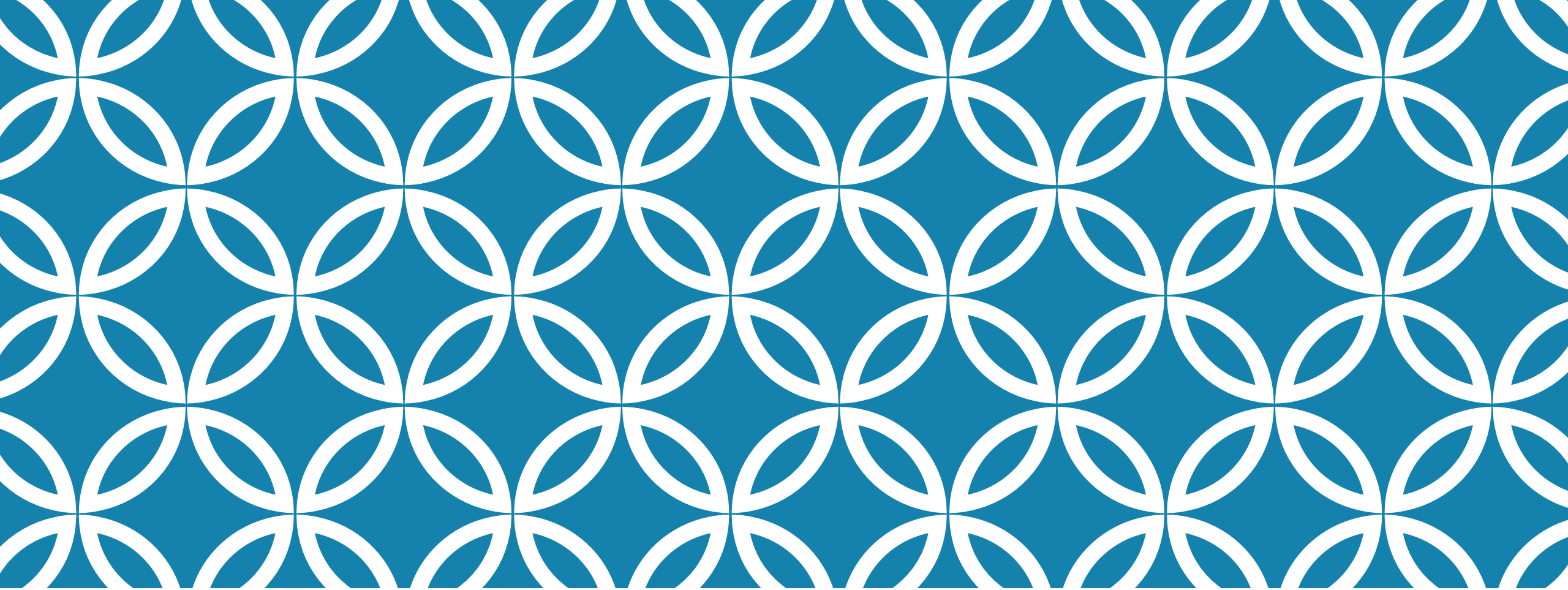


DEPENDENCY STRUCTURE-HYDROINFORMATION

INTRODUCTION

- Rainfall – f (climate variables, surface variables, topography, vegetation, etc.)
- Runoff – f (Rainfall, Evapotranspiration, climate variables, soil properties, vegetation, basin properties, etc.)
- Dissolved Oxygen – f (streamflow, water temperature, pollutant concentration, etc.)
- Drought – f (Rainfall, Streamflow, soil moisture)

INTRODUCTION

- Determining the dependency structure between a set of variables
- Hydrological variables:
  - Rainfall – Streamflow
  - Air temperature – Water temperature
  - Rainfall – Drought
  - Predictions and forecasting

INTRODUCTION

- Quantification of dependence structure is vital to understand the probability of occurrence of joint occurrence of events
- Spatial dependence structure is required to understand the variation in occurrence of events
- Various approaches to estimate the dependence measures

INTRODUCTION

- Dependence – Conceptual relationship between two random variables or two data sets
- Dependence – Statistical relationship between two random variables or two data sets
- Correlation – statistical relationships involving dependence
- “dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence”

DEPENDENCE MEASURES – PEARSON'S CORRELATION COEFFICIENT

- One of the primary measure of dependence is the Pearson's product moment correlation coefficient
- It is a linear dependence related to observations at pairwise points
- It is a measure of the distance between two patterns
- Measures the extent to which two patterns are similar with each other.

CORRELATION COEFFICIENT

$$\rho_{xy} = \frac{\sigma\, xy}{\sigma\, x \sigma\, y} \qquad \sigma\, xy = E(XY) - \mu_X \mu_Y$$

$\sigma\, xy$ is the covariance between x and y

$\sigma\, x$ is the standard deviation of x

$\sigma\, y$ is the standard deviation of y

μ_x is the mean of x

μ_y is the mean of y

$-1 \le \rho_{xy} \le 1$ Range of correlation coefficient

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$cov_{x,y}$ = covariance between variable x and y

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

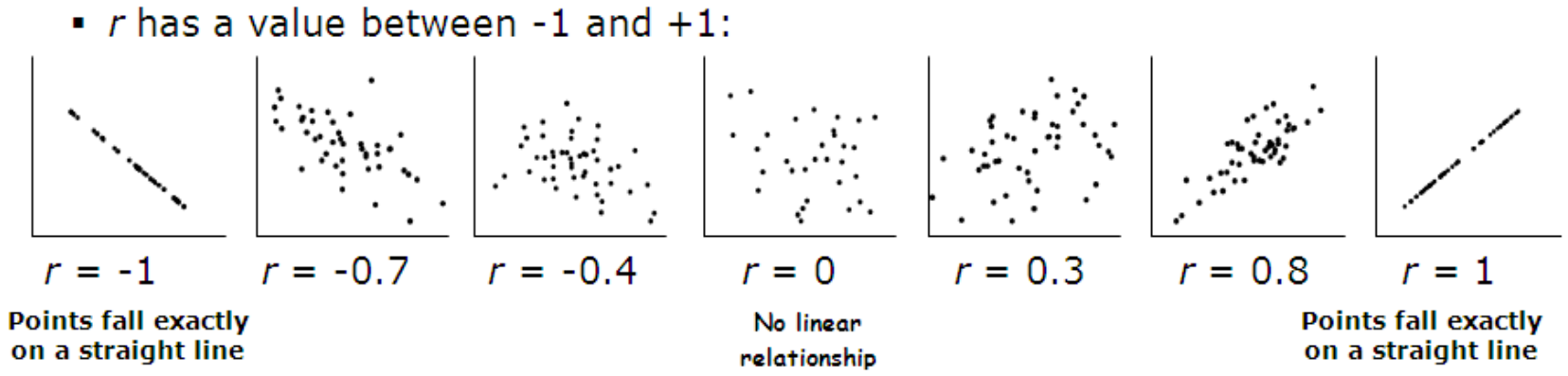
$\bar{y}$ = mean of y

N = number of data values

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

PEARSON CORRELATION COEFFICIENTS

- It represent the strength and nature of linear association between two variables
- Pearson's "r" is dimensionless and ranges between -1 and +1.
- A value of +1 indicates perfect positive association
- A value of -1 indicates perfect negative association
- A value of 0 indicates that there is no association between two variables
- Pearson's r provides an accurate measure of the strength of linear association



ASSUMPTIONS

- Both variables have an approximately Normal distribution
- The true association among the variables X and Y is linear
- Pearson's correlation may not be promising for random variables above a certain extreme threshold

RANK CORRELATION COEFFICIENT

- Spearman's rank correlation coefficient
- Kendall's rank correlation coefficient
- Non-parametric
- They measure the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship
- Rank correlation coefficients reduce the amount of calculation or to make the coefficient less sensitive to non-normality distribution
- If, one variable increases, the other decreases, the rank correlation coefficient will be negative.

SPEARMAN'S RANK CORRELATION COEFFICIENT

- The Spearman rank correlation coefficient, also known as Spearman's rho, is a nonparametric (distribution-free) rank statistic proposed by Spearman in 1904 as a measure of the strength of the associations between two variables
- The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables.
- After raw data, x and y, are converted to ranked data, the Spearman correlation coefficient is defined

SPEARMAN'S RANK CORRELATION COEFFICIENT

The Spearman correlation coefficient is defined as the Pearson correlation coefficient using the rank variables. After raw data, x and y, are converted to ranked data rx and ry, The Spearman correlation coefficient is defined as:

For a sample of size *n*, the *n* raw scores X_i, Y_i are converted to ranks $\operatorname{rg} X_i, \operatorname{rg} Y_i$, and r_s is computed as

$$r_s = \rho_{\operatorname{rg}_X, \operatorname{rg}_Y} = \frac{\operatorname{cov}(\operatorname{rg}_X, \operatorname{rg}_Y)}{\sigma_{\operatorname{rg}_X} \sigma_{\operatorname{rg}_Y}},$$

where

ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,

$\operatorname{cov}(\operatorname{rg}_X, \operatorname{rg}_Y)$ is the covariance of the rank variables,

$\sigma_{\operatorname{rg}_X}$ and $\sigma_{\operatorname{rg}_Y}$ are the standard deviations of the rank variables.

SPEARMAN'S RANK CORRELATION COEFFICIENT

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

ρ= Spearman rank correlation

di= the difference between the ranks of corresponding variables

n= number of observations

This formula is an alternative to Pearson's correlation if the data are ordinal (categorical data with a set order or scale to it) and monotonic (function which is either entirely nonincreasing or nondecreasing) and if there are no ties in data (ties in the data: for instance if you have 1,2,3,3,4 as the dataset then the two 3's are tied data)

KENDALL'S RANK CORRELATION COEFFICIENT

- Non-parametric rank based correlation metric, which is defined based on the number of concordant and discordant pairs in the two hydroclimate extreme events
- Test the similarity in the ordering of data when it is ranked by quantities, instead of observations as the basis
- It uses pairs of observations and determine the strength of association based on the pattern of concordance and discordance between the pairs

KENDALL'S RANK CORRELATION COEFFICIENT

- Let (x1, y1), (x2, y2), …, (xn, yn) be a set of joint observations from two random variables X and Y respectively, such that all the values of (xi) and (yi) are unique.
- Any pair of observations (xi, yi) and (xj, yj) are said to be concordant if the ranks for both elements agree: that is, if both xi > xj and yi > yj or if both xi < xj and yi < yj.
- They are said to be discordant, if xi > xj and yi < yj or if xi < xj and yi > yj.
- If xi = xj or yi = yj, the pair is neither concordant nor discordant.

KENDALL'S CORRELATION COEFFICIENTS

- **Concordant:** Ordered in the same way (consistency). A pair of observations is considered concordant if (x2 — x1) and (y2 — y1) have the same sign.
- **Discordant:** Ordered differently (inconsistency). A pair of observations is considered concordant if (x2 — x1) and (y2 — y1) have opposite signs.

Kendall's Tau = (C – D / C + D)

Where C is the number of concordant pairs and D is the number of discordant pairs.

KENDALL'S CORRELATION COEFFICIENTS

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

The denominator is the total number of pairs, so the coefficient must be in the range $-1 \leq \tau \leq 1$.

HYPOTHESIS TESTING FOR POPULATION CORRELATION COEFFICIENT

Sample correlation coefficient $= r$

Population correlation coefficient $= \rho$

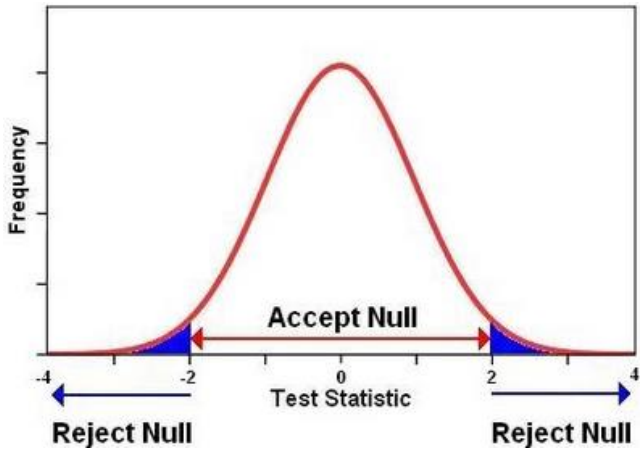
Hypothesis test for the population correlation to learn of a linear association between two variables.

$$\text{Test statistic: } t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

P – Value is estimated using t- distribution with n-2 degrees of freedom.

t- distribution looks almost identical to the normal distribution, used when one has small samples. The larger the sample size, the more the t-distribution looks like the normal distribution.

TESTING THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT



Performing the Hypothesis Test

- Null hypothesis: H_0: $\rho = 0$
- Alternate hypothesis: H_a: $\rho \neq 0$

What the Hypothesis Means in Words:

- **Null hypothesis H_0:** The population correlation coefficient *is not* significantly different from zero. There *is not* a significant linear relationship (correlation) between *x* and *y* in the population.
- **Alternate hypothesis H_a:** The population correlation coefficient *is* significantly different from zero. There *is* a significant linear relationship (correlation) between *x* and *y* in the population.

If the *p*-value is less than the significance level ($\alpha = 0.05$),

- Decision: Reject the null hypothesis.
- Conclusion: There is sufficient evidence to conclude there is a significant linear relationship between *x* and *y* because the correlation coefficient is significantly different from zero.

If the *p*-value is *not* less than the significance level ($\alpha = 0.05$),

- Decision: Do not reject the null hypothesis.
- Conclusion: There is insufficient evidence to conclude there is a significant linear relationship between *x* and *y* because the correlation coefficient is not significantly different from zero.

JOINT PROBABILITY DISTRIBUTIONS

If X and Y are two hydrological variables, the probability distribution for their simultaneous occurrence can be represented by a function called as joint probability distribution of X and Y.

The function, $f(x, y)$ is a joint probability distribution or probability mass function of the discrete random variables X and Y, if

1. $f(x, y) \geq 0$ for all (x, y)

2. $\sum_x \quad \sum_y f(x, y) = 1$

3. $\mathrm{P}(X = x, Y = y) = f(x, y)$

The function, $f(x, y)$ is a joint probability density function of the continuous random variables X and Y, if

1. $f(x, y) \geq 0$ for all (x, y)

2. $\int_{-\infty}^{\infty} \quad \int_{-\infty}^{\infty} . f(x, y)\, dx\, dy = 1$

3. $\mathrm{P}(X, Y) \in A) = \iint_A \quad f(x, y)\mathrm{dx\ dy}$

Ex: A storm event occurring at a point in space is characterized by two variables, namely the duration X of the storm and the depth of rainfall Y. Find the joint pdf of both rainfall duration and depth.

The variables X and Y follow following distribution, respectively:

$$F_x(x) = 1 - e^{-x} \quad x \geq 0$$

$$F_Y(y) = 1 - e^{-2y} \quad y \geq 0$$

The joint CDF of X and Y is following the bivariate distribution:

$$F_{x,y}(x,y) = 1 - e^{-x} - e^{-2y} + e^{-x-2y-xy} \qquad x, y \geq 0$$

Hint: Cumulative distribution function, univariate distribution function, F(x) of a continuous random variable X with probability density function $f(x)$ is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx \quad \text{for } -\infty < x < \infty$$

$$f(x) = \frac{dF(x)}{dx}$$

We know, $f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$

Differentiating the joint CDF w.r.t x, we get

$$\frac{\partial F}{\partial x} = \frac{\partial \left(1 - e^{-x} - e^{-2y} + e^{-x-2y-xy}\right)}{\partial x} = e^{-x} - (1 + y)\, e^{-x-2y-xy}$$

Again differentiating the above equation w.r.t y

$$f_{X,Y}(x, y) = \frac{\partial^2 F}{\partial x \partial y} = \frac{\partial \left(e^{-x} - (1 + y)\, e^{-x-2y-xy}\right)}{\partial y}$$
$$= [(1 + y)(2 + x) - 1]\, e^{-x-2y-xy}$$

Hence, joint *pdf* of X and Y is

$$f_{X,Y}(x, y) = [(1 + y)(2 + x) - 1]\, e^{-x-2y-xy} \qquad x, y \geq 0$$