Intro to multivariate analytics

1. What is modeling science?
   1. Development and use of interactive, customizable, decision models for analyzing, planning, and implementing marketing strategies
   2. Utilization of business concepts and theories for interpreting data and making decisions
2. What is modeling and analytics?
   1. Analytics tools for decision making
      1. Break down problem into single and manageable components
      2. Solving each component and identify solutions
      3. Analyze a set of solutions from components to arrive cohesive conclusions
3. Example of analytics and analytical thinking
   1. Profit = revenue – cost
   2. Revenue = price \* sales = p1\*s1 + p2\*s2
   3. Cost = fixed cost + variable costs
   4. Fixed costs = salaries
   5. Variable costs = running costs of operation(eg., travelling costs, etc)

Data source

* **Zero party source**: data is given directly and is proactively shared by the customer for personalization.
* **First party source**: inferred customer data or self-reported. This could typically come from Point of Sales in offline stores, where customers share their name, address, mobile number, etc.
* **Second party source**: another company’s first-party data that can be directly purchased from the source. Typically, it is collected from trusted partners who have agreed to share the data, which might mutually benefit both businesses.
* **Third party source**: This data is any information collected about users that may not be directly related to customers or their transactions. It usually helps to understand the behavior and demographic of customers. But a major disadvantage of third-party data is its statistical and aggregated nature and the fact that it has not been directly received from the user.

What are models

* A model is a stylized representation of reality that is easier to deal than reality itself
* Models allow us to explore and simulate for specific purposes

*= p +*

Statstatistical Model

f(t) f(t) = pdf of adoption at time *t*

F(t) F(t) = cdf of adoption at time *t*

*N(t) N(t) = Total number of people who have adopted product by time t*

*N N = Population size*

*p, q p,q = Constants to be determined. Impact the actual path of the curve*

Multivariate analytics- business

* A closer approximation to model real world scenario than univariate or bivariate methods
* Eg. Differentiate between high risk customers and low risk customers for mortgage loans

Interdependence methods(exploratory, descriptive)

* No distinction between independent and dependent variables
* Analyzing information contained in a large set of variables without modeling causality
* Eg, PCA, FA, clustering

Dependence methods(causal, predictive)

* Modeling how two or more variables are related
* Eg, LDA, Logistic regression, structural equation modeling(SEM), MLR

Variate: a linear combination of all variables

Measurement scale:

* non- metric(nominal/ categorical, ordinal) ---- frequency table
* metric(interval, ratio) – mean, standard deviation

**reliability**: consistency of results

* test- retest
* split- half
* coefficient alpha(0.80 - 0.96: **very good** reliability; 0.70 - 0.80: **good** reliability; 0.60 - 0.70: **fair** reliability; Below 0.60: **poor** reliability)

**validity**: degree to which the test actually measures what it claims to measure

* content validity
* criterion validity
* convergent validity
* discriminant validity
* fit validity
* construct validity

hypothesis

* a statement, made on the basis of limited evidence and often as a starting point for further investigation
* H0: null hypothesis
* H1: alternative hypothesis
* Type I error: false positive
* Type II error false negative

Power of hypothesis testing

Power = 1- type II error

Graphical user interface

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| **No of Population\*** | **Variables** | **Hypothesis** | **Technique** |
| L=1 (e.g., Canadians) | P=1 (e.g., height) | H0: µ=µ0  H1: µ≠µ0 | T-test |
| L=1 (e.g., Canadians) | P=2 (e.g., height, weight) | H0: µ=µ0  H1: µ≠µ0  µ= vector (2x1) | Hoteling T2 |
| L=2 (e.g., Canadians, US) | P=1 (e.g., height) | H0: µCan=µUS  H1: µCan≠µUS | T-test |
| L=2 (e.g., Canadians, US) | P=2 (e.g., height, weight) | H0: µCan=µUS  H1: µCan≠µUS  µCan, µUS each vector (2x1) | Hoteling T2 |
| L ≥ 3 (e.g., Canadians, US, Mexico) | P=1 (e.g., height) | H0: µCan=µUS=µMex  H1: ~H0 | ANOVA |
| L ≥ 3 (e.g., Canadians, US, Mexico) | P=2 (e.g., height, weight) | H0: µCan=µUS=µMex  H1: ~H0  µCan, µUS, µMex each vector (2x1) | MANOVA |

Partial correlations

* Correlation between two variables from which the effects of the rest of the other variables have been removed
* Nth order partial correlation is when n variables are partial out from(y,x1)

**𝛽1 =**

* Limitation:
  + Misinterpretation/ meaningless results
  + Desperation to fit the model(use higher order model)

Semi- partial correlations – when variable is partial out from either y or x1, but not from both

Confounding – when true impact of variable is not measured due to presence of other variables

Suppression – when true impact of a variable is suppressed in absence of other variable

Mediation: effect of an independent variable on y is transmitted both directly and indirectly

Complete mediation: X1 cannot impact y in absence of m1

Moderation: strength of the relationship between y and x1

Linear regression:

* Observations tend to move around population mean(regression towards mean)
* CLT is based on the regression towards mean
* Regression deals with dependence of 1 variable on other variables. It does not imply causation

Correlation

* Strength of relationship between two or more variables
* Corr(x1,x2) = corr(x2,x1)

Regression

* Regression(X1,x2) != regression(x2,x1)

Dependent variable(DV)

* Explained variable
* Predictand
* Regressand
* Response
* Endogenous

Independent variable(IV)

* Explanatory variable
* Predictor
* Control variable
* Exogenous
* Features

Regression assumptions

* Linearity of relationship between(x,y)
* Error terms have equal variance
* Error terms are uncorrelated
* Error terms follow normal distribution

Interpretation of coefficient and error terms

* Omitted variables: e is a substitute for all OTHER variables which can affect DV but not included in the model
* Unavailability of data on some variable leads is to treat them all under e
* Intrinsic randomness: these are things which cannot be explained, not matter how hard you try

Coefficient ‘b’: average change in y for a given change in x (y=ax+b+e)

Multicollinearity:

* When X’s are correlated, it is difficult to understand the unique impact of each variable on y

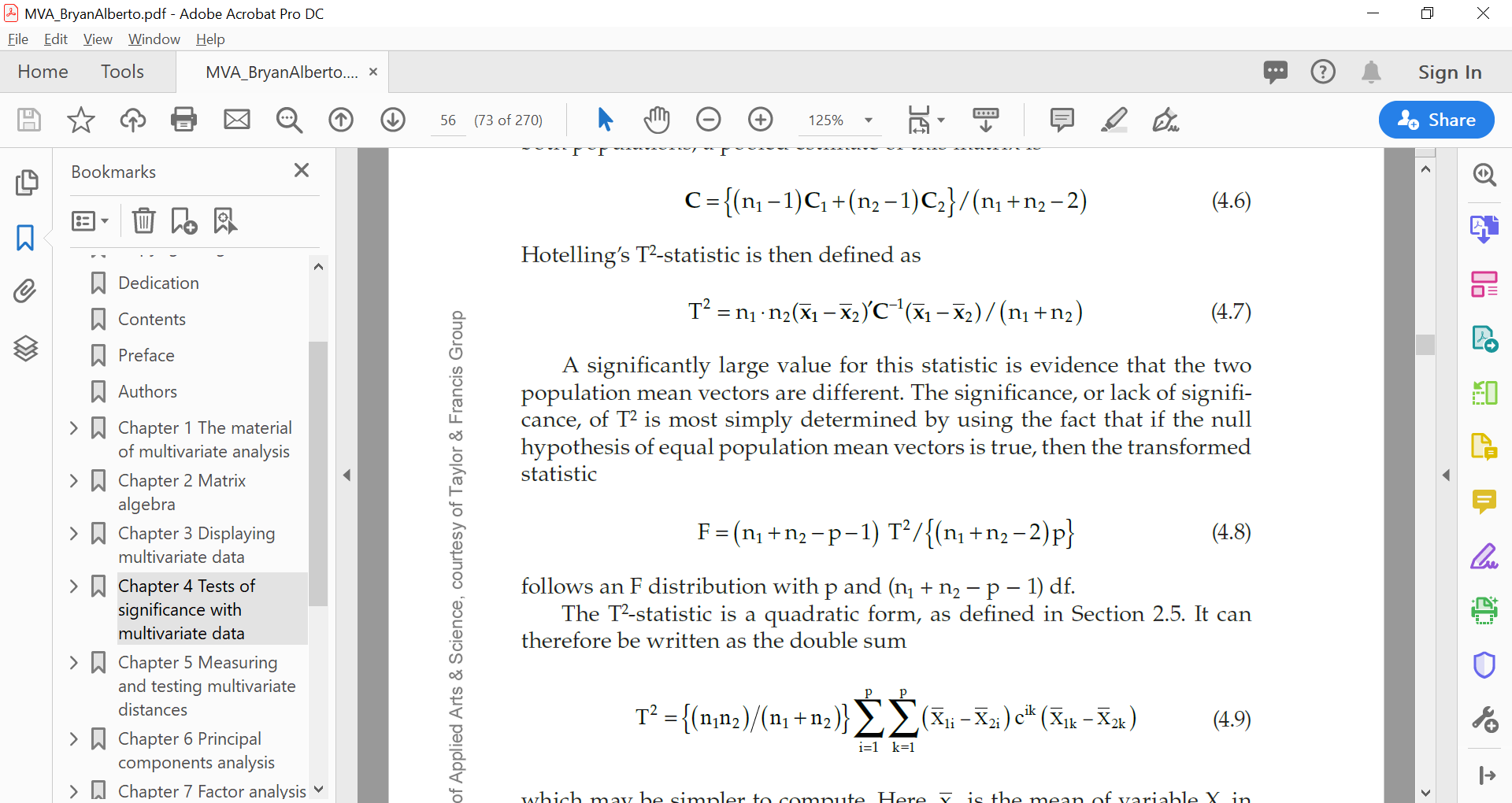
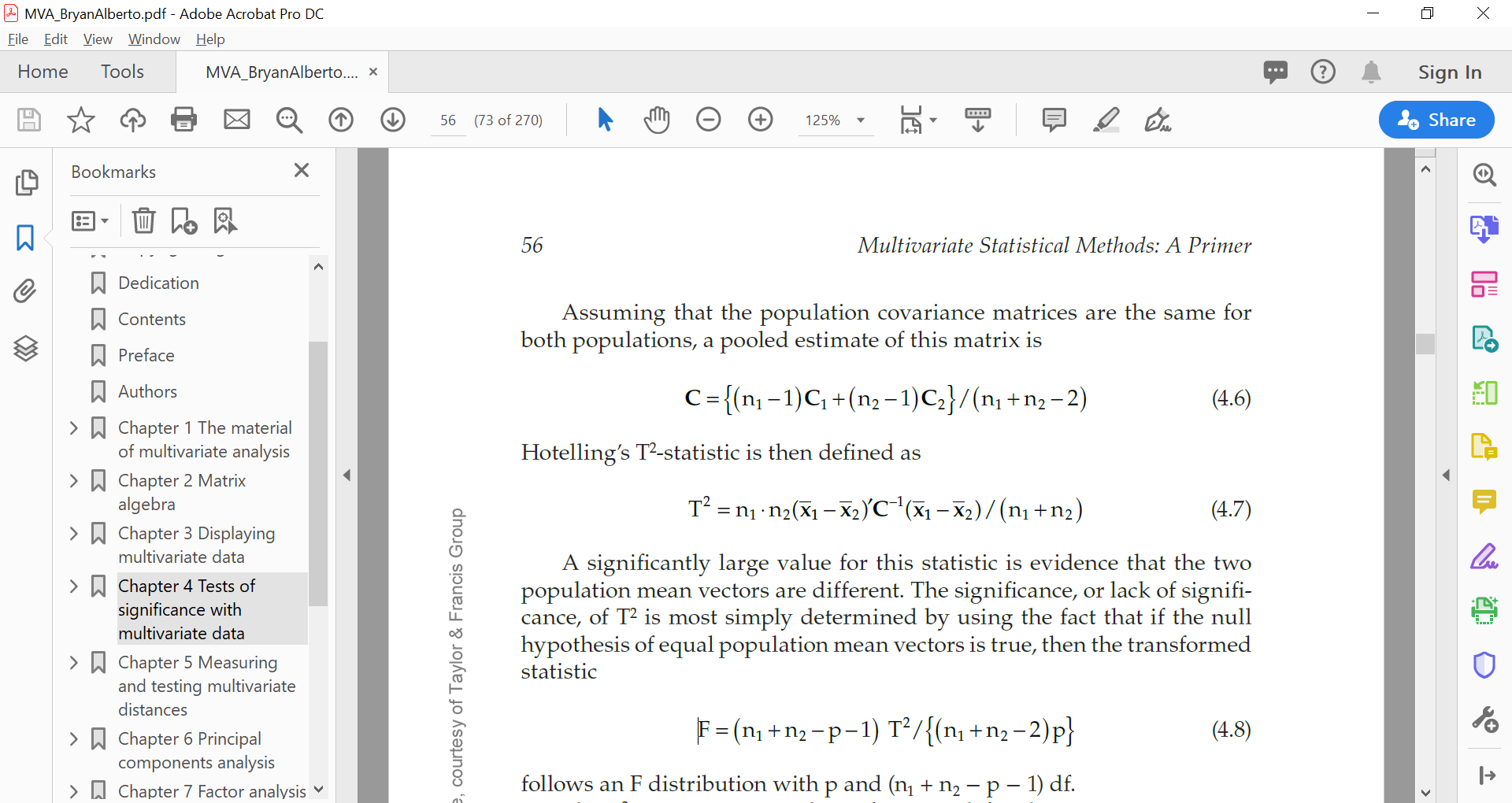
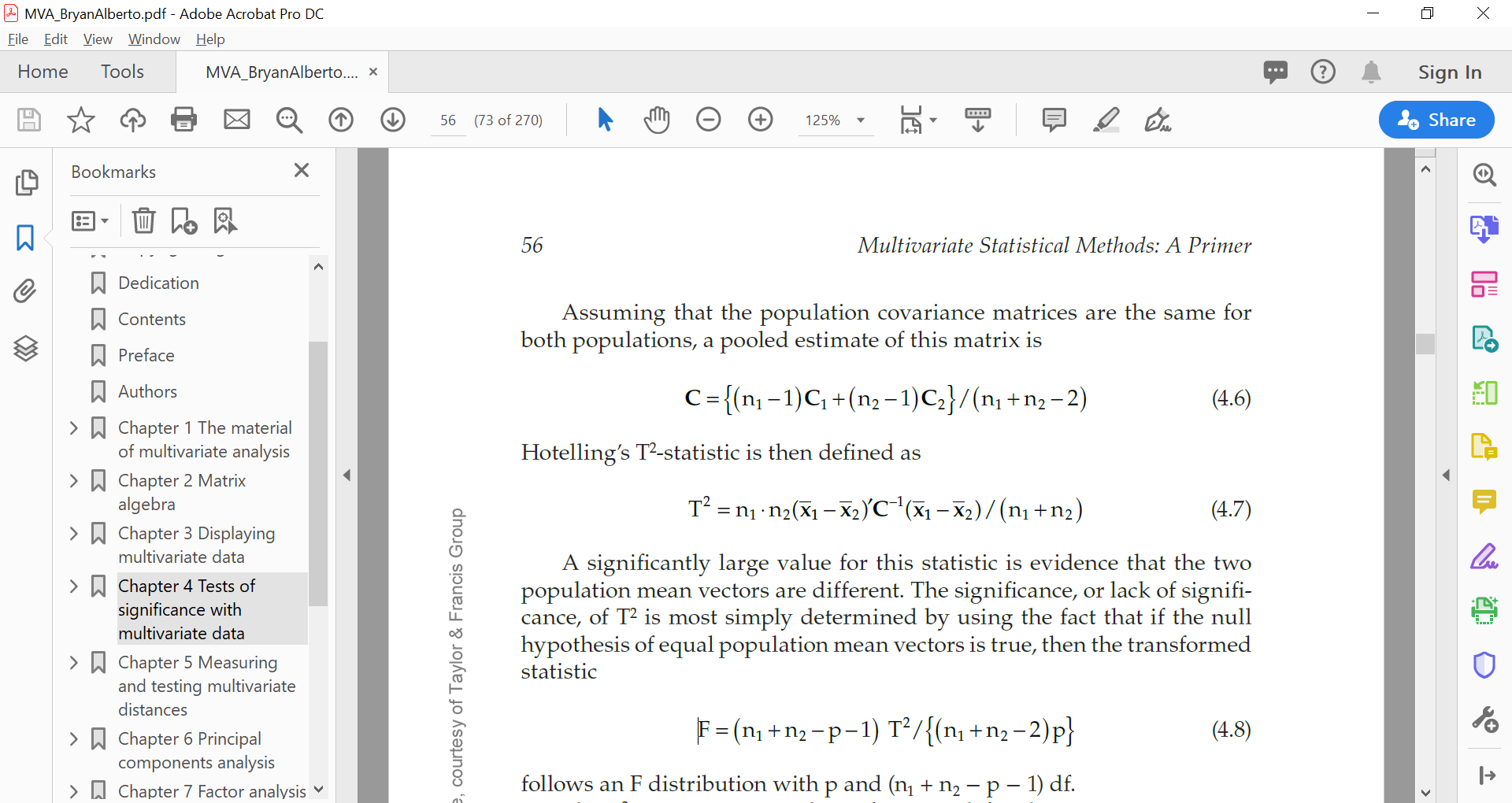
Unbiasedness of parameters

Univariate tests for means

Text

Description automatically generated

Multivariate tests for means



Variable coding

Manipulated:

* Low, medium, high
* Lotalty

Non- manipulated

* Different flavours
* Color

Variables must be at least on interval scale

Dummy coding(1,0)

Effect coding (-1,0,1)

Contrast coding

* Weighted linear combination of g means

Orthogonal coding

* Products of the weights for 2 contrasts sum to 0
* Intercept represents the unweighted mean of the g groups
* If group sizes are equal then intercept represents the grand mean

Dummy coding: useful when key object is to compare of an experimental group with that of control group

Effect coding: useful when key object is to compare treatment effects of different groups

Orthogonal coding: used when key object is orthogonal comparison of means

Variable coding does not change the model outcome

* R-sq, regression sum of squares, residual sum of squares, and the F ratio are **the same with any coding method**
* Predictions based on the regression equations resulting from the different coding methods are also **identical**

Experiment

* Manipulating an independent variable to see how it affects a dependent variable, while controlling the effects of additional extraneous variables

Experimental design

* Systematic procedure such that a change in a dependent variable may be attributed solely to the change in an independent variable or a set of procedures to investigate a relationship between variables

Type of experiments and experimental design

* Laboratory experiment: complete control over independent variables
* Field experiment: IVs are manipulated in a natural setting
* Natural experiments: IVs are not actively manipulated by researchers but the natural outcomes allow for possibility of measuring their impact on DV

Control group: the group that has not been exposed to change in independent variable

Experiment group: the group that has been exposed to change in independent variable

Completely randomized vs randomized block(stratified random) design

* Completely randomized: every subject is assigned to a treatment group at random. E.g. with cell phone usage low / high, each subject is assigned to one of the groups randomly
* Randomized block: first subjects are grouped by, for example, age and then in the group they are randomly assigned between low/high cell phone usage

Quasi- experiment design

* Subjects are assigned to groups based on non- random criteria

Experimental design process:

1. identify specific research questions
2. Define variables
3. List hypothesis
4. Design experimental treatments
5. Assigning subjects to treatment groups
6. Measurement of dependent variable
   1. Pretest
   2. Posttest
7. Validity and reliability

Factorial design: An experimental design that consists of two or more factors (IVs), with each factor having multiple discrete possible values or “levels”

* Full FD: all possible factor levels combinations are investigated
* Partial FD: only selected factor- level combinations are investigated

Main effects- the effects of just 1 independent variable on dependent variable, ignoring the effects on the rest of the independent variables

Interaction effects: when effect of an independent variable(on DV) depends on the level of another independent variable

Advantage: studying impacts of 2 or more variables

Assessing if IVs interact or not in their influence on DV

Greater control over model

Individual and joint effects are studied in the same dataset with same observations(efficiency)

Effect of a treatment is studied across different conditions of other treatment

|  |  |  |
| --- | --- | --- |
| **Moderator variable**  **Independent**  **Variable** | **Categorical** | **Continuous** |
| **Categorical** | Difference between group means vary by group membership of moderator variable | Difference between group means vary by level of moderator |
| **Continuous** | Slope of relationship (y,x1) differs by groups represented by categorical variable | Slope of regression line (or relationship between y, x1)  varies by level of moderator |