Математическая статистика

Основные понятия

Вариационные ряды

Множество всех объектов, подлежащих исследованию, называют генеральной совокупностью. Множество объектов, случайным образом отобранных из генеральной совокупности, называется выборкой. Объемом совокупности (генеральной или выборочной) называют число объектов этой совокупности.

Последовательность результатов наблюдения $x_1, x_2, \ldots, x_m$

записанных в порядке неубывания, т.е. $x_1 \le x_2 \le \ldots \le x_m$

называется вариационным рядом.

Если варианты $x_1, x_2, \ldots, x_m$

при наблюдении встретились соответственно $n_1, n_2, \ldots, n_m$ раз,

то числа $n_1, n_2, \ldots, n_m$ называются частотами.

Если объем выборки равен *n*, то $n_1 + n_2 + \ldots + n_m = n$

Статистическая таблица частот				
Варианты x_i	x_1	x_2	…	x_m
Частоты n_i	n_1	n_2	…	n_m

Отношения частот к объему выборки

$\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_m}{n}$ называются относительными частотами.

Статистическая таблица относительных частот				
Варианты x_i	x_1	x_2	…	x_m
Относительные частоты $\frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_m}{n}$

Провели следующий эксперимент. Книгу открывали на случайной странице, где выбирали случайное слово. При этом фиксировали длину слова. В результате 20 опытов получена следующая выборка:
4, 1, 4, 5, 1, 13, 4, 10, 2, 4, 7, 2, 2, 4, 6, 4, 5, 6, 2, 4.

Ей соответствует вариационный ряд:
1, 1, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, 4, 5, 5, 6, 6, 7, 10, 13.

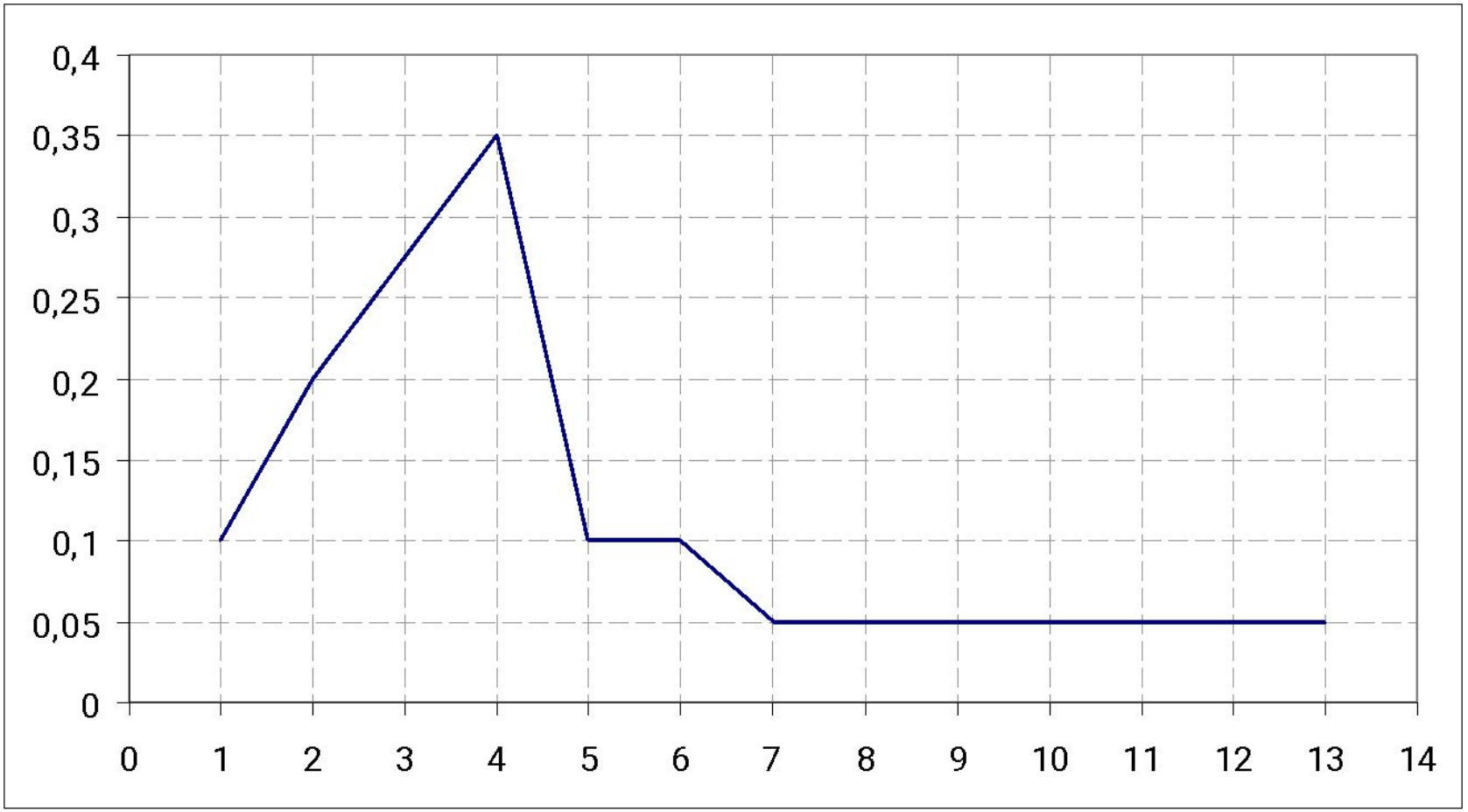
Статистическая таблица частот

x_i	1	2	4	5	6	7	10	13
n_i	2	4	7	2	2	1	1	1

Статистическая таблица относительных частот

x_i	1	2	4	5	6	7	10	13
$\frac{n_i}{n}$	0,1	0,2	0,35	0,1	0,1	0,05	0,05	0,05

Рассмотрим полигон относительных частот статистического распределения, приведенного в таблице.



Во многих задачах значения признака разбивают на группы. Статистическое распределение выборки задают в виде последовательности интервалов и соответствующих им частот. В качестве частоты, соответствующей интервалу, принимают сумму частот вариант, попавших в этот интервал.

Если каждое значение частоты разделить на длину l_i

соответствующего интервала, то полученные числа

$\frac{n_1}{l_1}, \frac{n_2}{l_2}, \ldots, \frac{n_m}{l_m}$ называют плотностями частот.

Если каждое значение относительной частоты разделить на длину l_i

соответствующего интервала, то полученные числа

$$\frac{n_1}{n \cdot l_1}, \frac{n_2}{n \cdot l_2}, \ldots, \frac{n_m}{n \cdot l_m}$$

называют плотностями относительных частот.

Для наглядности изображения статистической таблицы строят ступенчатую фигуру, состоящую из прямоугольников, в основании которых лежат интервалы, а высотами являются соответствующими им плотности частот или относительные плотности частот.

Гистограммой частот называется ступенчатая фигура, состоящая из прямоугольников с основанием

$$h = x_i - x_{i-1}$$ и высотами $$\frac{n_i}{h}$$

На оси абсцисс откладывают частичные интервалы длиной h, на i-м интервале строят прямоугольник высотой

$$\frac{n_i}{h}$$ (плотность частоты).

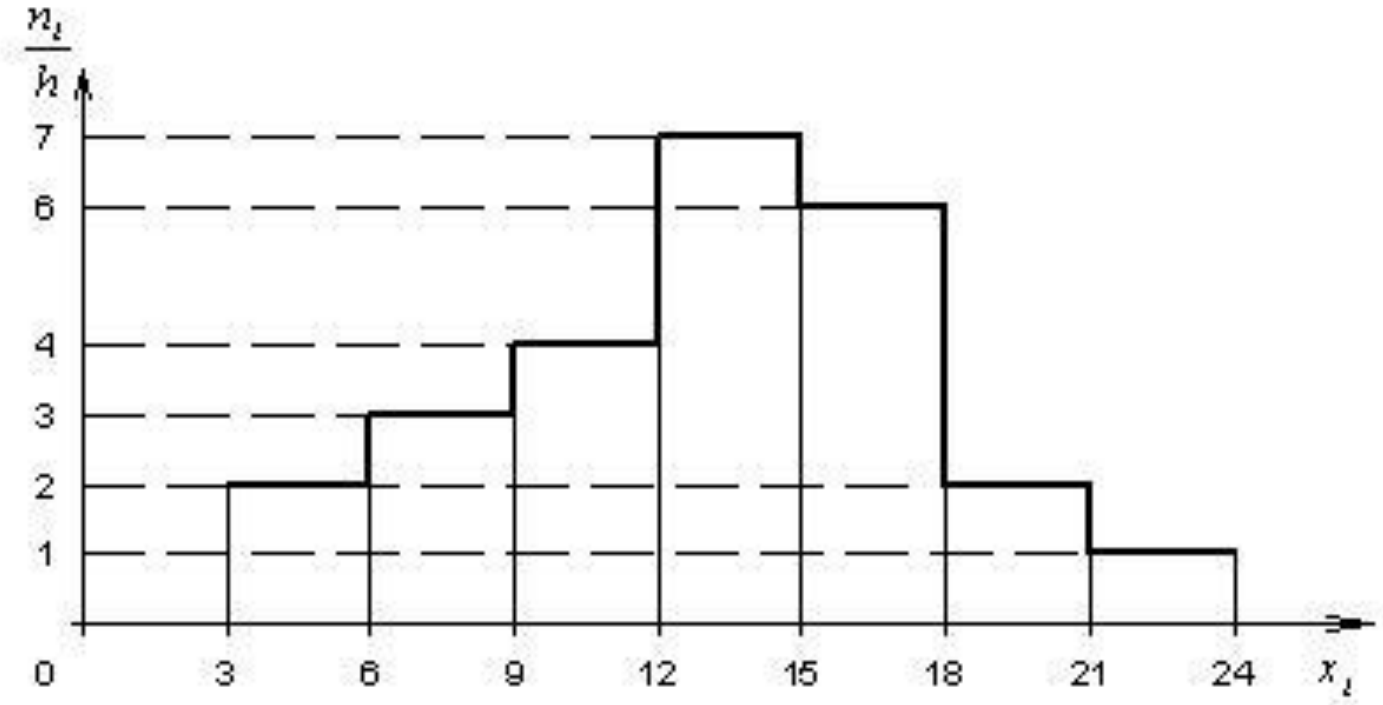
Площадь S гистограммы частот равна сумме всех частот, т.е. объему выборки.

Действительно, если S_i – площадь прямоугольника, то

$$S_i = h\frac{n_i}{h} = n_i \qquad S = \sum_{i=1}^{k} S_i = \sum_{i=1}^{k} n_i = n$$

Приведем гистограмму частот распределения объема n = 75, указанного в таблице.

Частичный интервал длины $h = 3$	[3; 6]	(6; 9]	(9; 12]	(12;15]	(15; 18]	(18; 21]	(21; 24]
Сумма частот частичного интервала n_i	6	9	12	21	18	6	3
Плотность частоты $\frac{n_i}{h}$	2	3	4	7	6	2	1


$\frac{n_i}{h}$
7
6
4
3
2
1
0
3
6
9
12
15
18
21
24
x_i

Эмпирическая функция распределения

Эмпирической функцией распределения (функцией распределения выборки) называется функция

$$F^*(x)$$

определяющая для каждого значения x частоту события $X < x$

Пусть n_x – число вариант, меньших x, n – объем выборки. Тогда

$$F^*(x) = \frac{n_x}{n}$$

Из определения эмпирической функции $F^*(x)$ следуют ее свойства:

1. Значения функции $F^*(x)$ принадлежат отрезку [0,1].

2. $F^*(x)$ – неубывающая функция.

3. Если a – наименьшая, b – наибольшая варианта, то

$F^*(x)=0$ при $x \le a$

$F^*(x)=1$ при $x \ge b$

4. Функция $F^*(x)$

непрерывна слева, так как она постоянна на полуинтервалах $(x_i, x_{i+1}]$

Пример 1. Построить эмпирическую функцию по данному распределению выборки

Варианты x_i	6	8	12	15
Частоты n_i	2	3	10	5

Объем выборки

$$n = 2 + 3 + 10 + 5 = 20$$

Наименьшая варианта $x_1 = 6$ поэтому $F^*(x) = 0$

если $x \le 6$

Значение $X < 8$ $x_1 = 6$ наблюдалось 2 раза, поэтому

$$F^*(x) = \frac{2}{20} = 0{,}1$$ если $6 < x \le 8$

Значения $X<12$ $x_1=6,\quad x_2=8$

наблюдались $2+3=5$ раз, поэтому $F^*(x)=\frac{5}{20}=0{,}25$

если $8<x\le 12$

Значения $X<15$ $x_1=6,\quad x_2=8,\quad x_3=12$

наблюдались $2+3+10=15$ раз, поэтому

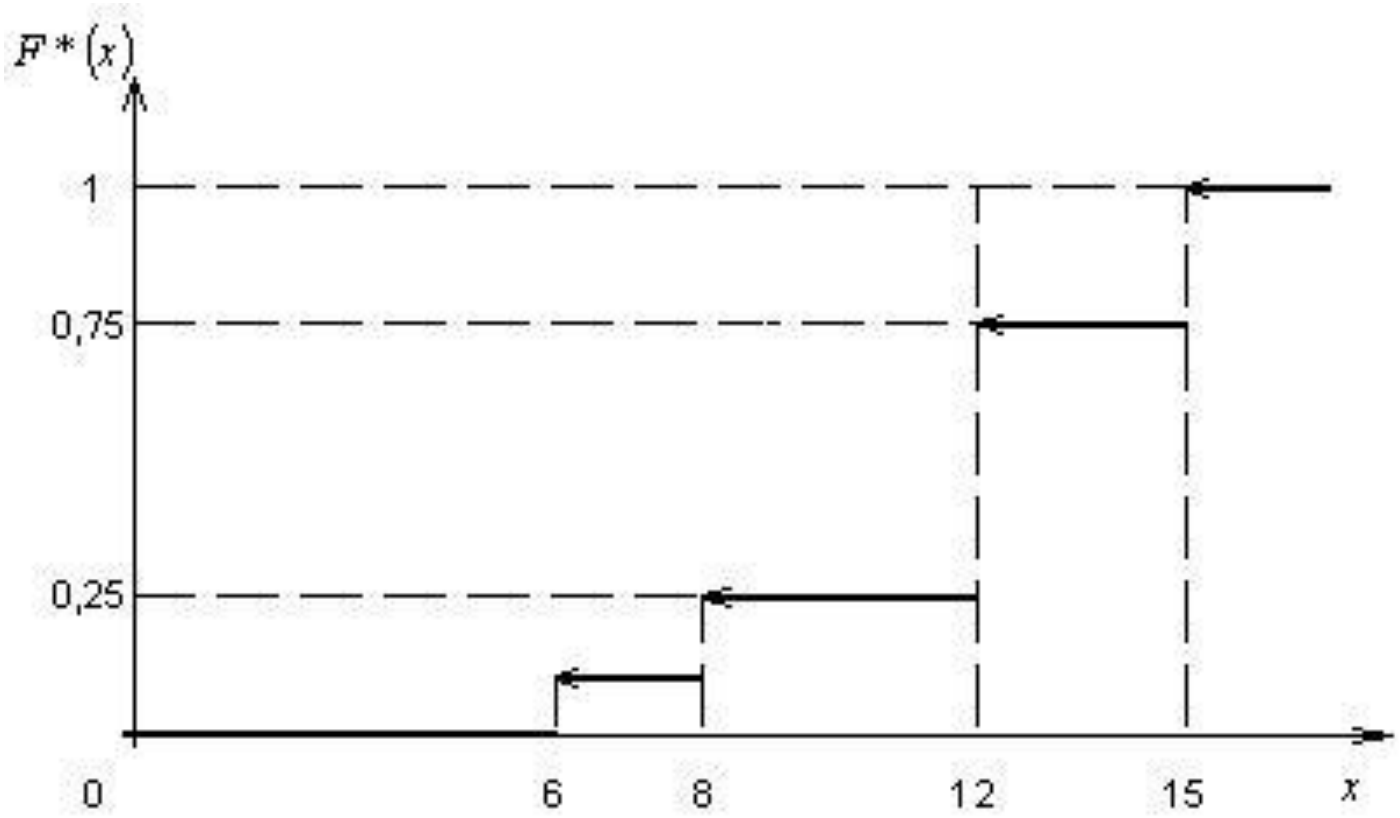
$F^*(x)=\frac{15}{20}=0{,}75$ если $12<x\le 15$

Поскольку $x_4=15$ – наибольшая варианта, то

$F^*(x)=1$ если $x>15$

Итак, искомая эмпирическая функция определяется формулами

$$F^*(x)=\begin{cases}0 & при\ x\le 6\\ 0{,}1 & при\ 6<x\le 8\\ 0{,}25 & при\ 8<x\le 12\\ 0{,}75 & при\ 12<x\le 15\\ 1 & при\ x>15\end{cases}$$


F*(x)
1
0,75
0,25
0
6
8
12
15
x

Числовые характеристики вариационных рядов

Средним арифметическим называется постоянная, равная сумме произведений значений признака на соответствующие значения относительных частот

$$\bar{x} = x_1 \cdot \frac{n_1}{n} + x_2 \cdot \frac{n_2}{n} + x_m \cdot \frac{n_m}{n} = \frac{\sum_{i=1}^{m} x_i n_i}{n}$$

Размахом вариации *R* называется разность между наибольшим и наименьшим значениями признака

$$R = x_{\max} - x_{\min}$$

Модой *Мо* называется значение признака, встречающееся с наибольшей частотой, т.е. наиболее типичное в данном вариационном ряду.

Медианой *Ме* называется значение признака, лежащее в середине вариационного ряда, если этот ряд имеет нечетное число членов, и среднее арифметическое двух значений признака, расположенных в середине ряда, если ряд состоит из четного числа членов.

Статистические оценки параметров распределения

Статистическая таблица частот					
Варианты	x_i	x_1	x_2	…	x_m
Частоты	N_i	N_1	N_2	…	N_m

$$N_1 + N_2 + \ldots + N_m = N$$

Генеральную среднюю подсчитывают по формуле

$$\overline{x}_{Г} = \frac{x_1 N_1 + x_2 N_2 + \square + x_m N_m}{N} = \frac{1}{N}\sum_{i=1}^{m} x_i N_i$$

а генеральную дисперсию по формулам:

$$D_{Г} = \frac{1}{N}\sum_{i=1}^{m}\left(x_i - \overline{x}_{Г}\right)^2 N_i$$

$$D_{Г} = \frac{1}{N}\sum_{i=1}^{m} x_i^2 \cdot N_i - \left[\frac{1}{N}\sum_{i=1}^{m} x_i \cdot N_i\right]^2$$

Выборочную среднюю подсчитывают по формуле

$$\bar{x}_B = \frac{x_1 n_1 + x_2 n_2 + \square + x_m n_m}{n} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i$$

а выборочную дисперсию по формулам:

$$D_B = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \bar{x}_B\right)^2 n_i$$

$$D_B = \frac{1}{n}\sum_{i=1}^{m} x_i^2 \cdot n_i - \left[\frac{1}{n}\sum_{i=1}^{m} x_i \cdot n_i\right]^2$$

Выборочная дисперсия является заниженной оценкой генеральной дисперсии. Несмещенной оценкой генеральной дисперсии является исправленная дисперсия.

$$s^2 = \frac{n}{n-1} D_B = \frac{1}{n-1} \sum_{i=1}^{m} \left(x_i - \bar{x}_B \right)^2 \cdot n_i$$

В супермаркете проводились наблюдения над числом покупателей, обратившихся в кассу за 1 час. Наблюдения проводились в течение 30 часов (15 дней в период с 9 до 10 и с 10 до 11 часов) дали следующие результаты:
70, 75, 100, 120, 75, 60, 100, 120, 70, 60, 65, 100, 65, 100, 70, 75, 60, 100, 100, 120, 70, 75, 70, 120, 65, 70, 75, 70, 100, 100.
Составить ряд распределения частот. Найти моду, медиану, размах выборки. Найти выборочное среднее и несмещенную оценку дисперсии.

Составим вариационный ряд

60, 60, 60, 65, 65, 65, 70, 70, 70, 70, 70, 70, 70, 75, 75, 75, 75, 75, 100, 100, 100, 100, 100, 100, 100, 100, 120, 120, 120, 120

Составим ряд распределения частот

Номер группы	i	1	2	3	4	5	6
Число обращений	x_i	60	65	70	75	100	120
Частота	n_i	3	3	7	5	8	4

Составим ряд распределения относительных частот $n = 30$

Номер группы	i	1	2	3	4	5	6
Число обращений	x_i	60	65	70	75	100	120
Частота	n_i	3	3	7	5	8	4
Относительная частота	$\frac{n_i}{n}$	$\frac{3}{30}$	$\frac{3}{30}$	$\frac{7}{30}$	$\frac{5}{30}$	$\frac{8}{30}$	$\frac{4}{30}$

60, 60, 60, 65, 65, 65, 70, 70, 70, 70, 70, 70, 70, 75, 75, 75, 75, 75,
100, 100, 100, 100, 100, 100, 100, 100, 120, 120, 120, 120

$$Mo = 100$$

60, 60, 60, 65, 65, 65, 70, 70, 70, 70, 70, 70, 70, 75, 75, 75, 75, 75,
100, 100, 100, 100, 100, 100, 100, 100, 120, 120, 120, 120

$$Me = \frac{75 + 75}{2} = 75$$

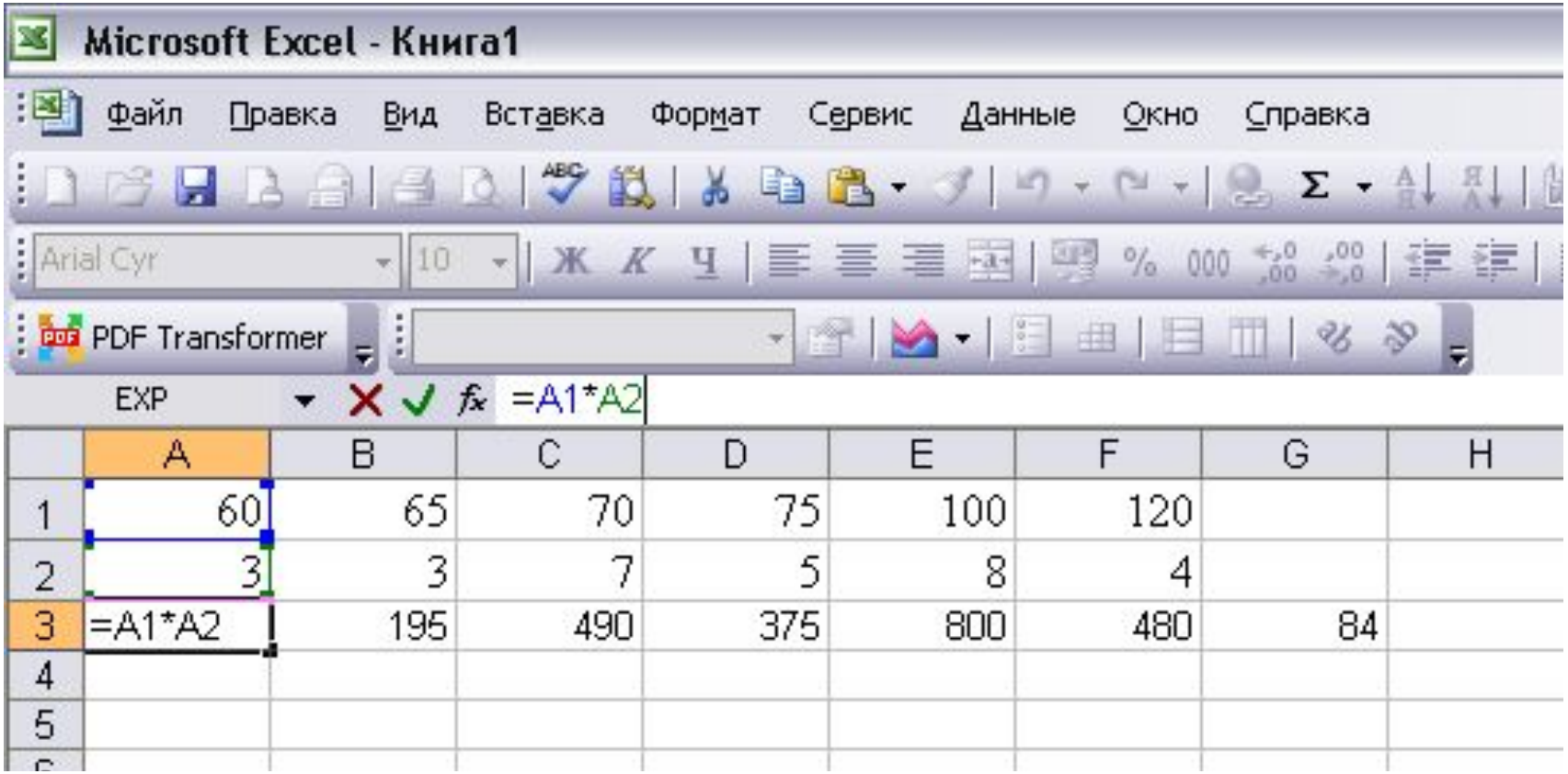
$$R = x_{\max} - x_{\min}$$

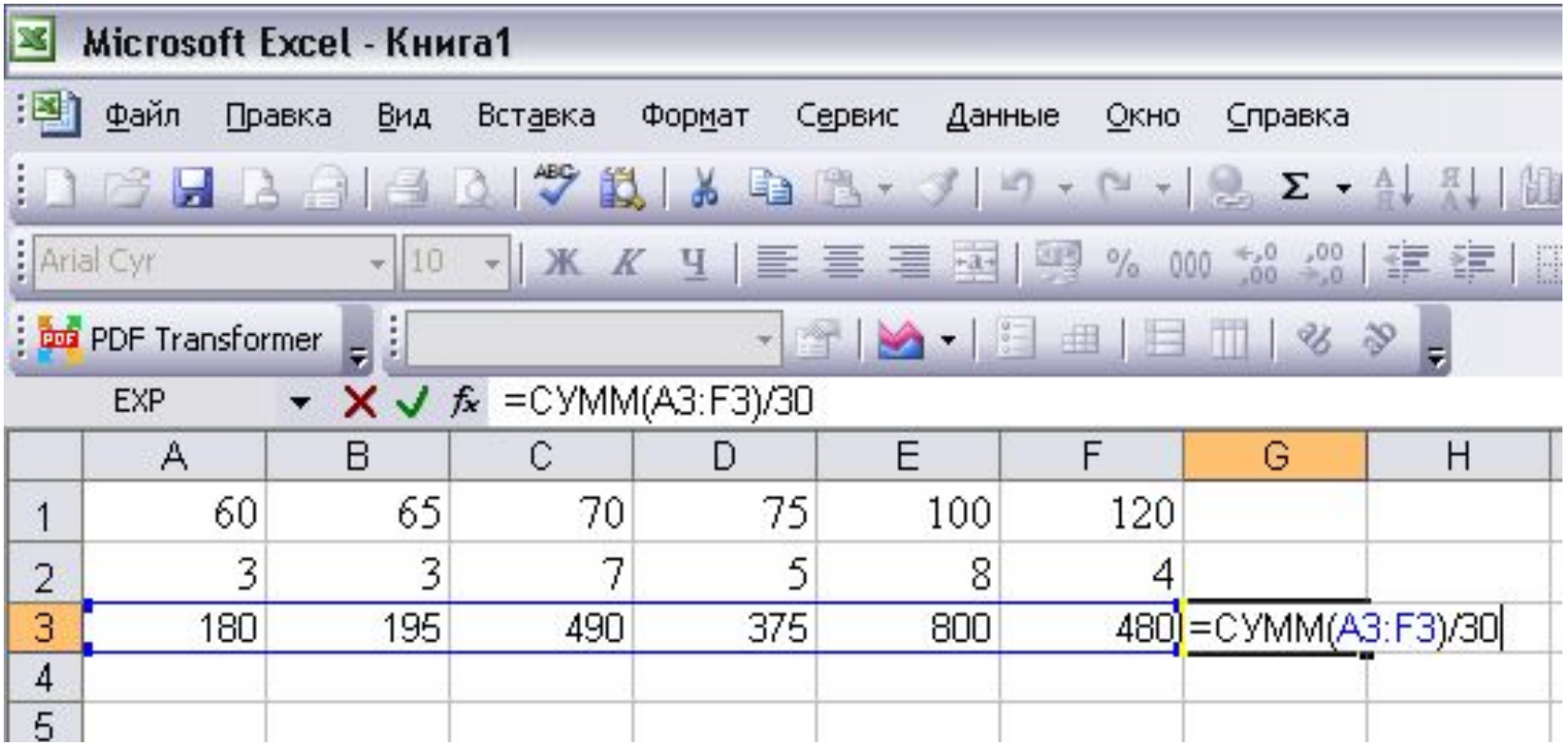
$$R = 120 - 60 = 60$$

Номер группы	i	1	2	3	4	5	6
Число обращений	x_i	60	65	70	75	100	120
Частота	n_i	3	3	7	5	8	4

$$\overline{x}_B = \frac{x_1 n_1 + x_2 n_2 + \square + x_m n_m}{n} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i$$

$$\overline{x}_B = \frac{60 \cdot 3 + 65 \cdot 3 + 70 \cdot 7 + 75 \cdot 5 + 100 \cdot 8 + 120 \cdot 4}{30} = 84$$


Microsoft Excel - Книга1
Файл Правка Вид Вставка Формат Сервис Данные Окно Справка
Arial Cyr
10
PDF Transformer
EXP
=A1*A2
A B C D E F G H
1 60 65 70 75 100 120
2 3 3 7 5 8 4
3 =A1*A2 195 490 375 800 480 84
4
5

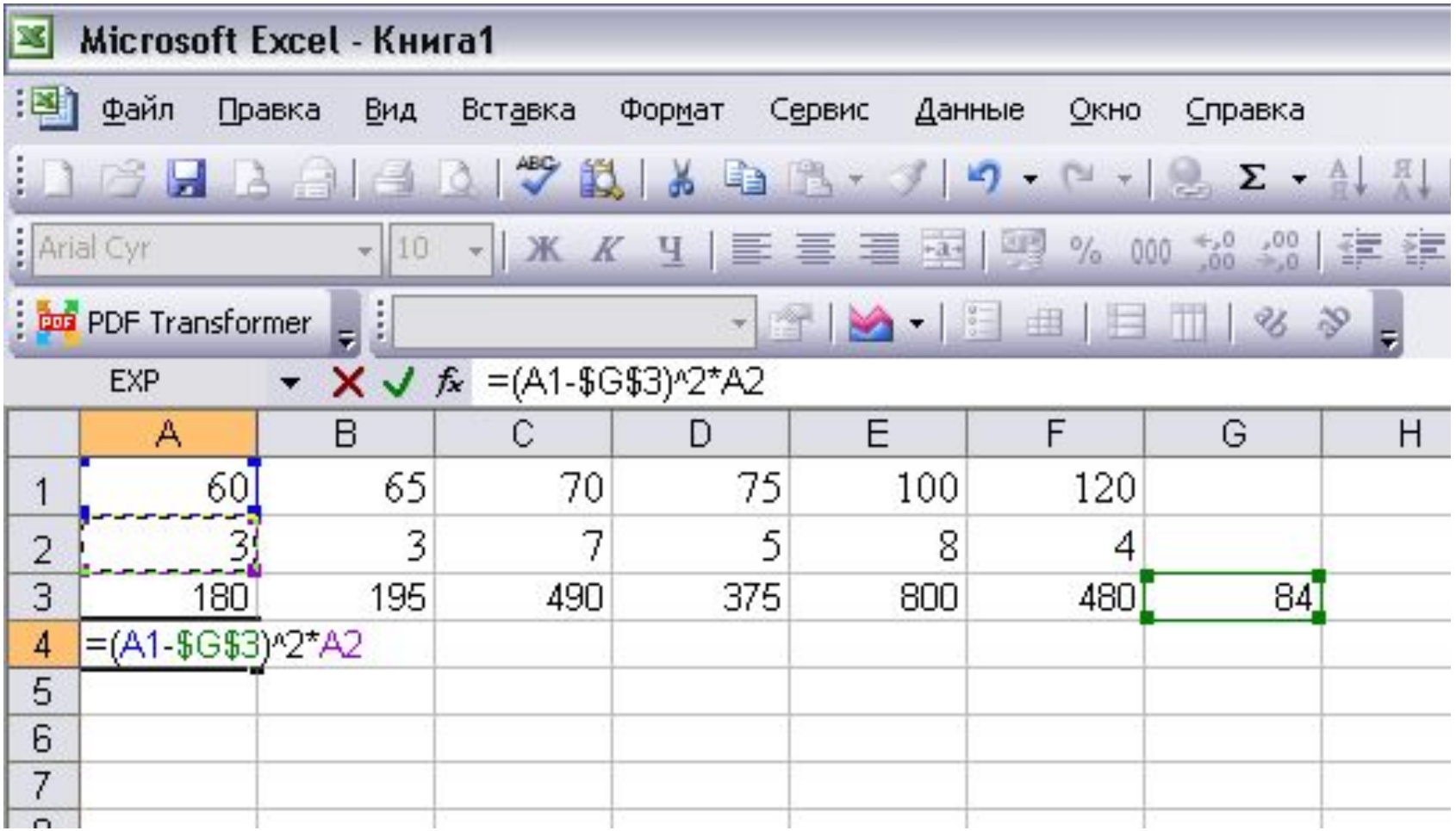

Microsoft Excel - Книга1
Файл Правка Вид Вставка Формат Сервис Данные Окно Справка
Arial Cyr
10
PDF Transformer
EXP
=СУММ(A3:F3)/30
A B C D E F G H
1 60 65 70 75 100 120
2 3 3 7 5 8 4
3 180 195 490 375 800 480 =СУММ(A3:F3)/30
4
5

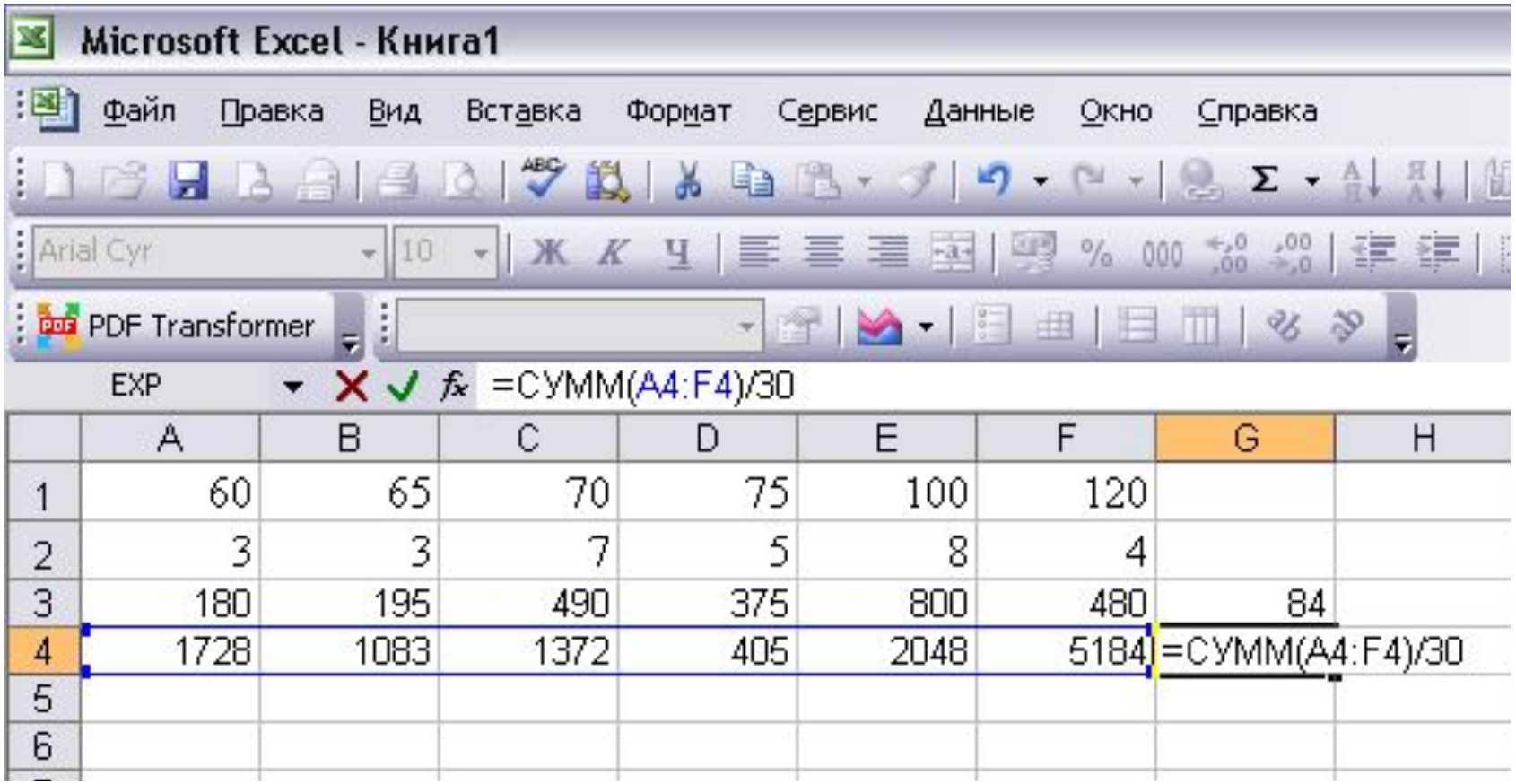
Номер группы	i	1	2	3	4	5	6
Число обращений	x_i	60	65	70	75	100	120
Частота	n_i	3	3	7	5	8	4

$$D_B = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \overline{x}_B\right)^2 n_i$$

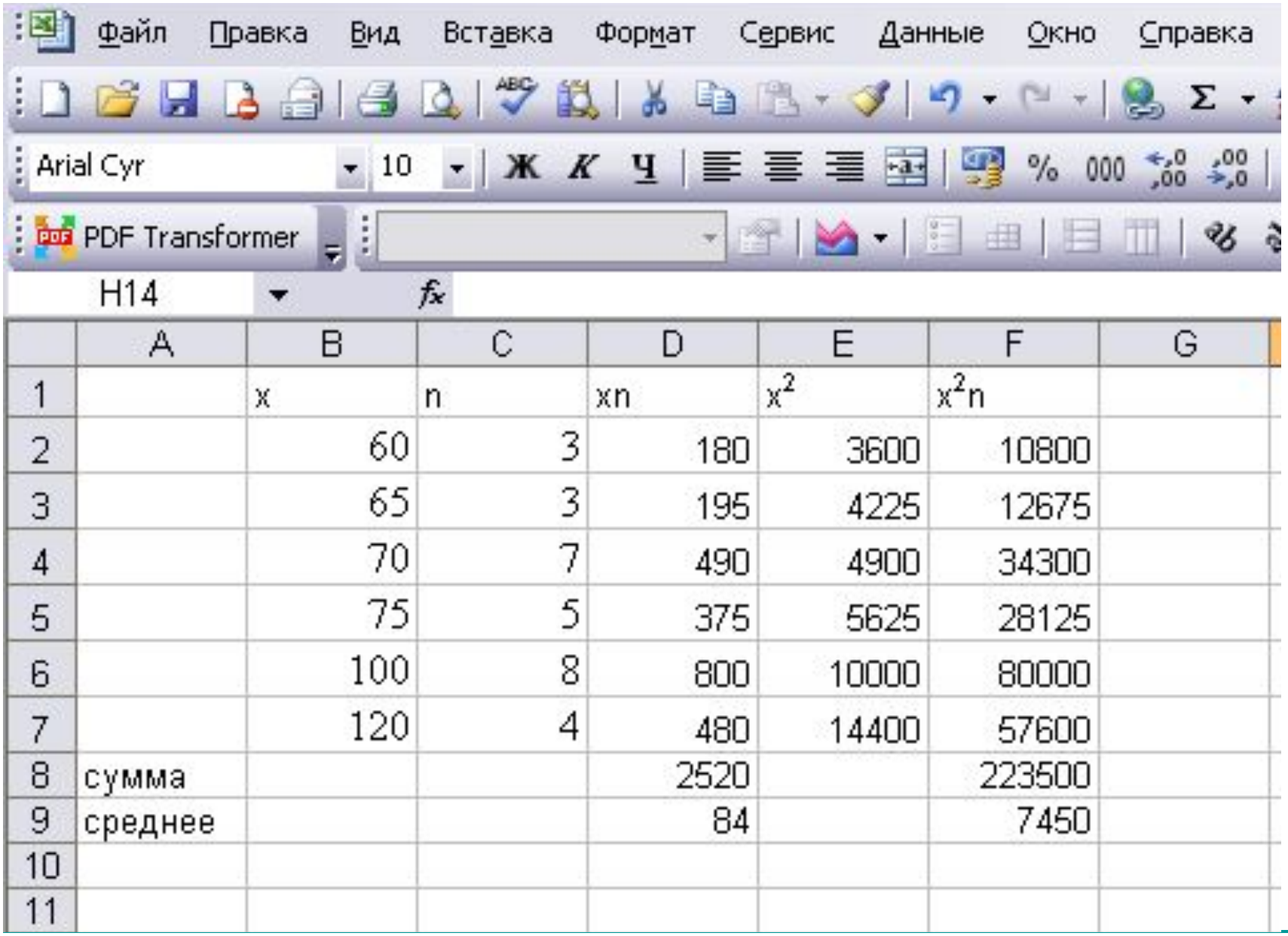
$$D_B = \frac{1}{30}\Big(\left(60-84\right)^2\cdot 3 + \left(65-84\right)^2\cdot 3 + \left(70-84\right)^2\cdot 7 +$$

$$+\left(75-84\right)^2\cdot 5 + \left(100-84\right)^2\cdot 8 + \left(120-84\right)^2\cdot 4\Big) = 394$$


Microsoft Excel - Книга1
Файл Правка Вид Вставка Формат Сервис Данные Окно Справка
Arial Cyr
10
PDF Transformer
EXP
=(A1-G3)^2*A2
A B C D E F G H
1 60 65 70 75 100 120
2 3 3 7 5 8 4
3 180 195 490 375 800 480 84
4 =(A1-G3)^2*A2
5
6
7


Microsoft Excel - Книга1
Файл Правка Вид Вставка Формат Сервис Данные Окно Справка
Arial Cyr
10
PDF Transformer
EXP
=СУММ(A4:F4)/30
A B C D E F G H
1 60 65 70 75 100 120
2 3 3 7 5 8 4
3 180 195 490 375 800 480 84
4 1728 1083 1372 405 2048 5184 =СУММ(A4:F4)/30
5
6

$$D_B = \frac{1}{n}\sum_{i=1}^{m} x_i^2 \cdot n_i - \left[\frac{1}{n}\sum_{i=1}^{m} x_i \cdot n_i\right]^2$$

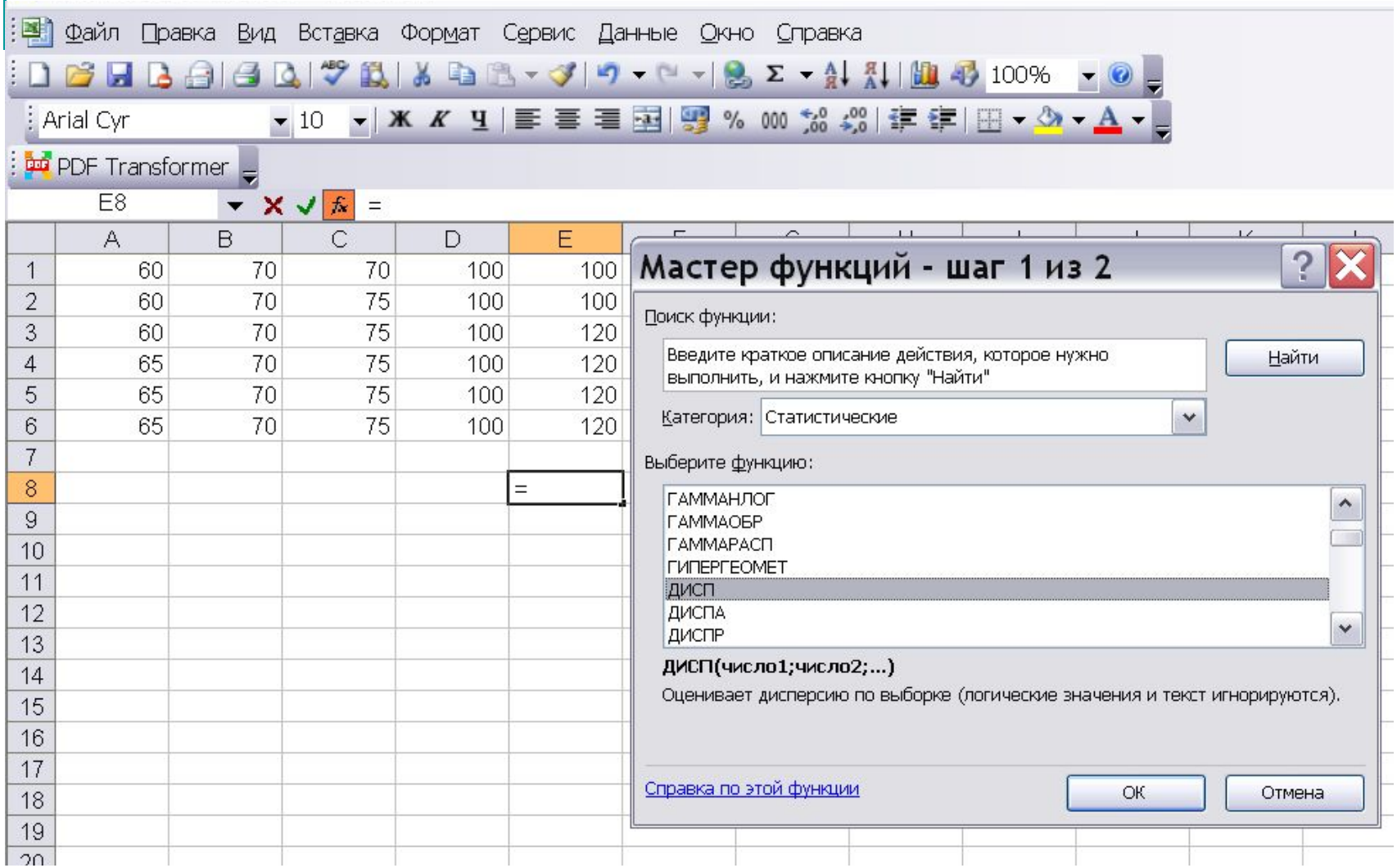


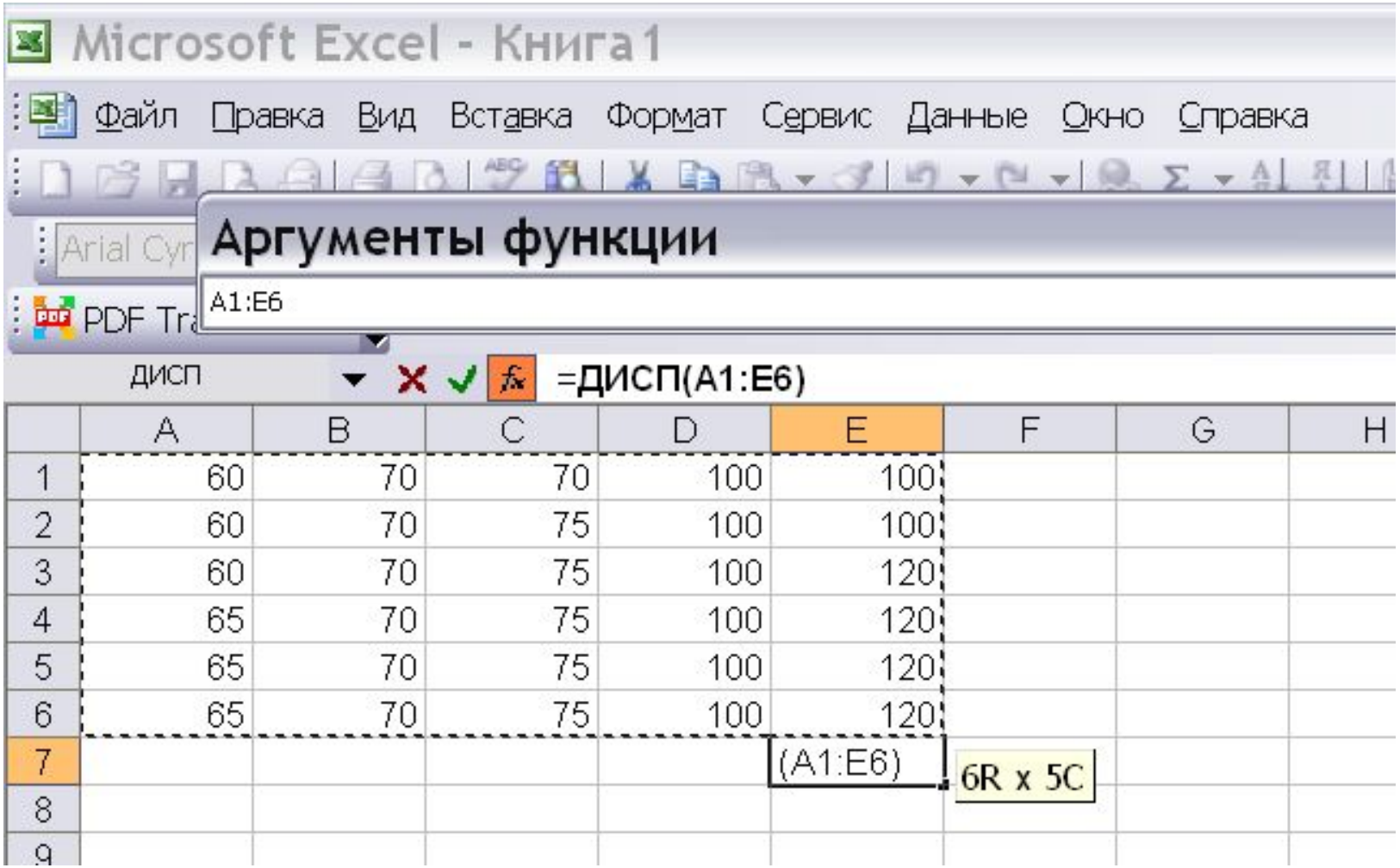
	A	B	C	D	E	F	G
1		x	n	xn	x^2	x^2n	
2		60	3	180	3600	10800	
3		65	3	195	4225	12675	
4		70	7	490	4900	34300	
5		75	5	375	5625	28125	
6		100	8	800	10000	80000	
7		120	4	480	14400	57600	
8	сумма			2520		223500	
9	среднее			84		7450	
10							
11							

$$D_B = \frac{1}{n}\sum_{i=1}^{m} x_i^2 \cdot n_i - \left[\frac{1}{n}\sum_{i=1}^{m} x_i \cdot n_i\right]^2$$

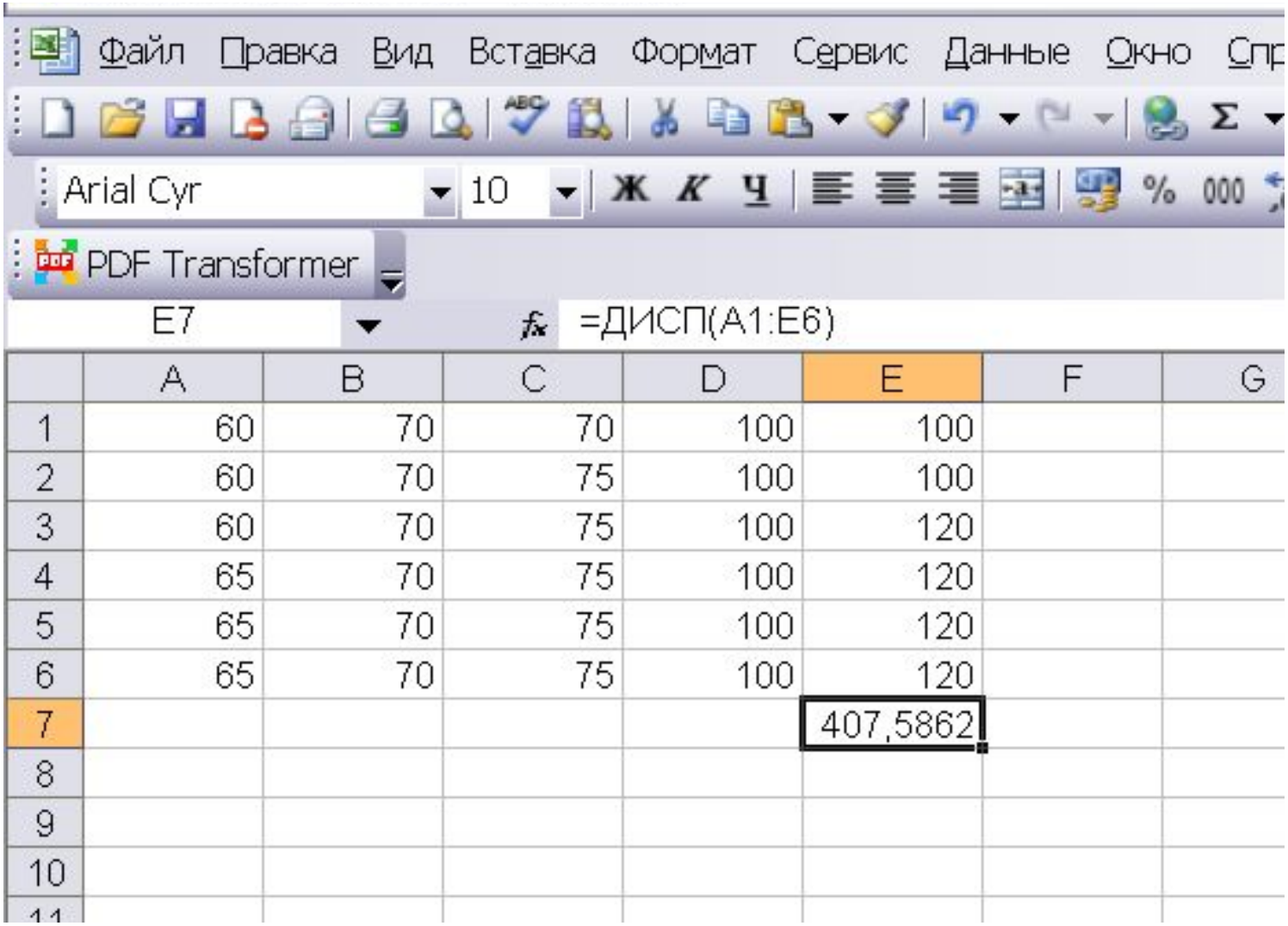
$$D_B = 7450 - 84^2 = 394$$

$$s^2 = \frac{n}{n-1} D_B = \frac{30}{29} \cdot 394 \approx 407{,}59$$


Файл Правка Вид Вставка Формат Сервис Данные Окно Справка
100%
Arial Cyr
10
PDF Transformer
E8
=
A B C D E
1 60 70 70 100 100
2 60 70 75 100 100
3 60 70 75 100 120
4 65 70 75 100 120
5 65 70 75 100 120
6 65 70 75 100 120
8 =
Мастер функций - шаг 1 из 2
Поиск функции:
Введите краткое описание действия, которое нужно выполнить, и нажмите кнопку "Найти"
Найти
Категория: Статистические
Выберите функцию:
ГАММАНЛОГ
ГАММАОБР
ГАММАРАСП
ГИПЕРГЕОМЕТ
ДИСП
ДИСПА
ДИСПР
ДИСП(число1;число2;...)
Оценивает дисперсию по выборке (логические значения и текст игнорируются).
Справка по этой функции
ОК
Отмена



	A	B	C	D	E	F	G	H
1	60	70	70	100	100			
2	60	70	75	100	100			
3	60	70	75	100	120			
4	65	70	75	100	120			
5	65	70	75	100	120			
6	65	70	75	100	120			
7								
8								
9								



	A	B	C	D	E	F	G
1	60	70	70	100	100		
2	60	70	75	100	100		
3	60	70	75	100	120		
4	65	70	75	100	120		
5	65	70	75	100	120		
6	65	70	75	100	120		
7					407,5862		
8							
9							
10							

Составить эмпирическую функцию распределения

x	$F^*(x)$
$x \le 60$	0
$60 < x \le 65$	$\frac{3}{30}$
$65 < x \le 70$	$\frac{3}{30}+\frac{3}{30}=\frac{6}{30}$
$70 < x \le 75$	$\frac{3}{30}+\frac{3}{30}+\frac{7}{30}=\frac{13}{30}$
$75 < x \le 100$	$\frac{3}{30}+\frac{3}{30}+\frac{7}{30}+\frac{5}{30}=\frac{18}{30}$
$100 < x \le 120$	$\frac{3}{30}+\frac{3}{30}+\frac{7}{30}+\frac{5}{30}+\frac{8}{30}=\frac{26}{30}$
$x > 120$	$\frac{3}{30}+\frac{3}{30}+\frac{7}{30}+\frac{5}{30}+\frac{8}{30}+\frac{4}{30}=1$

В таблице приведена выборка результатов измерения роста 105 студентов. Измерения проводились с точностью до 1 см. Требуется составить интервальный вариационный ряд

155	170	185	180	188	152	173	178	178	168	185
173	170	183	175	173	170	183	175	180	175	193
178	183	180	197	178	181	187	168	174	179	184
183	178	180	178	163	166	178	175	182	190	167
170	178	183	170	178	181	173	168	185	175	170
155	169	186	179	189	155	174	179	179	169	186
174	171	184	175	193	178	184	180	196	175	181
188	168	179	178	183	184	178	181	177	163	166
178	175	183	190	167	170	178	183	170	178	182
173	168	186	176	171	188					

n=105 R=197–152=45

Индекс интервала	Рост студентов	Частота	Относительная частота
1	150–155	4	$\frac{4}{105}$
2	155–160	0	0
3	160–165	2	$\frac{2}{105}$
4	165–170	19	$\frac{19}{105}$
5	170–175	18	$\frac{18}{105}$
6	175–180	27	$\frac{27}{105}$
7	180–185	21	$\frac{21}{105}$
8	185–190	10	$\frac{10}{105}$
9	190–195	2	$\frac{2}{105}$
10	195–200	2	$\frac{2}{105}$

Интервальные оценки

В каждом рассмотренном примере результат зависит от рассмотренных выборок. Вполне возможно, что для других выборок будет получен другой результат.
Возникает вопрос: на сколько статистические характеристики отличаются от соответствующих генеральных характеристик?
Для ответа на этот вопрос вводится понятие интервальных оценок генеральных характеристик

Интервальной называют оценку, которая определяется двумя числами – концами интервала

Пусть Θ^* - оценка неизвестного параметра Θ, полученная по данным выборки. Оценка тем точнее, чем меньше величина $|\Theta - \Theta^*|$

Если $\delta > 0$ и $|\Theta - \Theta^*| < \delta$, то чем меньше δ, тем точнее оценка Θ^*, т.е. число δ характеризует точность оценки

Доверительной вероятностью (надежностью) оценки Θ* параметра Θ называется вероятность γ, с которой осуществляется неравенство

$|\Theta - \Theta^*| < \delta$, т.е.

$$\gamma = P\left(\left|\Theta - \Theta^*\right| < \delta\right)$$

Обычно доверительная вероятность задается заранее, причем в качестве γ берут число, близкое к единице.

Наиболее часто надежность задается равной 0,95; 0,99; 0,999.

Так как неравенство $|\Theta - \Theta^*| < \delta$ равносильно неравенству $-\delta < \Theta - \Theta^* < \delta$, или $\Theta^* - \delta < \Theta < \Theta^* + \delta$, то формулу вероятности можно записать в виде

$$\gamma = P\left(\Theta^* - \delta < \Theta < \Theta^* + \delta\right)$$

Вероятность того, что интервал $(\Theta^* - \delta, \Theta^* + \delta)$ заключает в себе неизвестный параметр Θ, равна γ.

Интервал $(\Theta^* - \delta, \Theta^* + \delta)$, который покрывает неизвестный параметр Θ с заданной надежностью γ, называется ***доверительным интервалом***.

Концы доверительного интервала называются ***доверительными границами***.