ПРИЛОЖЕНИЕ 7

В 3/9 царстве, в 3/10 государстве все были помешаны на дробях

Результат наблюдения

Статистические оценки
для интервального ранжированного частотного ряда

В некоторых исследованиях исходные данные для статистической оценки задаются интервальным ранжированным частотным рядом. Например, в виде

Возраст особи		частота встречаемости
лет от	лет до	
18	21	1
21	24	3
24	27	6
27	30	10
30	33	5
33	36	3
36	39	2

Требуется вычислить статистические оценки данных (в списке стандартных функций электронных таблиц необходимые средства для данной формы входных данных отсутствуют). А именно, требуется определить среднее, медиану, моду и параметры вариации.

Медиана (*Me*) – это величина, которая соответствует варианту, находящемуся в середине ранжированного ряда.

Для ранжированного ряда с нечетным числом индивидуальных величин (например, 1, 2, 4, 4, 6, 7, 8, 8, 10) медианой будет величина, которая расположена в центре ряда, т.е. пятая величина.

Для ранжированного ряда с четным числом индивидуальных величин (например, 1, 5, 7, 10, 11, 14) медианой будет средняя арифметическая величина, которая рассчитывается из двух смежных величин. В данном случае медиана равна (7+10) : 2= 8,5.

Таким образом, для нахождения медианы сначала необходимо определить ее порядковый номер (ее положение в ранжированном ряду) по формуле

$$N_{Me} = \frac{n+1}{2},$$ где n – объем совокупности.

Численное значение медианы определяют по накопленным частотам дискретного вариационного ряда, для чего требуется указать интервал нахождения медианы в интервальном ряду распределения. Медианным называют первый интервал, где сумма накопленных частот превышает половину наблюдений от общего числа всех наблюдений.

Численное значение медианы определятся по формуле

$$Me = x_{Me} + \Delta \frac{\frac{n+1}{2} - S_{-1}}{f_{Me}},$$

где x_{Me} – нижняя граница медианного интервала;
Δ – величина интервала;
S_{-1} – накопленная частота интервала, которая предшествует медианному;
f_{Me} – частота медианного интервала.

Модой (Mo) называют значение признака, которое встречается наиболее часто у единиц совокупности. Для дискретного ряда модой будет являться вариант с наибольшей частотой. Для определения моды интервального ряда сначала определяют модальный интервал (интервал, имеющий наибольшую частоту). Затем в пределах этого интервала находят то значение признака, которое может являться модой.

Чтобы найти конкретное значение моды, необходимо использовать формулу

$$Mo = x_{Mo} + \Delta \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})}$$

где x_{Mo} – нижняя граница модального интервала;
f_{Mo} – частота модального интервала;
f_{Mo-1} – частота интервала, предшествующего модальному;
f_{Mo+1} – частота интервала, следующего за модальным.

Размах вариации R. Это самый доступный по простоте расчета абсолютный показатель, который определяется как разность между самым большим и самым малым значениями признака у единиц данной совокупности:

Вычисление размаха количественной вариации (выборки) производится по формуле:

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

где x_{max} – значение максимальной варианты,
x_{min} – значение минимальной варианты выборки.

Функции MS Excel: МИН(данные); МАКС(данные).
Функция Calc (Open Office): =MIN(данные) и =MAX(данные).

Размах вариации (размах разброса данных) – важный показатель, но только крайних отклонений. Для более точной характеристики рассеяния вариации признака используются другие показатели.

Среднее отклонение (выборочная оценка среднего отклонения), подобно стандартному отклонению, характеризует разброс эмпирической выборки относительно среднего значения и вычисляется по формуле

$$\overline{d} = \frac{1}{n}\sum_i |x_i - \overline{x}|,$$

n – численность выборки,

x_i – значения вариант выборки.

Выборочное среднее значение $\overline{x}$ определяется очевидным соотношением

$$\overline{x} = \frac{\sum x_i f_i}{\sum f_i}.$$

Среднее отклонение отражает так называемый модульный подход к вычислению меры отклонения между величинами в противоположность тому, что стандартное отклонение отражает квадратический подход.

Функция MS Excel: СРОТКЛ(данные);
Функция Calc (Open Office): AVEDEV(данные).

Частотно-взвешенная оценка среднего отклонения вычисляется соотношением

$$\overline{d} = \frac{\sum(|x_i - \overline{x}|f_i)}{\sum f_i}.$$

При использовании показателя среднего линейного отклонения возникают определенные неудобства, связанные с расчетом модуля разности. Широкое распространение получили обобщающие показатели, найденные с использованием вторых степеней отклонений. В предположении принадлежности выборки генеральной совокупности к таким показателям относятся несмещенная оценка дисперсии D^2

$$D^2 = \frac{\sum(x_i - \overline{x})^2}{n}$$

и среднее квадратическое отклонение D

$$D = \sqrt{D^2} = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}}.$$

Функция MS Excel: ДИСПР(данные);
Функция Calc (Open Office): VARP(данные).

Частотно-взвешенная оценка дисперсии вычисляется соотношением

$$D^2 = \frac{\sum[(x_i - \overline{x})^2 f_i]}{\sum f_i} = \frac{\sum x_i^2 f_i}{\sum f_i} - \overline{x}^2$$

Кроме показателей вариации, выраженных абсолютными величинами, в статистических исследованиях используются относительные показатели вариации V_*, в частности для сравнения разброса признаков нескольких совокупностей.

Данные показатели рассчитываются как отношение размаха вариации к средней величине признака (коэффициент осцилляции), отношение среднего линейного отклонения к средней величине признака (линейный коэффициент вариации), отношение среднего квадратического отклонения к средней величине признака (коэффициент вариации) и, как правило, выражаются в процентах.

Формулы расчета относительных показателей вариации:

– коэффициент осцилляции V_R $$V_R = \frac{R}{\overline{x}} \cdot 100\%,$$

– линейный коэффициент вариации V_A $$V_A = \frac{\overline{d}}{\overline{x}} \cdot 100\%,$$

– коэффициент вариации V_D $$V_D = \frac{D}{\overline{x}} \cdot 100\%,$$

Отметим следующее – из приведенных формул видно, что чем больше коэффициент V_* ближе к нулю, тем меньше вариация значений признака.

В статистической практике наиболее часто применяется коэффициент вариации V_D. Он используется не только для сравнительной оценки вариации, но и для характеристики однородности совокупности. Для распределений, близких к нормальному, совокупность считается однородной, если коэффициент вариации не превышает 33%.

Пример П7. Определить среднее, медиану, моду и параметры вариации, заданной таблицей справа.

Возраст особи		частота встречаемости
лет от	лет до	
18	21	1
21	24	3
24	27	6
27	30	10
30	33	5
33	36	3
36	39	2

Расчет описанных выше статистических параметров для интервального ранжированного частотного ряда можно выполнить по следующей схеме.

1. В соответствующие ячейки вносятся текстовые поясняющие данные (см. рис. П7.1).

2. Заносятся исходные данные в диапазоны A5:B11, D5:D11 (на рис. П7.1 выделены голубым фоном).

3. В ячейке C5 формулой =(A5+B5)/2 определяется средний по диапазону возраст. Далее автозаполнение на диапазон C6:C11.

4. Формулой =СУММ(D5:D11) в ячейке D12 находится сумма частот.

5. Столбец накопленных частот строится внесением в E1 формулы =D5, в E2 формулы =E5+D6 и тиражированием последней на диапазон E7:E11.

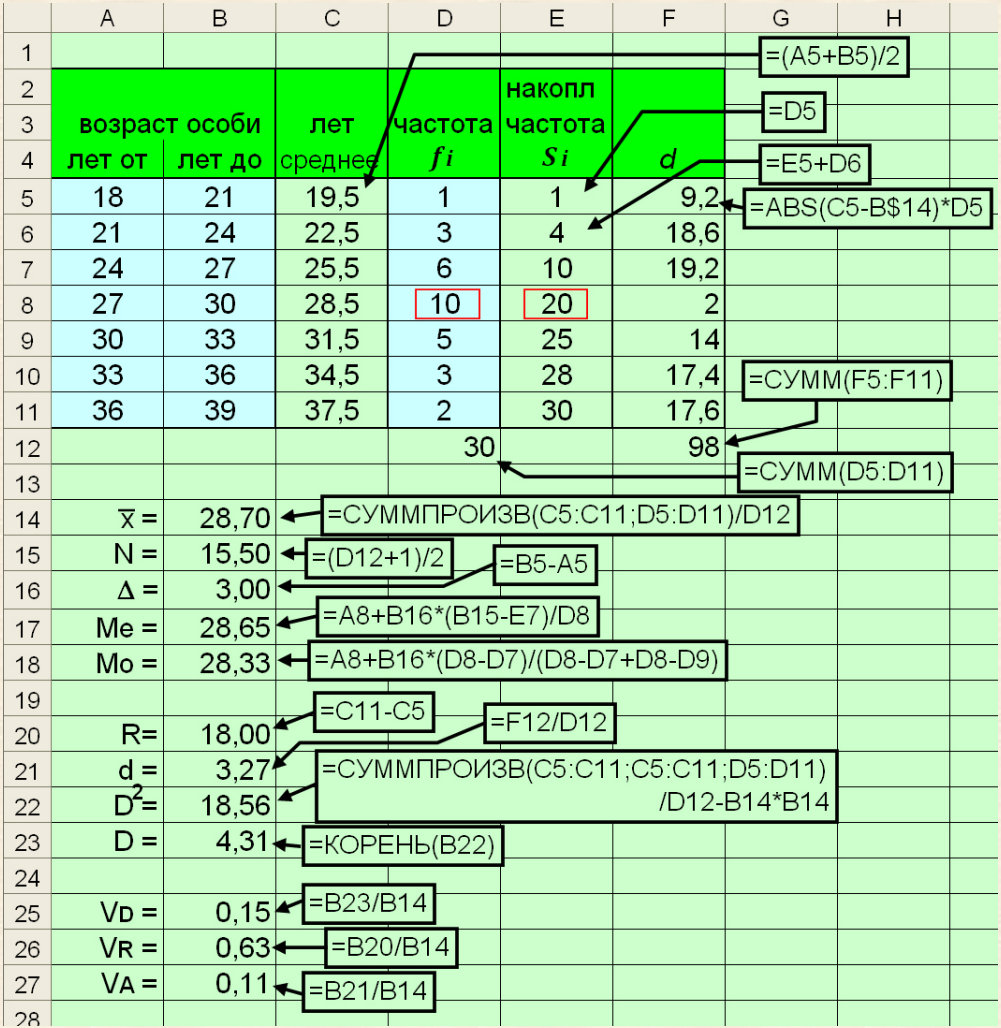

Рис. П7.1 Скриншот схемы вычислений

6. В столбце F5:F11 подсчитываются частичные суммы среднего отклонения: в F5 заносится формула =ABS(C5-B$14)*D5 и тиражируется автозаполнением на диапазон F6:F11. В ячейке F12 определяется сумма этих величин =СУММ(F5:F11).

7. В соответствии с нижеприведенной таблицей рассчитываются все требуемые параметры распределения.

адрес ячейки	Формула в ячейке	Пояснение
A14	=СУММПРОИЗВ(C5:C11;D5:D11) /D12	Подсчитывается среднее по формуле $\overline{x} = \frac{\sum x_i f_i}{\sum f_i}$.
A15	=(D12+1)/2	Определяется порядковый номер медианы $N_{Me} = \frac{n+1}{2}$ где $n = \sum x_i$
A16	=B5-A5	Определяется величина интервала Δ
A17	=A8+B16*(B15-E7)/D8	Рассчитывается значение медианы. По величине N_{Me} определяется номер строки, где значение накопленной частоты "содержит" N_{Me}. В данном примере это строка 8, соответствующее накопленная частота отмечена красной рамкой. $Me = x_{Me} + \Delta \frac{\frac{n+1}{2} - S_{-1}}{f_{Me}}$

A18	=A8+B16*(D8-D7)/(D8-D7+D8-D9)	Рассчитывается значение моды. По максимальному значению частоты в столбце D определяется номер строки, по положению которой в ряду распределения рассчитывается мода: $$Mo = x_{Mo} + \Delta \frac{f_{Mo} - f_{Mo-1}}{2f_{Mo} - f_{Mo-1} - f_{Mo+1}}.$$ Максимальное значение частоты отмечено красной рамкой.
A20	=C11-C5	Рассчитывается размах вариации. Для ранжированного ряда это весьма просто. $$R = x_{max} - x_{min}$$
A21	=F12/D12	Вычисляется частотно-взвешенная оценка среднего отклонения $$\overline{d} = \frac{\sum(\lvert x_i - \overline{x}\rvert f_i)}{\sum f_i}$$
A22	=СУММПРОИЗВ(C5:C11;C5:C11; D5:D11)/D12-B14*B14	Считается дисперсия выборки: $$D^2 = \frac{\sum x_i^2 f_i}{\sum f_i} - \overline{x}^2$$
A23	=КОРЕНЬ(B22)	Среднеквадратическое отклонение: $$D = \sqrt{D^2}$$

A25	=B23/B14	Коэффициент вариации $V_D = \frac{D}{\overline{x}} \cdot 100\%,$
A26	=B20/B14	Коэффициент осцилляции $V_R = \frac{R}{\overline{x}} \cdot 100\%,$
A27	=B21/B14	Линейный коэффициент вариации $V_A = \frac{\overline{d}}{\overline{x}} \cdot 100\%$

7. Окончательно:

Средневыборочное	28,70
Медиана	28,65
Мода	28,33
Размах вариации	18,00
Среднее отклонение	3,27
Дисперсия выборки	18,56
Среднеквадратическое отклонение	4,31
Коэффициент вариации	0,15
Линейный коэффициент вариации	0,11
Коэффициент осцилляции	0,63