АНАЛИЗ ДАННЫХ НА ОСНОВЕ КРИТЕРИЯ «ХИ-КВАДРАТ»

Пусть имеются фактические (т.е. полученные в результате наблюдений или экспериментов) значения некоторой исследуемой величины: $X_{Ф1}$, $X_{Ф2}$, …, $X_{Фm}$, где m – количество имеющихся значений. Об этой величине выдвинуто некоторое теоретическое предположение. Рассчитано, что в случае, если бы фактические данные *полностью соответствовали* этому предположению, значения исследуемой величины были бы следующими: X_{T1}, X_{T2}, …, X_{Tm}. Требуется выяснить, значимо ли различие между фактическими данными и теоретическими (ожидаемыми) величинами.

Для оценки значимости различий находится критерий χ^2:

$$\chi^2 = \sum_{j=1}^{m} \frac{(X_{\Phi j} - X_{Tj})^2}{X_{Tj}}. \qquad (1)$$

По значению критерия χ^2 и числу степеней свободы s (примеры определения числа степеней свободы см. ниже) определяют расчетный уровень значимости P, т.е., упрощенно говоря, вероятность *незначимости* различий между фактическими данными и теоретическими величинами. Если $P > \alpha$ (обычно принимают $\alpha = 0{,}05…0{,}2$), то такое различие можно считать статистически незначимым. В этом случае можно считать, что теоретическое предположение, выдвинутое в отношении исследуемой величины, соответствует фактическим данным (т.е., как минимум, не опровергается ими).

Пример 1 – Компания продает свою продукцию в пяти городах примерно одинакового размера. За некоторый период выручка компании в этих городах составила 110, 130, 70, 90, 100 ден.ед. Определить, можно ли считать эти величины выручки примерно одинаковыми (т.е. считать, что выручка не зависит от конкретного города).

Здесь 110, 130, 70, 90, 100 – фактические данные ($X_{Фj}$, j=1,…,5). Суммарная выручка составляет 110 + 130 + 70 + 90 + 100 = 500 ден.ед. Если предположение о независимости выручки от города абсолютно верно, то выручка в городах должна быть одинаковой: 500 / 5 = 100 ден.ед. в каждом городе. Это ожидаемые (теоретические) величины. Вычислим критерий χ^2 по формуле (1):

$$\chi^2 = \frac{(110-100)^2}{100} + \frac{(130-100)^2}{100} + \frac{(70-100)^2}{100} + \frac{(90-100)^2}{100} + \frac{(100-100)^2}{100} = 20.$$

Для задач такого типа (где некоторая величина подразделяется на несколько величин) число степеней свободы определяется как s=m-1. Для данного примера s=5-1=4.

Для вычисления уровня значимости P в MS Excel воспользуемся функцией **ХИ2.РАСП.ПХ** со следующими параметрами: **X**: критерий χ^2 (для данного примера – 20); **Степени свободы**: s = m-1 (для данного примера – 4). В данном случае P = 0,0005, т.е. $P < \alpha$ (пусть, например, α=0,05). Таким образом, вероят-

ность незначимости различия между фактическими данными (110, 130, 70, 90, 100) и теоретическими величинами (по 100 в каждом городе) очень мала. Это означает, что предположение об одинаковой выручке во всех городах не подтверждается фактическими данными. Другими словами, выручка в разных городах существенно различается.

Пример 2 – Университет принимает 150 студентов. Всего подано 320 заявлений, в том числе 80 из Северного региона, 90 – из Южного, 50 – из Западного, 100 – из Восточного. По результатам экзаменов принято, соответственно 36, 40, 21 и 53 студента из этих регионов. Требуется определить, является ли доля принятых примерно одинаковой для всех регионов.

Если предположить, что доля принятых одинакова, то среди принятых студентов доля каждого региона должна быть такой же, как и доля этого региона среди подавших заявления. Например, заявления из Северного региона составляют 80/320 = 0,25. Значит, количество принятых студентов из этого региона должно составлять 0,25·150 = 37,5 (так как эта величина – теоретическая, она может быть и дробной). Аналогичные расчеты по всем регионам приведены в таблице ниже.

Регион	Северный	Южный	Западный	Восточный	Всего
Заявления	80	90	50	100	320
Доля заявлений	0,25	0,28	0,16	0,31	1
Должно быть принято (если доля принятых одинакова по всем регионам)	37,5	42,19	23,44	46,88	150
Фактически принято	36	40	21	53	150

Вычислим критерий χ^2:

$$\chi^2 = \frac{(36-37.5)^2}{37.5} + \frac{(40-42.19)^2}{42.19} + \frac{(21-23.44)^2}{23.44} + \frac{(53-46.88)^2}{46.88} = 1{,}23.$$

Здесь имеется разделение на четыре величины. Поэтому число степеней свободы равно s=4-1=3.

Чтобы найти уровень значимости P, используем функцию **ХИ2.РАСП.ПХ** со следующими параметрами: **X**: 1,23; **Степени свободы**: 3. Уровень значимости равен P = 0,746. Таким образом, $P > \alpha$ (для любого возможного значения α). Это значит, что фактические данные (количество принятых студентов из разных регионов) существенно не отличаются от теоретических данных (рассчитанных в предположении, что доли принятых студентов по регионам соответствуют количеству заявлений). Таким образом, доля принятых студентов примерно одинакова по всем регионам.

Пример 3 – Испытано 105 изделий типа А и 120 изделий типа В. Результаты испытаний следующие:

Тип	Нормальные	Мелкий дефект	Серьезный дефект	Полный отказ
А	77	22	2	4

B	96	11	7	6

Определить, значимо ли различаются результаты испытаний для изделий разных видов (или, другими словами, различаются ли изделия A и B по надежности).

Предположим, что надежность изделий одинакова, и при этом условии вычислим ожидаемые результаты испытаний.

Всего испытано 225 изделий. Из них изделия типа A составляют 105/225 = 0,47, а изделия типа B – 120/225 = 0,53.

Всего нормальными оказались 77+96=173 изделий. Если предположить, что изделия A и B одинаково надежны, нормальных изделий A должно быть 0,47·173=80,73, а нормальных изделий B – 0,53·173=92,27. Результаты аналогичных расчетов для других случаев (мелкий дефект, серьезный дефект, полный отказ) приведены в следующей таблице.

Тип	Нормальные	Мелкий дефект	Серьезный дефект	Полный отказ
A	80,73	15,40	4,20	4,67
B	92,27	17,60	4,80	5,33

Таким образом, получены теоретические величины. Фактические данные приведены в первой таблице.

Вычислим критерий χ^2:

$$\chi^2 = \frac{(82-80.73)^2}{80.73} + \frac{(22-15.40)^2}{15.40} + ... + \frac{(6-5,33)^2}{5.33} = 7,97.$$

Для задач, где имеются таблицы фактических и теоретических величин, количество степеней свободы определяется как $s=(a-1)\cdot(b-1)$, где a и b – количество строк и столбцов таблицы. Для данной задачи $s=(2-1)\cdot(4-1)=3$.

Найдем уровень значимости P, используя функцию **ХИ2.РАСП.ПХ** с параметрами **X**: 7,97; **Степени свободы**: 3. Уровень значимости составит P = 0,047. Таким образом, $P < \alpha$. Это значит, что фактические данные (результаты испытаний) существенно противоречат теоретическим величинам, рассчитанным в предположении об одинаковой надежности изделий A и B. Это означает, что результаты испытаний для A и B существенно различаются. Чтобы проанализировать, в чем именно состоят различия, построим диаграмму, отражающую результаты испытаний (см. рисунок 1). Из нее видно, что изделия типа A более подвержены мелким дефектам, чем изделия типа B, но несколько менее подвержены серьезным дефектам и полным отказам.

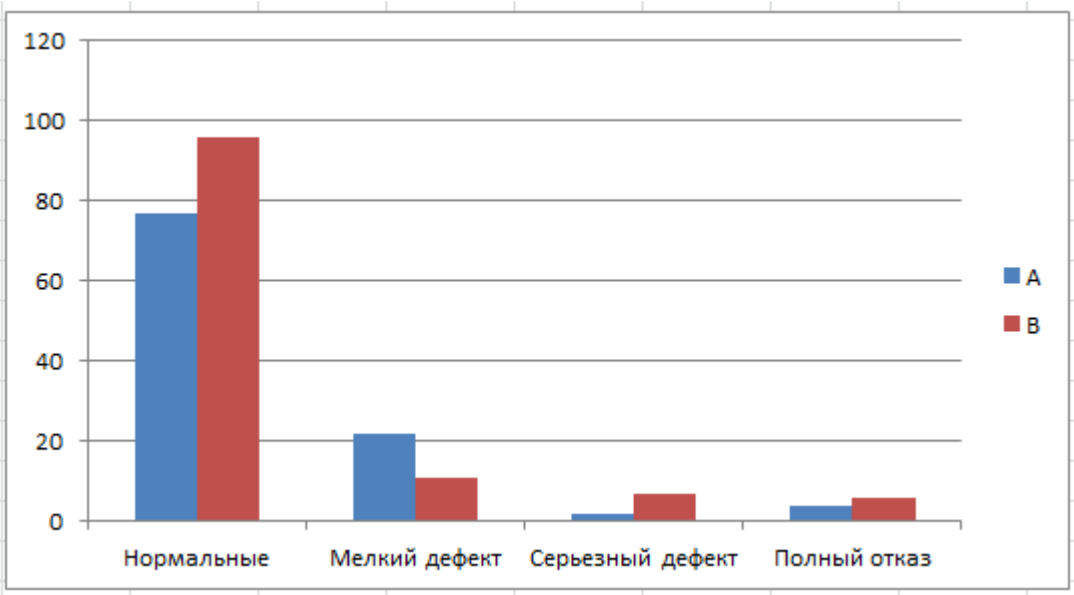

Рисунок 1 – Результаты испытаний изделий двух типов