Comprehensive Review

DS705

Statistics in the Data Science Context

- Statistical methods were not designed with Big Data in mind
- If Big Data constitute a population, then inference procedures are not necessary, only descriptive statistics
- Good to be able to classify variables as quantitative or categorical
- Good to think about how variables may be related, to make comparisons
- Good to know which statistical methods are used to answer various types of questions

Statistics in the Data Science Context (Continued)

- While you likely won't be designing experiments, you will be collecting data in some way, so it is important to think about how that sampling is done and what population it might represent
- Concepts of estimation (confidence intervals) and comparison (hypothesis testing) are useful
- Hypothesis test mentality - formulate hypotheses, use data to draw conclusions
- Statistical applications in Data Science are more likely to be exploratory (data-driven) rather than confirmatory (model-driven)

Picking a Procedure

Variables		Procedures
Quantitative response; categorical explanatory	→	*t*; Wilcoxon Rank Sum; ANOVA; KW
Quantitative response; paired data	→	*t*; Wilcoxon Signed Rank; Sign
Quantitative response; quantitative explanatory or mix of quantitative and categorical explanatory	→	Linear regression
Binary categorical response; quantitative and/or categorical explanatory	→	Logistic regression
One categorical variable	→	Chi-square GOF; *z*-procedures for proportion
Two categorical variables	→	Chi-square test for Independence; logistic regression; *z* for difference of proportions
Large number of quantitative variables	→	Exploratory factor analysis

Comparing Population Central Values

	Independent		Dependent
	Equal variance	No equal var. assumption	Paired
Normal	Pooled *t* ANOVA Wilcoxon Rank Sum Kruskal-Wallis Bootstrap	Welch *t* Bootstrap	Paired *t* Wilcoxon Signed Rank Sign Test Bootstrap
Non-normal, but same shape	Wilcoxon Rank Sum Kruskal-Wallis Bootstrap	Bootstrap	Wilcoxon Signed Rank (if sym) Sign Test Bootstrap

Using Statistical Procedures Wisely

- Graphs and common sense go a long way
- Don't apply inferences beyond the population that a given sample represents
- Keep in mind how a very small or very large sample size may affect results
- Sometimes more than one procedure can be correctly and effectively used for the same situation
- Be clear about what the variables are actually measuring
- Be sure to understand and evaluate the data conditions for the procedures you use
  - If you have to assume any data conditions, be sure it is reasonable to do so
  - Normality conditions often can be relaxed a little
  - Independence conditions must be met