Paper I

Distribution Theory and Stochastic Process

Unit I: Bivariate Normal Distribution

What is a Bivariate Normal Distribution? The "regular" normal distribution has one random variable; A bivariate normal distribution is made up of two independent random variables. The two variables in a bivariate normal are both are normally distributed, and they have a normal distribution when both are added together. Visually, the bivariate normal distribution is a three-dimensional bell curve.

Francis Galton (1822-1911) was one of the first mathematicians to study the bivariate normal distribution in depth, during his study on the heights of parents and their adult children. Bravais, Gauss, Laplace, Plana also studied the distribution in the early nineteenth century (Balakrishnan & Lai, 2009).

Definition 1. A two-dimensional RV (X, Y) is said to have a bivariate normal

$$(1)\quad f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-Q(x,y)/2},\quad -\infty<x<\infty,\quad -\infty<y<\infty,$$

where $\sigma_1>0, \sigma_2>0, |\rho|<1$, and Q is the positive definite quadratic form

$$(2)\quad Q(x,y)=\frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2}+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

and

$$\rho\equiv \operatorname{cor}(x_1,x_2)=\frac{V_{12}}{\sigma_1\sigma_2} \qquad (3)$$

is the correlation of x_1 and x_2 and V_{12} is the covariance.

Marginal Distribution of X and Y

The moment generating function of (X, Y) can be given by

$$M(t_1, t_2) = \exp\left(\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2}{2}\right).$$

The moment generating function of X can be given by

$$M_X(t_1) = M(t_1, 0) = exp\left[\mu_1 t_1 + \frac{1}{2}\sigma_1^2 t_1^2\right].$$

Similarly, the moment generating function of Y can be given by

$$M_Y(t_2) = M(0, t_2) = exp\left[\mu_2 t_2 + \frac{1}{2}\sigma_2^2 t_2^2\right].$$

Thus, X and Y are both marginally normal distributed, i.e.,

$X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim N(\mu_2, \sigma_2^2)$.

The pdf of X is

$$f_X(x) = f_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right].$$

The pdf of Y is

$$f_Y(y) = f_2(y) = \frac{1}{\sqrt{2\pi}\sigma_2} exp\left[-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right].$$

Theorem: X and Y have Bivariate Normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2 and correlation coefficient ρ. Then X and Y are independent if $\rho = 0$.

Proof:

$$M(t_1, t_2) = \exp\left(\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2}{2}\right).$$

If $\rho = 0$, then

$M(t_1, t_2) = \exp[\mu_1 t_1 \ \mu_2 t_2 + \frac{1}{2}(\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2)]$

MGF of X is

$M_X(t_1) = M(t_1, 0) = exp\left[\mu_1 t_1 + \frac{1}{2}\sigma_1^2 t_1^2\right]$

MGF of Y is

$M_Y(t_2) = M(0, t_2) = exp\left[\mu_2 t_2 + \frac{1}{2}\sigma_2^2 t_2^2\right]$

So if $\rho = 0$, then

$M(t_1, t_2) = \exp[\mu_1 t_1 \ \mu_2 t_2 + \frac{1}{2}(\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2)] = M(t_1, 0)M(t_2, 0)$

Therefore, X and Y are independent.

Conversely If X and Y are independent, then

$M(t_1, t_2) = M(t_1, 0)M(t_2, 0) = \exp[\mu_1 t_1 \ \mu_2 t_2 + \frac{1}{2}(\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2)]$

Therefore, $\rho = 0$

Accordingly, we have the following theorem:

Let X and Y have Bivariate Normal distribution with means $\mu_1 \ and \ \mu_2$, variances $\sigma_1^2 and \ \sigma_2^2 \ and \ correlation \ coefficient \ \rho$. Then X and Y are independent if $\rho = 0$.

Theorem: X and Y have Bivariate Normal distribution with means $\mu_1 \ and \ \mu_2$, variances $\sigma_1^2 and \ \sigma_2^2 \ and \ correlation \ coefficient \ \rho$. Then aX + bY ~ N ($a\mu_x + b\mu_y, a^2\sigma_x^2 + 2ab\rho\sigma_x\sigma_y + b^2\sigma_y^2$)

Proof: $M_{X+Y}(t) = E\left[e^{t(X+Y)}\right] = E\left[e^{tX+tY}\right]$

Recall that $M(t_1, t_2) = E\left[e^{t_1 X + t_2 Y}\right]$, therefore we can obtain $M_{X+Y}(t)$ by setting $t_1 = t_2 = t$ in $M(t_1, t_2)$

That is,

$$M_{X+Y}(t) = M(t,t) = \exp\left[\mu_X t + \mu_Y t + \frac{1}{2}\left(\sigma_X^2 t^2 + 2\rho\sigma_X\sigma_Y t^2 + \sigma_Y^2 t^2\right)\right]$$

$$= \exp\left[\left(\mu_X + \mu_Y\right)t + \frac{1}{2}\left(\sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2\right)t^2 \right]$$

$$\therefore X + Y \sim N(\mu = \mu_X + \mu_Y, \sigma^2 = \sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2)$$

$\Rightarrow$ aX + bY ~ N ($a\mu_x + b\mu_y$, $a^2\sigma_x^2 + 2ab\rho\sigma_x\sigma_y + b^2\sigma_y^2$)

The conditional distribution of X given Y=y is given by

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} exp\left\{ -\frac{\left(x - \mu_1 - \frac{\sigma_1}{\sigma_2}\rho(y-\mu_2)\right)^2}{2(1-\rho^2)\sigma_1^2} \right\}.$$

Similarly, we have the conditional distribution of Y given X=x is

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} exp\left\{ -\frac{\left(y - \mu_2 - \frac{\sigma_2}{\sigma_1}\rho(x-\mu_1)\right)^2}{2(1-\rho^2)\sigma_2^2} \right\}.$$

Therefore:

$$X|Y = y \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2), (1-\rho^2)\sigma_1^2\right)$$

$$Y|X = x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), (1-\rho^2)\sigma_2^2\right)$$

Proof of MGF of (X, Y)

The m.g.f. of a bivariate normal distribution can be determined as follows. We have

$$M(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 x + t_2 y} f(x, y)\ dx\, dy$$

$$= \int_{-\infty}^{\infty} e^{t_1 x} f_1(x) \left[\int_{-\infty}^{\infty} e^{t_2 y} f_{2|1}(y|x)\, dy \right] dx$$

for all real values of t_1 and t_2. The integral within the brackets is the m.g.f. of the conditional p.d.f. $f_{2|1}(y|x)$. Since $f_{2|1}(y|x)$ is a normal p.d.f. with mean $\mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$ and variance $\sigma_2^2(1 - \rho^2)$, then

$$\int_{-\infty}^{\infty} e^{t_2 y} f_{2|1}(y|x)\, dy = \exp\left\{ t_2 \left[\mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1) \right] + \frac{t_2^2 \sigma_2^2 (1 - \rho^2)}{2} \right\}.$$

Accordingly, $M(t_1, t_2)$ can be written in the form

$$\exp\left\{t_2\mu_2 - t_2\rho\frac{\sigma_2}{\sigma_1}\mu_1 + \frac{t_2^2\sigma_2^2(1-\rho^2)}{2}\right\}\int_{-\infty}^{\infty}\exp\left[\left(t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right)x\right]f_1(x)\,dx.$$

But $E(e^{tX}) = \exp[\mu_1 t + (\sigma_1^2 t^2)/2]$ for all real values of t. Accordingly, if we set $t = t_1 + t_2\rho(\sigma_2/\sigma_1)$, we see that $M(t_1, t_2)$ is given by

$$\exp\left\{t_2\mu_2 - t_2\rho\frac{\sigma_2}{\sigma_1}\mu_1 + \frac{t_2^2\sigma_2^2(1-\rho^2)}{2} + \mu_1\left(t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right) + \sigma_1^2\frac{\left(t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right)^2}{2}\right\}$$

or, equivalently,

$$M(t_1, t_2) = \exp\left(\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2}{2}\right).$$

Let X and Y be jointly normal random variables with parameters $\mu_X = 1$, $\sigma_X^2 = 1$, $\mu_Y = 0$, $\sigma_Y^2 = 4$, and $\rho = \frac{1}{2}$.

a. Find $P(2X + Y \leq 3)$.
b. Find $\text{Cov}(X + Y, 2X - Y)$.
c. Find $P(Y > 1 | X = 2)$.

Solution

a. Since X and Y are jointly normal, the random variable $V = 2X + Y$ is normal. We have

$$EV = 2EX + EY = 2,$$

$$\begin{aligned} Var(V) &= 4Var(X) + Var(Y) + 4\text{Cov}(X, Y) \\ &= 4 + 4 + 4\sigma_X \sigma_Y \rho(X, Y) \\ &= 8 + 4 \times 1 \times 2 \times \frac{1}{2} \\ &= 12. \end{aligned}$$

Thus, $V \sim N(2, 12)$. Therefore,

$$P(V \leq 3) = \Phi\left(\frac{3-2}{\sqrt{12}}\right) = \Phi\left(\frac{1}{\sqrt{12}}\right) = 0.6136$$

b. Note that $\text{Cov}(X, Y) = \sigma_X \sigma_Y \rho(X, Y) = 1$. We have

$$\begin{aligned} \text{Cov}(X + Y, 2X - Y) &= 2\text{Cov}(X, X) - \text{Cov}(X, Y) + 2\text{Cov}(Y, X) - \text{Cov}(Y, Y) \\ &= 2 - 1 + 2 - 4 = -1. \end{aligned}$$

c. Using Theorem 5.4, we conclude that given $X = 2$, Y is normally distributed with

$$\begin{aligned} E[Y | X = 2] &= \mu_Y + \rho \sigma_Y \frac{2 - \mu_X}{\sigma_X} = 1 \\ Var(Y | X = x) &= (1 - \rho^2)\sigma_Y^2 = 3. \end{aligned}$$

Thus

$$P(Y > 1 | X = 2) = 1 - \Phi\left(\frac{1-1}{\sqrt{3}}\right) = \frac{1}{2}.$$

Let X and Y be jointly normal random variables with parameters $\mu_X = 0$, $\sigma_X^2 = 1$, $\mu_Y = -1$, $\sigma_Y^2 = 4$, and $\rho = -\frac{1}{2}$.

1. Find $P(X+Y>0)$.
2. Find the constant a if we know $aX+Y$ and $X+2Y$ are independent.
3. Find $P(X+Y>0|2X-Y=0)$.

Solution

1. Since X and Y are jointly normal, the random variable $U = X+Y$ is normal. We have

$$EU = EX + EY = -1,$$

$$\begin{aligned} Var(U) &= Var(X) + Var(Y) + 2\text{Cov}(X,Y) \\ &= 1 + 4 + 2\sigma_X\sigma_Y\rho(X,Y) \\ &= 5 - 2 \times 1 \times 2 \times \frac{1}{2} \\ &= 3. \end{aligned}$$

Thus, $U \sim N(-1,3)$. Therefore,

$$P(U>0) = 1 - \Phi\left(\frac{0-(-1)}{\sqrt{3}}\right) = 1 - \Phi\left(\frac{1}{\sqrt{3}}\right) = 0.2819$$

2. Note that $aX+Y$ and $X+2Y$ are jointly normal. Thus, for them, independence is

Let X be the height of the father, Y the height of the son, in a sample of father-son pairs. Assume X and Y bivariate normal, as found by Karl Pearson around 1900. Assume $E(X) = 68$ (inches), $E(Y) = 69$, $\sigma_X = \sigma_Y = 2, \rho = .5$. (We expect ρ to be positive because on the average, the taller the father, the taller the son.)

Given $X = 80$ (6 feet 8 inches), Y is normal with mean

$$\mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X) = 69 + .5(80 - 68) = 75$$

which is 6 feet 3 inches. The variance of Y given $X = 80$ is

$$\sigma_Y^2(1 - \rho^2) = 4(3/4) = 3.$$

Thus the son will tend to be of above average height, but not as tall as the father. This phenomenon is often called *regression*, and the line $y = \mu_Y + (\rho\sigma_Y/\sigma_X)(x - \mu_X)$ is called the *line of regression* or the *regression line*.

It determines the joint distribution of X and Y uniquely and it also yields the moments:

$\frac{\partial^{m+n}}{(\partial t_1)^m(\partial t_2)^n} \mathrm{M}(t_1, t_2)|t_1{=}t_2{=}0 \quad = \mathrm{E}(\mathrm{X}^m \mathrm{Y}^n).$

$$\frac{\partial M(0,0)}{\partial t_1} = E(X) \qquad \frac{\partial M(0,0)}{\partial t_2} = E(Y)$$

$$\frac{\partial^2 M(0,0)}{\partial t_1^2} = E(X^2) \qquad \frac{\partial^2 M(0,0)}{\partial t_2^2} = E(Y^2)$$

$$\frac{\partial^2 M(0,0)}{\partial t_1 \partial t_2} = E(XY)$$

$$\frac{\partial M(t_1,t_2)}{\partial t_1} = \frac{\partial}{\partial t_1} \exp[\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2 t_1^2 + 2\rho\, \sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2}{2}]$$

$$= \exp[\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2 t_1^2 + 2\rho\, \sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2}{2}] \, [\mu_1 + \frac{2t_1\sigma_1^2 + 2\rho\sigma_1\sigma_2 t_2}{2}]$$

$$\frac{\partial M(0,0)}{\partial t_1} = E(X) = \mu_1$$

Similarly, $\frac{\partial M(0,0)}{\partial t_2} = E(Y) = \mu_2$

$$\frac{\partial^2 M(t_1,t_2)}{\partial t_1^2} = \frac{\partial}{\partial t_1} \exp[\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2\, t_1^2 + 2\rho\, \sigma_1\sigma_2 t_1 t_2 + \sigma_2^2\, t_2^2}{2}\,]\,[\mu_1 + \frac{2t_1\sigma_1^2 + 2\rho\sigma_1\sigma_2 t_2}{2}]$$

$$= \exp[\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2\, t_1^2 + 2\rho\, \sigma_1\sigma_2 t_1 t_2 + \sigma_2^2\, t_2^2}{2}\,][\sigma_1^2]$$

$$+ [\mu_1 + \frac{2t_1\sigma_1^2 + 2\rho\sigma_1\sigma_2 t_2}{2}] \exp[\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2\, t_1^2 + 2\rho\, \sigma_1\sigma_2 t_1 t_2 + \sigma_2^2\, t_2^2}{2}\,]\,[\mu_1 + \frac{2t_1\sigma_1^2 + 2\rho\sigma_1\sigma_2 t_2}{2}]$$

$$\frac{\partial^2 M(0,0)}{\partial t_1^2} = E(X^2) \quad = \sigma_1^2 + \mu_1^2$$

$\mathrm{V(X)} = E(X^2) - \big(E(X)\big)^2 = \sigma_1^2 + \mu_1^2 - \ \mu_1^2 = \sigma_1^2$

Similarly, $\mathrm{E(Y)} = \mu_2 \ and \ V(Y) = \sigma_2^2$

$$\frac{\partial^2 M(t_1,t_2)}{\partial t_1 \partial t_2} =$$

$$\frac{\partial}{\partial t_2} \exp[\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2\, t_1^2 + 2\rho\, \sigma_1\sigma_2 t_1 t_2 + \sigma_2^2\, t_2^2}{2}\,]\,[\mu_1 + \frac{2t_1\sigma_1^2 + 2\rho\sigma_1\sigma_2 t_2}{2}]$$

$$= \exp[\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2\, t_1^2 + 2\rho\, \sigma_1\sigma_2 t_1 t_2 + \sigma_2^2\, t_2^2}{2}\,][\,\rho\sigma_1\sigma_2]$$

$$+ \ [\mu_1 + \frac{2t_1\sigma_1^2 + 2\rho\sigma_1\sigma_2 t_2}{2}] \exp[\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2\, t_1^2 + 2\rho\, \sigma_1\sigma_2 t_1 t_2 + \sigma_2^2\, t_2^2}{2}\,][\mu_2 + \frac{2t_2\sigma_2^2 + 2\rho\sigma_1\sigma_2 t_1}{2}]$$

$$\frac{\partial^2 M(0,0)}{\partial t_1 \partial t_2} = E(XY) = \rho\sigma_1\sigma_2 + \mu_1\mu_2$$

$\text{Cov}(X, Y) = E(XY) - (EX)(EY) = \rho\sigma_1\sigma_2 + \mu_1\mu_2 - \mu_1\mu_2 = \rho\sigma_1\sigma_2$

and therefore $\rho(X, Y) = \frac{\text{Cov}(x,y)}{\sigma_1\sigma_2} = \rho$

Thus, we have: $\text{Cov}(X, Y) = \rho\sigma_1\sigma_2$ and $\rho(X, Y) = \rho$.

Let X and Y be two jointly continuous random variables with joint PDF

$$f_{XY}(x,y) = \begin{cases} 2 & y+x \leq 1, x > 0, y > 0 \\ \\ 0 & \text{otherwise} \end{cases}$$

Find $\text{Cov}(X,Y)$ and $\rho(X,Y)$.

Solution

For $0 \leq x \leq 1$, we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dy \\ &= \int_0^{1-x} 2dy \\ &= 2(1-x). \end{aligned}$$

Thus,

$$f_X(x) = \begin{cases} 2(1-x) & 0 \leq x \leq 1 \end{cases}$$

Similarly, we obtain

$$f_Y(y) = \begin{cases} 2(1-y) & 0 \leq y \leq 1 \\ \\ 0 & \text{otherwise} \end{cases}$$

Thus, we have

$$\begin{aligned} EX &= \int_0^1 2x(1-x)dx \\ &= \frac{1}{3} = EY, \end{aligned}$$

$$\begin{aligned} EX^2 &= \int_0^1 2x^2(1-x)dx \\ &= \frac{1}{6} = EY^2. \end{aligned}$$

Thus,

$$Var(X) = Var(Y) = \frac{1}{18}.$$

We also have

$$\begin{aligned} EXY &= \int_0^1 \int_0^{1-x} 2xydydx \\ &= \int_0^1 x(1-x)^2 dx \\ &= \frac{1}{12}. \end{aligned}$$

Now, we can find $\mathrm{Cov}(X,Y)$ and $\rho(X,Y)$:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= EXY - EXEY \\ &= \frac{1}{12} - \left(\frac{1}{3}\right)^2 \\ &= -\frac{1}{36}, \end{aligned}$$

$$\begin{aligned} \rho(X,Y) &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{Var(X)Var(Y)}} \\ &= -\frac{1}{2}. \end{aligned}$$

Correlation is a measure of the strength of relationship between random variables. The population correlation between two variables X and Y is defined as:

$$\rho(X, Y) = \text{Covariance}(X, Y) / \{\text{Variance}(X) * \text{Variance}(Y)\}^{1/2}$$

ρ is called the Product Moment Correlation Coefficient or simply the Correlation Coefficient. It is a number that summarizes the direction and closeness of linear relations between two variables. The sample value is called *r*, and the population value is called ρ (rho). The correlation coefficient can take values between -1 through 0 to +1. The sign (+ or -) of the correlation defines the direction of the relationship. When the correlation is positive ($r > 0$), it means that as the value of one variable increases, so does the other. For example, as the dose amount of an oncology medicine increases, so does the survival time, in a certain range. If a correlation is negative ($r < 0$),, it indicates that when one variable **increases**, the other variable **decreases**. This means there is an inverse relationship between the two variables. For example, as the dose amount of an anti-hypertensive medicine increases, the diastolic blood pressure decreases.

The formula for the population Pearson product-moment correlation, denoted by ρ_{xy}, is

The formula for the population Pearson product-moment correlation, denoted by ρ_{xy}, is

$$\rho_{xy} = \frac{\operatorname{Cov}(x, y)}{\sqrt{V(x)V(y)}} = \frac{\mathrm{E}(\,(x - \mathrm{E}(x))(y - \mathrm{E}(y))\,)}{\sqrt{\mathrm{E}(x - \mathrm{E}(x))^2\,\mathrm{E}(y - \mathrm{E}(y))^2}}$$

The formula for the sample Pearson product-moment correlation is

$$r_{xy} = \frac{\sum_i (\,(x_i - \bar{x})(y_i - \bar{y})\,)}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Testing using Student's *t*-distribution

For pairs from an uncorrelated bivariate normal distribution, the sampling distribution of a certain function of Pearson's correlation coefficient follows

Student's *t*-distribution with degrees of freedom $n - 2$. Specifically, for a bivariate normal distribution, the variable

$t = r\sqrt{\frac{n-2}{1-r^2}}$ has a student's *t*-distribution under the null Hypothesis

H_0: $\rho = 0$ (zero correlation).

If the value of t is greater than critical value we reject H_0 at given level of significance and conclude that $\rho \neq 0$This holds approximately in case of non-normal observed values if sample sizes are large enough. For determining the critical values for *r* the inverse function is needed:

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

Alternatively, large sample, asymptotic approaches can be used.

The **Fisher Z-Transformation** is a way to transform the sampling distribution of Pearson's r (i.e. the correlation coefficient) so that it becomes normally distributed. The "z" in Fisher Z stands for a z-score. ... Fisher's z' is used to find confidence intervals for both r and differences between correlations.

Fisher's z-transformation of *r* is defined as

$$Z = Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

where "ln" is the natural logarithm function

If (*X*, *Y*) has a bivariate normal distribution with correlation ρ and the pairs (X_i, Y_i) are independent and identically distributed, then *z* is approximately normally distributed with mean

$Z_\rho = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ and standard error $\frac{1}{\sqrt{N-3}}$

where *N* is the sample size, and ρ is the true correlation coefficient.

The applications of Fisher *Z* transformation are provided below:

1. To test whether a provided value is equal to the population correlation or not.
2. To test the equality of two population correlations.
3. To combine correlation estimates obtained from different samples.

Significance Testing of Correlation Coefficients:

- Inference about a population correlation coefficient:
  - Testing H_0: $\rho = 0$ or some specific value
  - Testing H_0: $\rho \neq 0$ for two or more correlations based on the same sample

- H_0: ρ = some specified value
- H_1: $\rho \neq$ some specified value (< or > than some specified value)

Method of Testing: Transform sample and population correlation coefficients to Z_r and Z_ρ

- Calculate $Z_{observed} = \frac{Z_r - Z_\rho}{\frac{1}{\sqrt{N-3}}}$ using the formula.
- Test against $z_{critical}$ (determined from table for chosen level of significance)

Confidence Interval:

Under H_0: $\rho = 0$, $Z_r \sim N(Z_\rho, \frac{1}{N-3})$

$\Rightarrow \frac{Z_r - Z_\rho}{\frac{1}{\sqrt{N-3}}} \sim N(0,1)$

95% confidence interval for ρ on the basis of the sample is

$| Z_r - Z_\rho | < 1.96 \frac{1}{\sqrt{N-3}}$

$\Rightarrow -1.96 \frac{1}{\sqrt{N-3}} < Z_r - Z_\rho < +1.96 \frac{1}{\sqrt{N-3}}$

$\Rightarrow - Z_r - 1.96 \frac{1}{\sqrt{N-3}} < Z_\rho < - Z_r + 1.96 \frac{1}{\sqrt{N-3}}$

$$\Rightarrow Z_r - 1.96\frac{1}{\sqrt{N-3}} < Z_\rho < Z_r + 1.96\frac{1}{\sqrt{N-3}}$$

$$\Rightarrow Z_r - 1.96\frac{1}{\sqrt{N-3}} < \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) < Z_r + 1.96\frac{1}{\sqrt{N-3}}$$

$$\Rightarrow 2(Z_r - 1.96\frac{1}{\sqrt{N-3}}) < \ln\left(\frac{1+\rho}{1-\rho}\right) < 2(Z_r + 1.96\frac{1}{\sqrt{N-3}})$$

$$\Rightarrow exp\,[2(Z_r - 1.96\frac{1}{\sqrt{N-3}})] < \frac{1+\rho}{1-\rho} < \exp[2(Z_r + 1.96\frac{1}{\sqrt{N-3}})]$$

$$\frac{exp\,[2(Z_r - 1.96\frac{1}{\sqrt{N-3}})]-1}{exp\,[2(Z_r - 1.96\frac{1}{\sqrt{N-3}})]+1} < \rho < \frac{exp\,[2(Z_r + 1.96\frac{1}{\sqrt{N-3}})]-1}{exp\,[2(Z_r + 1.96\frac{1}{\sqrt{N-3}})]+1}$$

- Inference about a difference between population correlation coefficients
  - Testing H_0: $\rho_1 - \rho_2 = 0$ (or $\rho_1 = \rho_2$)

Method of Testing: Transform sample and population correlation coefficients to Z_r and Z_ρ

Let $Zr_1 = \frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right)$ $\quad$ $Zr_2 = \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right)$

- Under H_0, $\frac{Zr_1 - Zr_1}{\sqrt{(\frac{1}{n_1-3}+\frac{1}{n_2-3})}} \sim N(0, 1)$
- Test against $z_{critical}$ (determined from table for chosen level of significance)

Example:

Study of relationship between achievement motivation and performance in school (grade point average). Theory and prior research suggests that the correlation between these two variables is positive and moderately high (.50)

- The observed correlation in this study was .75 based on N=63

- H_0: $\rho = .50$
- H_1: $\rho \neq .50$

- Level of Significance: .05
- Verify Assumptions
  - Independence of score pairs
  - Bivariate Normality
  - n >= 30
- Find Fisher Z transformation for r_{xy} and ρ_{xy} (from a Table I)
  - r = .75 so Z_r = .973
  - ρ = .50 so Z_ρ = .549
- Set up $Z_{robserved}$: Z_r-Z_ρ/s_Z to get distance of Z_r from Z_ρ in standard error points
- Computation formula for $Z_{robserved}$:
  - (Z_r-Z_ρ) (sqrt n-3) =
  - (.973 - .549)/7.75 =
  - (.424)(7.75) = 3.29
- Find $z_{critical}$ (from table or memory) = 1.96
- Decision Rule:
  - Reject H_0 if absolute value of $z_{robserved}$ >= 1.96 (3.29 is greater than 1.96)
  - Do not reject H_0 if absolute value of $z_{robserved}$ < 1.96
- Conclusion: the relationship between achievement motivation and school performance (grade point average) is greater than the specified value of .50

Select all of the following choices that are possible confidence intervals on the population value of Pearson's correlation:

- ☐ (-0.4, 0.6)
- ☐ (0.3, 0.5)
- ☐ (-0.85, -0.47)
- ☐ (0.72, 1.2)

A sample of 28 was taken from a population, and r = .45. What is the 95% confidence interval for the population correlation?

- ○ (.058, .842)
- ○ (.093, .877)
- ○ (.058, .687)
- ○ (.093, .705)

\

The sample correlation is -0.8. If the sample size was 40, then the 99% confidence interval states that the population correlation lies between -.909 and _____