Пример 2.3. Сравнивается (одинаков или нет?) средний размер листьев у двух ярусов кроны акации. Данные эксперимента приведены в таблице.

Результаты эксперимента

ЯРУС 1 (выборка X)							
11,4	11,9	11,5	11,6	12,0	11,5	11,1	11,3
12,4	12,1	12,6	12,1	12,5	12,2	14,1	14,8
8,2	10,1	10,7	10,4				

ЯРУС 2 (выборка Y)							
14,3	14,4	14,9	14,3	17,5	17,5	17,7	11,4
10,8	11,4	16,3	16,1	11,4	11,9	15,8	12,1
12,5	12,2	17,0	16,6	12,3	17,3	13,2	13,9
13,0	14,4	14,1	13,9	12,0	13,5	12,0	15,5
10,0	10,0	10,0	10,0	10,0	10,0		

Алгоритм решения предусматривает следующие этапы (см. представленную в разд. 2 схему исследования):

- ✓ расчет основных характеристик (объем, среднее, дисперсия, отклонение) выборок X$\{x_i\}$ и Y$\{y_i\}$;
- ✓ проверку применимости критерия Стьюдента, а именно
  - a) проверку равенства (однородности) дисперсий выборок;
  - b) проверку нормальности распределений для обеих выборок;

- ✓ при выполнении условия b) выполняется анализ нулевой гипотезы о равенстве выборок по критерию Стьдента. Если дисперсии статистически одинаковы (выполнилось условие a)), то используется форма критерия для равных дисперсий (гомоскедастический тест)

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{(n_x - 1)\sigma_x^2 + (n_y - 1)\sigma_y^2}} \sqrt{\frac{n_x n_y}{n_x + n_y} df},$$

если нет, то применяется формула гетероскедастического теста

$$t = \frac{|\overline{x} - \overline{y}|}{\sqrt{\Omega_x + \Omega_y}}, где\ \Omega = \frac{\sigma^2}{n}.$$

Проверка нормальности распределений проводится сравнением по критерию $\chi^2_{экс} = \sum_{i=1}^{M} \frac{(n_i - n_i^{теор})^2}{n_i^{теор}}$ частот экспериментальных данных n_i и теоретических $n_i^{теор}$ (нормального закона распределения) для выборки такого же объема:

Для нормального закона распределения

$$n_i^{теор} = n\left[\Phi(x_i^{кон}) - \Phi(x_i^{нач})\right], \quad \Phi(x) = \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x - \overline{x})^2}{2\sigma^2}\right) dx,$$

где $x_i^{нач}$, $x_i^{кон}$ – начало и конец *i*-того частотного интервала.

Этап 1. Расчет основных характеристик выборок.

1. В диапазон A5:H7 заносятся исходные данные по выборке X. Для удобства данный диапазон идентифицируется (выделяет диапазон, далее ВСТАВКА-ИМЯ-ПРИСВОИТЬ) {ВСТАВКА-НАЗВАНИЯ-ОПРЕДЕЛИТЬ} именем X (см. рис. А). Исходные данные по выборке Y заносятся в ячейки A11:H15. Диапазон именуется именем Y.

2. В требуемые ячейки (рис. А) заносятся поясняющие данные.

3. В ячейку K3 заносится величина уровня значимости. Далее для выборки X заполняются формулами следующие ячейки.

адрес	формула	пояснение
K5	=СЧЁТ(X) {=COUNT(X)}	подсчет n_x – количества элементов выборки
K6	=K5-1	расчет df – числа степеней свободы
K7	=СРЗНАЧ(X) {=AVERAGE(X)}	расчет среднеарифметического значения $\overline{x}$ по выборке X
K8	=СТАНДОТКЛОН(X) {=STDEV(X)}	расчет стандартного отклонения σ_x
K9	=ДИСП(X) {=VAR(X)}	расчет выборочной дисперсии σ_x^2

Аналогично приведенному выше для выборки Y заполняются ячейки K11:K15.

Этап 2. Проверка однородности – равенства дисперсий выборок. Проверка проводится с использованием критерия Фишера (см. п.1.3, пример 1.8).

	A	B	C	D	E	F	G	H	I	J	K
1	Нулевая гипотеза: средняя длина листьев ярусов 1 и 2 одинакова										
2	*(сравнение двух независимых выборок методом Стьюдента)*										
3										α =	0,05
4	**ЯРУС 1**			***выборка X***							
5	11,4	11,9	11,5	11,6	12,0	11,5	11,1	11,3		*n x* =	20
6	12,4	12,1	12,6	12,1	12,5	12,2	14,1	14,8		*df=n x - 1* =	19
7	8,2	10,1	10,7	10,4						$\bar{x}$ =	11,73
8										σx =	1,38
9										σ_x^2 =	1,91
10	**ЯРУС 2**			***выборка Y***							
11	14,3	14,4	14,9	14,3	17,5	17,5	17,7	11,4		*n y* =	38
12	10,8	11,4	16,3	16,1	11,4	11,9	15,8	12,1		*df=n y - 1* =	37
13	12,5	12,2	17,0	16,6	12,3	17,3	13,2	13,9		$\bar{y}$ =	13,45
14	13,0	14,4	14,1	13,9	12,0	13,5	12,0	15,5		σy =	2,45
15	10,0	10,0	10,0	10,0	10,0	10,0				σ_y^2 =	6,02
16											
17										*df max* =	37
18				Критерий Фишера *F*						*Fэмп* =	3,16
19		Поскольку *Fэмп* > *Fкр* , то ***выборки разнодисперсны***								*Fкр* =	2,04

Рис. А. Скриншот расчетного листа MS Excel для примера 2.3.

1. В ячейку K18 заносится формула подсчета критерия Фишера =МАКС (K9;K15) / МИН(K9; K15) {=MAX(K9;K15)/MIN(K9;K15)} в соответствии с уравнением

$$F_{\text{эмп}} = \frac{max\{\sigma_x^2; \sigma_y^2\}}{min\{\sigma_x^2; \sigma_y^2\}}.$$

2. В ячейку K17 заносится формула подсчета числа степеней свободы выборки с максимальной дисперсией (ЕСЛИ(K9>K15;K6;K12)) {IF(K9>K15; K6;K12)}.

3. С помощью встроенной функции MS Excel в ячейке K19 формулой = FРАСПОБР(K3; K17; K6+K12-K17) {=FINV(K3; K17; K6+K12-K17)} определяется критическое значение статистики.

Поскольку $F_{\text{эмп}} > F_{\text{кр}}$ (3,16 > 2,04), то в терминах статистических гипотез это значит, что H_0 (гипотеза о равенстве дисперсий) на уровне 5% отвергается – можно утверждать, что выборки разнодисперсны.

Этап 3. Проверка нормальности распределений для обеих выборок.

Проверка проводится с использованием критерия χ^2.

1. Сначала определяется диапазон изменения параметров. Для удобства определяется массив XY данных объединенных выборок – выделяется диапазон A5:H15, ВСТАВКА–ИМЯ–ПРИСВОИТЬ {ВСТАВКА – НАЗВАНИЯ – ОПРЕДЕЛИТЬ} имя XY (см. рис. В).

Заполняются формулами следующие ячейки.

адрес	формула	пояснение
K21	=СЧЁТ(XY) {=COUNT(XY)}	подсчет n – объема объединенной выборки
K22	=ОКРУГЛВНИЗ(1+3,3*LOG10(K21);0) {=ROUNDDOWN(1+3,3*LOG10(K21);0)}	подсчет M – количества диапазонов по формуле Sturgess'a $M = 1 + 3{,}3 \cdot \lg n$
H21	=МИН(XY) {=MIN(XY)}	расчет x_{min} – минимального в выборке
H22	=МАКС(XY) {=MAX(XY)}	расчет x_{max} – максимального в выборке
H23	=ОКРУГЛВВЕРХ((H22-H21)/K22;1) {=ROUNDUP((H22-H21)/K22;1)}	расчет ширины диапазона $\Delta x/M$
K23	=СЧЁТ(B27:B32)-3 {=COUNT(B27:B32)-3}	число степеней свободы для критерия χ^2
K24	=ХИ2ОБР(K3;K23) {=CHIINV(K3;K23)}	критическое значение χ^2

2. В ячейку B27 заносится формула =H21 – начальное значение первого диапазона, равное минимальному в выборке. В ячейку B28 заносится формула =B27+$H$23 – начальное значение второго диапазона, равное началу предыдущего плюс ширина диапазона. Далее по B28 производится автозаполнение ячеек B29:B32 начальных значений для всех остальных диапазонов (интервалов, классов).

	A	B	C	D	E	F	G	H	I	J	K
21		Проверка нормальности распределения					x_{min} =	8,2	$n_x + n_y$ =		58
22							x_{max} =	17,7		M=	6
23							dx=	1,6	df для $\chi 2$ =		3
24									критич $\chi 2$ =		7,8
25											
26		диапазон		n(X) экс	n(Y) экс	n(X) теор	n(Y) теор	$\chi 2$(X)	$\chi 2$(Y)		
27		8,2	9,8	1	0	1,63	2,60	0,24	2,60		
28		9,8	11,4	5	7	6,51	5,06	0,35	0,74		
29		11,4	13	12	10	8,31	8,56	1,64	0,24		
30		13	14,6	1	10	3,18	9,62	1,50	0,02		
31		14,6	16,2	1	4	0,36	7,17	1,13	1,40		
32		16,2	17,8	0	7	0,01	4,99	0,01	0,81		
33			сумма:	20	38	20,0	38,0	4,88	5,80		
34											
35		Поскольку для первой и второй выборки $\chi 2 < \chi 2_{крит}$, то									
36		обе ***выборки имеют нормальное распределение***									
37							=(K9/K5+K15/K11)^2/((K9/K5)^2/K6+(K15/K11)^2/K12)				
38							=ABS(K7-K13)/КОРЕНЬ(K9/K5+K15/K11)				
39											
40				Критерий Стьюдента t						df=	55,68
41		Поскольку $t > t_{кр}$, то нулевая гипотеза отвергается								t=	3,43
42		на выбранном уровне значимости								$t_{кр}$ =	2,00
43		***различия между выборками признаются статистически значимыми***									

Рис. В. Скриншот расчетного листа MS Excel для примера 2.3.

3. В ячейку C27 заносится формула =B27+$H$23 – конечное значение первого диапазона, равное его началу плюс ширина диапазона. Далее по C27 производится автозаполнение ячеек C28:C32 конечных значений для всех остальных диапазонов (классов).

4. В ячейки D27:D32 механизмом введения формул для массивов (использование F2 затем Ctrl+Shift+Enter) заносится формула =ЧАСТОТА (X; B27:B32) {=FREQUENCY(X; B27:B32), формула вводится в выделенные ячейки и активируется через Ctrl+Shift+Enter }. Данной операцией определяются количества элементов выборки X, относящиеся к каждому классу (диапазону). Аналогично заполняется диапазон ячеек E27:E32 подсчета количества элементов выборки Y.

5. В ячейку F27 заносится формула =K$5*НОРМРАСП(C27; K$7; K$8; ИСТИНА) {=K$5*NORMDIST (C27; K$7; K$8; TRUE())}, которая определяет для нормального закона распределения (при соответствующих значениях $\overline{x}$ и σ) теоретически ожидаемое число элементов выборки $n_i^{\text{теор}}$ для отрезка изменения переменной от -∞ до конца первого диапазона данных. Аналогично для выборки Y в ячейку G27 заносится формула =K$11*НОРМРАСП (C27; K$13; K$14; ИСТИНА) {=K$11*NORMDIST (C27; K$13; K$14; TRUE()) }.

6. В ячейку F28 заносится формула =K$5*(НОРМРАСП(C28; K$7; K$8; ИСТИНА) -НОРМРАСП(B28; K$7; K$8; ИСТИНА)) {=K$5*(NORMDIST (C28; K$7; K$8; TRUE()) -NORMDIST(B B28; K$7; K$8; TRUE()))}, которая определяет для нормального закона распределения теоретически ожидаемое число элементов выборки $n_i^{\text{теор}}$ для соответствующего диапазона данных. Далее по содержимому F28 производится автозаполнение ячеек F29:F31 ожидаемого числа элементов выборки X для всех остальных диапазонов (классов).

Аналогично для выборки Y в ячейку G28 заносится формула =K$11*(НОРМРАСП(C28; K$13; K$14; ИСТИНА) -НОРМРАСП(B28; K$13; K$14; ИСТИНА)) {K$11*(NORMDIST (C28; K$13; K$14; TRUE()) - NORMDIST (B28; K$13; K$14; TRUE())) } и тиражируется до ячейки G31.

7. В ячейку F32 заносится формула =K$5*(1-НОРМРАСП(B32; K$7; K$8; ИСТИНА)) {= K$5*(1-NORMDIST(B32; K$7; K$8; TRUE()))}, которая определяет для нормального закона распределения теоретически ожидаемое число $n_i^{\text{теор}}$ элементов выборки X для отрезка от начала последнего диапазона данных до + ∞. Для выборки Y подобная формула =K$11*(1-НОРМРАСП (B32; K$13; K$14; ИСТИНА)) {=K$11*(1 - NORMDIST (B32; K$13; K$14; TRUE()))} заносится в ячейку G32.

8. Для подсчета сумм в ячейку D33 заносится формула =СУММ(D27:D32) {=SUM(D27:D32)} и механизмом автозаполнения тиражируется в диапазоне E33:I33.

9. В ячейку H27 заносится формула =(D27-F27)^2/F27, определяющая "частное" значение χ^2 для пары частот первого интервала первой выборки. Механизмом автозаполнения по содержимому H27 вводятся значения H28:H32 для остальных интервалов. Аналогично для второй выборки формула =(E27-G27)^2/G27 ячейки I27 тиражируется на I28:I32.

10. Сравнивая критическое значение $\chi^2_{\text{крит}}$ (ячейка K24) и экспериментальные значения критерия для выборок X и Y (ячейки H33,I33) приходим к выводу: нулевые гипотезы о принадлежности выборок к нормальному закону распределения принимаются, обе выборки подчиняются нормальному закону распределения.

Этап 4. Анализ нулевой гипотезы о равенстве выборок.

1. Поскольку обе выборки соответствуют нормальному закону распределения, но статистически отличаются их дисперсии, то для сравнения средних этих выборок используется гетероскедастический тест.

2. В ячейку K40 (рис. В) заносится формула подсчета условных степеней свободы =(K9/K5+K15/K11)^2 / ((K9/K5)^2 /K6+(K15/K11)^2 /K12), реализующая соотношение для df,

$$df = \frac{\left(\Omega_x + \Omega_y\right)^2}{\frac{\Omega_x^2}{n_x - 1} + \frac{\Omega_y^2}{n_y - 1}},$$

а в ячейку K41 формула вычисления критерия Стьюдента =ABS(K7-K13) / КОРЕНЬ (K9/K5 +K15/K11) {=ABS (K7-K13) / SQRT(K9/K5+K15/K11)}, реализующая вычисление статистики

$$t = \frac{|\bar{x} - \bar{y}|}{\sqrt{\Omega_x + \Omega_y}}, \text{ где } \Omega = \frac{\sigma^2}{n}.$$

3. В ячейку K42 заносится формула =СТЬЮДРАСПОБР(K3;K40) {=TINV (K3; K40)}, определяющая критическое значение $t_{кр}$ для заданного уровня значимости.

4. Сравнение величин t и $t_{кр}$ позволяет сделать вывод: различия между выборками являются статистически значимыми.