Описательная статистика в Excel

1 Оценивание числовых характеристик случайной величины.

1.1 Выборочные числовые характеристики

Пусть случайная величина X исследуется на основании случайной выборки $\mathrm{X_1}$, $\mathrm{X_2}$,…, $\mathrm{X_n}$, реализация которой x_1, x_2,…, x_n. Пусть θ – некоторый параметр распределения случайной величины X, значение которого неизвестно. Например, предполагается, что X имеет показательный закон распределения $f(x) = \lambda e^{-\lambda x}, x \geq 0$, но значение параметра $\theta = \lambda$ неизвестно. Задача оценивания параметра θ состоит в нахождении его приближенного значения по результатам наблюдений $\mathrm{X_1}$, $\mathrm{X_2}$,…, $\mathrm{X_n}$.

Функцию от случайной выборки $\theta_n^* = \theta_n^*(\mathrm{X_1},\ \mathrm{X_2},\ldots,\mathrm{X_n})$ называют *статистикой*. Очевидно, что для оценивания параметра θ следует рассматривать только те статистики θ_n^*, которые в определенном смысле близки к истинному значению параметра θ. Статистику θ_n^*, принимаемую в качестве приближенного значения θ, называют *оценкой* параметра θ. Например, для оценивания неизвестного математического ожидания $\mathrm{E(X)} = m$ используется оценка $\overline{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$, называемая *выборочным средним.* Оценкой для дисперсии $\mathrm{D(X)} = \sigma^2$ является *выборочная дисперсия* $\mathrm{S^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. На реализации x_1, x_2,…, x_n случайной выборки оценки $\overline{\mathrm{X}}$ и $\mathrm{S^2}$ принимают числовые значения $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ и $\mathrm{s^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Согласно теории, когда n велико, с большой вероятностью $\overline{x}$ и $\mathrm{s^2}$ будут близки к m и σ^2 соответственно.

Приведем основные числовые характеристики генеральной совокупности и соответствующие оценки.

1. Характеристики положения

Числовая характеристика случайной величины X	Оценка числовой характеристики
Математическое ожидание $E(X) = \sum_i x_i p_i$ – для дискретной случайной величины. $E(X) = \int xf(x)dx$ – для непрерывной случайной величины.	*Выборочное среднее* $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Медиана непрерывной случайной величины – это такое ее значение M_e, для которого $P\{X \le M_e\} = P\{X > M_e\} = \frac{1}{2}$.	*Выборочная медиана* определяется как такое значение M_e^*, меньше и больше которого оказывается одинаковое число наблюдений. $M_e^* = \begin{cases} \frac{1}{2}\left(X_{(n/2)} + X_{((n+2)/2)}\right), & \text{если } n - \text{четное} \\ X_{((n+1)/2)}, & \text{если } n - \text{нечетное.} \end{cases}$ $X_{(k)} - k$*-й* элемент вариационного ряда.
Мода M_o – значение случайной величины X, соответствующее локальному максимуму плотности $f(x)$ для непрерывной случайной величины или максимуму вероятности для дискретной случайной величины.	*Выборочная мода* – это наиболее часто встречающееся значение в выборке. Если данные сгруппированы и построено распределение частот, модой является значение, имеющее наибольшую частоту.

2. Основные характеристики разброса

Дисперсия $D(X) = E(X - EX)^2$, среднее квадратическое отклонение $\sigma(X) = \sqrt{D(X)}$	*Выборочная дисперсия* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ *S* – выборочное стандартное отклонение (выборочное среднее квадратическое отклонение).
p- квантиль функции распределения $F(x)$ –это значение x_p случайной величины X, задаваемое уравнением $x_p = \inf\{x: F(x) > p\}$.	*Выборочная p- квантиль* $\hat{x}_p$ определяется равенством $\hat{x}_p = \begin{cases} \frac{1}{2}\left(X_{(np)} + X_{(np+1)}\right), & \text{если } np - \text{целое} \\ X_{([np]+1)}, & \text{если } np - \text{не целое число.} \end{cases}$

3. Характеристики формы распределения

Мерой отклонения от симметрии плотности $f(x)$ являются 3-й момент $\mu_3 = \int (x - EX)^3 f(x)dx$ и коэффициент асимметрии $A = \mu_3/\sigma^3$	*Выборочный коэффициент асимметрии* $a_s = \sum_{i=1}^{n}(X_i - \overline{X})^3/nS^3$
Мерой отклонения плотности $f(x)$ от нормального распределения является эксцесс $E = \mu_4/\sigma^4 - 3;$	*Выборочный эксцесс* $E^* = \sum_{i=1}^{n}(X_i - \overline{X})^4/nS^4 - 3$ – мера отклонения эмпирического распределения от нормального.

Для нахождения выборочных характеристик могут быть использованы встроенные функции категории «статистические» из электронных таблиц Excel. Средство Excel ОПИСАТЕЛЬНАЯ СТАТИСТИКА позволяет вычислить важнейшие числовые характеристики выборки и представить их в виде таблицы.

*Пример***.** Случайная величина X, характеризующая уровень воды в реке по отношению к номиналу, измерялась в течение 36 весенних паводков. Результаты измерений приведены в табл. 8.

Таблица 8

№ измерения	Уровень (в см)	№ измерения	Уровень (в см)	№ измерения	Уровень (в см)	№ измерения	Уровень (в см)
1	47	10	164	19	93	28	115
2	151	11	158	20	121	29	171
3	52	12	243	21	118	30	205
4	163	13	190	22	110	31	61
5	77	14	85	23	173	32	174
6	156	15	139	24	243	33	148
7	205	16	179	25	254	34	217
8	181	17	257	26	307	35	149
9	311	18	143	27	99	36	187

а. Найдем оценки среднего значения, дисперсии и среднего квадратического отклонения случайной величины X по формулам:

выборочное среднее $\overline{x} = \frac{1}{36}\sum_{i=1}^{36} x_i \approx 162{,}4;$

выборочная дисперсия $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\overline{x}^2 = \frac{1}{35}\sum_{i=1}^{36} x_i^2 - \frac{36}{35}\overline{x}^2 \approx 4295{,}2;$

выборочное среднее квадратическое отклонение $s \approx 65{,}5.$

b. Оценим эти же характеристики с помощью функций пакета Excel. Предположим, что элементы выборки $x_1, x_2, \ldots, x_{36}$ находятся в ячейках с адресами A1:A36 . В ячейки B1, B2, B3, предназначенные для результатов $\overline{x}$, s^2, s, нужно ввести формулы: =СРЗНАЧ(A1:A36), =ДИСП.В(A1:A36), = СТАНДОТКЛОН.В(A1:A36) соответственно.

c. Определим основные числовые характеристики выборки с помощью процедуры ДАННЫЕ – АНАЛИЗ ДАННЫХ – ОПИСАТЕЛЬНАЯ СТАТИСТИКА. В качестве ВХОДНОГО ИНТЕРВАЛА следует указать адреса ячеек, содержащих выборку $x_1, x_2, \ldots, x_{36}$. Результат получим в виде таблицы

Среднее	162,4
Стандартная ошибка	10,9
Медиана	160,5
Мода	205,0
Стандартное отклонение	65,5
Дисперсия выборки	4295,2
Эксцесс	0,02
Асимметричность	0,4
Интервал	264,0
Минимум	47,0
Максимум	311,0
Сумма	5846,0
Счет	36,0

Здесь *стандартная ошибка* – это оценка параметра $\sigma(\overline{X})$, равная $S/\sqrt{n}$.

Оценивание генеральных числовых характеристик по выборке, представленной в виде группированного статистического ряда

Если в группированной выборке (1) все выборочные элементы из интервала $(d_{i-1}, d_i]$ положить равными величине $x_i^* = \frac{d_{i-1}+d_i}{2}$, то распределение выборки принимает вид

x_1^*	x_2^*	$\ldots$	x_k^*
ν_1/n	ν_2/n	$\ldots$	ν_k/n

(2)

С помощью группированного статистического ряда (2) приближенное вычисление выборочных моментов порядка l выполняется по формуле $\overline{a_l} = \frac{1}{n}\sum_{i=1}^{k} \nu_i \, (x_i^*)^l$. Формулы, определяющие выборочное среднее и выборочную дисперсию, принимают вид:

$$\overline{x} = \overline{a}_1 \approx \frac{1}{n}\sum_{i=1}^{k} \nu_i \, x_i^*, \;\; s^2 = \frac{n}{n-1}\left(\overline{a}_2 - \overline{a}_1{}^2\right) = \frac{1}{n-1}\sum_{i=1}^{k} \; (x_i^*)^2 \nu_i - \frac{n}{n-1}(\overline{x})^2 .$$

(2. 1)

Усреднение по интервалам вносит ошибку, особенно заметную при малом числе интервалов. Для уменьшения ошибок, вносимых подобной группировкой, применяют «поправки Шеппарда».

Пример. Распределение средних температур июня в Стокгольме в течение 100 лет дано в виде интервального статистического ряда в табл. 5 (столбцы 2-4). Найти оценки математического ожидания и дисперсии случайной величины X, характеризующей среднюю июньскую температуру. (Данные взяты из книги «Математические методы статистики», Г. Крамер.)

В примере по условию дано: объем выборки $n = 100$, наименьший и наибольший из элементов данной выборки равны: $x_{min} = 12{,}07$ и $x_{max} = 16{,}94$, $k = 10$. Отрезок [12, 17], содержащий все выборочные значения, разбит на 10 интервалов $\Delta_i = (d_{i-1}, d_i]$ длины $h = 0.5$ с границами: $d_0 = 12$, $d_i = d_0 + ih, i = 1{,}2, \ldots 10$.

Вычислим значения выборочного среднего и выборочной дисперсии, используя формулы (2.1).

Значения величин $x_i^* \gamma_i$ и $(x_i^*)^2 \gamma_i$, необходимых для нахождения оценок, получим в столбцах 6 и 7 таблицы 9.

Таблица 9

Номер интервала Δ_i	Левая граница d_{i-1}	Правая граница d_i	Частота ν_i	$x_i^* = \frac{d_{i-1} + d_i}{2}$	$x_i^* \nu_i$	$(x_i^*)^2 \nu_i$
1	12	12,5	10	12,25	122,5	1500,625
2	12,5	13	12	12,75	153	1950,75

3	13	13,5	9	13,25	119,25	1580,063
4	13,5	14	10	13,75	137,5	1890,625
5	14	14,5	19	14,25	270,75	3858,188
6	14,5	15	10	14,75	147,5	2175,625
7	15	15,5	9	15,25	137,25	2093,063
8	15,5	16	6	15,75	94,5	1488,375
9	16	16,5	7	16,25	113,75	1848,438
10	16,5	17	8	16,75	134	2244,500
			∑=100		∑= 1430	∑= 20630,252

Выборочное среднее $\overline{x} = \frac{1}{100}\sum_{i=1}^{10} x_i^* \nu_i$ =14,30;

выборочная дисперсия $s^2 = \frac{1}{99}\sum_{i=1}^{10} (x_i^*)^2 \nu_i - \frac{100}{99}(\overline{x})^2 \approx 1{,}83;\ s \approx 1{,}35.$

Задачи для решения

Задача 2.1. В таблице 9 приводятся данные об ожидаемой продолжительности жизни (2017 г.)

Таблица 9

№	Субъект РФ	Ожидаемая продолжительность жизни при рождении, лет	№	Субъект РФ	Ожидаемая продолжительность жизни при рождении, лет
1	Алтайский край	71	43	Пермский край	71
2	Амурская область	69	44	Приморский край	70
3	Архангельская обл.	72	45	Псковская область	70
4	Астраханская обл.	73	46	Республика Адыгея	73
5	Белгородская обл.	74	47	Республика Алтай	71
6	Брянская обл.	71	48	Башкортостан	72
7	Владимирская обл.	71	49	Республика Бурятия	71
9	Волгоградская обл.	74	50	Республика Дагестан	78

9	Вологодская обл.	71	51	Республика Ингушетия	82
10	Воронежская обл.	73	52	Республика Калмыкия	74
11	г. Москва	78	53	Республика Карелия	71
12	Санкт-Петербург	75	54	Республика Коми	71
13	г. Севастополь	73	55	Республика Крым	72
14	Еврейская авт. обл.	69	56	Республика Марий Эл	72
15	Забайкальский край	70	57	Республика Мордовия	73
16	Ивановская обл.	71	58	Республика Саха	72
17	Иркутская обл.	69	59	Республика Северная Осетия	76
18	Кабардино-Балкарская Респ.	76	60	Республика Татарстан	74
19	Калининградская обл.	73	61	Республика Тыва	66
20	Калужская обл.	72	62	Республика Хакасия	70
21	Камчатский край	70	63	Ростовская обл.	73
22	Карачаево-Черкесская Респ.	76	64	Рязанская область	73
23	Кемеровская обл.	69	65	Самарская обл.	72
24	Кировская обл.	73	66	Саратовская обл.	73
25	Костромская обл.	72	67	Сахалинская обл.	70
26	Краснодарский край	73	68	Свердловская область	71
27	Красноярский край	71	69	Смоленская область	71
28	Курганская обл.	71	70	Ставропольский край	74
29	Курская область	72	71	Тамбовская обл.	73
30	Ленинградская обл.	73	72	Тверская область	70
31	Липецкая обл.	72	73	Томская область	72
32	Магаданская область	69	74	Тульская область	71
33	Московская обл.	73	75	Тюменская обл.	72
34	Мурманская обл.	72	76	Удмуртская Республика	72
35	Ненецкий авт. окр.	72	77	Ульяновская область	72
36	Нижегородская обл.	72	78	Хабаровский край	70
37	Новгородская область	70	79	Ханты-Мансийский ао	74

38	Новосибирская обл.	72	80	Челябинская область	72
39	Омская область	71	81	Чеченская Республика	75
40	Оренбургская область	71	82	Чувашская Республика	73
41	Орловская обл.	72	83	Чукотский авт. округ	66
42	Пензенская обл.	73	84	Ямало-Ненецкий	74
			85	Ярославская обл.	72

Вычислить важнейшие числовые характеристики выборки с помощью процедуры ДАННЫЕ – АНАЛИЗ ДАННЫХ – ОПИСАТЕЛЬНАЯ СТАТИСТИКА.

Задача 2.2. По данным, приведенным в задаче 1.2, оценить важнейшие числовые характеристики случайной величины X, характеризующей денежные доходы населения.

Задача 2.3. Проведено исследование коммерческих фирм по затратам на рекламу в год. Для этого случайным образом выбраны 70 фирм. Результаты представлены в таблице.

Расходы на рекламу млн. руб.	0—20	20—40	40—60	60—80	80-100	120 и более
Количество фирм	2	8	12	27	16	5

Оценить среднее и дисперсию случайной величины X, выражающей затраты фирм на рекламу в год.

Задача 2.4. В воде мелководного озера в течение года были измерены концентрации общего фосфора (в мкг/л):

46	41	153	98	140	95	208	88	65	108
60	42	179	320	176	118	191	108	62	91
90	66	189	274	170	95	62	108	45	58
90	83	202	134	166	82	117	62	91	37
80	45	111	83	120	108	91	241	90	66
163	110	117	91	180	104	91	134	92	83

По полученной выборке оценить важнейшие числовые характеристики случайной величины X, выражающей концентрацию общего фосфора в озере.

Указание. Воспользоваться средством Excel ОПИСАТЕЛЬНАЯ СТАТИСТИКА.

Проверка выборки на соответствие нормальному закону при помощи экспресс-метода

При проведении статистического анализа важно знать, насколько близок закон распределения выборки к нормальному закону. Выборочные асимметрия и эксцесс характеризуют степень отличия эмпирического распределения от нормального. Коэффициент асимметрии и эксцесс нормального распределения равны нулю. Поэтому, достаточно малые значения соответствующих выборочных величин дают основание предполагать, что генеральная совокупность распределена по нормальному закону. Для первоначальной проверки выборки на соответствие нормальному закону можно применить *экспресс-метод:* предположение о близости выборочного распределения к нормальному отвергается при условии $a_s \geq 4s/\overline{x}$. Здесь $a_s -$ в*ыборочный коэффициент асимметрии,* $s -$ выборочное стандартное отклонение, $\overline{x} -$ выборочное среднее. Величина $s/\overline{x}$ называется *коэффициентом вариации*.

Задания

По каждой выборке, приведенной ниже в задачах 3.1-3.4, требуется:

a) построить гистограмму частот;

b) найти числовые характеристики выборки, используя средство Excel ОПИСАТЕЛЬНАЯ СТАТИСТИКА;

c) проверить выборку на соответствие нормальному закону с помощью *экспресс-метода***.** Сделать заключение.

Задача 1. Данные о площади лесов в странах мира (% от земельной площади) указаны в таблице:

1	Бразилия	59,2	21	Австралия	16,2
2	Канада	38,0	22	Португалия	34,7
3	Япония	68,5	23	Испания	36,8
4	Швеция	68,9	24	Словакия	40,3
5	Австрия	46,9	25	Россия	49,8
6	Дания	14,6	26	Литва	34,8
7	Финляндия	73,1	27	Польша	30,8
8	Германия	32,7	28	Греция	31,5
9	США	33,9	29	Болгария	35,2
10	Мексика	34,0	30	Бельгия	22,6
11	Франция	31,0	31	Турция	15,2
12	Швейцария	31,7	32	Словения	62,0
13	Китай	22,2	33	Латвия	54,0
14	Голландия	11,2	34	Украина	16,7
15	Чехия	34,5	35	Беларусь	42,5
16	Норвегия	33,2	36	Грузия	40,6
17	Великобритания	13,0	37	Аргентина	9,9
18	Эстония	52,7	38	Монголия	8,8
19	Венгрия	22,9	39	Киргизия	3,3
20	Италия	31,6	40	Румыния	29,8

Задача 2. В таблице 9 приводятся данные об ожидаемой продолжительности жизни (2017 г.)

Задача 3. Средняя температура июня в Ярославле измерялась в течение 60 лет. Данные измерений приведены в таблице:

12,4	14,1	15,2	13,8	17,2	14,9
11,3	13,8	16	15	18	16
12	13,7	16,2	16	17,8	17,8
13,1	15	17,2	13,6	19,2	13,8
14,9	15,7	16,8	14,8	19,3	15,1
14	14,9	16,6	14,9	20,1	17
13,5	15,9	17	16	20	17,7
13	16	16,7	17	14	14,5
14,1	15,5	17,3	16,1	13,6	14,8
15	16	17	16	14,1	15,2

Задача 4. Проведено испытание нового сорта зерновой культуры на 56 участках одинаковой площади и получены следующие значения урожайности (ц/г):

46	46	44	55	39	58	49	51
48	47	46	49	40	50	44	47
46	46	43	42	47	48	42	43
49	49	47	43	50	49	48	49
47	47	46	48	52	44	49	46