What Statistical Method is best for your Study?

Derek Chiu, MSc.

August 3, 2016

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data
- Important component of Methods & Results sections of a paper

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data
- Important component of Methods & Results sections of a paper
- Need to adequately understand how and why things work

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data
- Important component of Methods & Results sections of a paper
- Need to adequately understand how and why things work
- The “best” statistical method

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data
- Important component of Methods & Results sections of a paper
- Need to adequately understand how and why things work
- The “best” statistical method
    - Appropriate given the data structure

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data
- Important component of Methods & Results sections of a paper
- Need to adequately understand how and why things work
- The “best” statistical method
  - Appropriate given the data structure
  - Address complexities of data

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data
- Important component of Methods & Results sections of a paper
- Need to adequately understand how and why things work
- The “best” statistical method
  - Appropriate given the data structure
  - Address complexities of data
  - Answer the research question

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data
- Important component of Methods & Results sections of a paper
- Need to adequately understand how and why things work
- The "best" statistical method
  - Appropriate given the data structure
  - Address complexities of data
  - Answer the research question
- Assume rectangular data where rows are observations and columns are variables

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data
- Important component of Methods & Results sections of a paper
- Need to adequately understand how and why things work
- The "best" statistical method
  - Appropriate given the data structure
  - Address complexities of data
  - Answer the research question
- Assume rectangular data where rows are observations and columns are variables
- Analyzing using software Excel, R, SPSS, etc.

Introduction

- **Statistics** is the discipline of collecting, analyzing, interpreting, and reporting data
- Important component of Methods & Results sections of a paper
- Need to adequately understand how and why things work
- The “best” statistical method
  - Appropriate given the data structure
  - Address complexities of data
  - Answer the research question
- Assume rectangular data where rows are observations and columns are variables
- Analyzing using software Excel, R, SPSS, etc.
- Goal: Provide an analytical framework so you understand what aspects to consider in a study

1. Missing Data

- Percent of missingness

1. Missing Data

- Percent of missingness
- Reason for missingness: MCAR, MAR, MNAR

1. Missing Data

- Percent of missingness
- Reason for missingness: MCAR, MAR, MNAR
- Complete case analysis: keep only observations with no missing values across all variables

1. Missing Data

- Percent of missingness
- Reason for missingness: MCAR, MAR, MNAR
- Complete case analysis: keep only observations with no missing values across all variables
- Complete variable analysis: keep only variables with no missing values across all observations

1. Missing Data

- Percent of missingness
- Reason for missingness: MCAR, MAR, MNAR
- Complete case analysis: keep only observations with no missing values across all variables
- Complete variable analysis: keep only variables with no missing values across all observations
- Mean Imputation: fill in missing value with mean of non-missing values

1. Missing Data

- Percent of missingness
- Reason for missingness: MCAR, MAR, MNAR
- Complete case analysis: keep only observations with no missing values across all variables
- Complete variable analysis: keep only variables with no missing values across all observations
- Mean Imputation: fill in missing value with mean of non-missing values
- **Multiple Imputation**: more complex but widely used method

2. Data Structure

- Data cleaning is often a tedious yet necessary first step before analysis

2. Data Structure

- Data cleaning is often a tedious yet necessary first step before analysis
- Do you have qualitative and quantitative variables correctly defined in your data?

2. Data Structure

- Data cleaning is often a tedious yet necessary first step before analysis
- Do you have qualitative and quantitative variables correctly defined in your data?
- Example: dichotomization

2. Data Structure

- Data cleaning is often a tedious yet necessary first step before analysis
- Do you have qualitative and quantitative variables correctly defined in your data?
- Example: dichotomization
    - Transform age into two groups: <60 and $>=60$

2. Data Structure

- Data cleaning is often a tedious yet necessary first step before analysis
- Do you have qualitative and quantitative variables correctly defined in your data?
- Example: dichotomization
    - Transform age into two groups: <60 and $>=60$
- Do you even have a response variable?

2. Data Structure

- Data cleaning is often a tedious yet necessary first step before analysis
- Do you have qualitative and quantitative variables correctly defined in your data?
- Example: dichotomization
    - Transform age into two groups: $<$60 and $>=$60
- Do you even have a response variable?
    - Supervised or Unsupervised Learning

Levels of Measurement

- **Nominal** scale: unordered categories

Levels of Measurement

- **Nominal** scale: unordered categories
    - e.g. continents, countries, fruits

Levels of Measurement

- **Nominal** scale: unordered categories
  - e.g. continents, countries, fruits
- **Ordinal** scale: ordered categories

Levels of Measurement

- **Nominal** scale: unordered categories
  - e.g. continents, countries, fruits
- **Ordinal** scale: ordered categories
  - e.g. Cup Sizes: Tall, Grande, Venti

Levels of Measurement

- **Nominal** scale: unordered categories
  - e.g. continents, countries, fruits
- **Ordinal** scale: ordered categories
  - e.g. Cup Sizes: Tall, Grande, Venti
  - e.g. Likert Scale: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree

Levels of Measurement

- **Nominal** scale: unordered categories
  - e.g. continents, countries, fruits
- **Ordinal** scale: ordered categories
  - e.g. Cup Sizes: Tall, Grande, Venti
  - e.g. Likert Scale: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
- **Interval** scale: quantitative numbers

Levels of Measurement

- **Nominal** scale: unordered categories
  - e.g. continents, countries, fruits
- **Ordinal** scale: ordered categories
  - e.g. Cup Sizes: Tall, Grande, Venti
  - e.g. Likert Scale: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
- **Interval** scale: quantitative numbers
  - e.g. Odds Ratio, Relative Risk, Age, BMI

3. Complexity

- How large is your dataset?

3. Complexity

- How large is your dataset?
- Ratio of observations to variables?

3. Complexity

- How large is your dataset?
- Ratio of observations to variables?
    - High-dimensional setting

3. Complexity

- How large is your dataset?
- Ratio of observations to variables?
  - High-dimensional setting
- How many response variables are we interested in?

3. Complexity

- How large is your dataset?
- Ratio of observations to variables?
    - High-dimensional setting
- How many response variables are we interested in?
- How many explanatory variables are we interested in?

Regression: One Response, One Predictor

- Simple Linear Regression (SLR)

Regression: One Response, One Predictor

- Simple Linear Regression (SLR)
  - Fits straight line through scatterplot

Regression: One Response, One Predictor

- Simple Linear Regression (SLR)
  - Fits straight line through scatterplot
  - Tests whether two quantitative variables have linear association

Regression: One Response, One Predictor

- Simple Linear Regression (SLR)
  - Fits straight line through scatterplot
  - Tests whether two quantitative variables have linear association
  - Can make transformations (e.g. log)

Regression: One Response, One Predictor

- Simple Linear Regression (SLR)
  - Fits straight line through scatterplot
  - Tests whether two quantitative variables have linear association
  - Can make transformations (e.g. log)
  - e.g. Are the **heights** of **male** students related to that of the **female** students in this class?

Regression: One Response, Many Predictors

- Multiple Linear Regression (MLR)

Regression: One Response, Many Predictors

- Multiple Linear Regression (MLR)
  - Fits hyperplane through multi-dimensional scatterplot

Regression: One Response, Many Predictors

- Multiple Linear Regression (MLR)
  - Fits hyperplane through multi-dimensional scatterplot
  - Tests whether a quantitative response has linear association with any numerical predictors

Regression: One Response, Many Predictors

- Multiple Linear Regression (MLR)
  - Fits hyperplane through multi-dimensional scatterplot
  - Tests whether a quantitative response has linear association with any numerical predictors
  - Can build interactions (e.g. Weight × Height)

Regression: One Response, Many Predictors

- Multiple Linear Regression (MLR)
  - Fits hyperplane through multi-dimensional scatterplot
  - Tests whether a quantitative response has linear association with any numerical predictors
  - Can build interactions (e.g. Weight × Height)
  - e.g. Are the **BMIs** associated with **height**, **weight**, or **age**?

ANOVA

- Test if k groups have the same mean

ANOVA

- Test if k groups have the same mean
  - if $k = 2$, this is a two sample t-test

ANOVA

- Test if k groups have the same mean
    - if k = 2, this is a two sample t-test
- Can test across one or more factors (One-Way or Multi-Way)

ANOVA

- Test if k groups have the same mean
  - if k = 2, this is a two sample t-test
- Can test across one or more factors (One-Way or Multi-Way)
  - e.g. Are the mean **weights** of men the same across **countries** (Canada, USA, UK) and **income** (Low, Medium, High)?

ANOVA

- Test if k groups have the same mean
  - if k $=$ 2, this is a two sample t-test
- Can test across one or more factors (One-Way or Multi-Way)
  - e.g. Are the mean **weights** of men the same across **countries** (Canada, USA, UK) and **income** (Low, Medium, High)?
  - 9 combinations

Linear Models

- Can solve many statistical problems

Linear Models

- Can solve many statistical problems
- Generalization of regression and ANOVA

Linear Models

- Can solve many statistical problems
- Generalization of regression and ANOVA
- Predictors can be a mix of quantitative and qualitative types

Linear Models

- Can solve many statistical problems
- Generalization of regression and ANOVA
- Predictors can be a mix of quantitative and qualitative types
- e.g. Are the **final exam scores** associated with **midterm scores**, **instructor**, or **gender**?

Log-Linear Models

- Generalization of

Log-Linear Models

- Generalization of
  - Poisson regression: model counts

Log-Linear Models

- Generalization of
    - Poisson regression: model counts
        - e.g. Number of cancer cases in a year

Log-Linear Models

- Generalization of
    - Poisson regression: model counts
        - e.g. Number of cancer cases in a year
    - Chi-squared tests: tests of independence

Log-Linear Models

- Generalization of
  - Poisson regression: model counts
    - e.g. Number of cancer cases in a year
  - Chi-squared tests: tests of independence
    - e.g. The **HIV test result** for a patient is independent of the **HIV status**.

Logistic Regression

- Binary response

Logistic Regression

- Binary response
- Log odds of outcome is modelled

Logistic Regression

- Binary response
- Log odds of outcome is modelled
- Multinomial logistic regression: more than two categories

Logistic Regression

- Binary response
- Log odds of outcome is modelled
- Multinomial logistic regression: more than two categories
- Ordinal logistic regression: categories have an intrinsic order

Logistic Regression

- Binary response
- Log odds of outcome is modelled
- Multinomial logistic regression: more than two categories
- Ordinal logistic regression: categories have an intrinsic order
- e.g. Model **presence of cancer cells** as a function of **mutations**

Generalized Linear Models

- Generalization of linear, log-linear, and logistic models

Generalized Linear Mixed Models

- GLMs which can handle mixed models

Generalized Linear Mixed Models

- GLMs which can handle mixed models
- Repeated measures designs

Generalized Linear Mixed Models

- GLMs which can handle mixed models
- Repeated measures designs
    - Matched pairs t-test

Generalized Linear Mixed Models

- GLMs which can handle mixed models
- Repeated measures designs
    - Matched pairs t-test
    - Longitudinal studies

Generalized Linear Mixed Models

- GLMs which can handle mixed models
- Repeated measures designs
  - Matched pairs t-test
  - Longitudinal studies
  - Spatial data

Generalized Linear Mixed Models

- GLMs which can handle mixed models
- Repeated measures designs
  - Matched pairs t-test
  - Longitudinal studies
  - Spatial data
- Models have **fixed** and **random** effects

Generalized Linear Mixed Models

- GLMs which can handle mixed models
- Repeated measures designs
  - Matched pairs t-test
  - Longitudinal studies
  - Spatial data
- Models have **fixed** and **random** effects
- e.g. Are **operation times** associated with the patient's **age** or **hospital**?

Time-To-Event Data

- Survival Analysis

Time-To-Event Data

- Survival Analysis
  - Parametric models (Exponential, Weibull)

Time-To-Event Data

- Survival Analysis
  - Parametric models (Exponential, Weibull)
  - Cox Proportional hazards model

Time-To-Event Data

- Survival Analysis
  - Parametric models (Exponential, Weibull)
  - Cox Proportional hazards model
  - Kaplan-Meier

Time-To-Event Data

- Survival Analysis
    - Parametric models (Exponential, Weibull)
    - Cox Proportional hazards model
    - Kaplan-Meier
- Important to understand **censoring**

Time-To-Event Data

- Survival Analysis
  - Parametric models (Exponential, Weibull)
  - Cox Proportional hazards model
  - Kaplan-Meier
- Important to understand **censoring**
- e.g. Is **breast cancer specific survival** associated with **gene A**?

Multivariate World

- Multiple responses

Multivariate World

- Multiple responses
- Non-statisticians often mistakenly refer to this as more than one *predictor*

Multivariate World

- Multiple responses
- Non-statisticians often mistakenly refer to this as more than one *predictor*
    - Multivariable means more than one predictor

Multivariate World

- Multiple responses
- Non-statisticians often mistakenly refer to this as more than one *predictor*
  - Multivariable means more than one predictor
- Everything we have talked about thus far have multivariate versions

Multivariate World

- Multiple responses
- Non-statisticians often mistakenly refer to this as more than one *predictor*
  - Multivariable means more than one predictor
- Everything we have talked about thus far have multivariate versions
- Ask if they are really interested in all the different responses

Multivariate World

- Multiple responses
- Non-statisticians often mistakenly refer to this as more than one *predictor*
    - Multivariable means more than one predictor
- Everything we have talked about thus far have multivariate versions
- Ask if they are really interested in all the different responses
- Difficult to visualize

Non-parametric Statistics

- Does not assume make any distributional assumptions (e.g. Kaplan-Meier estimator)

Non-parametric Statistics

- Does not assume make any distributional assumptions (e.g. Kaplan-Meier estimator)
- Use when assumptions for parametric models are violated (e.g. Normality, Constant variance, etc.)

Non-parametric Statistics

- Does not assume make any distributional assumptions (e.g. Kaplan-Meier estimator)
- Use when assumptions for parametric models are violated (e.g. Normality, Constant variance, etc.)
- More conservative p-values

Non-parametric Statistics

- Does not assume make any distributional assumptions (e.g. Kaplan-Meier estimator)
- Use when assumptions for parametric models are violated (e.g. Normality, Constant variance, etc.)
- More conservative p-values
- A number of parametric methods have non-parametric variants

Unsupervised Learning

- No response variable

Unsupervised Learning

- No response variable
- Usually for exploratory purposes

Unsupervised Learning

- No response variable
- Usually for exploratory purposes
- Examples:

Unsupervised Learning

- No response variable
- Usually for exploratory purposes
- Examples:
  - Principal Components Analysis

Unsupervised Learning

- No response variable
- Usually for exploratory purposes
- Examples:
  - Principal Components Analysis
  - Factor Analysis

Unsupervised Learning

- No response variable
- Usually for exploratory purposes
- Examples:
  - Principal Components Analysis
  - Factor Analysis
  - Cluster Analysis

Unsupervised Learning

- No response variable
- Usually for exploratory purposes
- Examples:
    - Principal Components Analysis
    - Factor Analysis
    - Cluster Analysis
        - e.g. Do these **tumour samples** cluster into clinically relevant biological subgroups?

Conclusion

- Learn to transform a research question into a statistical method

Conclusion

- Learn to transform a research question into a statistical method
- The “best” statistical method for your study is a balance between ease of application and complexity of data

Conclusion

- Learn to transform a research question into a statistical method
- The "best" statistical method for your study is a balance between ease of application and complexity of data
- Statisticians often try several different methods for a single problem and then compare results

Conclusion

- Learn to transform a research question into a statistical method
- The "best" statistical method for your study is a balance between ease of application and complexity of data
- Statisticians often try several different methods for a single problem and then compare results
- Make sure to document all work for reproducible research

Thank You!

Questions?