Тема 1.

Создание модуля для эффективной обработки генетических данных с использованием классических категориальных тестов

Предлагаются к обработке табличные данные следующего формата:

`id`	`phenotype`	`genotype`
$\{1, 2, \ldots\}$	$\{0, 1, 2, \ldots, k-1\}$	$\{0, 1, 2, 3\}$

Требуется выяснить зависимости фенотипа от каждого из генотипов. Иными словами, пусть случайная величина ξ принимает значение на множестве фенотипов, а η — на множестве генотипов, и выборка объема n имеет вид $\{(x_i, y_i)\}_{i=1}^n$; тогда требуется проверить гипотезу $H_0 : \xi \perp\!\!\!\perp \eta$ что ξ и η независимы.

Поскольку признаки качественные, уместно использовать критерий независимости χ^2[1]. Для его построения группируют выборку (пусть всего при этом обнаружилось K и S уникальных фенотипов и генотипов) и составляют таблицу сопряженности, сопоставляющую каждой уникальной паре (x_i^*, y_j^*) абсолютную частоту n_{ij}:

$$\begin{array}{c|ccc} & y_1^* & \cdots & y_S^* \\ x_1^* & n_{11} & \cdots & n_{1S} \\ \vdots & \vdots & \ddots & \vdots \\ x_K^* & n_{K1} & \cdots & n_{KS} \end{array}$$

Статистикой критерия является

$$T = \sum_{i=1}^{K} \sum_{j=1}^{S} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n)^2}{n_{i\cdot} n_{\cdot j}/n} \xrightarrow{\sim} \chi^2((K-1)(S-1)), \quad \text{где } n_{i\cdot} = \sum_{\ell=1}^{S} n_{i\ell},\ n_{\cdot j} = \sum_{\ell=1}^{K} n_{\ell j}.$$

Замечание. Критерий может быть использован только если $n > 40$ и $np_{ij} > 5\ \forall i, j$; в противном случае следует воспользоваться критерием с перестановкой[2] (FIXME) или, в случае, если в выборке встречаются только 2 генотипа и 2 фенотипа, точным критерием Фишера.

Проверка гипотезы происходит по обычному плану:

1. Фиксируют уровень значимости α (обычно 0.05 или 0.01).
2. Определяются с разбиением области значений статистики на критическую и доверительную область (поскольку при независимых ξ, η $T = 0$, определить её на правом конце $[0, +\infty)$).

[1] Детали о критерии можно найти в файле `statistics-manual.doc`.
[2] https://en.wikipedia.org/wiki/Resampling_(statistics)$Permutation_tests

3. Считают значение T на данных, после чего получают p-value как $1-\mathrm{cdf}_{\chi^2((K-1)(S-1))}(T)$, где $\mathrm{cdf}_{\chi^2((K-1)(S-1))}$ — функция распределения распределения $\chi^2((K-1)(S-1))$. После чего, если p-value $< \alpha$, то это означает попадание в критическую область и отвержение гипотезы о независимости ξ, η.

Замечание. Для проверки реализации можно воспользоваться[3] функцией `chisq.test` из R.

[3] http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence