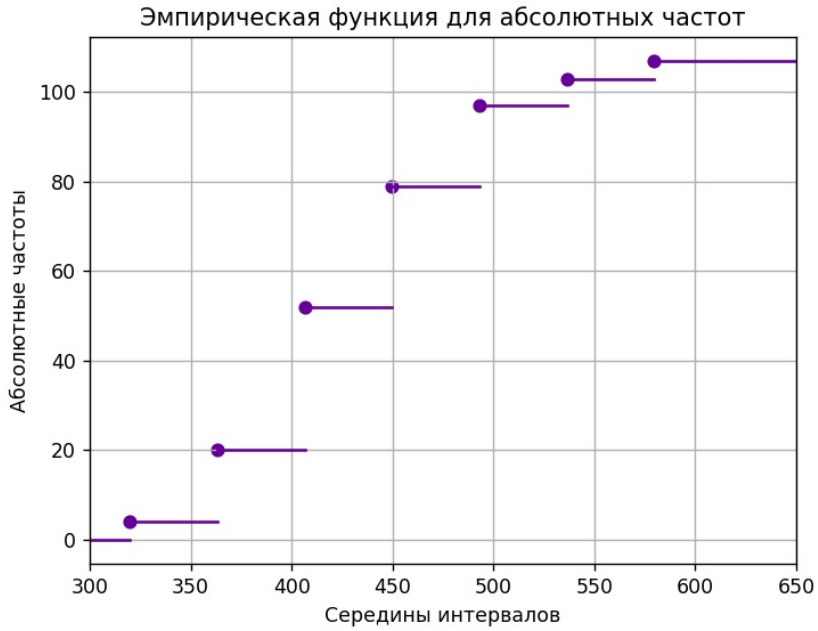

Рис. 3 – Эмпирическая фукнция для абсолютных частот

Имея размер интервала 43.2857 и представленные выше графики, можно сделать вывод, что в интервале 406.571;449.8571] находится больше всего элементов выборки. □

Задание 4. *Для интервального ряда относительных частот построить и отобразить графически полигон, гистрограмму и эмпирическую функцию. Сделать выводы.*

Решение. Были простроены графики:

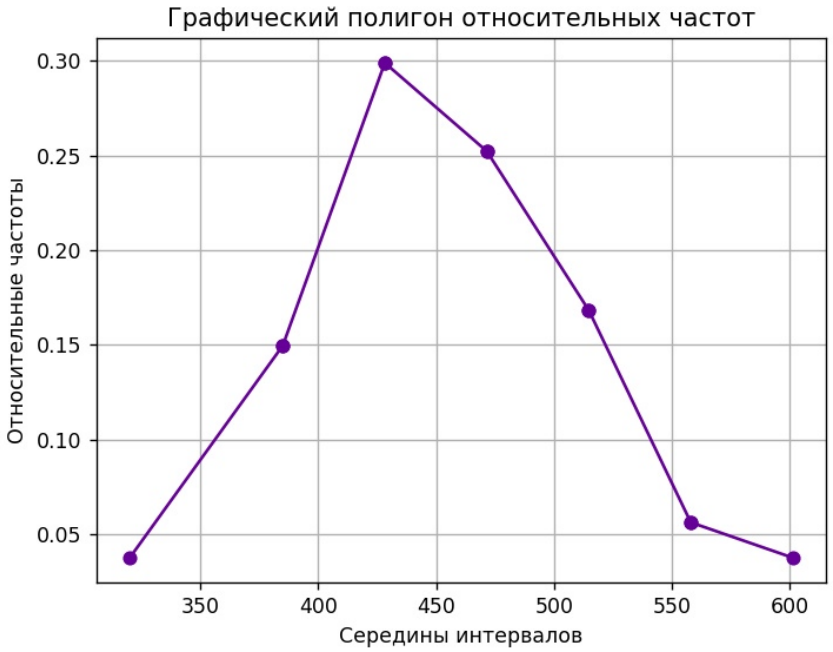

Рис. 4 – Графический полигон для относительных частот

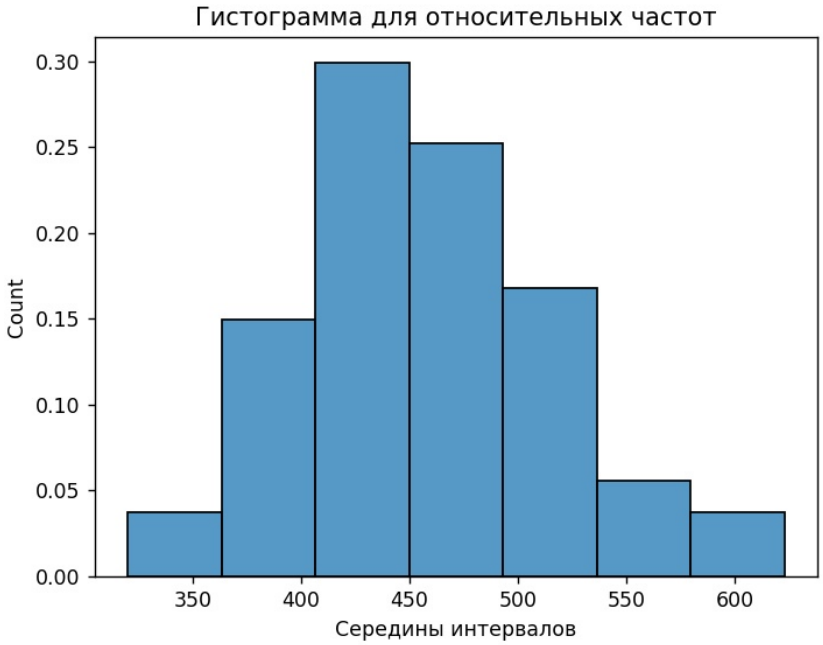

Рис. 5 – Гистрограмма для относительных частот

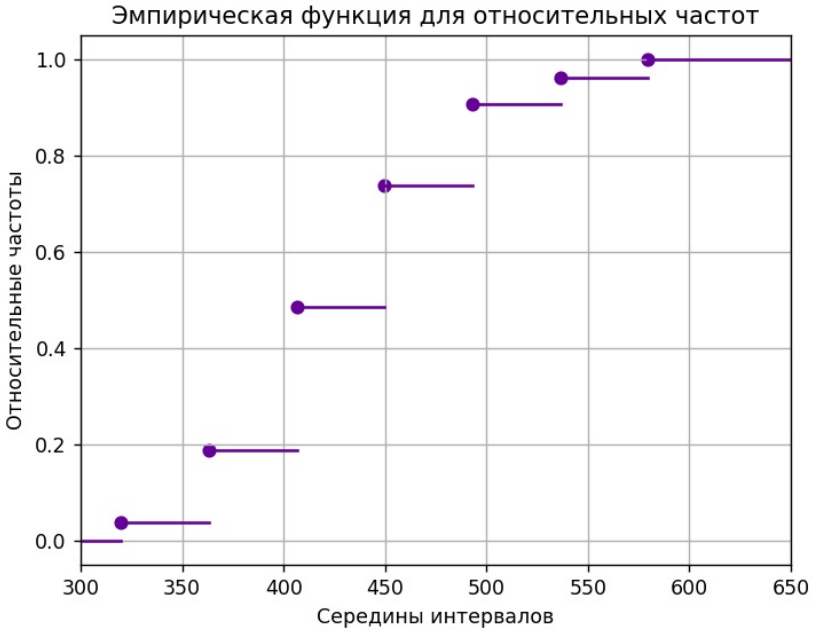

Рис. 6 – Эмпирическая фукнция для относительных частот

Можно сделать вывод, что графики относительных и абсолютных частот различаются только масштабированием вертикальной числовой оси. □

Задание 5. *Для интервального ряда найти середины интервалов, а также накопленные частоты. Результаты представить в виде таблицы.*

Решение. Середины интервалов можно найти сложив крайние точки интервала и разделив пополам. Накопленные частоты же являются просто суммой всех абсолютных частот, предшествующих интервалов. В таком случае таблица с результаты вычислений будет иметь следующий вид:

Интервал	Середина	n_i	p_i	n_s
320-363.2857	341.6429	4	4/107	4
363.2857-406.5714	384.9286	16	16/107	20
406.5714-449.8571	428.2143	32	32/107	52
449.8571-493.1429	471.5	27	27/107	79
493.1429-536.4286	514.785	18	18/107	97
536.4286-579.7142	558.0714	6	6/107	103
579.7143-623	601.357	4	4/107	107

□

Задание 6. *Вычсилить выборочное среднее и дисперсию. Вычислить исправленную выборочную дисперсию и исправленное СКО. Сравнить данные оценки с смещенными оценками дисперсии и СКО.*

Решение. Для начала вычислим выборочное среднее:

$$\overline{x_v} = \frac{1}{n}\sum_{i=1}^{k} x_{si} n_i \approx 457.7457$$

Далее выборочную диспресию:

$$\sigma_v^2 = \frac{1}{n}\sum_{i=1}^{k} (x_{si} - \overline{x_v})^2 n_i \approx 3488.0812$$

Найдем исправленную выборочную дисперсию:

$$s^2 = \frac{n \cdot \sigma_v^2}{n-1} \approx 3520.9876$$

СКО:

$$s^2 = \sqrt{\sigma_v^2} \approx 59.06$$

Исправленное СКО:

$$s = \sqrt{s^2} \approx 59.3379$$

Найдем прогрешность выборочных дисперсий и СКО:

$$|\sigma_v^2 - s^2 \approx 32.9064$$

$$|\sigma - s| \approx 0.2779$$

□

Задание 7. *Найти статистическую оценку коэффициентов асимметрии и эксцесса. Сделать выводы*

Решение. Найдем a_s^*:

$$a_s^* = \frac{\mu_3^*}{\sigma_v^3}$$

где μ_3^* - центральный эмперический момент третьего порядка равен:

$$\mu_3^* = v_3 - 3v_2v_1 + 2v_1^3$$

$$v_r = \sum_{i=1}^{k} x_i^r p_i$$

тогда коэффициент ассиметрии a_s^* =0.3654: то есть расперделение немного смещено влево.

Вычислим коэффициент эксцесса ξ_k^*:

$$\xi_k^* = \frac{\mu_4^*}{\sigma_v^4} - 3$$

где μ_4^* - центральный эмперический момент четвертого порядка равен:

$$\mu_4^* = v_4 - 4v_3v_1 + 6v_2v_1^2 - 3v_1^4$$

тогда коэффициент эксцесса ξ_k^* = -0.1565, что говорит о немного пологой вершине графика распределния, относительно нормального. □

Задание 8. *Вычислить моду, медиану и коэффициент вариации для заданного распределения. Сделать выводы.*

Решение. Найдем моду M_0^*:

$$M_0^* = x_{M_0}^{(0)} + h \cdot \frac{n_M - n_{M-1}}{(n_M - n_{M-1}) + (n_M - n_{M+1})} \approx 439.551$$

А также медиану m_e:

$$m_e = x_0 + \frac{0.5n - n_{m-1}^{\sum}}{n_m} h \approx 450.679$$

Тогда коэффициент варианции V^* равен:

$$V^* = \frac{\sigma_v}{|\overline{x_v}|} \cdot 100\% \approx 12.9024$$

Отсюда можно сдлеать вывод, что мода и медиана находятся чуть левее центра выборки а также что степень рассеивания данных средняя. □

Задание 9. *Вычислить точность и доверительный интервал для математического ожидания при неизвестном среднеквадратичном отклонении при заданном обьёме выборки для доверительной точности $\gamma \in 0.95, 0.99$. Сделать выводы.*

Решение. Найдем точность доверительного интервала при γ =0,95 и γ = 0,99:

$$\delta = \frac{t_{0.95}s}{\sqrt{n}} \approx 11.373$$

$$\delta = \frac{t_{0.99}s}{\sqrt{n}} \approx 15.0467$$

Тогда найдем доверительные интервалы с помощью формулы:

$$(\overline{x_v} - \frac{t_\gamma s}{\sqrt{n}}; \overline{x_v} + \frac{t_\gamma s}{\sqrt{n}})$$

Доверительный интервал при γ = 0,95:

$$(\overline{x_v} - \frac{t_{0.95}s}{\sqrt{n}}; \overline{x_v} + \frac{t_{0.95}s}{\sqrt{n}}) = (446.3727; 469.1187)$$

Доверительный интерва при γ = 0.99: □

$$(\overline{x_v} - \frac{t_{0.99}s}{\sqrt{n}}; \overline{x_v} + \frac{t_{0.99}s}{\sqrt{n}}) = (442.699; 472.7923)$$

Задание 10. *Для вычисления границ доверительного интервала для среднеквадратичного отклонения определить значение q при заданных γ и n. Построить доверительные интервалы, сделать выводы.*

Решение. Значение q при заданных γ и n можно найти из таблицы:

$$q(0.95, 107) \approx 0.142$$

$$q(0.99, 107) \approx 0.197$$

В обоих случаях q меньше единцы, тогда доверительный интервал вычисляется как:

$$(s - sq; s + sq)$$

При $\gamma = 0{,}95$ доверительный интервал СКО равен (50.9119; 67.7639), а при $\gamma = 0.99$ - (47.6483; 71.0275). □

Задание 11. *Проверить гипотезу о нормальности заданного распределения с помощью критерия X^2 (Пирсона). Для этого необходимо найти теоретические частоты и вычислить наблюдаемое значение критерия. Далее по заданному уровню значимости $\alpha = 0.05$ и числу степеней свободы найти критическую точку и сравнить с наблюдаемым значением. Сделать выводы.*

Решение. Необходимо проверить гипотезу о нормальности распределения. В таком случае неоьходимо определить количество степеней свободы, где r, k - количество оцениваемых параметров и количество интервало соответственно:

$$df = k - r - 1 = 7 - 3 = 4$$

По таблице найдем значение критической точки при $\alpha = 0.05$ и $df = 4$:

$$X^2_{crit}(0.05, 4) = 9.488$$

Далее необходимо найти теоретические частоты n'_i:

$$n_i = n \cdot p^*_i$$

где $p^*_i = \Phi(\frac{x_i - \overline{x_v}}{s}) - \Phi(\frac{x_{i-1} - \overline{x_v}}{s})$ - вероятность попадания в интервал.

После чего наблюдаемое значения критерия X^2_{obs} вычисляется по формуле:

$$X^2_{obs} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i} \approx 17.97$$

Т.к. $X^2_{obs} > X^2_{crit}$, то можно отвергнуть гипотезу о нормальности заданной выборки. Сделать выборку нормальной можно с помощью увеличения ее размеров. □

Ссылка на код https://github.com/dart-mih/tv$_5$