В изложении доцента БГТУ Сахарова В.Ю.

Часть 1

Пусть произведено n взаимно независимых экспериментов с нормальной двумерной случайной величиной (X, Y). Получены пары чисел (X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n). Ранее было показано, что точечной оценкой корреляции ρ_{XY} является случайная величина

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)\left(Y_i - \frac{1}{n}\sum_{j=1}^{n}Y_j\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right)\cdot\sum_{i=1}^{n}\left(\left(Y_i - \frac{1}{n}\sum_{j=1}^{n}Y_j\right)^2\right)}}$$

Математики (Фишер) доказали, что случайная величина

$$U = \frac{1}{2}\ln\frac{1+r_{XY}}{1-r_{XY}}$$

Имеет распределение близкое к нормальному.

Причём её математическое ожидание

$$EU = \frac{1}{2}\ln\frac{1+\rho_{XY}}{1-\rho_{XY}} + \frac{\rho_{XY}}{2(n-1)}$$

И дисперсия

$$DU = \frac{1}{n-3}$$

Тогда случайная величина

$$Z = \frac{U - EU}{\sqrt{DU}}$$

имеет стандартное нормальное распределение.

Выберем достаточно большую вероятность β (по мнению исследователя равную вероятности практически достоверного события) и построим интервал $(-c;c)$ практически гарантировано содержащий случайную величину Z. Очевидно, что чем больше будет эта вероятность, тем длиннее окажется этот интервал. Такую вероятность просто находить как разность значений функции распределения $\Phi(x)$ стандартного нормального распределения

$$P\left(-c < \frac{U - EU}{\sqrt{DU}} < c\right) = P(-c < Z < c) = \Phi(c) - \Phi(-c) = \Phi(c) - (1 - \Phi(c)) = 2\Phi(c) - 1$$

По нашему условию эта вероятность равна β.

$$2\Phi(c) - 1 = \beta$$

$$2\Phi(c) = 1 + \beta$$

$$\Phi(c) = \frac{1+\beta}{2}$$

В программе Microsoft Office Excel встроена функция НОРМСТОБР(вероятность), с помощью которой можно находить значения функции, обратной к функции распределения $\Phi(x)$. Обозначим здесь эту функцию $G(x)$. Тогда

$$c = G\left(\frac{1+\beta}{2}\right)$$

Итак, c это известное число и после подстановок можно попытаться разрешить относительно корреляции ρ_{XY} неравенство

$$-c < \frac{U - EU}{\sqrt{DU}} < c$$

$$-c\sqrt{DU} < U - EU < c\sqrt{DU}$$

$$-\frac{c}{\sqrt{n-3}} < \frac{1}{2}\ln\frac{1+r_{XY}}{1-r_{XY}} - \left(\frac{1}{2}\ln\frac{1+\rho_{XY}}{1-\rho_{XY}} + \frac{\rho_{XY}}{2(n-1)}\right) < \frac{c}{\sqrt{n-3}}$$

К сожалению, функция относительно ρ_{XY} оказалась слишком сложной. Неизвестное ρ_{XY} находится и в первой степени и под знаком логарифма. Арифметически такое неравенство не решается. Для того, чтобы получить хоть какое-то (пусть не точное, а приближённое) решение можно пренебречь слагаемым $\frac{\rho_{XY}}{2(n-1)}$, поскольку при стремлении n к бесконечности оно стремится к нулю.

$$-\frac{c}{\sqrt{n-3}} < \frac{1}{2}\ln\frac{1+r_{XY}}{1-r_{XY}} - \frac{1}{2}\ln\frac{1+\rho_{XY}}{1-\rho_{XY}} < \frac{c}{\sqrt{n-3}}$$

$$-\frac{2c}{\sqrt{n-3}} < \ln\frac{1+r_{XY}}{1-r_{XY}} - \ln\frac{1+\rho_{XY}}{1-\rho_{XY}} < \frac{2c}{\sqrt{n-3}}$$

$$-\frac{2c}{\sqrt{n-3}} < \ln\left(\left(\frac{1+r_{XY}}{1-r_{XY}}\right)/\left(\frac{1+\rho_{XY}}{1-\rho_{XY}}\right)\right) < \frac{2c}{\sqrt{n-3}}$$

$$-\frac{2c}{\sqrt{n-3}} < \ln\left(\frac{1+r_{XY}}{1-r_{XY}}\cdot\frac{1-\rho_{XY}}{1+\rho_{XY}}\right) < \frac{2c}{\sqrt{n-3}}$$

$$e^{-\frac{2c}{\sqrt{n-3}}} < \frac{1+r_{XY}}{1-r_{XY}}\cdot\frac{1-\rho_{XY}}{1+\rho_{XY}} < e^{\frac{2c}{\sqrt{n-3}}}$$

$$e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY})(1+\rho_{XY}) < (1+r_{XY})(1-\rho_{XY}) < e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY})(1+\rho_{XY})$$

$$e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY}) + e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY})\rho_{XY} < 1 + r_{XY} - (1+r_{XY})\rho_{XY} < e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY}) + e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY})\rho_{XY}$$

Запишем двойное неравенство в виде системы

$$\begin{cases} e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY}) + e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY})\rho_{XY} < 1 + r_{XY} - (1+r_{XY})\rho_{XY} \\ 1 + r_{XY} - (1+r_{XY})\rho_{XY} < e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY}) + e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY})\rho_{XY} \end{cases}$$

$$\begin{cases} \left(e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY}) + 1 + r_{XY}\right)\rho_{XY} < 1 + r_{XY} - e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY}) \\ 1 + r_{XY} - e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY}) < \left(e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY}) + 1 + r_{XY}\right)\rho_{XY} \end{cases}$$

Так как корреляция меньше единицы то коэффициенты при ρ_{XY} положительные и на них можно разделить

$$\begin{cases} \rho_{XY} < \dfrac{1+r_{XY}-e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY})}{e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY})+1+r_{XY}} \\ \dfrac{1+r_{XY}-e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY})}{e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY})+1+r_{XY}} < \rho_{XY} \end{cases}$$

Получаем доверительный интервал

$$\frac{1+r_{XY}-e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY})}{e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY})+1+r_{XY}} < \rho_{XY} < \frac{1+r_{XY}-e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY})}{e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY})+1+r_{XY}}$$

Или в виде промежутка:

$$\left(\frac{1+r_{XY}-e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY})}{e^{\frac{2c}{\sqrt{n-3}}}(1-r_{XY})+1+r_{XY}}; \frac{1+r_{XY}-e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY})}{e^{-\frac{2c}{\sqrt{n-3}}}(1-r_{XY})+1+r_{XY}} \right)$$

Часть 2

Чаще всего интерес представляет вопрос не столько о величине корреляции, сколько о её наличии или отсутствии. В этом случае важно равен коэффициент корреляции нулю или нет. В силу парадокса нулевой вероятности выборочный коэффициент корреляции r_{XY} всегда окажется отличным от нуля. И важно знать, значимо ли это отличие. Можно построить критерий для проверки гипотезы H_0: $\rho_{XY}=0$, при альтернативной гипотезе H_1: $\rho_{XY}\neq 0$. Математики доказали, что для нормальной двумерной случайной величины (X, Y) при $\rho_{XY}=0$ случайная величина

$$T = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}}}$$

имеет распределение Стьюдента с (n-2) степенями свободы.

Зададимся уровнем значимости α (это вероятность отвергнуть гипотезу H_0, если на самом деле она верная). Тогда вероятность принять (точнее: не найти причины отвергнуть) гипотезу H_0: $\rho_{XY}=0$, если на самом деле она верная, будет равна (1-α). Это приводит к равенству

$$1-\alpha = P\left(-c < \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}}} < c\right) = P\left(\left|\frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}}}\right| < c\right) = P(|T|<c) = 1-P(|T|\geq c)$$

$$\alpha = P(|T|\geq c)$$

В программе Microsoft Office Excel встроена функция СТЬЮДРАСПОБР(вероятность; число степеней свободы), с помощью которой можно находить значения функции, обратной к функции распределения Стьюдента. Согласно тексту справки она возвращает такое значение $t(x,n)$, для которого верно равенство $P(|S_n|>t)=x$.

Поэтому $c = t(\alpha, n-2)$

Итого, если неравенство

$$\left|\frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}}}\right| < t(\alpha, n-2)$$

выполняется, то нет оснований отвергать гипотезу H_0: $\rho_{XY}=0$. Если неравенство не выполняется, то принимается гипотеза, что $\rho_{XY}\neq 0$. Ещё раз обратите внимание на вид

формулировки. Гипотезы не подтверждают, а говорят, что критерий не выявил оснований для отвержения гипотезы (возможно, что это не гипотеза правильная, а критерий не достаточно хороший).