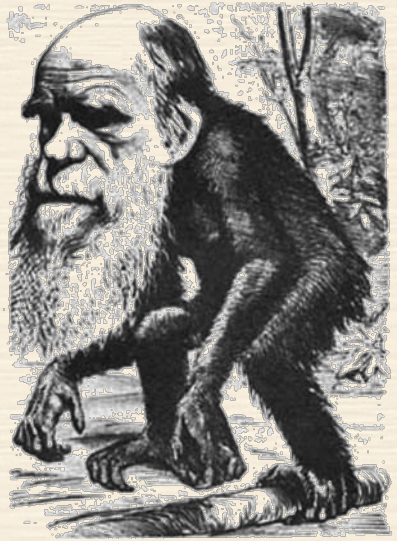
ПРИЛОЖЕНИЕ 5

Сложные проблемы всегда имеют простые, легкие для понимания неправильные решения.
Мерфология

Проверка гипотезы о независимости двух выборок



Для проверки гипотезы о независимости двух выборок X и Y используется таблица частот, которые можно было бы ожидать в том случае, *если переменные оказались бы независимыми*. В общем случае критерий χ^2 независимости принято применять следующим образом.

1. Множество значений выборки X разбивается на s непересекающихся классов (интервалов, разрядов), а множество значений Y на r непересекающихся интервалов.

2. Составляется таблица исходных данных в виде списка экспериментальных частот Z_{ij} всех комбинаций категорий двух качественных переменных.

Таблица дополняется строкой и столбцом сумм "попаданий" компонент выборки в конкретный класс другой выборки:

$$Z_{Yk} = \sum_i Z_{ik}\,, k = 1, \dots, r \qquad Z_{Xk} = \sum_j Z_{kj}\; k = 1, \dots, s$$

Переменные (выборки), разделенные по классам
(экспериментальные частоты)

интервалы i=1,2…,s выборки X	интервалы j=1,2,…,r выборки Y				
	элементы Y_1	элементы Y_2	* * *	элементы Y_r	
элементы X_1	Z_{11}	Z_{12}	* * *	Z_{1r}	$Z_{X1} = \sum_j Z_{1j}$
элементы X_2	Z_{21}	Z_{22}	* * *	Z_{2r}	$Z_{X2} = \sum_j Z_{2j}$
* * *	* * *	* * *	* * *	* * *	* * *
элементы X_s	Z_{s1}	Z_{s2}	* * *	Z_{sr}	$Z_{Xs} = \sum_j Z_{sj}$
	$Z_{Y1} = \sum_i Z_{i1}$	$Z_{Y2} = \sum_i Z_{i2}$	* * *	$Z_{Y1} = \sum_i Z_{ir}$	$n = \sum_i \sum_j Z_{ij}$

3. Составляется таблица "относительных частот" по соотношению

$$T_{ij} = \frac{Z_{ij}}{Z_{Yj}\, Z_{Xi}}, i = 1, \dots, s, j = 1, \dots, r\ .$$

Переменные (выборки), разделенные по классам

	элементы Y_1	элементы Y_2	* * *	элементы Y_r
элементы X_1	T_{11}	T_{12}	* * *	T_{1r}
элементы X_2	T_{21}	T_{22}	* * *	T_{2r}
* * *	* * *	* * *	* * *	* * *
элементы X_s	T_{s1}	T_{s2}	* * *	T_{sr}

4. Определяется расчетное значение χ^2 по формуле

$$\chi^2_{\text{расч}} = n\left(\sum_i \sum_j T_{ij} - 1\right).$$

5. Критическое значение $\chi^2_{\text{кр}}$ (s категорий X и r категорий Y) определяется для числа степеней свободы $df = (s-1)(r-1)$.

6. Проводится сравнение расчетного значения $\chi^2_{\text{расч}}$ с критическим $\chi^2_{\text{кр}}$, определенному по "обычным" уровням значимостиα, равному 0,05 или 0,01 или другому выбранному значению.

7. Строится заключение: при $\chi^2_{\text{расч}} > \chi^2_{\text{кр}}$ гипотеза об отсутствии связи между признаками и параметрами отвергается, при $\chi^2_{\text{расч}} < \chi^2_{\text{кр}}$ – подтверждается.

Пример П5. По переписи населения Швеции 1936 г. из совокупности всех супружеских пар была получена выборка (данные в таблице), содержащая распределение годовых доходов (в тыс. крон) и количества детей в семье[1] .

Распределение годовых доходов (в тыс. крон) и количества детей в семье

число детей	доходы, тыс. крон			
	0-1	1-2	2-3	3 и выше
0	2161	3577	2184	1636
1	2755	5081	2222	1052
2	936	1753	640	306
3	225	419	96	38
4 и более	39	98	31	14

Требуется установить на уровне значимости α=0,05 (с доверительной вероятностью 0,95) являются ли зависимыми количество детей в семье (величина X) и уровень годового дохода этой семьи (величина Y).

Для проверки гипотезы H_0 о независимости X и Y используется критерий χ^2. Множество значений X (количество детей в семье) разбивается на 5 разрядов, множество значений Y (годовой доход семьи) разбито на 4 разряда (см. таблицу доходов). Алгоритм построения решения следующий.

[1] Гланц, С. Медико-биологическая статистика / С. Гланц, –М.: Практика, 1998. –459 с.

1. В диапазон B5:T9 заносятся исходные данные по выборкам X и Y. В ячейку F1 вводится значение величины уровня значимости. В требуемые ячейки (рис. П5.1) заносятся поясняющие данные.
2. В ячейку B10 вводится формула =СУММ(B5:B9) {=SUM(B5:B9)} подсчета суммы первой группы выборки Y по всей X. Формула тиражируется автозаполнением на диапазон C10:F10. В ячейку F5 вводится формула =СУММ(B5:E5) {=SUM(B5:E5) } подсчета суммы первой группы выборки X по всей Y. Формула тиражируется автозаполнением на диапазон F6:F9.
3. В ячейку B13 вводится формула =B5/B$10*B5/$F5 подсчета "относительной частоты" суммарной выборки. Формула тиражируется автозаполнением на диапазон B13:E17.
4. В ячейку B19 вводится формула =F10*(СУММ(B13:E17)-1) {=F10*(SUM(B13:E17)-1) }, реализующая подсчет критерия χ^2 в соответствии с соотношением

$$\chi^2_{\text{расч}} = n\left(\sum_i \sum_j T_{ij} - 1\right).$$

5. В ячейке B20 формулой =F10*(СУММ(B13:E17)-1) {=F10*(SUM(B13:E17)-1)} подсчитывается число степеней свободы, а формулой =ХИ2ОБР(F1;B21) {=CHIINV(F1;B21) } определяется критическое значение статистики $\chi^2_{\text{кр}}$.

6. Поскольку $\chi^2_{\text{расч}} > \chi^2_{\text{кр}}$, то гипотеза H_0 о независимости X (количество детей в семье) и Y (уровень годового дохода семьи) отвергается с доверительной вероятность 95%.

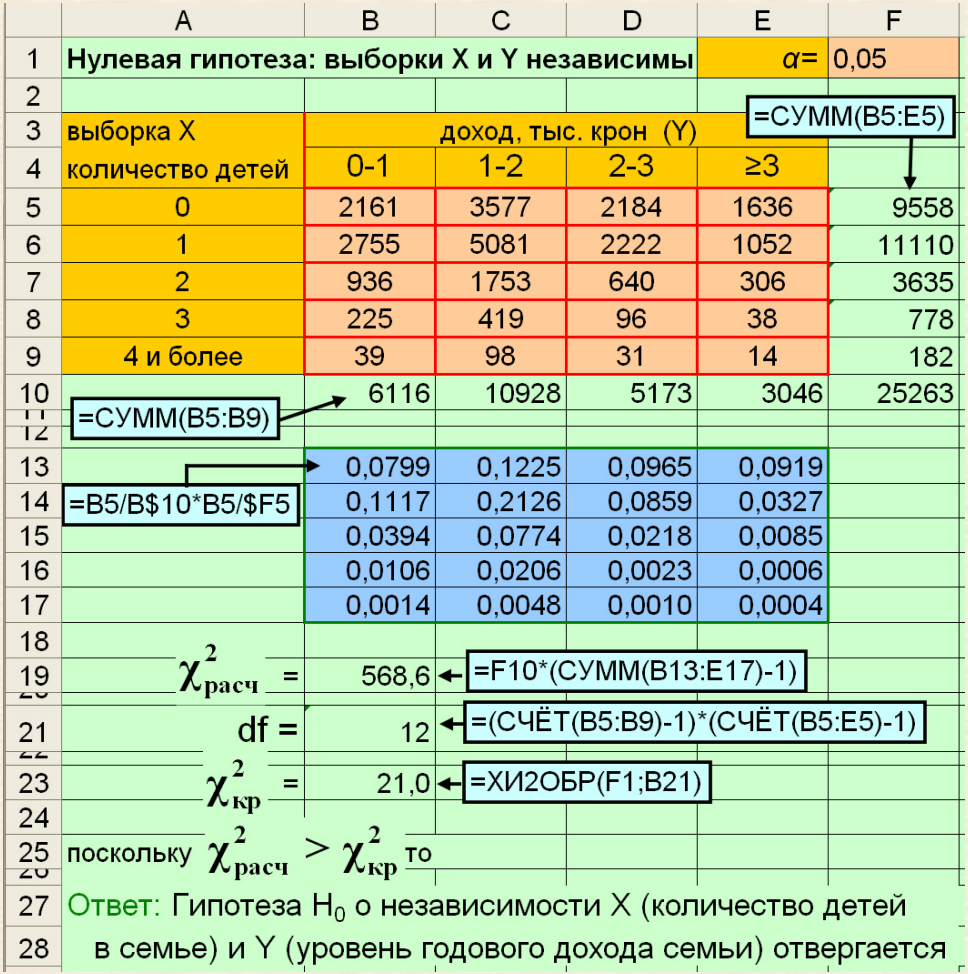


Рис. П5.1. Скриншот расчетного листа MS Excel