Семинар №13.
Коэффициенты корреляции.

Во многих задачах математической статистики (кто не верит, спросите у экономистов) требуется проверить гипотезу о том, что две выборки являются независимыми. Иначе говоря, пусть $X_1, \dots, X_n \sim F_X$ и $Y_1, \dots, Y_n \sim F_Y$ — две выборки, хотим проверить гипотезу $H_0 : F_{X,Y}(s,t) = F_X(s)F_Y(t)$. Ясно, что доподлинно убедиться, что две конечные выборки независимы, в общем случае нереально (мы же не знаем точных функций распределения выборок), а вот отвергнуть гипотезу о независимости реальнее, например, посчитав коэффициент корреляции двух выборок и убедившись, что он далеко отстоит от 0.

1. Коэффициент корреляции Пирсона и нормальные выборки.

Это тот случай, когда мы можем проверить гипотезу о независимости, так как для нормальных случайных величин равенство их коэффициента корреляции нулю эквивалентно их независимости.
Определение. Коэффициентом корреляции Пирсона, или обычным коэффициентом корреляции, называется следующая статистика:

$$\widehat{\rho} = \frac{\sum\limits_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum\limits_{i=1}^{n}(X_i - \overline{X})^2 \sum\limits_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

Свойства коэффициента корреляции Пирсона.

1. $\widehat{\rho} \xrightarrow{P} \rho(X_1, Y_1) = \frac{cov(X_1,Y_1)}{\sqrt{DX_1 DY_1}}$, т.е. можно сказать, что коэффициент корреляции Пирсона соответствует стандартному коэффициенту корреляции двух случайных величин.

2. Если гипотеза H_0 о независимости выборок верна, то статистика $T = \frac{\widehat{\rho}\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2}} \sim T_{n-2}$, т.е. распределена по Стьюденту с $n-2$ степенями свободы.
   Таким образом, критерий для проверки гипотезы H_0 для нормальных выборок выглядит так: если $T \notin (z_\alpha, z_{1-\alpha})$, где z_α и $z_{1-\alpha}$ — квантили уровня α и $1-\alpha$ из распределения Стьюдента с $n-2$ степенями свободы, то отвергаем гипотезу H_0.

Но, к сожалению, как и большинство параметрических методов, коэффициент корреляции Пирсона не является устойчивым к выбросам (не является робастной статистикой), т.е. если в выборке есть далеко отстоящие данные, связанные с ошибками измерения или ещё чем, то коэффициент корреляции может выдавать не совсем адекватные значения. Широко известен пример (так называемый "квартет Энскомба"), демонстрирующий, насколько неробастные методы обработки статистических данных способны «врать», даже если выброс всего один на 10 «обычных» результатов.

2. Коэффициент корреляции Спирмэна.

Пусть имеется выборка $X_1, \dots, X_n$ из некого непрерывного распределения (т.е. элементы выборки не совпадают почти наверное). Упорядочим элементы выборки по возрастанию (т.е. построим вариационный ряд выборки).

Определение. Номера, которые получили элементы выборки при таком упорядочивании, называются их рангами.

Будем обозначать ранги выборки X_i как $R(X_i)$ (получается, что ранг $R(X_i)$ – это номер наблюдения X_i в вариационном ряде выборки $X_1, \dots, X_n$).

Основное свойство рангов следующее: $P(R(X_1) = r_1, \dots R(X_n) = r_n) = \frac{1}{n!}$, где $(r_1, \dots, r_n)$ — произвольная перестановка чисел $(1, \dots, n)$. Это верно хотя бы потому, что $P(X_i > X_j) = P(X_i < X_j)$, так как члены выборки одинаково распределены. Из основного свойства рангов также следует, что их распределение не зависит от первоначального, неизвестного нам распределения, из которого бралась выборка.

Рассмотрим теперь, как и ранее, две выборки $X_1, \dots, X_n \sim F_X$ и $Y_1, \dots, Y_n \sim F_Y$. Обозначим ранг наблюдения X_i в выборке $(X_1, \dots, X_n)$ как R_i и ранг наблюдения Y_j в выборке $(Y_1, \dots, Y_n)$ как S_j, где эти ранги, как легко видеть, могут принимать значения от 1 до n.

Определение. Коэффициентом корреляции Спирмэна называется следующая статистика:

$$\rho_S = \frac{\sum\limits_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum\limits_{i=1}^{n}(R_i - \overline{R})^2 \sum\limits_{i=1}^{n}(S_i - \overline{S})^2}}.$$

Свойства коэффициента корреляции Спирмэна.

1. Можно доказать, что $\rho_S = 1 - \frac{6}{n^3 - n}\sum\limits_{i=1}^{n}(R_i - S_i)^2$.
2. При верной гипотезе H_0 $E\rho_S = 0$ и $D\rho_S = \frac{1}{n-1}$.
3. $-1 \le \rho_S \le 1$, причём обе границы достигаются, т.е. именование ρ_S коэффициентом корреляции оправданно.
4. При верной гипотезе H_0 $\frac{\rho_S}{\sqrt{D\rho_S}} \xrightarrow{d} N(0,1)$. Этим нормальным приближением для построения критерия можно пользоваться при $n \ge 50$, при меньших значениях n рекомендуется использовать исправленную статистику $\widetilde{\rho_S} = \frac{1}{2}\rho_S\left(\sqrt{n-1} + \sqrt{\frac{n-2}{1-\rho_S^2}}\right)$. Тогда критерий будет выглядеть так: отвергать гипотезу независимости H_0, если $\widetilde{\rho_S} \notin (z_\alpha, z_{1-\alpha})$, где $z_\alpha = \frac{1}{2}(x_\alpha + y_\alpha)$, x_α — $\alpha-$квантиль $N(0,1)$, а y_α — $\alpha-$квантиль распределения Стьюдента с $n-2$ степенями свободы.

3. Коэффициент корреляции Кэндалла (τ Кэндалла).

Ещё одним примером робастного коэффициента корреляции, построенного с помощью ранговых методов, является коэффициент корреляции Кэндалла.

Определение. Пары (X_i, Y_i) и (X_j, Y_j) называются согласованными, если $sign(X_i - X_j)sign(Y_i - Y_j) = 1$.

Пусть S — число согласованных пар, R — число несогласованных. По-прежнему считаем, что внутри выборок нет одинаковых элементов. Определим $T = S - R = \sum\limits_{i<j} sign(X_i - X_j)sign(Y_i - Y_j)$, легко видеть, что T может меняться от $-\frac{n(n-1)}{2}$ до $\frac{n(n-1)}{2}$, т.к. $S + R = \{\text{количество всех пар } (i,j), i \ne j\} = \frac{n(n-1)}{2}$.

Определение. Коэффициентом корреляции Кэндалла называется статистика $\tau = \frac{2}{n(n-1)}T$.

Свойства коэффициента корреляции Кэндалла.

1. Можно также получить следующее представление: $\tau = 1 - \frac{4}{n(n-1)}R$.
2. τ является коэффициентом корреляции, т.е. $-1 \le \tau \le 1$ и границы достигаются, и при верной гипотезе H_0 $E\tau = 0$. Кроме того, $D(\tau|H_0) = \frac{2(2n+5)}{9n(n-1)}$.

3. При верной гипотезе H_0 $\frac{\tau}{\sqrt{D\tau}} \xrightarrow{d} N(0,1)$.

4. При верной гипотезе H_0 коэффициенты корреляции Спирмэна и Кэндалла сильно коррелированы ($\rho(\rho_S, \tau) > 0,99$ при $n > 5$). Но коэффициент корреляции Спирмэна более чувствителен к количеству несогласованных пар.

4. Обобщённый коэффициент корреляции.

Для удобства реализации на компьютере системы алгоритмов корреляционного анализа полезно вывести обобщённую формулу для вычисления различных парных корреляционных характеристик (таких, как $\widehat{\rho}$, ρ_S, τ).
Определим $C : (X_i, X_j) \to c_{ij}(X)$, где $c_{ij} = -c_{ji}$ и $c_{ii} = 0$.

Определение. Обобщённым коэффициентом корреляции называется следующая статистика:

$$\widehat{r} = \frac{\sum\limits_{i<j} c_{ij}(X)c_{ij}(Y)}{\sqrt{\sum\limits_{i<j} c_{ij}(X)^2 \sum\limits_{i<j} c_{ij}(Y)^2}}.$$