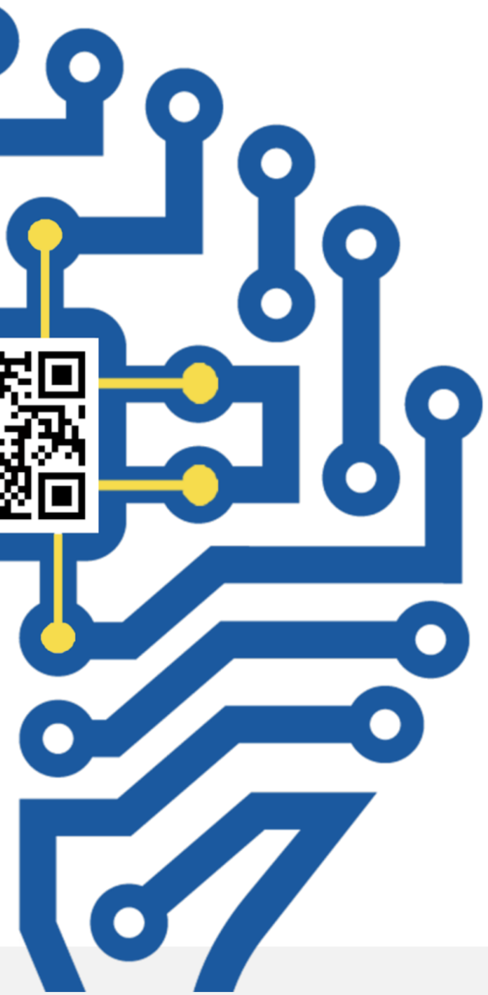
UGANDA CHRISTIAN UNIVERSITY

A Center of Excellence in the Heart of Africa



MSDS Data Science Lifecycle

COVARIANCE AND CORRELATION

Dr. Daphne Nyachaki Bitalo
(PhD Genetics and Bioinformatics)
Department of Computing & Technology
Faculty of Engineering, Design & Technology

25TH February 2023

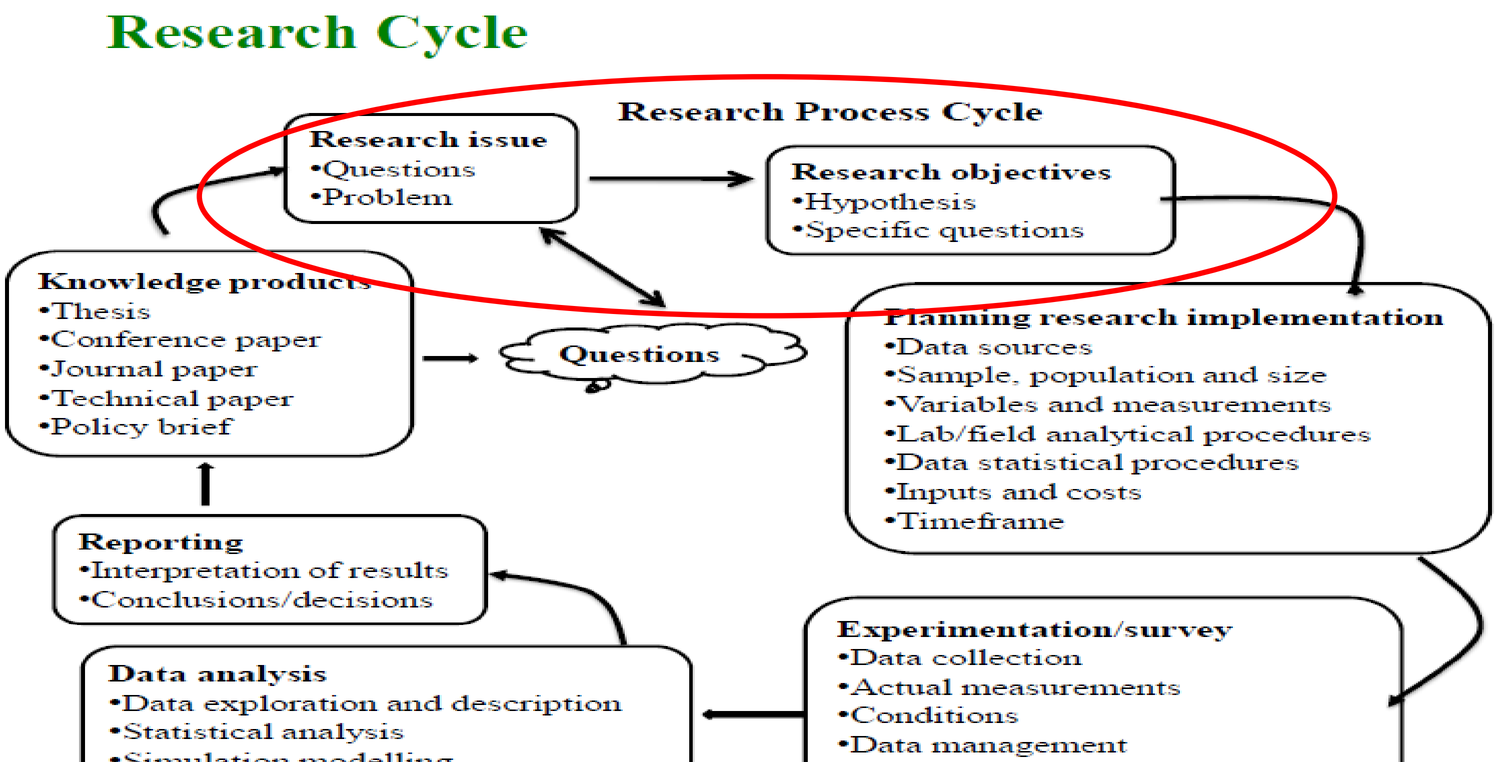
Hypotheses

Hypothesis: A research question posed in sentence format

$H0: \mu 1 = u2$

$Ha: \mu 1 \neq u2$

Null hypothesis: There is a significant difference between A and B
Alternative hypothesis: There is no significant difference between A and B



Research Cycle

Research Process Cycle

Research issue
- Questions
- Problem

Research objectives
- Hypothesis
- Specific questions

Planning research implementation
- Data sources
- Sample, population and size
- Variables and measurements
- Lab/field analytical procedures
- Data statistical procedures
- Inputs and costs
- Timeframe

Experimentation/survey
- Data collection
- Actual measurements
- Conditions
- Data management

Data analysis
- Data exploration and description
- Statistical analysis
- Simulation modelling

Reporting
- Interpretation of results
- Conclusions/decisions

Knowledge products
- Thesis
- Conference paper
- Journal paper
- Technical paper
- Policy brief

Questions

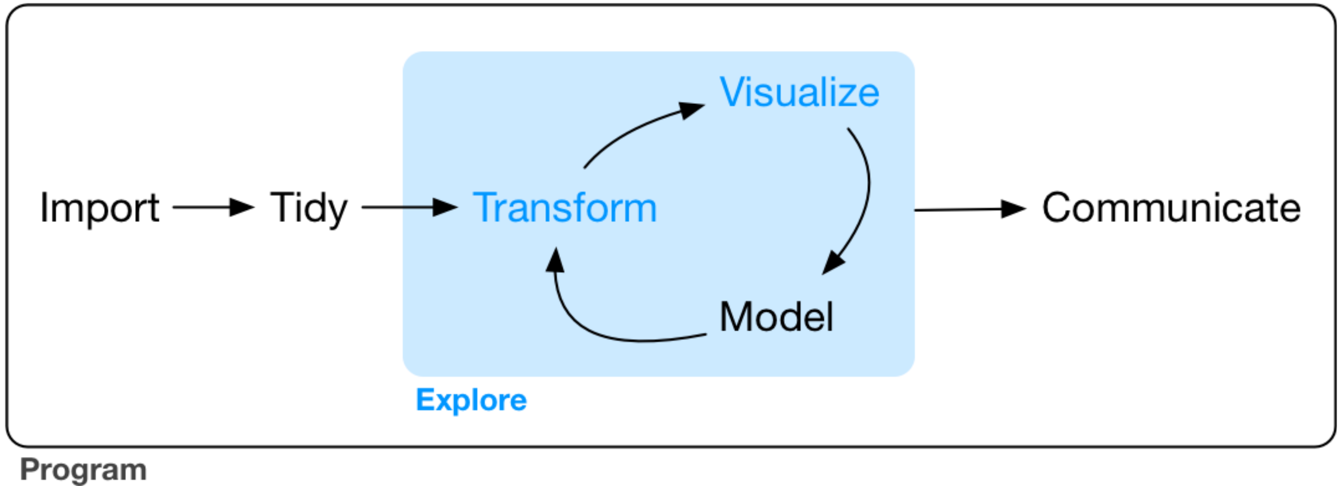
Exploratory Data Analysis (EDA)

An interactive cycle;

1. Generate questions about your data.
2. Search for answers by visualizing, transforming, and modeling your data.
3. Use what you learn to refine your questions and/or generate new questions.

There is no rule about which questions you should ask to guide your research. However, two types of questions will always be useful for making discoveries within your data. You can loosely word these questions as:

1. What type of variation occurs within my variables?
2. What type of covariation occurs between my variables?



Statistical significance

1. p-value adds statistical credence to the tests for null/alternative hypotheses stated
2. Works for normally distributed data (parametric testing)
3. t-test for continuous variables/data
4. chi-squared test for categorical/qualitative data
5. ANOVA for more in-depth analyses
6. Stats range from 0-1
7. Thresholds usually at 0.05
8. p =<0.05 reject null hypothesis
9. p>= 0.05 fail to reject the null hypothesis

T-test

1. p-value adds statistical credence to the tests for null/alternative hypotheses stated
2. Works for normally distributed data (parametric testing)
3. t-test for continuous variables/data
4. chi-squared test for categorical/qualitative data

Covariance

1 SAMUEL 7:12

1. Shows how two variables vary/differ
2. Used in regression analyses
3. Measures change in one variable associated to change in another variable
4. Ranges from -∞ to ∞. Higher values indicates stronger change association between variables

Correlation

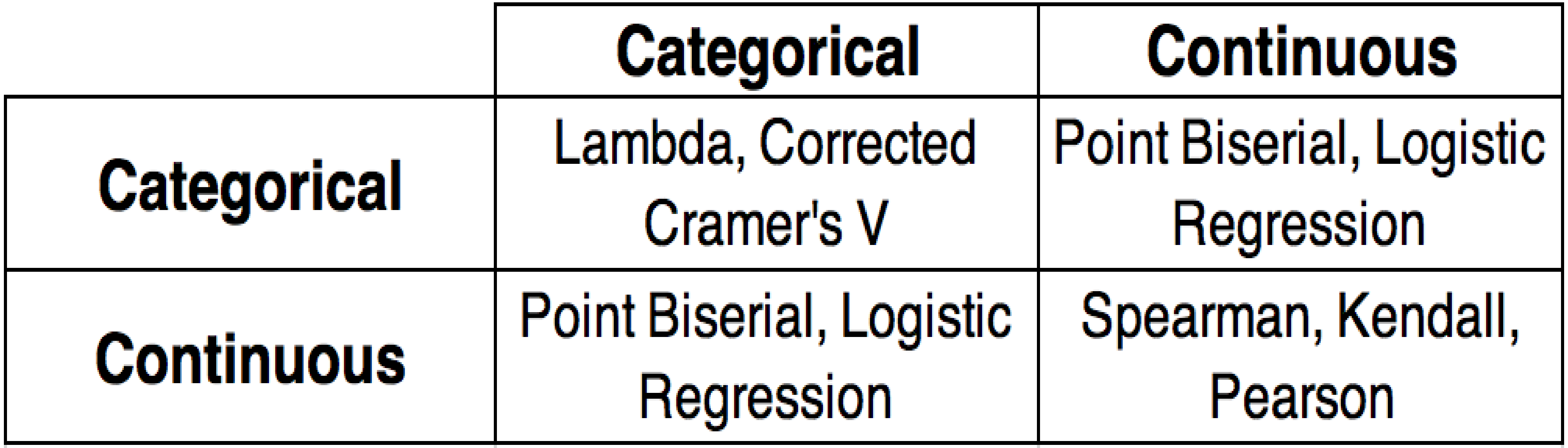
1. Shows relationship between variables
2. Used in regression analyses
3. Standardizes covariance on a scale of -1 to $+1$

$r<0.3$, weak correlation

$0.3<r<0.7$, moderate correlation

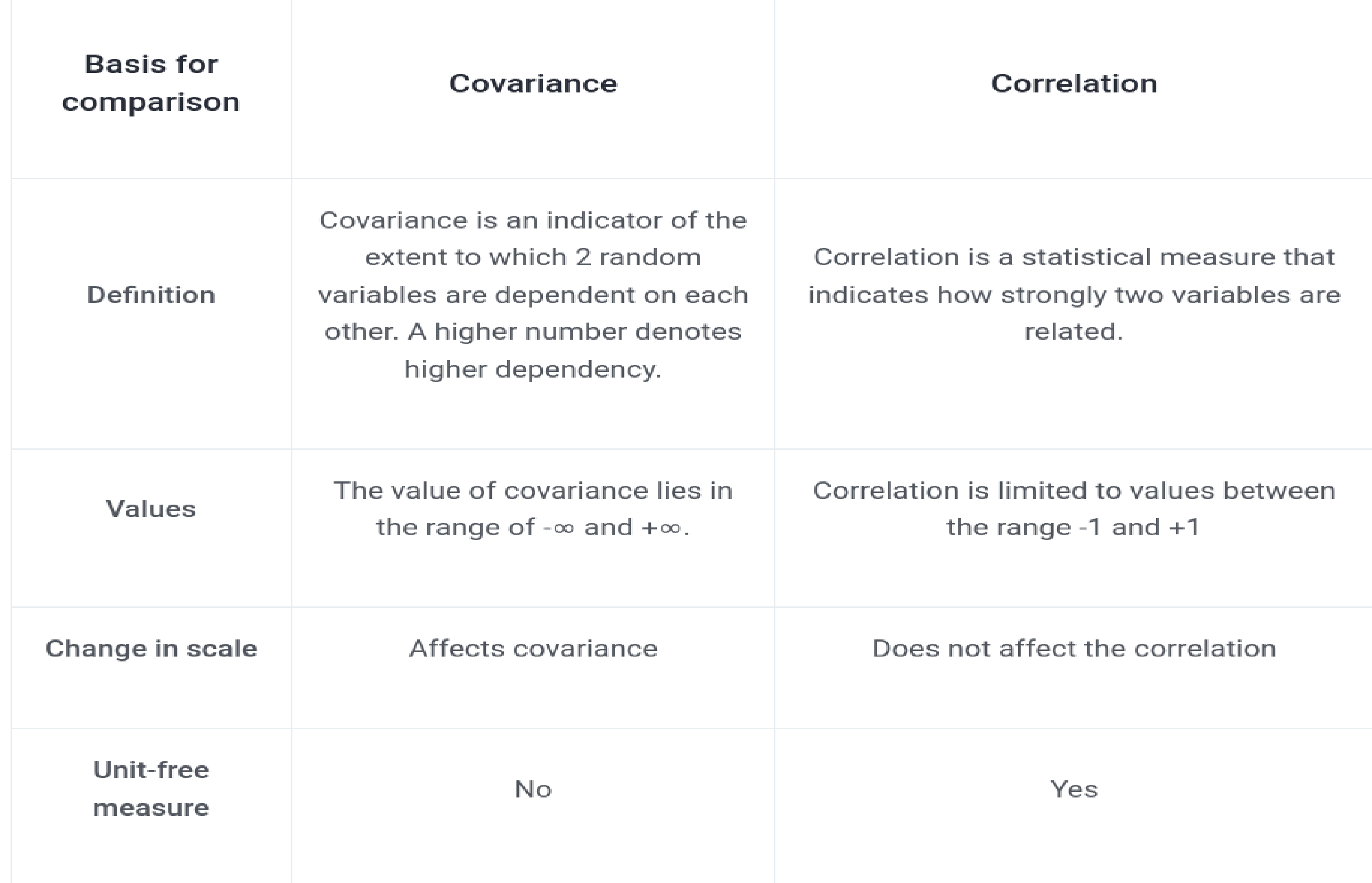
$r>0.7$, high correlation

Correlation Tests Between Variables



	Categorical	Continuous
Categorical	Lambda, Corrected Cramer's V	Point Biserial, Logistic Regression
Continuous	Point Biserial, Logistic Regression	Spearman, Kendall, Pearson

Differences between covariance and correlation



Basis for comparison	Covariance	Correlation
Definition	Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.	Correlation is a statistical measure that indicates how strongly two variables are related.
Values	The value of covariance lies in the range of -∞ and +∞.	Correlation is limited to values between the range -1 and +1
Change in scale	Affects covariance	Does not affect the correlation
Unit-free measure	No	Yes

1 SAMUEL 7:12

Questions ????