# *Data Roadmap*

*Tom Bruning*

*2018-04-11*

### *Learning Objectives*

In this chapter, you learn:

- The questions to ask when choosing which statistical methods to use to conduct data analysis
- Rules for applying statistics in future studies and analyses

### *Good Data Analysis Requires Choosing The Proper Technique(s)*

Choosing the proper technique(s) to use requires the consideration of:

- The purpose of the analysis
  - The type of variable being analyzed
    * Numerical
    * Categorical
- The assumptions about the variable ou are willing to make

### *Questions To Ask When Analyzing Numerical Variables*

Do you seek to:

- Describe the characteristics of the variable (possibly broken into several groups)
- Draw conclusions about the mean and standard deviation of the variable in a population
- Determine whether the mean and standard deviation of the variable differs depending on the group
- Determine which factors affect the value of the variable
- Predict the value of the variable based on the value of other variables
- Determine whether the values of the variable are stable over time

### *How to Describe the Characteristics of a Numerical Variable*

Develop tables and charts and compute descriptive statistics to describe the variable's characteristics:

- Tables and charts
  Stem-and-leaf display, percentage distribution, histogram, polygon, boxplot, normal probability plot

- Statistics
  Mean, median, mode, quartiles, range, interquartile range, standard deviation, variance, and coefficient of variation

*How to draw conclusions about the population mean or standard deviation*

- Confidence interval for the mean based on the t-distribution
- Hypothesis test for the mean (t-test)
- Hypothesis test for the standard deviation or variance ( $\chi^2$ test)

*How to determine whether the mean or standard deviation differs by group*

Two independent groups studying central tendency

- Normally distributed numerical variables
  - Pooled t-test if you can assume variances are equal
  - Separate-variance t-test if you cannot assume variances are equal
    Both tests assume the variables are normally distributed and you can examine this assumption by developing boxplots and normal probability plots
    To decide if the variances are equal you can conduct an F-test for the ratio of two variances
- Numerical variables not normally distributed
  Wilcoxon rank sum test
- Two groups of matched items or repeated measures studying central tendency
  - Paired differences normally distributed
    Paired t-test
- Two independent groups studying variability
  - Numerical variables normally distributed
    F-test
- Three or more independent groups and studying central tendency
  - Numerical variables normally distributed
    One Way Analysis of Variance

*How to determine which factors affect the value of the variable*

- Two factors to be examined
  - Two-factor factorial design

*How to predict the value of a variable based on the value of other variables*

- One independent variable
  - Simple linear regression model
- Two or more independent variables
  - Multiple regression model
  - Regression tree
  - Neural network
- Data taken over a period of time and you want to forecast future time periods
  - Moving averages
  - Exponential smoothing
  - Least-squares forecasting

*Questions To Ask When Analyzing Categorical Variables*

Do you seek to:

- Describe the proportion of items of interest in each category (possibly broken into several groups)
- Draw conclusions about the proportion of items of interest in a population
- Determine whether the proportion of items of interest differs depending on the group
- Predict the proportion of items of interest based on the value of other variables
- Determine whether the proportion of items of interest is stable over time

*How to describe the proportion of items of interest in each category*

- Summary tables
- Charts
  - Bar chart
  - Pie chart
  - Pareto chart
  - Side-by-side bar chart

*How to draw conclusions about the proportion of items of interest*

- Confidence interval for proportion of items of interest
- Hypothesis test for the proportion of items of interest (Z-test)

*How to determine whether the proportion of items of interest differs depending on the group*

Categorical variable has two categories

- Two independent groups
  - Two proportion Z-test
  - $\chi^2 - Test$ for the difference between two proportions
- More than two independent groups
  - $\chi^2 - Test$ for the difference among several proportions
    More than two categories and more than two groups
  - $\chi^2 - Test$ of independence

*How To Predict The Proportion Of Items Of Interest Based On The Value Of Other Variables*

- Logistic regression

*How to determine whether the proportion of items of interest is stable over time*

- Studying a process and data is taken over time
  - Collected items of interest over time