Chapter 1. Comparisons of Batches

Note. In this class on Applied Multivariate Statistical Analysis (STAT 5730) we consider data in several dimensions. That is, we consider a set of data

$$\{x_i\}_{i=1}^n = \{x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p \mid i = 1, 2, \ldots, n\}$$

where x_i is an observed value of a random variable $X \in \mathbb{R}^p$. So random variable X is composed of p random variables $X_1, X_2, \ldots, X_p$ and $X = (X_1, X_2, \ldots, X_p)$ where each X_j for $j = 1, 2, \ldots, p$ is a one-dimensional random variable. For a discussion of random variables, see my online notes for Mathematical Statistics 1 (STAT 4047/5057) on Section 1.5. Random Variables and Section 2.6. Extension to Several Random Variables.

Note. In describing multivariate data some questions we can address, according to Hädle and Simar are (quoting from page 2):

- Are there components of X that are more spread out than others?
- Are there some elements of X that indicate sub-groups of the data?
- Are there outliers in the components of X?
- How "normal" is the distribution of the data?
- Are there "low-dimensional" linear combinations of X that show "non-normal" behavior?

When dealing with a single random variable, we can visualize data using histograms, stem plots, or box plots (see online notes for Introduction to Probability

and Statistics [MATH 1530] on Chapter 1. Picturing Distributions with Graphs and Chapter 2. Describing Distributions with Numbers).

Note. In this chapter, we consider visual presentations of date from random variables. We start by reviewing the tools mentioned above for a single ("univariate") random variable (in Sections 1.1 to 1.4) and then consider techniques for visualizing several ("multivariate") random variables such as Chernoff-Flury faces (Section 1.5), Andrew's curves (Section 1.6), and parallel coordinate plots (Section 1.7). We end this chapter with a discussion of Boston Housing data which we will consider again in following chapters.

Revised: 4/25/2021