Mengyao Hu, 2016

Introduction In recent years, the number and scope of multinational, multicultural, or multiregional surveys, which to as '3MC' surveys, has increased dramatically. With the increased availability of large datasets covering multiple c such as the European Social Survey (ESS) and the Survey of Health, Ageing and Retirement in Europe (SHARE), n researchers have become engaged in analyzing these data . Not surprisingly, there has been increased interest in the development of statistical tests appropriate for cross-cultural survey data analysis. This chapter aims to provide a comprehensive introduction of different statistical methods, from basic statistics to advanced modeling approaches. that this chapter does not aim to teach statistics, but rather to provide an overview of what statistical tests are availal when to apply them in 3MC research. We also provide links and references to each statistical method for those who like additional detail.

↑ Ba

Guidelines

1. Exploring the variables.

1.1 The classification of variable types is important, because it will help to determine which statistical procedur be used. For example, when the dependent variable is continuous, a linear regression can be applied (see Guic 2.2); when it is categorical (binary), a logistic model can be applied (see Guideline 3.2); when it is categorical (nominal or ordinal), multinomial, or ordinal, logistic regressions may be used (see Guideline 3.3). If, in laten variable models (see Guideline 6), the latent variable is continuous, confirmatory factor analysis (CFA) or iter response theory (IRT) models can be used (see Guidelines 6.1 and 6.5). Figures 1 and 2 below list the choices regression and latent variable measurement models as regard the variable types of the dependent and indepen variables. Several commonly used variable types are listed as below:

- **Nominal variables**: variable values assigned to different groups. For example, respondent gender may be 'r 'female.'
- **Ordinal variables**: categorical variables with ordered categories. For example, 'agree,' 'neither agree nor di or 'disagree.'
- **Continuous variables**: variables which take on numerical values that measure something. “If a variable car any value between two specified values, it is called a continuous variable; otherwise, it is called a discrete v . Continuous variables are understood to have equal intervals between each adjacent pair of values in the distribution. Income is an example of a continuous variable.
- **Discrete (ratio) variable**: “a discrete variable can only take on a finite value, typically reflected as a whole number” . The variables have an absolute ‘0’ value. One example is the number of children a person has.

Figure 1: Variable types and choices of regression models.

(image)

Figure 2: Variable types and choices of latent variable measurement models (see Guideline 6 for more detai

(image)

1.2 Distribution of variables:

1.2.1 **Graphical illustrations of distributions**: it is commonly recommended to look at graphical summari both continuous and categorical distributions before fitting any models. Details of the graphical options l below can be found in this online statistics book: Online Statistics Education: An Interactive Multimedia of Study.

- For categorical variables:
  - Bar graphs
  - Pie charts
- For continuous or discrete variables:
  - Stem and Leaf Plots
  - Histograms
  - Box plots
- For any type of variable:
  - Frequency distributions

In 3MC data analysis, to get a direct visual comparison, researchers can plot distributions by country or raci group.

1.2.2 **Numerical summaries of distributions**: a distribution can be summarized with various descriptive st The mean and median capture the center of a distribution (central tendency), while the variance describe distribution spread or variability (see online book material).

- **Mean**: the average of a number of values. It is calculated by adding up the values and dividing by the n values added.
- **Median**: the "…median is the number separating the higher half of a data sample, a population, or a pr distribution, from the lower half" . For a highly skewed distribution, the median may be a more appropr measure of central tendency than the mean. For example, the median is more widely used to characteriz income, since potential outliers (e.g., those with very high incomes) have much more impact on the mea
- **Variance**: a measure of the extent to which a set of numbers are 'spread out.'
- **Precision**: the reciprocal of the variance; most commonly seen in Bayesian analysis (see Guideline 9).

1.3 Suggested reading:

- Tests of the equality of two means (see online material)
-
-

1.4 Potential uses in 3MC research:

- A good starting point for analysis is to look at the distributions of variables of interest and at graphical illust of the variables in each cultural group.
- One way of comparing survey estimates across various cultures is to directly compare mean estimates. A tw sample t-test can be used to evaluate the equality of two means (see Guideline 1.3). However, researchers n

aware that the observed mean differences are not necessarily equal to the latent construct mean differences (Guideline 6), and direct comparison using observed mean differences may lead to invalid results (see). In a factors irrelevant to the question content, such as response style differences in different cultures, may influe comparability across cultures. More advanced models (such as latent variable models) can be used to evalua control for these factors.

↑ Ba

2. Simple and multiple linear regression models.

2.1 A bivariate relationship is the relationship between two variables. For example, one may be interested in kn how height is associated with weight (i.e., whether those who are taller tend to weigh more). Basic informatio bivariate relationships can be found here.

- **Scatterplots**: before running any models, a scatterplot is essential to explore the associations (negative or p between variables.
- **Correlations between variables**: Pearson's correlation is the most commonly used method of evaluating th relationship between two variables. Refer to this website for more information.

2.2 Linear regression models can allow researchers to predict one variable using other variable(s). The depende variable in linear regression models is a continuous variable. Basic information about simple linear and multi regression models can be found here.

- **ANOVA table**: in the output of regression model results, an analysis of variance (ANOVA) table is usually provided, consisting "of calculations that provide information about levels of variability within a regression and form a basis for tests of significance" .
  - Resource 1
  - Example of a regression model results output using Stata
- **Dummy predictor variables**: as described by , a dummy variable or *indicator variable* is an artificial varia created to represent an attribute with two or more distinct categories or levels. If a categorical variable is ad directly to the regression models without being specially specified, the software will treat it as continuous. I the differences between the categories (e.g., category 2 minus category 1) do not have any actual meaning. variables are usually created in these situations to make sure that such categorical variables are correctly sp the model. For example, in 3MC data analysis, to compare Country A to Country B on the level of the depe variable, one can create a country dummy variable using one of the countries as a reference group and add i independent variable to the model. When multiple countries exist, one can use one of the countries as the re category, and treat the variable as categorical in the model .
  - For information on dummy variables and how they are created and used, see .
  - For information on regression models with categorical predicators using SAS, see here.
- **Interactions of predictor variables**: sometimes a regression model is used to test whether the relationship the dependent variable (DV) and one specific independent variable (IV) depends on another IV. To test this, interaction term between the two IVs can be added to the model.
  - Resource 1
- **Transformations of variables**: when non-linearity is found for predictors, transformations may be conside 'normalize' a variable which has a skewed distribution. For more detail, see .
- **Lack of fit testing**: various techniques are available to test for the lack of fit in regression models, includin (e.g., plots) and numerical methods (e.g., R^2 and F-tests).
- **Model diagnostics**: techniques are available to test the appropriateness of the model and whether the mode assumptions hold.
  - Resource 1
  - Resource 2 (using R)

- **Selecting reduced regression models** (variable selection): techniques for determining the model which cor most appropriate independent variables, giving the maximum \(R^2\) value.
  - Resource 1 (Including SAS code)
  -

2.3 Suggested reading:

- Applied Statistical Analysis and Data Display: An Intermediate Course with Examples in S-PLUS, R, and S
- Statistical Methods, 8th ed
- The Little SAS Book, 4th ed

2.4 Potential uses in 3MC research:

- As in linear regression models, a country variable/indicator can be added to the regression model as a covar (e.g.,).

↑ Ba

3. Categorical data analysis.

3.1 **Analysis of two-way tables**: categorical data are often displayed in a two-way table. Sometimes, one or bo variables are continuous. If so, the continuous variable(s) can be categorized into groups. A two-way table ca constructed using the new variables. Note that this approach may lead to a loss of information on the continuc variables. See also online marterial.

3.1.1 The Pearson chi-square test evaluates whether the row and column variables in a two-way table are as

- Resource 1

3.1.2 Odds ratios (OR) and relative risks (RR) describe the proportions in contingency tables. See for a comprehensive introduction.

- Resource 1
-

3.1.3 Log-linear models are commonly used to model the cell counts of contingency tables such as two-way

- Resource 1

3.2 **Logistic regression models**: these can be used when the dependent variable is a binary categorical variable technique allows researchers to model or predict the probability that an individual will fall into one specific c given other independent variables. Logistic regression is a type of generalized linear model where the logit fu selecting one category is expressed through a linear function of the predictors. Thus, as in other linear regress models, the predictors can include both continuous and categorical variables.

- (link)
- Resource 2

3.3 **Multinomial and ordinal logistic regressions**: when the DV is a nominal variable, a multinomial logistic regression model can be used. If the DV is an ordinal variable, an ordinal logistic regression can be used.

- Resource 1

3.4 Suggested reading:

-
-
-

3.5 Potential uses in 3MC research:

- To evaluate responses to a categorical variable across two different cultures, one can construct a two-way ta using the categorical variable and the country indicator as the rows and columns. A Pearson chi-square test used to evaluate whether the variable differs by cultures.
- As in logistic regression models, a country variable/indicator can be added to the logistic regression model covariate.

↑ Ba

4. Multilevel models.

Multilevel models are usually used when there is a hierarchical structure, such as when sampling units are nested in geographical areas (e.g., cluster sampling) and when they are selected in longitudinal studies. Multilevel models are known as hierarchical linear models, mixed models, random effects models, and variance components models. The for Multilevel modeling at the University of Bristol offers a free online course on multilevel modeling. Additional information on multilevel modeling can be found here and in . When many cultural groups are present, a multilevel framework can be used, with country being treated as a random variable. Multilevel models with latent variable, suc multilevel structural equation models (MLSEM), can also be run, as discussed by and . See Guideline 6.4 for more information on SEM models.

4.1 Suggested reading:

-
-
-
-

4.2 Potential uses in 3MC research:

- Many 3MC studies have a multilevel data structure with respondents nested within countries. Recent resear multilevel cross-cultural research has emerged in last several decades. For more information, see .

↑ Ba

5. Longitudinal analysis.

Longitudinal data analysis refers to techniques used to evaluate data collected through repeated measures.

- Resource 1

5.1 **Modeling longitudinal/panel data**: in panel surveys, respondents are interviewed at multiple points in tim producing 'panel' or 'longitudinal' data. The first step in analyzing longitudinal data is to look at the descriptiv then, select one of several possible methods of analysis. The traditional technique is the repeated measures an variance (rmANOVA), although this has several limitations. More commonly used approaches include multile models and marginal models.

5.1.1 **Descriptive plots**: The 'spaghetti' plot “involves plotting a subject’s values for the repeated outcome n (vertical axis) vs. time (horizontal axis) and connecting the dots chronologically” . Plots can be created a the individual data level and the mean level. For binary outcomes, proportions can be used to generate th for different population groups. In 3MC studies, the plots can be generated for different cultural or count groups.

- Resource 1

5.1.2 Repeated measures analysis of variance (rmANOVA):

- For more information, see this online example on rmANOVA.
- The rmANOVA approach is not recommended due to the limitations as mentioned below:
  - Subjects missing any data will not be included in the analysis.
  - A limited number of covariance structures are allowed.
  - Time-varying covariates are not allowed.

5.1.3 **Multilevel models for longitudinal data**: multilevel models account for between respondent variance including random effects in the model, such as random slope and random intercept.

- Resource 1
- Resource 2
- (presentation slides available here)

5.1.4 **Marginal modeling approaches**: If the between-subject variation is not of interest, the marginal mod approach, where only the correlated error terms are included in the model, can be used—no random effe added to the model.

-

5.2 Suggested reading:

-
-
-
-
-
-
-
-

5.3 Potential uses in 3MC research:

- A country variable/indicator can be added to the marginal models as a covariate, or it can be added in a mul model as a fixed effect.

↑ Ba

6. Latent variable models.

Latent variable models include both observed variables (the data) and latent variables. A latent variable is unobserve which represents hypothetical constructs or factors . A latent variable can be measured by several observed variable example of latent variable provided by describes the construct of intelligence. As mentioned by , “there is no single, definitive measure of intelligence. Instead, researchers use different types of observed variables, such as tasks of ver

reasoning or memory capacity, to assess various facets of intelligence." Examples of such latent variables are usuall measured in a measurement model which evaluates the relationship between latent variables and their indicators.

6.1 **Exploratory factor analysis** (EFA) and **confirmatory factor analysis** (CFA): these are two types of measu models where latent variables are indicated by multiple observed variables. The difference between EFA and related to the existence of a hypothesis about the measurement model before doing the analysis. As mentioned "CFA attempts to confirm hypotheses and uses path analysis diagrams to represent variables and factors, whe tries to uncover complex patterns by exploring the dataset and testing predictions." As seen in Figure 3, since indicators and the latent variables are all continuous, both EFA and CFA are based on linear functions. Figure the differences between EFA and CFA. Since EFA is purely data-driven and may be arbitrary in nature, it is su by some literature to always use CFA, which is theory-driven . However, as mentioned by , in selecting items, more appropriate to use EFA rather than CFA when the theory is not well established.

See for a comprehensive overview on EFA and for CFA. The code for conducting EFA and CFA are include Appendix A.

Figure 3: Comparisons between EFA and CFA.*

(image)

**Adapted from Exploratory and Confirmatory Factor Analysis presentation .*

Multigroup CFA (MCFA) is commonly used in 3MC research for measurement equivalence testing. The ba is to start with the same model, but allow the coefficients to differ by groups (assuming configural equivalenc then start introducing constrains in the model coefficients—such as to make them equal across the groups. Th model fit of the previously run models can be compared. Among all the models, the parsimonious model with fit solution will be selected to evaluate the data. If the model reveals no violations of scalar equivalence, the c means can be compared directly. In a panel study, with data available at different time points, one can also eva measurement equivalence across cultures over time. See Guideline 6.2 below for more information on measu equivalence testing. For more information on MCFA, see:

- Resource 1
-
-

6.2 **Measurement equivalence in 3MC research**: as mentioned by , "measurement equivalence implies that a measurement instrument used in different cultures measures the same construct." There are different levels of equivalence. Three of the most widely discussed levels are configural, metric, and scalar equivalence. These t levels are hierarchical, where the higher ones have higher requirements of equivalence and require the achieve the lower ones .

Configural equivalence refers to similar construction of the latent variable. In other words, the same indicat associated with the latent concepts in each culture. It does not require each culture to view the concept in the s way. For example, it allows the strength (i.e., loadings) to be different across cultures. *Metric equivalence* req same slope across cultures, which captures the associations between indicator and the latent variable. In other it implies "the equality of the measurement units or intervals of the scale on which the latent concept is measu across cultural groups" . *Scalar equivalence* implies that on the basis of equality of the measurement units, the of the latent variable also have the same origin across cultures . Under this equivalence level, the model achie measurement equivalence, and researchers can compare the country scores directly.

In situations where full equivalence is difficult to achieve, researchers also evaluate the conditions under wh different cultures achieve partial equivalence. An example of partial equivalence is when most of the indicato equivalent across cultures, but only one has a different slope and threshold across cultures. One can then conc

the different cultures achieve partial equivalence where they differ on one specific indicator. As mentioned by “partial equivalence enables a researcher to control for a limited number of violations of the equivalence requ and to proceed with substantive analysis of cross-cultural data” .

The aforementioned approaches to assessing measurement equivalence have been widely used in 3MC surv analysis. However, it has recently been criticized for being overly strict. As mentioned by , it is difficult to ach scalar equivalence, or even metric equivalence, in surveys with many countries or cultural groups. A Bayesian approximate equivalence testing approach has been recently proposed by . This approach allows for “small va in parameters across different cultural groups; thus, when approximate scalar measurement equivalence is rea one can compare across cultures meaningfully, even though the traditional method may indicate scalar inequiv point out that approximate measurement inequivalence or invariance are favorable for cases with large-scale c containing many groups or repeated measurements, small differences in intercepts and factor loadings, and m differences which are cancelled out both within and between groups. In contrast, approximate measurement in testing seems to lead to bias in latent mean estimates if large differences in intercepts or factor loadings or sys differences between groups exist . For introductions and references of Bayesian methods, see Guideline 9.

6.3 **Latent class analysis** (LCA): unlike the previously mentioned approach, such as CFA and SEM, where the variables are continuous, LCA treats the latent variables as categorical (nominal or ordinal) (see Figure 3). Th categories of the latent variable in LCA are referred to as classes, which represent “a mixture of subpopulation membership is not known but is inferred from the data” . That is to say, LCA can classify respondents into dif groups based on their attitudes or behaviors, such as classifying respondents by their drinking behavior. Respo in the same group are similar to each other, regarding the behavior/attitudes, and they differ from those in oth —i.e., heavy drinkers vs. non-drinkers. One can also add covariates to the model if those measures can influe class membership. In a second step, the class membership from the model can then be used for followup analy example, to better understand the differences between respondents, a logistic (or multinomial logistic, if more two groups) regression model can be run in which selected covariates are used to predict class membership. Alternatively, to evaluate the influence of class membership on other variables, LCA can also be used in regre models as a covariate to predict other outcomes. For more information on LCA, please see:

- Resource 1
-

As mentioned by , when testing for measurement invariance with latent class analysis, “the model selection procedure usually starts by determining the required number of latent classes or discrete latent factors for each … If the number of classes is the same across groups, then the heterogeneous model is fitted to the data; follo series of nested, restricted models which are evaluated in terms of model fit.” That is to say, unlike multigrou the multigroup LCA will need to identify whether the number of classes are the same across groups before tes models at different invariance levels. See , , and for more information.

6.4 **Structural equation modeling** (SEM) is a multivariate analysis technique used in many disciplines which test the causal relationship hypothesis between variables . It usually includes two components: 1) the measure model, which summarizes several observed variables using their latent construct (e.g., CFA as discussed in G 6.1), and 2) the structural model, which describes the relationship between multiple constructs (e.g., relationsl among both latent and observed variables).

6.4.1 Similar to previously discussed latent variable models, SEM can have both observed and latent variab where observed variables are the data collected from respondents and latent variables represent unobserv construct and factors . The observed variables which are used as measures of a construct are indicators o latent variable. In other words, the latent variable is indicated by these observed variables. Besides obser latent variables, SEM models also include error terms, similar to the error terms in a regression analysis. mentioned by , “a residual term represents variance unexplained by the factor that the corresponding ind

supposed to measure. Part of this unexplained variance is due to random measurement error, or score unreliability."

6.4.2 In SEM analysis, parameter estimation is done by comparing the model-based covariance matrix with based covariance matrix. The goal of this approach is to evaluate whether the model with best fit is supp the data—that is, whether the two covariance matrices are consistent with each other and whether the mo explain as much of the variance of the data.

6.4.3 Structural equation models can also estimate the means of latent variables. It also allows researchers to between- and within-group mean differences . In 3MC analysis, one can estimate the group mean differe latent variables, such as those between two cultures.

6.4.4 Suggested reading:

-
-
-

6.5 **Item response theory** (IRT) is commonly used for psychometric and educational testing, and is becoming m popular in 3MC analysis. It begins with the idea that when answering a specific question, the response provid individual depends on the ability/qualities of the individual and the qualities of the question item. As mention "the mathematical foundation of IRT is a function that relates the probability of a person responding to an iter specific manner to the standing of that person on the trait that the item is measuring. In other words, the funct describes, in probabilistic terms, how a person with a higher standing on a trait (i.e., more of the trait) is likely provide a response in a different response category to a person with a low standing on the trait." Therefore, IR researchers to model the probability of a specific response to a question item, given the item and the individua level.

The simplest IRT model is the Rasch model, also called the one-parameter IRT model, which assumes equa discrimination ("the extent to which the item is able to distinguish between individuals on the latent construct model starts from the premise that the probability of giving a 'positive' answer to a yes/no question is "a logist function of the distance between the item's location, also referred to as item difficulty, and the person's locati construct being measured," also known as the person's latent trait level . There are other types of IRT models available as well, which can be categorized by the number of parameters and the question response option for (e.g., binary or multiple response options and whether ordinal or nominal). Table 1 below summarizes the diff types of IRT models.

In a two-parameter (2PL) IRT model, an item discrimination parameter is also included in the model. The parameters are "analogous" to the factor loadings in CFA and EFA, since they all represent "the relationship b the latent trait and item responses" . A three-parameter (3PL) IRT model also includes a 'guessing' parameter. describes the situation wherein when a question can be answered by guessing, the probability of giving a corr answer is higher than zero even for those with low latent trait level.

For items with multiple response options (ordinal or nominal variables), polytomous IRT models can be use Table 1 for details. In this chapter, we will not discuss these models in detail. See suggested readings on IRT m for more information.

Table 1: Variable type and IRT model choices.

Type of observed variable	Model
Binary	1 parameter-logistic (1PL) model/Rasch model

	2PL model
	3PL model
Multiple response options (ordinal)	Graded response model/Thurstone/Samejima polytomous models
	Partial credit model (PCM)/Graded PCM
Multiple response options (nominal)	Rating scale model
	Nominal response model/Bock's model

6.5.1 Suggested reading:

-
-
-
-
-
-

6.6 Besides what is discussed above, other types of latent variable models are available. Some examples are list below:

- The **latent transition model** is "a special kind of latent class factor model that represents the shift from one different states, such as from nonmastery to mastery of a skill, is a latent transition model" . See these slides more information.
- In **latent profile models**, the latent variable is categorical and the indicators are continuous. It is commonly cluster analysis. See for more information.
- The **mixed Rasch model** is "a combination of the polytomous Rasch model with latent class analysis" . See more information.
- When we have data where the population of individuals are divided into different groups, such as in a 3MC **multilevel structural equation modeling** (MLSEM) can be used. This model decomposes individual data within-group and between-group components, and can simultaneously estimate within- and between-group For more information on MLSEM, see .

6.7 Potential uses in 3MC research:

- As mentioned by , the observed mean does not equal the latent mean, where the observed mean is a functio intercepts, factor loadings, and the latent mean. Similarly, "observed mean differences between two or more (e.g., cultures) do not necessarily indicate latent mean differences as unequal intercepts and/or factor loadin also lead to observed differences" . To conduct more valid comparisons across different groups (e.g., cultur measurement invariance testing is a widely used method which aims to evaluate whether the latent means o groups are comparable. In other words, it evaluates whether the different groups differ in factor loadings an intercepts of the measures. See for more information.
- Measurement invariance testing is usually conducted within the multigroup analysis (MGA) framework. Th commonly used method is multigroup confirmatory factor analysis (MGCFA) (e.g.,). Other types of MGA multigroup structural equation modeling (MGSEM) analysis (e.g.,), multigroup latent class analysis (e.g.,) multigroup IRT model (e.g.,) and multigroup mixed Rasch model (e.g.,). See for more information.
- A recent paper by discusses misconceptions in measurement equivalence analysis. Using data from the Wor Values Survey, they show that "constructs can entirely lack convergence at the individual level and neverthe

exhibit powerful and important linkages at the aggregate level".

- examine new approaches to measurement invariance evaluation in the context of international large-scale assessments (ILSAs), which aim to produce cross-national comparisons of student outcome measures (e.g., achievement, behaviors, and attitudes). begin by reviewing the extent to which exact measurement invarian be confirmed; then, using data from the International Civic Citizenship Education Study 2009, they extend t model to include partial invariance, indirect/covariate effects, and a Bayesian approach to evaluating the approximate invariance assumption. They find that the Bayesian approach, assuming a parsimonious varian procedure, can play a key role in evaluating measurement invariance across national samples used in ILSAs that the integration of a strong model with less stringent assumptions is sufficient for estimating comparable of the latent constructs and allows the reporting of a table charting mean comparisons across a large numbe groups. Yet, a high degree of consistency between 's results and models assuming exact invariance indicates previously published results are not invalidated in any substantial way.

↑ Ba

7. Differential item functioning.

Differential item functioning (DIF) is a statistical concept developed to identify to what extent a question item migh measuring different properties for individuals of separate groups (i.e., ethnicity, culture, region, language, sex, etc.). used as indicators for 'item bias' if the items function in a systematically different way across cultures. To detect DI different methods can be used, as listed below:

- Mantel-Haenszel (MH) statistic is regarded as a "reference" technique of detecting DIF due to its ease of use an fact that it can be applied to small samples . The disadvantage of MH statistic is that it does not allow for statist significance testing.
- Logistic regression can be used as an alternative method to detect DIF. For more information, see .
- DIF can be detected using an IRT framework. Item characteristic curves (ICCs) of the same item but from diffe groups can be compared to evaluate whether there is DIF. For more information, see and .

↑ Ba

8. Machine learning.

Machine learning is "a general term for a diverse number of classification and prediction algorithms" which have applications in many different fields. Unlike statistical modeling approaches, machine learning evaluates the relatio between outcome variables and predictors using a "learning algorithm without an a priori model" . Below, we introc several machine learning methods.

8.1 The **classification tree** is a data-driven method which allows researchers to evaluate the complex interactio between variables when there are many predictor variables present. In binary trees, the nodes of the tree are d into two branches. To reasonably construct and prune a given tree, deviance measure is used to choose the spl the 'rpart' package is used for classification tree analysis (see Appendix A for more information). The classific tree result can be evaluated through apparent error rate and true error rate. The former is the error rate when t applied to a training data set, and the latter is when it is applied to a new data set or to test data. In evaluating error rate, researchers usually divide the data into two parts—training data and test data—and validate the tre on the test data set.

8.2 Random forest is an algorithm for classification which uses an "ensemble" of classification trees . Through averaging over a large ensemble of "low-bias, high-variance but low correlation trees," the algorithm yields a ensemble that can achieve both "low bias and low variance" .

8.3 Suggested reading:

- Resource 1
- Resource 2
-
-
-
-

8.4 Potential uses in 3MC research:

- Classification tree analysis in cross-cultural research allows researchers to evaluate 1) the important factors culture and 2) how the factor interactions differ across cultures. One study used classification tree to evalua college student alcohol consumptions across American and Greek students, and found that “student attitude drinking were important in the classification of American and Greek drinkers” .

↑ Ba

9. Incorporate complex survey data features.

It is usually difficult to draw a simple random sample from the population due to cost and practical considerations s comprehensive sampling frame being available. As discussed in Sample Design, complex samples, such as surveys involving stratified/cluster sample design, are commonly used in surveys. In a simple random sample, one can assu observations are independent from one another. However, in a complex sample design such as multi-stage samples schools, classes, and students, students from one classroom are likely to be more correlated than those from another classroom. Therefore, as described in Sample Design, in the analysis phase, we need to compensate for complex su designs with features including, but not limited to, unequal likelihoods of selection, differences in response rates ac subgroups, and deviations from distributions on critical variables found in the target population from external sourc as a national census), most commonly through the development of survey weights for statistical adjustment. If comp sample designs are implemented in data collection but the analysis assumes simple random sampling, the variances survey estimates can be underestimated, and the confidence interval and test statistics are likely to be biased .

In a recent meta-analysis of 150 sampled research papers analyzing several surveys with complex sampling designs found that analytic errors caused by ignorance or incorrect use of the complex sample design features were frequent analytic errors define an important component of the larger total survey error framework, produce misleading descri populations, and ultimately yield misleading inferences . It is thus of critical importance to incorporate the complex design features in statistical analysis.

For many of the aforementioned statistical models, various statistical software programs, such as the 'svy' statement and SURVEY procedures in SAS, have enabled the analysis of complex survey data features. See Appendix A for n information.

9.1 Suggested reading:

-
-
-
-
-

↑ Ba

10. Introduction to Bayesian inference.

This section presents an overview of Bayesian theory, which follows closely the overview of , , and . In surveys, respondents' answers, denoted as y, reflect our measure of the true population's Y—a random variable takes c realized value y. In other words, Y is unoberseved, and the probability distribution Y is of researchers' inte We use θ to denote a parameter which reflects the characteristics of the distribution of Y. For example, $($ can be the mean of the distribution. The goal is to estimate the unknown parameter θ based on the data, whic $(p(\theta|y)$. Based on Bayes' theorem ($p(\theta|y)=\frac{p(\theta,y)}{p(y)}=\frac{p(y|\theta)p(\theta)}{p(y)}$), v $(p(y)$ is the probability distribution of the data, which is known for researchers; $(p(y|\theta)$ refers to the probabil the data given the unknown parameter θ; and $(p(\theta)$ is the *prior distribution* of the parameters. $(p(\thet thus referred to as the *posterior distribution* of the parameter θ given the data, which is also the results of the

In summary, Bayesian methods use both the prior information (which indicates the distribution of parameters) and t distribution of data to estimate the model results—the posterior distributions of the parameters. The key difference t between Bayesian and frequentist approaches relates to the unknown parameter θ. In the frequentist approac $(\theta$ is viewed as unknown but fixed. On the other hand, in the Bayesian approach, θ is random, which h posterior distribution taking into account the uncertainty of θ.

10.1 There are generally two types of priors: **noninformative** and **informative priors**. The choice between the t depends on our confidence about how much information we have about the priors and how accurate they are. Noninformative priors are also referred to as 'vague' or 'diffuse' priors. They are used when there is little infor about the priors, and thus their influence on the posterior distribution of θ is minimal . An example of noninformative prior can be a density with a huge variance, so that the Bayeisan estimation is mainly affected data. Informative priors are used when we have sufficient information about the priors, such as from expert kn and similar datasets.

10.2 There are multiple Bayesian model comparison statistics. Two of the most commonly used are the **Bayes fa** the **deviance information criterion** (DIC). The Bayes factor quantifies the odds that the data favor one hypo over another. As discussed in Guideline 5, Bayes factors are not well defined when using noninformative prio the evaluations can be computationally difficult . DIC compromises both goodness of fit and model complexi practical applications, the model with the smaller DIC value is preferred.

10.3 When we have estimated the posterior distributions of the parameters, we would like summaries of the distr (such as mean and variance) for hypothesis testing. One important way to evaluate the distribution is based or **credible interval**, which is often considered to be a similar measure to the 'confidence interval' in the frequen approach. Credible interval is based the quantiles of the posterior distributions. Based on the quantiles, we can evaluate the probability that the parameter lies in a particular interval. When this probability is 0.95, it is refer 95% credible interval. If the credible intervals from two models do not overlap, we say that the two posterior distributions of this parameter differ.

10.4 **Markov chain Monte Carlo** (MCMC) is the most common computational algorithm for Bayesian methods generates Markov chains, which simulate the posterior distribution. The basic idea is that by simulating a suff large number of observations from the posterior distribution, $(p(\theta|y)$, we can approximate the mean and summary statistics of the distribution. The use of MCMC for posterior simulation in latent variable models is the latent variables as missing data, which enables the augmentation of the observed variables. The most com MCMC algorithm is the Gibbs sampler, which performs on alternating conditional sampling at each of its iter More specifically, it draws each component conditional on the values of all the other components . In a Marko early proportion of the chain which may not converge to target distribution is called burn-in.

10.5 Multiple **convergence diagnostics** exist. In practice, it is common to inspect several different diagnostics, s there is no single adequate assessment. One of the most common statistics in a multiple chain condition is the diagnostic, which compares the within-chain and between-chain variance. A value above 1.1 is an indication of convergence. The common diagnostics for single chain condition include the convergence diagnostic and t

convergence diagnostic, which can help to decide how many iterations are needed and how many can be treat burn-in in a long enough chain.

10.6 Suggested reading:

-
-
-
-
-

10.7 Potential uses in 3MC research:

- As previously mentioned in Guideline 6, approximate Bayesian measurement equivalence approach can be cross-cultural comparison research (e.g., or). See also the suggested readings in Guideline 9.1 above for mo information.

↑ Ba

References

↑ Ba