Тема: ИСПЫТАНИЕ ГИПОТЕЗ НА ОСНОВЕ ВЫБОРОЧНОЙ СРЕДНЕЙ ПРИ НЕИЗВЕСТНОЙ ГЕНЕРАЛЬНОЙ ДИСПЕРСИИ
(сравнение средних по двум выборкам)

Пример 1[1]. Известны результаты измерения производительности труда рабочих двух разных возрастных групп. Необходимо установить: отличается ли производительность труда рабочих в двух возрастных группах.

Данные выборочного исследования

№	Группа 1	Группа 2	№	Группа 1	Группа 2	№	Группа 1	Группа 2
1	59	61	11	56	80	21	58	40
2	56	80	12	37	29	22	71	60
3	85	40	13	56	51	23	31	31
4	92	71	14	42	43	24	42	40
5	48	39	15	49	47	25	34	65
6	59	20	16	70	40	26	51	61
7	42	40	17	30	41	27	70	40
8	87	80	18	58	37	28	42	59
9	42	80	19	79	41	29	67	60
10	73	60	20	84	60	30	56	80

Формулируем нулевую и альтернативную гипотезы:

$H_0: a_1 = a_2$ - нулевая гипотеза (производительность труда не отличается в двух возрастных группах);

$H_1: a_1 \neq a_2$ - альтернативная гипотеза (производительность труда отличается в двух возрастных группах).

Дисперсии исследуемого показателя в двух группах не известны, но будем предполагать, что они одинаковы ($s_1^2 = s_2^2$). В этом случае, наблюдаемое значение критерия рассчитывают по формуле:

$$t_{\text{набл}} = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)s_1^2 + (m-1)s_2^2}} \cdot \sqrt{\frac{nm(n+m-2)}{n+m}}$$

$\bar{X} = 57{,}53$ – выборочное среднее, функция =СРЗНАЧ(массив)

$\bar{Y} = 52{,}53$ – выборочное среднее, функция =СРЗНАЧ(массив)

$s_1^2 = 298{,}53$ – исправленная выборочная дисперсия первой выборки, функция =ДИСП(массив)

$s_2^2 = 293{,}43$ – исправленная выборочная дисперсия второй выборки, функция =ДИСП(массив)

$n = m = 30$ – объем каждой выборки

Наблюдаемое значение критерия:

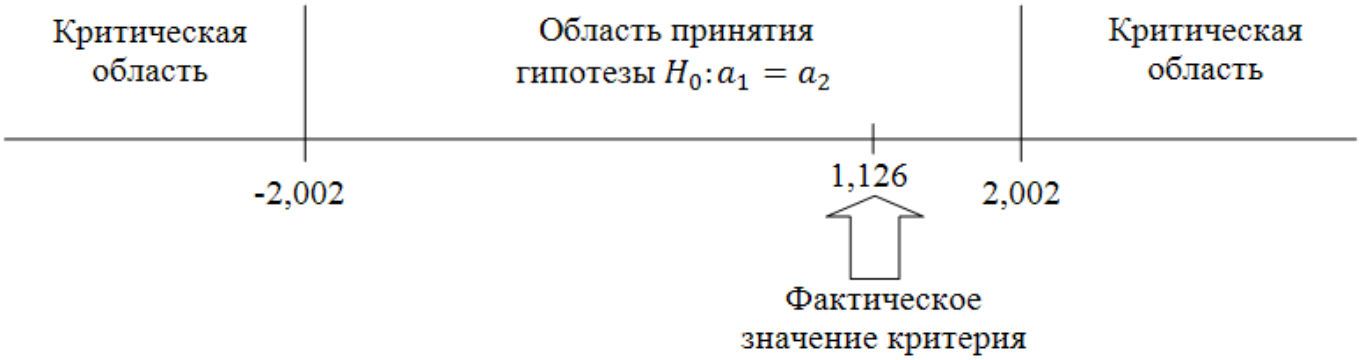
$$t_{\text{набл}} = 1{,}126.$$

Для двусторонней проверки ($H_1: a_1 \neq a_2$) находим табличное значение критерия с помощью функции: =СТЬЮДРАСПОБР(α;$n-1$).

$$t_{\text{табл}}(\alpha = 0{,}05; df = n + m - 2) = 2{,}002.$$

[1] Источник: Дубина И.Н. Математико-статистические методы в эмпирических социально-экономических исследованиях. – М.: Финансы и статистика, 2010. – С.179

Принятие решения:



Поскольку фактическое значение критерия попало в область принятия гипотезы, в данном случае $|t_{табл}| > |t_{набл}|$, то гипотезу H_0 отменить нельзя.

Таким образом, производительность труда рабочих двух разных возрастных групп не отличается.

Дополнение 1

Определим уровень значимости, на котором для двусторонней проверки можно отклонить нулевую гипотезу:

α = СТЬЮДРАСП($t_{факт}; n + m - 2; 2$) = 0,26 = 26%.

Это существенно больше 5% и даже 10%, поэтому нулевую гипотезу не отклоняем.

Примечание: параметр «хвосты»=2 в функции =СТЬЮДРАСП означает двустороннюю проверку.

Дополнение 2

Испытание гипотезы можно провести с помощью функции ТТЕСТ(массив1;массив2;хвосты;тип). При этом:

Массив1 – первое множество данных.

Массив2 – второе множество данных.

Хвосты – число хвостов распределения. Если хвосты = 1, то функция ТТЕСТ использует одностороннее распределение. Если хвосты = 2, то функция ТТЕСТ использует двустороннее распределение.

Тип – вид выполняемого t-теста. Мы проводим двухвыборочный тест с равными дисперсиями, поэтому Тип = 2.

Формула (для данных, расположенных как показано на рис.1):

=ТТЕСТ(A2:A31;B2:B31;2;2)

Результат: α = = 0,26.

Рис. 1. Организация расчета в MS Excel

	A	B	C	D	E	F	G	H
1	**Группа 1**	**Группа 2**						
2	59	61		57,53	средний X (по группе 1)			
3	56	80		52,53	средний Y (по группе 2)			
4	85	40		298,53	s(1) - стандартное отклонение по группе 1			
5	92	71		293,43	s(2) - стандартное отклонение по группе 2			
6	48	39		30	n - объем выборки первой группы			
7	59	20		30	m - объем выборки второй группы			
8	42	40						
9	87	80		1,126	наблюдаемое значение критерия			
10	42	80		2,002	табличное значение критерия			
11	73	60						
12	56	80						
13	37	29						
14	56	51						
15	42	43						
16	49	47						
17	70	40						
18	30	41						
19	58	37						
20	79	41						
21	84	60						
22	58	40						
23	71	60						
24	31	31						
25	42	40						
26	34	65						
27	51	61						
28	70	40						
29	42	59						
30	67	60						
31	56	80						

Рис. 2. Организация расчета в MS Excel (режим отображения формул)

	A	B	C	D	E
1	**Группа 1**	**Группа 2**			
2	59	61		=СРЗНАЧ(A2:A31)	средний X (по группе 1)
3	56	80		=СРЗНАЧ(B2:B31)	средний Y (по группе 2)
4	85	40		=ДИСП(A2:A31)	s(1) - стандартное отклонение по группе 1
5	92	71		=ДИСП(B2:B31)	s(2) - стандартное отклонение по группе 2
6	48	39		30	n - объем выборки первой группы
7	59	20		30	m - объем выборки второй группы
8	42	40			
9	87	80		=((D2-D3)/КОРЕНЬ((D6-1)*D4+(D7-1)*D5))*КОРЕНЬ(D6*D7*(D6+D7-2)/(D6+D7))	наблюдаемое значение критерия
10	42	80		=СТЬЮДРАСПОБР(0,05;D6+D7-2)	табличное значение критерия
11	73	60			
12	56	80			
13	37	29			
14	56	51			
15	42	43			
16	49	47			
17	70	40			
18	30	41			
19	58	37			
20	79	41			
21	84	60			
22	58	40			
23	71	60			
24	31	31			
25	42	40			
26	34	65			
27	51	61			
28	70	40			
29	42	59			
30	67	60			
31	56	80			

Для проведения испытания гипотезы о равенстве средних можно использовать инструмент «Двухвыборочный t-тест с одинаковыми дисперсиями» надстройки «Анализ данных».

Пример 2. Получена выборка цен на мужскую летнюю обувь в одном магазине и другом магазине. Проверить гипотезу о равенстве средних цен на обувь в этих магазинах.

Выборка цен в магазине №1:

6,5; 6,7; 2,3; 1,5; 6,4; 6,6; 6,2; 5,3; 5,5; 4,1; 5,2; 6,8; 2,5; 5,8; 4,4

Выборка цен в магазине №2:

5,4; 4,2; 4,1; 6,5; 7,1; 5,8; 4,9; 8,1; 6; 6,5; 8,8; 8,6; 3,7; 6,4; 6,9; 3,2; 3,2; 8; 4; 7,9; 2,7

Решение

Воспользуемся для проверки гипотезы инструментом «Двухвыборочный t-тест с одинаковыми дисперсиями» надстройки «Анализ данных». Для этого введем данные в первые два столбика электронных таблиц и в диалоговом окне укажем эти диапазоны:

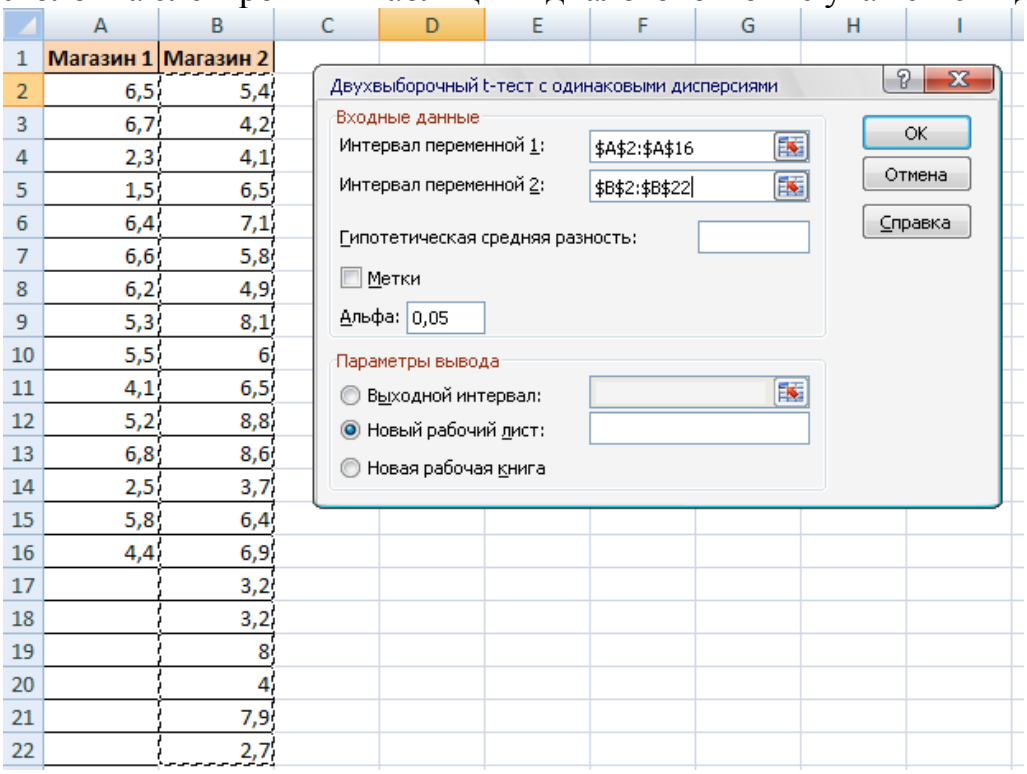


Рис. 3. Организация расчета с использованием инструмента «Двухвыборочный t-тест с одинаковыми дисперсиями»

Результаты работы надстройки:

	A	B	C
1	Двухвыборочный t-тест с одинаковыми дисперсиями		
2			
3		*Переменная 1*	*Переменная 2*
4	Среднее	5,053333333	5,80952381
5	Дисперсия	3,019809524	3,642904762
6	Наблюдения	15	21
7	Объединенная дисперсия	3,386336134	
8	Гипотетическая разность средних	0	
9	df	34	
10	t-статистика	-1,215542638	
11	P(T<=t) одностороннее	0,116266171	
12	t критическое одностороннее	1,690924198	
13	P(T<=t) двухстороннее	0,232532343	
14	t критическое двухстороннее	2,032244498	

Рис. 4. Результаты расчетов по сравнению средних цен в магазинах с использованием инструмента «Двухвыборочный t-тест с одинаковыми дисперсиями»

Получили:

$\bar{X} = 5{,}05$ – выборочное среднее

$\bar{Y} = 5{,}8$ – выборочное среднее

$s_1^2 = 3{,}02$ – исправленная выборочная дисперсия первой выборки

$s_2^2 = 3{,}6$ – исправленная выборочная дисперсия второй выборки

$n = 15$ – объем первой выборки

$m = 21$ – объем второй выборки

$t_{\text{набл}} = -1{,}22$ - наблюдаемое значение критерия.

$t_{\text{табл}}(\alpha = 0{,}05; df = 34) = 2{,}03$ - табличное значение критерия (для двухсторонней проверки).

Т.к. $|t_{\text{табл}}| = 2{,}03 > 1{,}22 = |t_{\text{набл}}|$, то наблюдаемое значение критерия попадает в область принятия гипотезы H_0.

Вероятность ошибки при отклонении истинной нулевой гипотезы составляет 23% (для двухсторонней проверки)

Вывод: принимаем нулевую гипотезу, т.е. средние цены в магазинах одинаковые.

Пример 3. По данным сети интернет получить первую выборку стоимости 1 кв.м. квартиры на вторичном рынке жилья в одном районе города (объем выборки не менее 15), а вторую выборку - стоимость 1 кв.м. квартиры на вторичном рынке жилья в другом районе города (объем выборки не менее 15). Проверить гипотезу о равенстве соответствии средних цен на жилье в выбранных районах.

По результатам расчетов подготовить отчет.