

Тема. Математическая статистика

1. Необходимый теоретический материал

Генеральная совокупность—все множество имеющихся объектов.

Выборка—набор объектов, случайно отобранных из генеральной совокупности.

Объем генеральной совокупности N и объем выборки n —число объектов в рассматриваемой совокупности.

Например. Генеральная совокупность—все множество имеющихся студентов ВУЗа. Выборка — отобранные студенты для проверки знаний. Для получения реальных результатов при проверке необходимо обеспечить *случайность отбора* студентов. Это означает, что у студентов с разным уровнем знаний *возможность попасть в выборку была одинаковой*. При этом *объем выборки должен быть достаточно большой* — настолько большой, чтобы обеспечить нужную точность.

Пусть интересующая нас случайная величина X принимает в выборке значение x_1 — n_1 раз, x_2 — n_2 раз, ..., x_k — n_k раз, причем $\sum_{k=1}^k n_k = n$, где n —объем

выборки. Тогда наблюдаемые значения случайной величины $x_1, x_2, \dots, x_k$ называют **вариантами**, а $n_1, n_2, \dots, n_k$ — **частотами**. Если разделить каждую частоту на объем выборки, то получим **относительные частоты** $w_i = \frac{n_i}{n}$.

Последовательность вариантов, записанных в порядке возрастания, называют **вариационным рядом**, а перечень вариант и соответствующих им частот или относительных частот — **статистическим рядом**.

Если исследуется некоторый непрерывный признак, то вариационный ряд может состоять из очень большого количества чисел. В этом случае удобнее использовать **группированную выборку**. Для ее получения интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько равных частичных интервалов длиной h , а затем находят для каждого частичного интервала n_i —сумму частот вариантов, попавших в i -й интервал. Составленная по этим результатам таблица называется группированным (интервальным) статистическим рядом.

Выборочной (эмпирической) функцией распределения называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$.

Для непрерывного признака графической иллюстрацией служит **гистограмма**, то есть ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиной h , а высотами—отрезки длиной n_i/h (гистограмма частот) или w_i/h (гистограмма относительных частот).

Выборочным средним называется среднее арифметическое значений

случайной величины в выборке: $\bar{x}_B = \frac{1}{n} \sum_{i=1}^k n_i x_i$.

Выборочной дисперсией называется $D_B = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x}_e)^2$ или

$D_B = \overline{x^2} - (\bar{x})^2$, а **выборочным средним квадратичным отклонением** — $\sigma_B = \sqrt{D_B}$.

Мода Мо-варианта, имеющая наибольшую частоту.

Медиана Ме-варианта, которая делит вариационный ряд на две части, равные по числу вариантов.

Статистическая оценка Θ^* называется **несмещенной**, если ее математическое ожидание равно оцениваемому параметру Θ при любом объеме выборки:

$$M(\Theta^*) = \Theta.$$

Статистическая оценка называется **эффективной**, если она при заданном объеме выборки n имеет наименьшую возможную дисперсию.

Состоятельной называется статистическая оценка, которая при $n \rightarrow \infty$ стремится по вероятности к оцениваемому параметру.

Выборочное среднее является не только несмещенной, но и состоятельной оценкой математического ожидания.

Исправленная дисперсия s^2 является несмещенной оценкой дисперсии генеральной совокупности $s^2 = \frac{n}{n-1} D_B$.

Доверительная вероятность и доверительный интервал. Определение необходимого объема выборки.

Надежностью (доверительной вероятностью) оценки Θ^* параметра Θ называется вероятность γ того, что выполняется неравенство $|\Theta^* - \Theta| < \delta$. Таким образом, γ есть вероятность того, что Θ попадает в интервал $(\Theta^* - \delta, \Theta^* + \delta)$.

Доверительным называется интервал, в который попадает неизвестный параметр с заданной надежностью γ .

Доверительные интервалы строятся:

- а) для оценки математического ожидания нормального распределения при известной дисперсии;
- б) для оценки математического ожидания нормального распределения при неизвестной дисперсии;
- в) для оценки среднего квадратичного отклонения нормального распределения.

2. Решение типового варианта и методические указания

Задача № 1.

Проведено выборочное обследование 25 частных фирм по количеству занятых в них служащих, получены следующие результаты (чел.):

266, 278, 315, 336, 336, 354, 368, 369, 391, 408, 411, 416, 427, 437, 444, 444, 457, 462, 481, 483, 495, 512, 518, 536, 576

Требуется:

- 1) составить интервальный ряд распределения выборки;
- 2) построить гистограмму частот и выдвинуть гипотезу о неизвестном законе распределения эмпирических данных;
- 3) найти выборочную среднюю, выборочную дисперсию и выборочное среднее квадратичное отклонение;
- 4) найти доверительный интервал для оценки неизвестного математического ожидания генеральной средней;
- 5) дать экономическую интерпретацию полученных результатов.

Исходные данные : объем выборки $n = 25$; интервал, содержащего все значения выборки ($X_{\text{нм}}=250$ – наименьшее; $X_{\text{нб}}=600$ – наибольшее) ; число интервалов возьмем $k = 7$; надёжность оценки $\gamma = 0,95$; коэффициент Стьюдента $t = 2,05$.

Решение

1) Для составления интервального ряда распределения выборки объема $n = 25$ обозначим $(x_{\text{нм}}; x_{\text{нб}})$ интервал, содержащий все значения выборки, в котором за наименьшее значение примем $x_{\text{нм}} = 250$, а за наибольшее - $x_{\text{нб}} = 600$. Число интервалов возьмем $k = 7$. Тогда длину частичного интервала h (шаг) найдем по формуле

$$h = \frac{x_{\text{нб}} - x_{\text{нм}}}{k} = \frac{600 - 250}{7} = 50.$$

Составим интервальное распределение выборки с шагом $h=50$, взяв за начало первого интервала $x_1=250$. Результаты поместим в таблицу 1. В первой строке расположим в порядке возрастания интервалы. Во второй строке запишем частоту выборки n_i (количество значений признака в выборке, попавших в i -ый интервал). *Например.* В первый интервал попадают два значения 266, 278. Поэтому частота $n_1=2$. Во втором интервале три значения 315, 336, 336. Поэтому частота $n_2=3$ и т.д.

Интервальный статистический ряд таков:

Таблица 1

интервал	250–300	300–350	350–400	400–450	450–500	500–550	550–600
n_i	2	3	4	7	5	3	1

- 2) Для построения гистограммы частот на оси абсцисс откладываем частичные интервалы длиной $h=50$; на каждом из них строим прямоугольники высотой $\frac{n_i}{h}$ (рис.1).

Составим вспомогательную таблицу.

Таблица 2

интервал	250–300	300–350	350–400	400–450	450–500	500–550	550–600
n_i	2	3	4	7	5	3	1
$\frac{n_i}{h}$	0,04	0,06	0,08	0,14	0,1	0,06	0,02

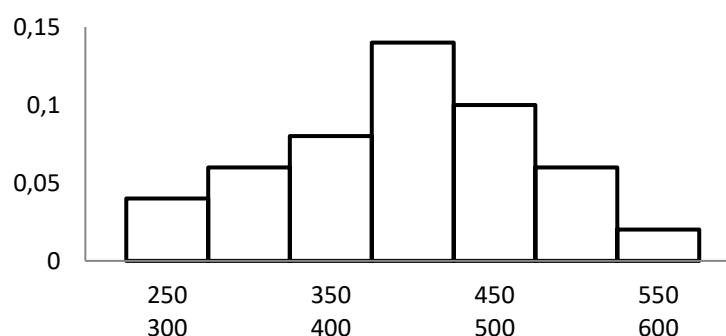


Рис.1. Гистограмма частот.

Форма гистограммы позволяет предположить, что закон распределения эмпирических данных близок к нормальному (наблюдается незначительная асимметрия.)

- 3) Выборочную среднюю найдём по формуле: $\bar{x}_B = \frac{1}{n} \sum_{i=1}^k n_i x_i$.

Для удобства вычислений построим вспомогательную таблицу 3, в которой в первых двух строках записаны полученные интервалы и их частоты. В третью строку поместим середины x_i соответствующих интервалов. В четвертой строке запишем произведения $x_i n_i$.

Используя 4-ую строчку таблицы 3, найдем выборочную среднюю:

$$\bar{x}_B = \frac{1}{n} \sum_{i=1}^k n_i x_i = \frac{1}{25} (275 \cdot 2 + 325 \cdot 3 + 375 \cdot 4 + 425 \cdot 7 + 525 \cdot 3 + 575) = 421.$$

Таблица 3

интервал	250–300	300–350	350–400	400–450	450–500	500–550	550–600
n_i	2	3	4	7	5	3	1
x_i	275	325	375	425	475	525	575
$x_i n_i$	275·2	325·3	375·4	425·7	475·5	525·3	575·1
$(x_i - \bar{x}_B)^2$	(275-421) ²	(325-421) ²	(375-421) ²	(425-421) ²	(475-421) ²	(525-421) ²	(575-421) ²
$n_i(x_i - \bar{x}_B)^2$	2(-146) ²	3(-96) ²	4(-46) ²	7(4) ²	5(34) ²	3(84) ²	1(134) ²

В пятой строке поместим квадраты отклонений середин интервалов от выборочной средней. В последней строке запишем произведения квадратов этих отклонений и соответствующих частот.

Используя 6-ую строчку таблицы 3, найдем выборочную дисперсию:

$$D_B = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x}_e)^2 = \frac{1}{25} (2(-146)^2 + 3(-96)^2 + 4(-46)^2 + 7(4)^2 + 5(34)^2 + 3(84)^2 + (134)^2) = 4950.$$

Найдем выборочное среднее квадратичное отклонение: $\sigma_B = \sqrt{D_B} = \sqrt{4950} \approx 70$.

4) Для определения доверительного интервала с надёжностью $\gamma = 0,95$ оценки неизвестного математического ожидания нормально распределённого признака X генеральной совокупности воспользуемся формулой

$$\bar{x}_B - \frac{s \cdot t}{\sqrt{n}} \leq a \leq \bar{x}_B + \frac{s \cdot t}{\sqrt{n}},$$

где коэффициент Стьюдента $t = 2,05$ берем по таблице, s - исправленное среднее квадратичное отклонение, a - математическое ожидание нормального распределения, объем выборки $n = 25$, выборочная средняя $\bar{x}_B = 421$.

Так как объем выборки мал ($n = 25 < 100$), исправим полученное

$$\sigma_B = 70 \text{ по формуле } s = \sqrt{\frac{n}{n-1}} \sigma_B = \sqrt{\frac{25}{25-1}} \cdot 70 \approx 71,44.$$

$$\text{Точность оценки } \Delta = \frac{s \cdot t}{\sqrt{n}} = \frac{71,44 \cdot 2,05}{\sqrt{25}} = \frac{146,452}{5} \approx 29,29.$$

Доверительный интервал для оценки неизвестного математического ожидания a генеральной средней примет вид:

$$421 - 29,29 \leq a \leq 421 + 29,29;$$

$$391,71 \leq a \leq 450,29.$$

5) Полученная интервальная оценка (с надёжностью $\gamma = 0,95$) математического ожидания количества занятых служащих позволяет с вероятностью 95% утверждать, что среднее количество занятых служащих в генеральной совокупности находится в пределах от 391,71 до 450,29 чел. Интервал $\pm 29,29$ составляет примерно $\pm 7\%$ среднего количества занятых в выборке ($\bar{x}_B = 421$). Это небольшое отклонение, поэтому среднее значение выборки можно считать надёжной оценкой среднего значения генеральной совокупности.

3. Варианты индивидуальных заданий

Задача №1

Проведено выборочное обследование 30 частных фирм по количеству занятых в них служащих. Результаты обследования помещены в таблице каждого варианта (30 человек в каждой таблице). В первом столбце таблицы

у каждого варианта заданы границы интервала, содержащего все значения выборки ($X_{\text{нм}}$ – наименьшее; $X_{\text{нб}}$ - наибольшее)

Требуется:

- 1) составить интервальный ряд распределение выборки;
- 2) построить гистограмму частот и выдвинуть гипотезу о неизвестном законе распределения эмпирических данных;
- 3) найти выборочную среднюю, выборочную дисперсию и выборочное среднее квадратичное отклонение;
- 4) найти доверительный интервал для оценки неизвестного математического ожидания генеральной средней;
- 5) дать экономическую интерпретацию полученных результатов.

Исходные данные (одинаковые у всех вариантов): объем выборки $n = 30$; число интервалов возьмем $k = 7$; надёжность оценки $\gamma = 0,95$; коэффициент Стьюдента $t = 2,05$.

Вариант 1															
$X_{\text{нм}}=250$	266	278	315	336	347	348	349	368	368	370	375	375	380	408	408
$X_{\text{нб}}=600$	411	416	427	444	444	444	457	462	481	483	512	518	536	560	576

Вариант 2															
$X_{\text{нм}}=350$	366	378	415	436	447	448	449	468	468	470	475	475	480	508	508
$X_{\text{нб}}=700$	511	516	527	544	544	544	557	562	581	583	612	618	636	660	676

Вариант 3															
$X_{\text{нм}}=450$	466	478	515	536	547	548	549	568	568	570	575	575	580	608	608
$X_{\text{нб}}=800$	611	616	627	644	644	644	657	662	681	683	712	718	736	760	776

Вариант 4															
$X_{\text{нм}}=550$	566	578	615	636	647	648	649	668	668	670	675	675	680	708	708
$X_{\text{нб}}=900$	711	716	727	744	744	744	757	762	781	783	812	818	836	860	876

Вариант 5															
$X_{\text{нм}}=650$	666	678	715	736	747	748	749	768	768	770	775	775	780	808	808
$X_{\text{нб}}=1000$	811	816	827	844	844	844	857	862	881	883	912	918	936	960	976

Вариант 6															
$X_{\text{нм}}=150$	166	178	215	236	247	248	249	268	268	270	275	275	280	308	308
$X_{\text{нб}}=500$	311	316	327	344	344	344	357	362	381	383	412	418	436	460	476

Вариант 7															
$X_{\text{нм}}=50$	66	78	115	136	147	148	149	168	168	170	175	175	180	208	208

X _{нб} =400	211	216	227	244	244	244	257	262	281	283	312	318	336	360	376
----------------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Вариант 8															
X _{нм} =100	116	128	165	186	197	198	199	218	218	220	225	225	230	258	258
X _{нб} =450	261	266	267	294	294	294	307	312	331	333	362	368	386	410	426

Вариант 9															
X _{нм} =200	216	228	265	286	297	298	299	318	318	320	325	325	330	358	358
X _{нб} =550	361	366	367	394	394	394	407	412	431	433	462	468	486	510	526

Вариант10															
X _{нм} =300	316	328	365	386	397	398	399	418	418	420	425	425	430	458	458
X _{нб} =650	461	466	467	494	494	494	507	512	531	533	562	568	586	610	626

Вариант11															
X _{нм} =400	416	428	465	486	497	498	499	518	518	520	525	525	530	558	558
X _{нб} =750	561	566	567	594	594	594	607	612	631	633	662	668	686	710	726

Вариант12															
X _{нм} =500	516	528	565	586	597	598	599	618	618	620	625	625	630	658	658
X _{нб} =850	661	666	667	694	694	694	707	712	731	733	762	768	786	810	826

Вариант13															
X _{нм} =600	616	628	665	686	697	698	699	718	718	720	725	725	730	758	758
X _{нб} =950	761	766	767	794	794	794	807	812	831	833	862	868	886	910	926

Вариант14															
X _{нм} =650	666	678	715	736	747	748	749	768	768	770	775	775	780	808	808
X _{нб} =1000	811	816	827	844	844	844	857	862	881	883	912	918	936	960	976

Вариант15															
X _{нм} =150	166	178	215	236	247	248	249	268	268	270	275	275	280	308	308
X _{нб} =500	311	316	327	344	344	344	357	362	381	383	412	418	436	460	476

Вариант16															
X _{нм} =50	66	78	115	136	147	148	149	168	168	170	175	175	180	208	208
X _{нб} =400	211	216	227	244	244	244	257	262	281	283	312	318	336	360	376

Вариант17															
X _{нм} =300	316	328	365	386	397	398	399	418	418	420	425	425	430	458	458
X _{нб} =650	461	466	467	494	494	494	507	512	531	533	562	568	586	610	626

Вариант18															
$X_{HM}=100$	116	128	165	186	197	198	199	218	218	220	225	225	230	258	258
$X_{H6}=450$	261	266	267	294	294	294	307	312	331	333	362	368	386	410	426
Вариант19															
$X_{HM}=350$	366	378	415	436	447	448	449	468	468	470	475	475	480	508	508
$X_{H6}=700$	511	516	527	544	544	544	557	562	581	583	612	618	636	660	676

Вариант20															
$X_{HM}=450$	466	478	515	536	547	548	549	568	568	570	575	575	580	608	608
$X_{H6}=800$	611	616	627	644	644	644	657	662	681	683	712	718	736	760	776