Пример 2.1. Сравнивается – одинаков ли средний размер листьев у двух ярусов кроны акации. Данные эксперимента приведены в таблице.

Результаты эксперимента

ярус 1 выборка x_i

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	10,7
13,7	13,9	13,2	13,8	16,3	16,2	14,7
14,3	14,8	14,8	15,2	15,6	15,5	14,7

ярус 2 выборка y_i

14,3	14,4	14,9	14,3	17,5	17,5	17,7	11,4
11,8	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	13,8	13,5	15,6	15,5
15,3	15,1	15,1	15,0	15,1	15,8		

Алгоритм решения предусматривает следующие этапы (см. представленную в разд. 2 схему исследования). Для сокращения объема примера будем считать, что для выборок уже предварительно проведен отсев грубых ошибок эксперимента каким-либо методом (см. Приложение 2).

Общая структура данного исследования предполагает

- ✓ расчет основных характеристик (объем, среднее, дисперсия, отклонение) выборок x_i и 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 Стьюдента. 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 $\chi^2_{эксп} = \sum_{i=1}^{M} \frac{(n_i - n_i^{теор})^2}{n_i^{теор}}$ 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.

1.1. В диапазон A5:H9 заносятся исходные данные по выборке x_i. Для удобства данный диапазон идентифицируется (выделяет диапазон, далее ВСТАВКА-ИМЯ-ПРИСВОИТЬ) {ВСТАВКА-НАЗВАНИЯ-ОПРЕДЕЛИТЬ} именем X (см. рис. А). Исходные данные по выборке y_i заносятся в ячейки A12:H16. Диапазон (массив) определяется именем Y.

1.2. В ячейки A4, D4, A11, D11, J3, J5:J9, J12:J20 заносятся поясняющие данные.

1.3. В ячейку K3 заносится величина уровня значимости. Далее для выборки x_i заполняются формулами следующие ячейки.

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

Аналогично приведенному по данным выборки y_i (массив Y) заполняются ячейки K12:K16.

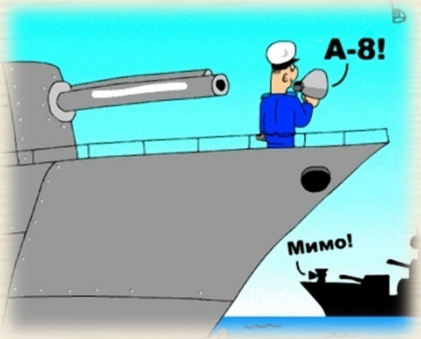
Этап 2. Проверка равенства (однородности) дисперсий выборок.

Проверка проводится использованием критерия Фишера (см. п.1.3, пример 1.8).

2.1. В ячейку K18 заносится формула подсчета числа степеней свободы выборки с максимальной дисперсией ЕСЛИ(K9>K16;K5;K12) {IF(K9>K16;K6;K13)}.

2.2. В ячейку K19 заносится формула подсчета критерия Фишера =МАКС (K9;K16) / МИН(K9;K16) {=MAX(K9;K16) / MIN(K9;K16)}.

2.3. С помощью встроенной функции MS Excel в ячейке K20 формулой =FРАСПОБР(K3; K18; K6+K13-K18) {=FINV(K3; K18; K6+K13-K18)} определяется критическое значение статистики Фишера. Поскольку ($F_{кр}$)=1,74 >1,08=($F_{эмп}$), то можно утверждать, что гипотеза о сходстве на уровне 5% подтверждается – дисперсия обеих выборок одинакова.



Этап 3. Проверка нормальности распределений для обеих выборок.

Проверка проводится с использованием критерия $\chi^2_{кр}$ (рис. В).

3.1. Сначала определяется диапазон изменения параметров. Для удобства определяется массив XY данных объединенных выборок – выделяется диапазон A5:H16, ВСТАВКА–ИМЯ–ПРИСВОИТЬ {ВСТАВКА – НАЗВАНИЯ – ОПРЕДЕЛИТЬ} имя XY.

	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* =	35
6	12,4	12,1	12,6	12,1	12,5	12,2	14,1	14,8		*df=n x -1*=	34
7	8,2	10,1	10,7	10,4	10,7	13,7	13,9	13,2		$\bar{x}$ =	12,91
8	13,8	16,3	16,2	14,7	14,3	14,8	14,8	15,2		σx =	1,95
9	15,6	15,5	14,7							σ_x^2 =	3,78
10											
11	**ЯРУС 2**			***выборка Y***							
12	14,3	14,4	14,9	14,3	17,5	17,5	17,7	11,4		*n y* =	38
13	11,8	11,4	16,3	16,1	11,4	11,9	15,8	12,1		*df=n y -1*=	37
14	12,5	12,2	17,0	16,6	12,3	17,3	13,2	13,9		$\bar{y}$ =	14,45
15	13,0	14,4	14,1	13,9	13,8	13,5	15,6	15,5		σy =	1,87
16	15,3	15,1	15,1	15,0	15,1	15,8				σ_y^2 =	3,50
17											
18										*df max* =	34
19				Критерий Фишера *F*						*Fэмп* =	1,08
20		Поскольку *Fэмп* < *Fкр* , то ***выборки однородны***								*Fкр* =	1,74

Рис. А. Скриншот расчетного листа MS Excel для примера 2.1.

Заполняются формулами следующие ячейки.

адрес	формула	пояснение
K22	=СЧЁТ(XY) {=COUNT(XY)}	подсчет n – объема объединенной выборки
K23	=ОКРУГЛВНИЗ(1+3,3* LOG10(K22);0) {=ROUNDDOWN(1+3,3*LOG10(K22);0)}	подсчет M – количества диапазонов по формуле Sturgess'a $M = 1 + 3{,}3 \cdot \lg n$
H22	=МИН(XY) {=MIN(XY)}	расчет x_{min} – минимального значения выборки x_i
H23	=МАКС(XY) {=MAX(XY)}	расчет x_{max}– максимального значения в выборке x_i
H24	=ОКРУГЛВВЕРХ((H23-H22)/K23;1) {=ROUNDUP((H23-H22)/K23;1)}	расчет ширины диапазона $\Delta x/M$
K25	=СЧЁТ(B29:B35)-3 {=COUNT(B29:B35)-3}	число степеней свободы для критерия χ^2
K26	=ХИ2ОБР(K3;K25) {=CHIINV(K3;K25)}	критическое значение χ^2

3.2. В ячейку B29 заносится формула =H22 – начальное значение первого диапазона, равное минимальному в выборке. В ячейку B30 заносится формула =B29+$H$24 – начало второго диапазона, равное началу предыдущего плюс ширина диапазона. Далее по B30 производится автозаполнение ячеек B31:B35 начальных значений остальных классов.

	A	B	C	D	E	F	G	H	I	J	K
21											
22		Проверка нормальности распределения					*x min* =	8,2		*n x* + *n y* =	73
23							*x max* =	17,7		*L* =	7
24							*dx* =	1,4			
25									*df для* χ^2 =		4
26									*критич* χ^2 =		9,5
27											
28		диапазон		*n(X) экс*	*n(Y) экс*	*n(X) теор*	*n(Y) теор*	$\chi^2(X)$	$\chi^2(Y)$		
29		8,2	9,6	1	0	1,6	0,2	0,20	0,18		
30		9,6	11	4	0	4,1	1,1	0,01	1,06		
31		11	12,4	12	8	8,2	4,0	1,80	4,13		
32		12,4	13,8	5	5	9,8	8,7	2,35	1,55		
33		13,8	15,2	9	12	7,1	11,1	0,48	0,08		
34		15,2	16,6	4	8	3,2	8,3	0,21	0,01		
35		16,6	18	0	5	1,0	4,7	1,01	0,01		
36			*сумма:*	35	38	35,0	38,0	6,05	7,01		
37											
38		Поскольку для первой и второй выборки $\chi^2 < \chi^2_{крит}$, то									
39		обе ***выборки имеют нормальное распределение***									

Рис. В. Скриншот расчетного листа MS Excel для примера 2.1.

3.3. В ячейку C29 заносится формула =B29+$H$24 – конечное значение первого диапазона, равное его началу плюс ширина диапазона. Далее по C29 производится автозаполнение ячеек C30:C35 конечных значений для всех остальных диапазонов (интервалов, классов).

3.4. В ячейки D29:D35 механизмом введения формул для массивов (использование F2 затем Ctrl+Shift+Enter) заносится формула =ЧАСТОТА (X; B30:B35) {=FREQUENCY(X;B30:B35), формула вводится в выделенные ячейки и активируется через Ctrl+Shift+Enter }. На этом этапе определяются количества элементов выборки x_i для каждого класса (диапазона). Аналогично заполняется диапазон ячеек E29:E35 подсчета количества элементов выборки y_i.

3.5. В ячейку F29 заносится формула =K$5 * НОРМРАСП (C29; K$7; K$8; ИСТИНА) {=K$5*NORMDIST (C29;K$7; K$8;TRUE())}, которая определяет для нормального закона распределения (при соответствующих значениях $\overline{x}$ и σ) теоретически ожидаемое число элементов выборки x_i для отрезка изменения переменной x от -∞ до конца первого диапазона данных (см. формулу для $n_i^{\text{теор}}$). Аналогично для выборки y_i в ячейку G29 заносится формула =K$12* НОРМРАСП(C29; K$14; K$15; ИСТИНА) {= K$12*NORMDIST (C29; K$14; K$15; TRUE())}.

3.6. В ячейку F30 заносится формула =K$5*(НОРМРАСП(C30; K$7; K$8; ИСТИНА) - НОРМРАСП(B30; K$7;K$8; ИСТИНА)) {K$5*(NORMDIST (C30; K$7; K$8; TRUE())-NORMDIST(B30; K$7;K$8; TRUE()))}, которая определяет для нормального закона распределения теоретически ожидаемое число элементов выборки для второго диапазона данных (см. формулу для $n_i^{\text{теор}}$). Далее по содержимому F30 производится автозаполнение ячеек F31:F34 ожидаемого числа элементов выборки x_i для всех остальных диапазонов (классов).

Аналогично для выборки y_i в ячейку G30 заносится формула =K$12*(НОРМРАСП(C30; K$14; K$15;ИСТИНА) -НОРМРАСП(B30; K$14; K$15; ИСТИНА)) { K$12*(NORMDIST (C30; K$14; K$15; TRUE()) - NORMDIST (B30; K$14; K$15; TRUE())) } и тиражируется до ячейки G34.

3.7. В ячейку F35 заносится формула =K$5*(1-НОРМРАСП(B35; K$7; K$8; ИСТИНА)) {= K$5*(1-NORMDIST(B35; K$7; K$8; TRUE()))}, которая определяет для нормального закона распределения теоретически ожидаемое число элементов выборки x_i для отрезка от начала последнего диапазона данных до ∞. Для выборки y_i подобная формула =K$12*(1-НОРМРАСП(B35; K$14; K$15; ИСТИНА)) {=K$12*(1- NORMDIST (B35; K$14; K$15; TRUE()))} заносится в ячейку G35.

3.8. Для подсчета сумм в ячейку D36 заносится формула =СУММ(D29:D35) {SUM(D29:D35)} и автозаполнением тиражируется на диапазон E36:I36.

3.9. В ячейку H29 заносится формула =(D29-F29)^2/F29, определяющая "частное" значение χ^2 для пары частот первого интервала первой выборки. Автозаполнением по содержимому H29 вводятся значения H30:H35 для остальных интервалов. Аналогично для второй выборки формула =(E29-G29)^2/G29 ячейки I29 тиражируется на I30:I35.

3.10. Сравнивая критическое значение $\chi^2_{\text{кр}}$ (ячейка K26) и экспериментальные значения критерия для выборок x_i и y_i (ячейки H36,I36) можно заключить: нулевые гипотезы о принадлежности выборок к нормальному закону распределения принимаются, выборки подчиняются нормальному закону распределения.

Этап 4. Анализ нулевой гипотезы о равенстве выборок.

4.1. В ячейку K42 (рис. C) заносится формула подсчета критерия Стьюдента (выборки нормальны и равнодисперсны, используется гомоскедастический тест)

=ABS(K7-K14)*КОРЕНЬ(K5*K12/(K5+K12)*(K6+K13)/(K9*K6+K16*K13))

{=ABS (K7-K14) *SQRT(K5*K12/(K5+K12) *(K6+K13) /(K9*K6 +K16*K13))}, реализующая уравнение

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

	A	B	C	D	E	F	G	H	I	J	K
41											
42				Критерий Стьюдента *t*						*t*=	3,44
43		Поскольку *t* > *tкр*, то нулевая гипотеза отвергается								*tкр*=	1,99
44		на выбранном уровне значимости									
45		***различия между выборками признаются статистически значимыми***									

Рис. C. Скриншот расчетного листа MS Excel для примера 2.1.

4.2. В ячейку K43 заносится формула =СТЬЮДРАСПОБР(K3;K6+K13) {=TINV (K3; K6+K13)}, определяющая критическое значение критерия Стьюдента для заданного уровня значимости.

4.3. Сравнение величин t и $t_{крит}$ позволяет сделать вывод:

различия между выборками являются статистически значимыми.