

Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
**«Уральский федеральный университет
имени первого президента России Б. Н. Ельцина»**

Итоговая расчётная работа по статистике

Дубовиков Н.Ю.
РИ-200003
Преподаватель Поторочина К.С.

Екатеринбург
2021

Содержание

1	Введение	2
1.1	Предмет исследования	2
1.2	Описание случайных величин	2
1.3	Актуальность исследования	2
1.4	Цель исследования	2
2	Часть первая. Исследование одномерной выборки	2
2.1	Вариационный ряд	2
2.2	Полигон и гистограмма частот	3
2.3	Эмпирическая функция распределения	4
2.4	Числовые характеристики	4
2.5	Доверительный интервал	5
2.6	Уровень значимости по критерию Пирсона	5
2.7	Выводы	6
3	Часть вторая. Исследование двумерной выборки	6
3.1	Первые начальные моменты	6
3.2	Вторые центральные моменты	6
3.3	Эмпирическая линия и прямая регрессии	8
A	Выборка из собранных данных	9

1 Введение

1.1 Предмет исследования

Предметом данного исследования является зависимость числа умышленных убийств (убийств первой степени в законодательстве Соединённых Штатов) от количества владельцев оружия в разных городах США.

1.2 Описание случайных величин

Величины берутся в единицах на сто тысяч человек. Исходные данные были загружены с сайта ФБР. Данные актуальны на 2019 год.

Количество умышленных убийств — количество известных ФБР правонарушений, связанных с умышленным причинением смерти, произошедших в 2019 году на территории определённого города.

Количество владельцев оружия — общее число граждан США, живущих на территории определённого города, получивших от ФБР разрешение на покупку оружия. Определяется в основном строгостью законов в штате или городе, потому может сильно отличаться на разных территориях страны.

Случайная выборка из ста значений исследуемых случайных величин представлена в таблице в приложении А (названия городов указаны как в оригинале — на английском).

1.3 Актуальность исследования

Исследование позволяет установить, помогает ли ограничение оборота оружия снизить количество преступлений, связанных с насилием, в частности убийств. Этим подтверждается её актуальность.

1.4 Цель исследования

Цель работы — установить, есть ли зависимость между количеством умышленных убийств и количеством владельцев оружия на душу населения.

2 Часть первая. Исследование одномерной выборки

При исследовании одномерной выборки будем использовать величину X — количество умышленных убийств на 100000 человек. Объём выборки $N = 100$.

2.1 Вариационный ряд

Найдем из выборки минимальное и максимальное значение. $X_{min} = 0$, $X_{max} = 636$. Далее во формуле Стерджиса находим оптимальное количество интер-

валов:

$$k = 1 + 3.332 \lg(N) \approx 7.35 \approx 7$$

Чтобы включить больше значений в крайние интервалы, возьмём $k = 6$ (при $k = 7$ в два крайних интервала попадёт 3 – 4 значения). Из размаха значений выборки и количества интервалов найдём длину одного интервала:

$$h = \frac{X_{max} - X_{min}}{k} = 106$$

Теперь составим вариационный ряд:

Интервал	[0; 106]	(106; 212]	(212; 318]	(318; 424]	(424; 530]	(530; 636]
n_i	31	25	14	10	8	6

2.2 Полигон и гистограмма частот

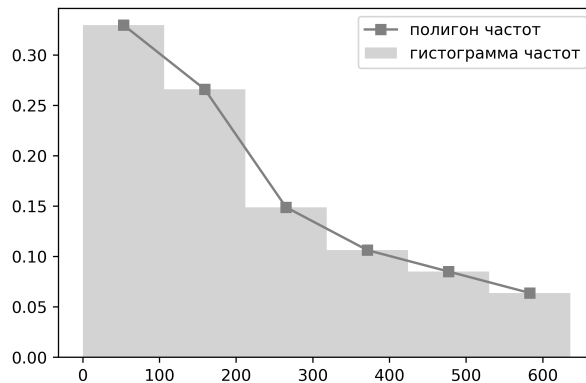
Построим полигон и гистограмму относительных частот. Для этого вычислим относительные частоты по формуле:

$$w_i = \frac{n_i}{N}$$

Также найдём середины всех интервалов и занесём все значения в таблицу:

Интервал	[0; 106]	(106; 212]	(212; 318]	(318; 424]	(424; 530]	(530; 636]
n_i	31	25	14	10	8	6
x_i	53	159	265	371	477	583
w_i	0.329	0.265	0.148	0.106	0.085	0.063

По этой таблице сделаем полигон и гистограмму частот:



2.3 Эмпирическая функция распределения

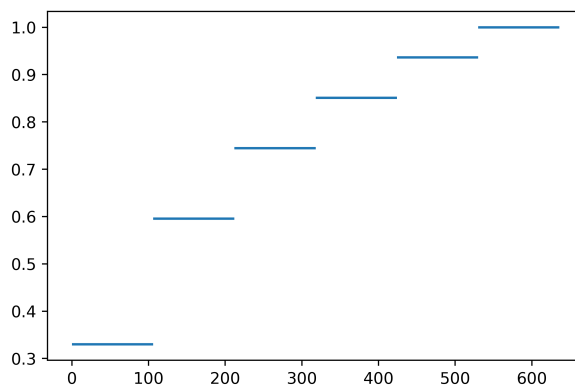
Для того, чтобы найти эмпирическую функцию распределения, необходимы накопительные частоты, они вычисляются по формуле:

$$W_i = \sum_{k=0}^i w_k$$

То есть это сумма всех предыдущих относительных частот. Так же внесём всё в таблицу:

Интервал	[0; 106]	(106; 212]	(212; 318]	(318; 424]	(424; 530]	(530; 636]
n_i	31	25	14	10	8	6
w_i	0.329	0.265	0.148	0.106	0.085	0.063
W_i	0.33	0.59	0.74	0.85	0.94	1

По получившимся значениям построим график:



2.4 Числовые характеристики

Рассчитаем выборочную среднюю, дисперсию, среднее квадратичное отклонение, асимметрию и эксцесс.

- Выборочная средняя (значения x_i и n_i возьмём из таблицы выше):

$$\bar{X}_B = \frac{1}{N} \sum_{i=1}^N x_i n_i \approx 216.51$$

- Выборочная дисперсия:

$$D_B = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X}_B)^2 n_i \approx 26934.1$$

- Исправленная дисперсия:

$$S^2 = \frac{N \cdot D_B}{N - 1} \approx 27223.7$$

- Среднее квадратическое отклонение:

$$\sigma = \sqrt{D_B} \approx 164.116$$

- Исправленное среднее квадратическое отклонение:

$$S = \sqrt{S^2} \approx 164.996$$

- Ассиметрия — мера смещения графика влево-вправо:

$$A = \frac{1}{NS^3} \sum_{i=1}^N (x_i - \bar{X}_B)^3 n_i \approx 0.793$$

- Эксцесс — мера высоты графика:

$$e_x = \frac{1}{NS^4} \sum_{i=1}^N (x_i - \bar{X}_B)^4 n_i \approx -0.512$$

2.5 Доверительный интервал

Определим доверительный интервал для оценки математического ожидания при надёжности $\gamma = 0.95$. Доверительный интервал — интервал, который покрывает неизвестный параметр с заданной надёжностью.

$$\bar{X}_B - \delta < a < \bar{X}_B + \delta$$

Где $\delta = \frac{t_\gamma S}{\sqrt{N}}$. Из уравнения $\Phi(t_\gamma) = \gamma$ вычислим, что $t_\gamma \approx 1.96$. Тогда $\delta \approx 33.355$.

Следовательно интервал $183.155 < a < 249.866$ покрывается с надёжностью $\gamma = 0.95$.

2.6 Уровень значимости по критерию Пирсона

Рассчитаем уровень значимости α , при котором распределение для выборки согласуется с нормальным законом по критерию согласия Пирсона (χ^2). Выдвинем гипотезы:

- H_0 — распределение является нормальным.
- H_1 — распределение нормальным не является.

Функция нормального распределения $F(x) = \frac{1}{2} + \Phi\left(\frac{x-a}{\sigma}\right)$.

$$P(X_i < x < X_{i+1}) = F(X_{i+1}) - F(X_i) = \Phi\left(\frac{x_{i+1} - \bar{X}_B}{S}\right) - \Phi\left(\frac{x_i - \bar{X}_B}{S}\right)$$

Занесём все значения, необходимые для расчётов, в таблицу:

X_i	X_{i+1}	$\frac{x_i - \bar{X}_B}{S}$	$\frac{x_{i+1} - \bar{X}_B}{S}$	$\Phi\left(\frac{x_{i+1} - \bar{X}_B}{S}\right)$	$\Phi\left(\frac{x_i - \bar{X}_B}{S}\right)$	P	nP	$\frac{(n-nP)^2}{nP}$
0	106	31	-1.312	-0.669	-0.405	-0.248	0.156	14.736
106	212	25	-0.669	-0.027	-0.248	-0.010	0.237	22.333
212	318	14	-0.027	0.615	-0.010	0.230	0.241	22.716
318	424	10	0.615	1.257	0.230	0.395	0.164	15.506
424	530	8	1.257	1.899	0.395	0.471	0.075	7.102
530	636	6	1.899	2.542	0.471	0.494	0.023	2.182

Откуда $\chi^2 = \sum \frac{(n_i - n_i P)^2}{n_i P} \approx 30.358$. В результате уровень значимости $\alpha \approx 1.16 \cdot 10^{-6}$. На вычисленном уровне значимости принимается нулевая гипотеза.

2.7 Выводы

Гипотеза о нормальном распределении была отклонена. Из эксцесса и коэффициента асимметрии следует, что распределение отклонено вправо и вниз относительно нормального.

3 Часть вторая. Исследование двумерной выборки

3.1 Первые начальные моменты

$$v_{10} = \bar{X}_B = \sum_{i=1}^N x_i n_{xi} \approx 216.51$$

$$v_{01} = \bar{Y}_B = \sum_{i=1}^N y_i n_{yi} \approx 41231.91$$

3.2 Вторые центральные моменты

$$\mu_{20} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X}_B)^2 n_{xi} \approx 26934.1$$

$$\mu_{02} = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{Y}_B)^2 n_{yi} \approx 106076853.77$$

Коэффициент ковариации:

$$K_{xy} = \overline{XY} - X_B \cdot Y_B$$

Составим корреляционную таблицу:

	[14700;23300]	(23300;31900]	(31900;40500]	(40500;49100]	(49100;57700]	(57700;66300]
[0;106]	0	5	4	16	6	0
(106;212]	1	2	4	13	5	0
(212;318]	0	4	2	6	2	0
(318;424]	2	2	0	3	1	2
(424;530]	0	4	0	2	2	0
(530;636]	2	0	2	0	2	0

И таблицу, в которой на пересечении i -й строки и j -го столбца находится значение $n_{ij} \cdot x_i \cdot y_j$:

	19000	27600	36200	44800	53400	62000
53.0	0	5035000	4028000	16112000	6042000	0
159.0	4388400	8776800	17553600	57049200	21942000	0
265.0	0	38372000	19186000	57558000	19186000	0
371.0	33241600	33241600	0	49862400	16620800	33241600
477.0	0	101887200	0	50943600	50943600	0
583.0	72292000	0	72292000	0	72292000	0

Используя эту таблицу можно посчитать \overline{XY} (сложить всё и поделить на объём выборки).

$$\overline{XY} = \frac{1}{N} \sum \sum x_i y_j n_{ij} \approx 9171142.55$$

Тогда $K_{xy} \approx 243994.34$. Откуда коэффициент корреляции:

$$r_{xy} = \frac{K_{xy}}{\sqrt{\mu_{20} \cdot \mu_{02}}} \approx 0.144$$

По таблице Чеддока, связь между величинами слабая. Проверим гипотезу H_0 о том, что $r_{xy} = 0$, а так же альтернативную гипотезу H_1 : $r_{xy} \neq 0$.

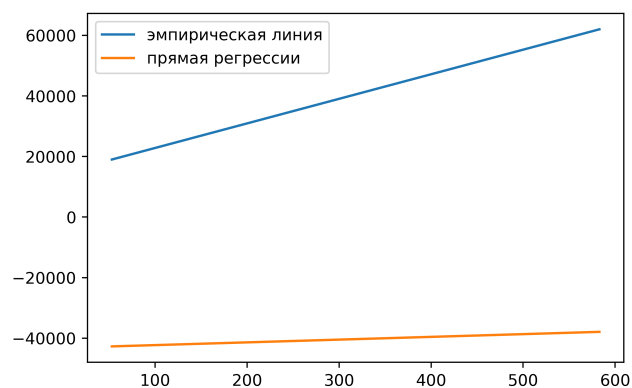
$$T = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \approx 1.399$$

При $N - 2 = 92$ степенях свободы коэффициент Стьюдента при уровне значимости 0.05 равен 1.98. $|T| < t_K$, поэтому гипотезу H_0 принимаем, то есть X и Y не коррелируют.

3.3 Эмпирическая линия и прямая регрессии

Регрессия Y по X: $y = r_{xy} \frac{\mu_{02}}{\mu_{20}}(x - v_{10}) - v_{01}$. Для вычисления составим таблицу произведений y_i на число появлений из корреляционной таблицы n_{ij} :

Y X	(0;106]	(106;212]	(212;318]	(318;424]	(424;530]	(530;636]
(14700;23300]	0	27600	0	89600	0	124000
(23300;31900]	95000	55200	144800	89600	213600	0
(31900;40500]	76000	110400	72400	0	0	124000
(40500;49100]	304000	358800	217200	134400	106800	0
(49100;57700]	114000	138000	72400	44800	106800	124000
(57700;66300]	0	0	0	89600	0	0
Сумма	117800.0	40588.235	42233.333	11200.0	23733.333	186000.0



А Выборка из собранных данных

Город	Количество убийств	Количество владельцев оружия
Buckeye	137	46300
Nevada	255	48800
Dilley	223	45700
GrossePointe	117	40200
NewHaven	297	44800
Harrisonville	277	48800
Boston,	607	14700
OakGrove	207	48800
Tomahawk	96	45300
Winchester	481	51600
Hulbert	171	54700
Emerson	0	45200
Winneconne	80	45300
VillaPark	157	27800
BradentonBeach	155	35300
Beatrice	180	45200
Northglenn	457	45100
Hutchinson	378	48900
WarminsterTownship	84	40700
Lyndhurst	104	40000
Bath	48	46800
Redmond	102	42100
RobinsonTownship	124	40700
Streetsboro	121	40000
Chickasha	433	54700
Murphy	43	45700
Melbourne	692	35300
Killeen	384	45700
Maumee	103	40000
Hanford	449	28300
Pembroke	92	44600
Elwood	133	27800
Ashburnham	126	14700
Athol	402	14700
HedwigVillage	224	45700
Elkton	103	44600
Clarkdale	68	46300
Auburn	146	46800
Bokchito	581	54700
Corning	372	28300

Город	Количество убийств	Количество владельцев оружия
Madison	84	45200
Hastings	246	40200
RanchoCordova	297	28300
Viola	0	27800
Jonesville	272	40200
Ackerman	139	55800
Mazon	103	27800
SouthHaven	555	40200
Fishers	53	44800
Orion	56	27800
Plymouth	327	14700
Federalsburg	528	30200
NewLondon	294	23600
Niles	774	40200
Rossville	0	51600
LaVergne	392	51600
Somerset	0	40000
Emporia	240	44600
Rosenberg	380	45700
EagleVillage	47	45300
Franklin	165	51600
Ellensburg	211	42100
Hemet	398	28300
RomanForest	148	45700
Perry	245	54700
Refugio	109	45700
Norfork	0	57200
OliverSprings	58	51600
Williamstown	0	54600
Wynona	0	54700
Dunwoody	128	49200
Verona	990	48800
Chicopee	615	14700
GenevaTown	40	45300
Avon	63	42800
Auburn	73	54600
Clinton	160	54600
Mathis	440	45700
Okmulgee	615	54700
Newport	806	50500
Halifax	164	44600
Kirtland	88	40000

Город	Количество убийств	Количество владельцев оружия
Adrian	689	40200
PalmSprings	636	35300
FrontRoyal	183	44600
MissouriCity	141	45700
Visalia	434	28300
PigeonForge	736	51600
Beebe	279	57200
Manhattan	375	66300
Williamston	176	40200
Berkeley	503	28300
BerwynHeights	214	30200
Muscatine	197	43600
WilkinsTownship	49	40700
LakeZurich	40	27800
Merrillville	144	44800
Buhl	360	60100
PalosHeights	40	27800
Martinez	215	28300