**Proposal for a New Edition**

**Title:** Statistical Inference via Data Science: A ModernDive into R and the Tidyverse

**Authors:** Chester Ismay, Albert Y. Kim, plus perhaps a third author, for example a Smith

College undergraduate, who would lay down code/groundwork for proposed changes.

**List of major potential changes and additions:**

1. **(Exists online, not in 1st ed)** Chapter 6 Multiple Regression: Augmented "6.3.1 Model selection" to emphasize both visualization/exploratory data analysis and numerical approaches to model selection. New format:
   1. "6.3.1 Model selection using visualizations"
   2. Added "6.3.2 Model selection using R-squared"
2. **(Exists online, not in 1st ed)** Chapter 7 Sampling: In order to make even clearer the chief definitions and concepts behind random sampling, which form the basis of statistical inference:
   1. In "7.3.1 Terminology & notation": we clustered definitions according to theme and connected back to sampling exercises
   2. In "7.3.2 Statistical definitions", we clarify at times confusing definitions even further
   3. Moved "7.5.2 Central Limit Theorem" to its own section, to make it more prominent and not just an after-thought
   4. Created a new "7.6.2 Theory-based standard errors" which split the original "8.7.2 Theory-based confidence intervals" into two parts and moved the earlier part to Chapter 7 Sampling.

That way all 4 statistical inference chapters (Ch 7-11) each have their own "theory-based X" subsection at the end, emphasizing the link between simulation based and traditional methods to inference.

1. **(New)** In either (or potentially both) Chapter 6.1 on Multiple Regression with 1 numerical and 1 categorical variable, or in Chapters 9-9.4 on Hypothesis Testing, replace dataset that uses gender as a binary variable.
   1. For example, consider using palmerpenguins dataset, which itself was popularized as an alternative to commonly-used iris dataset created by eugenicist Ronald Fisher.
   2. We would make a point of mentioning the above fact in textbook to emphasize that the field of statistics has its own share of sordid histories.
2. **(New)** Chapter 11 that picks up on use of infer package for inference for regression in Chapter 10. This chapter would preview the full potential of the infer package for simulation-based inference for
   1. Multiple regression
   2. Logistic regression

**List of minor potential changes and additions:**

1. **(Exists online, not in 1st ed)** Chapter 2 Data Visualization: Remove soft introduction to %>% operator (from Ch 3 Data Wrangling) since this only confused readers. Instead we now use preprocessed alaska\_flights and early\_january\_weather data frames loaded from moderndive package (v0.5.3).

**Questionnaire:**

|  |
| --- |
| **1) What is your general opinion of the book? What are its main strengths and weaknesses? Please give details.** |
|  |
| **2) Please explain how you have used the book for teaching in the past. What was the course title, and what students were on it? Did you use the whole book or selected chapters? Did you supplement with any other materials?** |
|  |
| **3) Please suggest any specific improvements to the book – material that could be added, revised, restructured, or deleted.** |
|  |
| **4) Do you have any comments on the proposed revisions outlined by the authors above?** |
|  |
| **5) Do you believe there is scope for a new edition of the book? Would you, for example, use the new edition for your teaching?** |
|  |
| **6) What are the key related and/or competing titles? How do they compare with this book?** |
|  |
| **7) Do you have any additional comments to help improve the book?** |
|  |