ЛАБОРАТОРНАЯ РАБОТА
«Корреляционный анализ»

1.1. Описание метода.

Пусть дана матрица данных $Z(N\times p)$.

$\bar{z}^j$ – среднее значение j–го признака $\bar{z}^j = \frac{1}{N}\sum_{i=1}^{N} z_{ij}$, $j=\overline{1,p}$;

$(S^2)^j = \frac{1}{N}\sum_{i=1}^{N}\left(z_{ij} - \bar{z}^j\right)^2$ – оценка дисперсии j-го столбца, $j=\overline{1,p}$.

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}$$ — ковариационная матрица, где $\sigma_{ij} = \frac{1}{N}\sum_{k=1}^{N}\left(z_{ki} - \bar{z}^i\right)\left(z_{kj} - \bar{z}^j\right)$.

$X(N\times p)$ — стандартизованная матрица: $X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}$, где $x_{ij} = \frac{z_{ij} - \bar{z}^j}{s^j}$.

$R(p\times p)$ — корреляционная матрица, $r_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ki}x_{kj}$

1.2. Оценка значимости коэффициента корреляции.

Пусть имеются статистические гипотезы:

H_0: $\rho(x,y) = 0$, связи между признаками x и y нет.

H_1: $\rho(x,y) \neq 0$, то есть связь есть. Здесь $\rho(x,y)$- коэффициент корреляции между x и y.

Действие / Состояние природы	H_0 принимаем	H_0 отвергаем
верна H_0	верное решение	α
верна H_1	β	верное решение

α — вероятность ошибки первого рода — вероятность отвергнуть верную гипотезу,
β — вероятность ошибки второго рода — вероятность принять неверную гипотезу.
Надо сформулировать такое правило , чтобы α и β были достаточно малыми. В математической статистике показано, что статистика

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

при условии, что H_0 справедлива, подчиняется закону распределения Стьюдента.

1.3. Алгоритм проверки статистической гипотезы о значимости коэффициента корреляции.

1) Пусть имеются экспериментальные данные

(x_1, y_1)
(x_2, y_2)
...
(x_n, y_n).

Вычисляем $r(x, y)$ — выборочный коэффициент корреляции.
2) Задаемся приемлемой для нас вероятностью ошибки α, пусть α=0,05.
3) Вычисляем статистику t.
4) По выбранному α и числу степеней свободы $f=n-2$ определяем $t_{\text{табличное}}$.
5) Правило вынесения решения: если $|t_{\text{расч}}| \geq t_{\text{табл}}$, то справедлива гипотеза H_1, в противном случае — H_0, а отличие от нуля коэффициента корреляции обусловлено случайными причинами.

1.4. Порядок выполнения работы.

Пусть дана Z — матрица данных размером N×p.
1) Составить программу для вычисления
   а) средних по столбцам, дисперсий по столбцам;
   б) стандартизованной матрицы;

в) ковариационной матрицы;

г) корреляционной матрицы.

2) Проверить гипотезу о значимости коэффициентов корреляции между столбцами матрицы данных.

1.5. Задание.

Выполнить работу для конкретной матрицы Z и результаты расчетов вывести на печать.