UNIVERSITY OF CAPE COAST

Cape Coast, Ghana

ENHANCING CAPACITY FOR POSTGRADUATE RESEARCH
BY
SCHOOL OF GRADUATE STUDIES

IN COLLABORATION WITH

COLLEGE OF EDUCATION STUDIES
COLLEGE OF AGRICULTURE AND NATURAL SCIENCES
COLLEGE OF HEALTH AND ALLIED SCIENCES
GRADUATE STUDENTS ASSOCIATION OF GHANA-UCC

STATISTICAL TOOLS, UNDERLYING ASSUMPTIONS AND PRESENTATION

BY

FRANCIS ENU-KWESI

November 21, 2017

OUTLINE OF THE PRESENTATION

- PRELIMINARIES
- DATA
- GENERAL REQUIREMENTS
- DESCRIPTIVE STATISTICS
- COMPARISONS
- RELATIONSHIPS
- SUPERVISORS AND EXAMINERS

- Theoretical and conceptual framework
- Data needs
- Sampling procedures
- Instruments for primary data collection

- Secondary data
  - Specifics-
    - demographic; health; agriculture; education; environment; climate; soils;
  - Sources
    - government documents; official statistics; technical reports; reference books; research institutions; universities; libraries; library search engines; computerized databases; international financial statistics; world development indicators; FAO, WHO etc…
  - Quality
    - check the original purpose and the potential bias; currency of the data;
    - Credentials of the sources; explanation of the methodology; references provided
    - Intended audience; experts' acceptance; disaggregated or composites
    - Agreement with other data sources

Primary data

- Issues
- Measurement scale
- Methods and instruments
- Data management
  - Editing, coding, inputting, cleaning
  - Creating composite variables/combining variables
  - Quantifying qualities

- Cautions
  - Indicators may represent different ideas and forcing them into one concept may not be correct
  - Using several items with different types of responses to represent one concept
  - Using a scale that is practically not a scale
  - Combining things that are heterogeneous by their measurement units, quantum meaning, and distributional qualities

- Examining the data
  - Distribution [defines the summary measures that should be used]
  - Type of variables
    - Scale
    - Restricted [limited] measurement ; unrestricted measurement
  - Need for reliability analysis

- Checking some common assumptions
  - Normality
    - Shapiro-Wilk statistic [preferred for N<50, but can be used for N up to 2000].....p-value<0.05 is an indication of normality violation.
    - Kolmogorov-Smirnov (K-S) -test statistic – Z-stat with p-value. If p-value<0.05, assumption is not met.
    - Skewness ±2 If sk <±2, then normality is met
    - Kurtosis <±7, normality is met
    - Graphical – Normal Q-Q Plots – points must lie close to diagonal.
      - Points departing from diagonal – an indication of violation of normality assumption

- Equality [homogeneity] of variance
  - Levene's test (F-test) for each continuous dependent/test variable. If p-value <0.05, then assumption is not met
  - Hartley's F-maximum test
  - Bartlett's test
  - Brown-Forsythe test
  - Plotting means and variances (scatter diagram)
- Linearity-examine scatter plots

- Data Transformation
  - Skewed continuous data – log; Box-Cox
  - Count data – Freeman Tukey square root
  - Proportions – Freeman Tukey arcsine [Poisson distributions]
  - Binary data – logit transformation
- Transformations
  - can be variance stabilizing [creates homogeneity of variance] and enable comparison of multiple samples
- Back-Transformation
  - always back-transform the results so that they are presented in the same units as the original data
  - Enables easier interpretation

- Choice of technique
  - What questions/objectives are you addressing? Which items address the objectives?
  - What are their scales of measurement? What is the nature of these items (their distributions)?
  - What is your understanding of the relevant technique?
- Knowledge of/Familiarity with Statistical Software
  - GenStat, Stata, R, SAS, SPSS, Eviews, Minitab, LISREL, Matlab; Statistica
  - Statistical Package for Agricultural Research; Statistical Package for Augmented Designs
  - Statistical Package for Factorial Experiments;
  - Statistical Package for Balanced Incomplete Block Designs
  - Statistical Package for Animal Breeding

Measurement scale	Central tendency	Measure of dispersion
Interval/ratio	Mean (arithmetic, harmonic, geometric) Trimmed mean–discarding extreme values Winsorized mean–replacing extreme values	Standard deviation, Coefficient of variation
Ordinal	Median	Quartile deviation
Nominal	Mode	Index of qualitative variation

Other outputs– range, mid-range; percentiles, Z-scores, cumulative frequencies/percentages

Discrete data – proportions

Categorical/nominal data – prevalence or risk

Count data – rate per space (Poisson issues)

All the above are for univariate analysis–one variable

- Theoretical basis; Empirical basis, and specification of alpha

Parametric tests	Non-Parametric tests
One sample t-test	Kolmogorov-Smirnov test for one sample (compares observed values for an ordinary scaled variable to some theoretical distribution-normal, uniform,Poisson) One sample chi-sq (uniform distribution)- (for nominal scaled variables
Two independent samples t-test	Mann-Whitney U-test K-S test for two unrelated samples- (distribution of values/frequencies in two groups, eg men/women to see if their ratings differ)
Multiple independent samples -ANOVA, ANCOVA	Kruskal-Wallis test, Median test
Paired samples t-test	Wilcoxon Signed Rank test, McNemar's test - cross-tab (categorical variables

RELATIONSHIPS

- Interval/Ratio scale

Applicability	Technique/coefficient
Continuous linear relationships	Pearson product moment correlation coefficient
One continuous, one dichotomous variable with an underlying normal distribution	Bi-serial correlation
Three + variables, relating some, with others' effects taken out	Partial correlation
Three +, relating all pairwise	Multiple correlation
2 constructs with latent variables (relationships between 2 multivariate sets of variables, measured on the same respondents)-canonical correlation	MANOVA (With) Correlation - canonical correlation

RELATIONSHIPS

- Non-Parametric

Measurement scale	Technique/coefficient
Ordinal (all the coefficients are from -1 to 1). They allow strength and direction of association, and are based concordant-discordant pairs.	1. Gamma (Goodman and Kruskal's) 2. Kendall's tau b (adjustment for tied ranks) 3. Kendall's tau c (adjustment for table dimensions) 4. Spearman's Rho
Nominal (strength of association) Flexible data and distribution assumptions How well the frequencies of one nominal variable offer predictive evidence about the frequencies of another. Proportional reduction in error-based interpr. Agreement measure (2 cat vars with equal resp categs)	1. Chi-sq based 2x2 tables Phi 0 to 1 2. Chi-sq dimension >2 Cramer's V 0 to 1 3. Chi-sq Contingency Coeff - varying Upper limit 1. Lambda (allows calculation for the direction of prediction 2. Goodman and Kruskal's tau (table marginals)

RELATIONSHIPS/DEPENDENCY TECHNIQUES

Dependent variable	Independent variables	Technique
Interval/ratio --1	Interval/ratio	Multiple regression
Limited – binary --1	Interval/ratio	Discriminant analysis
Limited – binary, multinomial -1, prevalence Count – rate per space	Varied scales	Logistic regression Poisson regression

Assumptions

1. Large samples (at least N> 50+8m) with probabilistic selection
2. Avoidance of multicollinearity (low Tolerance--<0.1, High VIF-->10; VIF = 1/Tolerance
3. Avoidance of Singularity
4. Dealing with Outliers and Influential points
5. Normality- residuals should be normally distributed about the predicted DV scores
6. Linearity – residuals should have a straight line relationship with predicted DV scores
7. Homoscedasticity – variance of the residuals about the predicted DV scores should be the same for all predicted scores

RELATIONSHIPS/DEPENDENCY TECHNIQUES

Dependent variable	Independent variables	Technique (Multivariate General Linear Models)
Interval/ratio – several	Several factors/classification variables	MANOVA, MANCOVA (bundles the continuous dependent variables into a weighted linear combination or composite variable) Statistical Tests: Hotelling's T square test; Wilks' lambda U Test Pillai's Trace Test

Additional Assumptions

1. Continuous dependent variables with moderate correlations; Categorical independent variables
2. Covariates related to dependent variable, but measured without error
3. Covariates can be dichotomous, ordinal, or continuous
4. Normality- residuals should be randomly distributed about the predicted DV scores
5. No outliers – MANCOVA is highly sensitive to outliers in the COVARIATES

RELATIONSHIPS/DEPENDENCY TECHNIQUES

- SEM- Dealing with "causality" among constructs that cannot be directly measured (combines MR and FA)
  - Derivation of a measurement model
  - Structuraling the model to show the causal relationships among the latent variables -path analysis
  - Using a set of linear structural equations to describe the model
- Time series, Panel data
  - Co-integration techniques
  - Autoregressive models
  - Panel regression

Survival Analysis – estimating the duration until event occurs.

- Event can be "cure', "recovery", "death", "divorce", "drought" ……
- We estimate risk of event at T(t)
- Procedures : Kaplan-Meier estimator (non-parametric estimator of the survival function)
- : Cox-proportional hazards model (Cox-Regression)- semi-parametric; median survival
- : Log-rank test (comparing two groups)

RELATIONSHIPS/INTERDEPENDENCY TECHNIQUES

Deals with several variables, with none viewed as dependent on another

- Factor analysis (data reduction)
  - Reducing many variables to a manageable number
  - Variables should have overlapping measurement characteristics
  - Identifying patterns in the reduced variables (extracted factors/components)
  - Using theory, experiences, intuition to thematize the patterns

RELATIONSHIPS/INTERDEPENDENCY TECHNIQUES

- Cluster analysis
  - Technique for grouping similar objects or people
  - Starts with an undifferentiated group (people, events, objects…)– attempt is to organize them into homogeneous subgroups
  - Steps
    - Selecting the sample
    - Defining the variables on which to measure the people, objects…
    - Computing the similarities, using correlation, Euclidean distances…
    - Selection of mutually exclusive clusters (maximized within cluster similarities, and between cluster differences) or hierarchically arranged clusters
    - Cluster comparison and validation

SUPERVISORS/EXAMINERS

- Supervisors
  - Consider students' ability when advising on the best technique
  - Timeliness of the advice
  - Seeking other opinions
  - Directing students to seek advice
  - Do not assume that the students know anything about the techniques that they want to use
  - Get students to explain their project and why they think that the technique is suitable

SUPERVISORS/EXAMINERS

- Examiners
  - Look out for empirical justification of the techniques
  - Applicability of the technique to the data
  - Theoretical relevance
  - Correct interpretation and reporting, evaluation of models
  - Dealing with transformed data
  - Usage of the conceptual framework
  - Application of theories in the interpretation

THANK YOU