Корреляция

Классическим инструментом для измерения линейной зависимости между двумя наборами данных является коэффициент корреляции. Коэффициент корреляции — это числовая величина, находящаяся в интервале от −1 до +1. Чем она больше по модулю (т. е. ближе к +1 или −1), тем выше линейная связь между наборами данных. Знак коэффициента корреляции показывает, в одном ли направлении изменяются наборы данных. Если один из наборов возрастает, а второй убывает, то коэффициент корреляции отрицателен, а если оба набора одновременно возрастают или убывают, то коэффициент корреляции положителен. Значение коэффициента корреляции по модулю равное 1 соответствует точной линейной зависимости между двумя наборами данных.

Обратите внимание, что значение коэффициента корреляции близкое к нулю не означает независимости наборов данных. Коэффициент корреляции — это мера линейной зависимости, поэтому этот факт означает лишь отсутствие линейной зависимости, но не исключает любой другой. Отсутствие линейной зависимости равносильно независимости только для нормально распределённых выборок, факт нормальности, естественно, надо проверять отдельно.

Для вычисления коэффициента корреляции в R реализована функция
cor(x,y)

Проверка гипотезы о значимости коэффициента корреляции равносильна проверке гипотезы о равенстве нулю коэффициента корреляции. Если гипотеза отвергается, то влияние одного набора данных на другой считается *значимым*. Для проверки гипотезы используется функция

```
z<- cor.test(x,y)
```

Выставить доверительный уровень для него можно с помощью опции conf.level.

```
cor.test(x,y, conf.level=0.9)
```

Кроме непосредственно p-значения (*p-value)*, достигнутого при проверке гипотезы о нулевом значении коэффициента корреляции, функция выводит оценку коэффициента корреляции и доверительный интервал для него.

Если *p-value <0.05, то нулевая гипотеза (*гипотеза о нулевом значении коэффициента корреляции*) отвергается* с заданным уровнем надёжности

Поэтому интервал $[\bar{x}-s,\ \bar{x}+s]$ называют 68-процентным доверительным интервалом. И говорят, что значение x, попадающее в этот интервал , значимо при 32-процентном уровне значимости (вероятность ошибки равна 32%).

Итак,

$[\bar{x}-s,\ \bar{x}+s]$ – 68% доверительный интервал, 32% уровень значимости;
$[\bar{x}-2s,\ \bar{x}+2s]$ – 95% доверительный интервал, 5% уровень значимости;
$[\bar{x}-3s,\ \bar{x}+3s]$ – 99,7% доверительный интервал, 0,3% уровень значимости.