В изложении доцента БГТУ Сахарова В.Ю.

Критерий согласия

Пусть произведено n взаимно независимых экспериментов со случайной величиной X. Получены числа X_1, X_2, … , X_n. Допустим, что высказана гипотеза, о виде распределения случайной величины X. Например, предполагается, что распределение нормальное. Тогда для проверки этой гипотезы можно использовать какой-нибудь «критерий согласия». Наиболее известным является «критерий хи-квадрат», который можно получить из следующей теоремы.

Теорема Пирсона (без доказательства). Пусть произведено n опытов со случайной величиной X. Пусть множество её возможных значений разбито на r непересекающихся промежутков, целиком её заполняющих. Обозначим ν_i число попаданий случайной величины в i-й промежуток, а p_i вероятность, с которой значение случайной величины X должно попасть в тот же промежуток. Тогда при стремлении n к бесконечности случайная величина $\sum_{i=1}^{r}\frac{(\nu_i - np_i)^2}{np_i}$ стремится к случайной величине, распределённой по закону хи-квадрат с (r–1) степенями свободы, то есть $\lim_{n\to\infty} P\left(\sum_{i=1}^{r}\frac{(\nu_i - np_i)^2}{np_i} < x\right) = F_{\chi^2_{r-1}}(x)$.

Если сравнивать результаты эксперимента не с каким-то априори данным распределением (для которого уже заданы конкретные значения параметров), а с тем, которое получено на основе результатов этого же эксперимента (то есть в качестве параметров используются их точечные оценки, например, для нормального распределения $a \approx \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ выборочное среднее и $\sigma^2 \approx (s_1)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2$ выборочная исправленная дисперсия), то несколько логичнее использовать теорему Фишера.

Не формулируя полностью теорему Фишера, отметим, что в ней доказывается, что при выполнении весьма общих условий, если параметры распределения оцениваются по выборке X_1, X_2, … , X_n, с помощью соответствующих методов, то при стремлении n к бесконечности случайная величина $\sum_{i=1}^{r}\frac{(\nu_i - np_i)^2}{np_i}$ стремится к случайной величине, распределённой по закону хи-квадрат с (r–1–s) степенями свободы, то есть $\lim_{n\to\infty} P\left(\sum_{i=1}^{r}\frac{(\nu_i - np_i)^2}{np_i} < x\right) = F_{\chi^2_{r-1-s}}(x)$, где s число оцениваемых параметров (например, для нормального распределения s=2, поскольку таких параметров два: математическое ожидание a и дисперсия σ^2).

На практике n не может быть равно бесконечности и поэтому остаётся только пользоваться приближённым равенством $P\left(\sum_{i=1}^{r}\frac{(\nu_i - np_i)^2}{np_i} < x\right) \approx F_{\chi^2_{r-1-s}}(x)$. Естественно, что это равенство будет тем более справедливым, чем больше n, но его точность ещё зависит от способа разбиения множества значений случайной величины (для нормального распределения это множество является всей числовой осью) на промежутки. В книгах по прикладной статистике можно найти рекомендации о том, как оптимально осуществлять это разбиение.

Поясним, как следует пользоваться критерием на примере теста, проверяющего, является ли случайная величина нормально распределённой. Разобьём всю числовую прямую на r промежутков $[a_i;b_i)$, где i=1,2,3,…,r. Поскольку числовая прямая бесконечна в обе стороны, то a_1=-∞, а b_r=+∞. Вероятности попадания нормально распределённой случайной величины равны $p_i = \Phi\left(\frac{b_i - a}{\sigma}\right) - \Phi\left(\frac{a_i - a}{\sigma}\right)$. Напомним, что Ф(−∞)=0, Ф(+∞)=1. Для остальных аргументов значения функции Ф(x) можно находить в Microsoft Office Excel с помощью встроенной функции НОРМСТРАСП(x). Вычислим выражение $\chi^{2*}_{r-1-s} = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$, фактически являющееся мерой отклонения случайной величины от нормальной. Если предположить, что случайная величина X действительно распределена нормально с параметрами a и σ^2, то число χ^{2*}_{r-1-s} не должно быть очень большим. Установим его критическое значение c_2, которое соответствует равенству $P(\chi^2_{r-1-s} \le c_2) = \beta$. Здесь и в дальнейшем β это вероятность практически достоверного события. Согласно справке Microsoft Office Excel встроенная в эту программу функция ХИ2ОБР(вероятность;число степеней свободы), возвращает такое число H(α,n), для которого верно равенство $P\left(\chi^2_n > H\right) = \alpha$. Немного преобразуем имеющееся равенство $P(\chi^2_{r-1-s} \le c_2) = \beta \Leftrightarrow 1 - P(\chi^2_{r-1-s} > c_2) = \beta \Leftrightarrow P(\chi^2_{r-1-s} > c_2) = 1 - \beta \Leftrightarrow c_2 = H(1 - \beta, r - 1 - s)$. Если $\chi^{2*}_{r-1-s} > H(1 - \beta, r - 1 - s)$ то гипотеза о том, что случайная величина имеет тестируемое распределение (в примере нормальное) отвергается. В таком случае рекомендуется построить гистограмму и по её виду (сравнивая с графиками известных плотностей вероятностей) догадаться, какой тип имеет распределение. Затем, почитав литературу по математической статистике или обратившись к специалисту, узнать формулы, по которым получаются точечные (и желательно интервальные) оценки параметров распределения этого типа. После чего снова протестировать результат критерием согласия. Если $\chi^{2*}_{r-1-s} \le H(1 - \beta, r - 1 - s)$, то, к сожалению, это ничего не доказывает. Просто это означает, что на основе данного критерия нет оснований отвергать гипотезу. Ещё раз напомним, что критерии не могут подтвердить гипотезы, а могут только отвергнуть.

Согласно теореме Фишера, при $r-1-s \ge 2$ имеется возможность проверить не подтасованы ли результаты экспериментов. Дело в том, что если величина СЛУЧАЙНАЯ, то она ОБЯЗАНА ОТКЛОНЯТЬСЯ от своего математического ожидания. Поэтому если число χ^{2*}_{r-1-s} окажется слишком маленьким, то случайная величина X имеет распределение подозрительно слишком хорошо совпадающее с предполагаемым. Найдём критическое значение c_1, которое соответствует равенству $P(\chi^2_{r-1-s} \le c_1) = 1 - \beta \Leftrightarrow 1 - P(\chi^2_{r-1-s} > c_1) = 1 - \beta \Leftrightarrow P(\chi^2_{r-1-s} > c_1) = \beta \Leftrightarrow c_1 = H(\beta, r - 1 - s)$. Если $\chi^{2*}_{r-1-s} < H(\beta, r - 1 - s)$, то это причина для того, чтобы утверждать, что данные экспериментов подтасованы. Если $\chi^{2*}_{r-1-s} \ge H(\beta, r - 1 - s)$ или $r-1-s<2$, то нет оснований подозревать экспериментатора в подтасовке результатов экспериментов.

Можно предложить более гибкий способ использования критерия согласия хи-квадрат. Предположим, что число χ^{2*}_{r-1-s} является одной из границ промежутка $[c_1; c_2)$. Вычислим $\alpha = P\left(\chi^2_{r-1-s} > \chi^{2*}_{r-1-s}\right)$. Это проще всего сделать с помощью функция ХИ2РАСП(x;число степеней свободы), встроенной в программу Microsoft Office Excel, которая вычисляет выражение, равное единице минус функции распределения случайной величины хи-квадрат. Если обозначить встроенную в Microsoft Office Excel функцию ХИ2РАСП за $F(x,n)$, то $\alpha = P\left(\chi^2_{r-1-s} > \chi^{2*}_{r-1-s}\right) = F\left(\chi^{2*}_{r-1-s}, r - 1 - s\right)$. Чтобы на основе критерия хи-квадрат не было оснований отвергать гипотезу о виде исследуемого распределения, экспериментатор должен согласиться с тем, что величина α не может считаться вероятностью практически невозможного события. А для того, чтобы не было оснований подозревать экспериментатора в подтасовке результатов экспериментов, экспериментатор должен согласиться с тем, что величина α не может считаться вероятностью практически достоверного события (имеет смысл проверять только при $r-1-s\geq 2$).