適合度検定，独立性の検定

母集団の属性が k 個の互いに排反なクラス $C_1, C_2, \cdots, C_k$ に分かれ、このクラスに入る確率が理論的に $p_1, p_2, \cdots, p_k, (\sum_i p_i = 1)$ であるものとしよう。しかし実際に母集団からの抽出した標本の値が $n_1, n_2, \cdots, n_k$ であれば、これらは，$(\sum_i n_i = n)$ であるが、その対応する理論値 $np_1, np_2, \cdots, np_k$ とは，必ずしも合致しないから, ずれが生じる。このずれに関して適当な尺度を考えて，理論と実際とが一致しているかどうかを検定する。

たとえば，ある大学での入学者１０００人について，男女の性別が５３５人と４６５人であったとき，これについて性差がないもとのと仮定するならば、当然入学の比率は 1/2 対 1/2 でなければならない。このとき、性差がないといえるかどうかを検定する問題である。すなわち比率が理論的には５００人対５００人になるはずであるが，この実際の標本値とのずれを適当な尺度で調べることを考える。

クラス	C_1	C_2	$\cdots$	C_k	計
観測度数（標本値）	n_1	n_2	$\cdots$	n_k	n
期待度数（理論値）	np_1	np_2	$\cdots$	np_k	n

検定統計量は $\chi^2 := \sum_{i=1}^{k} \dfrac{(n_i - np_i)^2}{np_i}$ を用いる。ここで分子は２乗をしているが，分母はそうでないことに注意。帰無仮説：母集団の各クラスに属する確率が $p_1, p_2, \cdots, p_k, (\sum_i p_i = 1)$ であるという仮説のもとでは近似的に自由度 $\nu = k - 1$ のカイ２乗分布に従う。近似であるから，この精度を高めるには各クラスの期待度数を５以上にする必要がある。この検定を**適合度検定**という。有意水準が α%の棄却域は，$\chi^2 > \chi^2_\alpha(\nu)$ でカイ 2 乗分布の右側（大きい値のほう）裾のパーセント点による。

つぎに属性が２つの場合 A, B を考える。縦と横のマス目の形でデータが与えられる場合である。一方の属性 A を縦 (第 i 行) ともう一方 B を横 (第 j 列) に並べれば，行列の形に枠が書ける。これを**分割表**とよぶ。このような２つの分類基準の間に関して、これらの関連性の有無があるかどうかを調べる。「関連がない」ということは，それぞれの枠に属する/属さないことが「独立である」とみなす。これを実際の標本値と理論値と比較する。前と同様、ずれの大きさで判断をおこなう。このような検定を**独立性の検定**という。各枠に対する確率の理論値はそれぞれの周辺度数から独立性の仮定のもとで計算できる。

周辺度数をそれぞれ横の合計，縦の合計から r_i, c_j とおくと，一つの枠 (i, j) には，$n \times (\frac{r_i}{n})(\frac{c_j}{n}) = \frac{r_i c_j}{n}$ が理論値による期待度数であり，実際の標本値 n_{ij} とを比較する。

検定統計量は

$$\chi^2 := \sum_i \sum_j \frac{\left(n_{ij} - \dfrac{r_i c_j}{n}\right)^2}{\dfrac{r_i c_j}{n}}$$

帰無仮説のもとでは，近似的に自由度が $\nu = \{$(行数－１)(列数－１)$\}$ のカイ 2 乗分布にしたがい，有意水準が α%の棄却域は，$\chi^2 > \chi^2_\alpha(\nu)$ でカイ 2 乗分布の右側（大きい値のほう）裾のパーセント点による。とくに 2 行 2 列の特別な場合には,

	B_1	B_2	計
A_1	n_{11}	n_{12}	r_1
A_2	n_{21}	n_{22}	r_2
計	c_1	c_2	n

期待度数が小さい場合には，近似をよくするために 0.5 だけずらす方法がイェーツの補正という。

$$\chi^2 = \sum_i \sum_j \frac{\left(\left|n_{ij} - \dfrac{r_i c_j}{n}\right| - 0.5\right)^2}{\dfrac{r_i c_j}{n}}$$