University of Waterloo

STAT 341 - Computational Statistics and Data Analysis

Winter 2024

Personal Course Notes

Brandon Zhou

BASIC INFO

Author	Brandon Zhou
Course Code	STAT 341
Course Name	Computational Statistics and Data Analysis
Course Instructor	Michael John Davis
Room	STC 0050
Days & Times	TTh 2:30PM - 3:50PM
Section	001
Date Created	January 05, 2024
Last Modified	January 10, 2024
Final Exam Date	TBA

DISCLAIMER

These course notes are intended to supplement primary instructional materials and facilitate learning. It's worth mentioning that some sections of these notes might have been influenced by ChatGPT, an OpenAI product. Segments sourced or influenced by ChatGPT, where present, will be clearly indicated for reference.

While I have made diligent efforts to ensure the accuracy of the content, there is a potential for errors, outdated information, or inaccuracies, especially in sections sourced from ChatGPT. I make no warranties regarding the completeness, reliability or accuracy of the notes contained in this notebook. It's crucial to view these notes as a supplementary reference and not a primary source.

Should any uncertainties or ambiguities arise from the material, I strongly advise consulting with your course instructors or the relevant course staff for comprehensive understanding. I apologize for any potential discrepancies or oversights.

Any alterations or modifications made to this notebook after its initial creation are neither endorsed nor recognized by me. For any doubts, always cross-reference with trusted academic resources.

TABLE OF CONTENTS

CHAPTER 1: Introduction

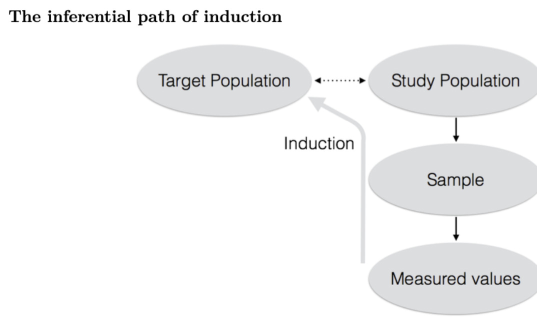

Figure 1: The inferential path of induction

CHAPTER 2: Populations

2.1 Populations

Definition 2.1

Here we aim to describe a population using attributes.

- A population is a finite (though possibly huge) set $\mathcal{P}$ of elements.
  - Elements of a population are called units $u \in \mathcal{P}$
  - Variates are functions $x(u), y(u)$, etc. on the individual units $u \in \mathcal{P}$. For simplicity we will more often use the notation x_u, y_u, etc. when referring to the realized values of these variates for the unit $u = 1, \ldots, N$.
- We will define and explore interesting population attributes, denoted generally as $a(\mathcal{P})$.

2.2 Explicitly Defined Population Attributes

2.2.1 Population Attributes

Definition 2.2

Some definitions we need to know:

- The population is typically a set or collection of units, each with one or more variates that we can measure.
- Variates are characteristics of each unit in the population, and they can take on numerical or categorical values.
  - The values of variates typically differ from unit to unit.
  - If we are only interested in the variate y's we might write the population as

$$\mathcal{P} = \{y_1, y_2, \ldots, y_N\}$$

- Population attributes are summaries describing characteristics of the population.
  - Formally, an attribute is a function applied to the entire population and determined by the variate values observed for each of the population's units.

$$\mathcal{P} = f(y_1, y_2, \ldots, y_N)$$

- Some examples of attributes are
  - the population total:

$$a(\mathcal{P}) = \sum_{u \in \mathcal{P}} y_u$$

- or various counts over the population

$$a(\mathcal{P}) = \sum_{u \in \mathcal{P}} I_A(y_u)$$

where $I_A(y)$ is the indicator function

$$I_A(y) = \begin{cases} 1 & \text{if } y \in A \\ 0 & \text{if } y \notin A \end{cases}$$

Definition 2.3

Location Attributes measure or describe the centre of the distribution of variate values in a dataset.

- the population average:

$$a(\mathcal{P}) = \bar{y} = \frac{1}{N} \sum_{u \in \mathcal{P}} y_u$$

- the population proportion:

$$a(\mathcal{P}) = \frac{1}{N} \sum_{u \in \mathcal{P}} I_A(y_u)$$

- Other examples include the mode, the median, etc.

Spread Attributes measure variability or spread of the variate values in a data set. Some are

- the population variance:

$$a(\mathcal{P}) = \frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \bar{y})^2$$

- the population standard deviation:

$$a(\mathcal{P}) = SD_{\mathcal{P}}(y) = \sqrt{\frac{1}{N} \sum_{u \in \mathcal{P}} (y_u - \bar{y})^2}$$

- coefficient of variation:

$$a(\mathcal{P}) = \frac{SD_{\mathcal{P}}(y)}{\bar{y}}$$

- *Note*: the population variance or standard deviation could also be defined using $N - 1$ in the denominator.
- Other examples include the range, the inter-quartile range, etc.

Order Statistics

- Population attributes can also be based on an indexed collection of values,

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}$$

which are the variate values $y_u \in \mathcal{P}$ ordered from smallest to largest (including ties).

Location Attributes based on Order Statistics

These attributes measure or describe the centre of the distribution of variate values in a data set.

- the population minimum:
$$a(\mathcal{P}) = \min_{u \in \mathcal{P}} y_u = y_{(1)}$$

- the population maximum:
$$a(\mathcal{P}) = \max_{u \in \mathcal{P}} y_u = y_{(N)}$$

- the population mid-range:
$$a(\mathcal{P}) = \frac{1}{2}\left[\min_{u \in \mathcal{P}} y_u + \max_{u \in \mathcal{P}} y_u\right] = \frac{y_{(1)} + y_{(N)}}{2}$$

- the population median:
$$a(\mathcal{P}) = \text{median}_{u \in \mathcal{P}} y_u = \begin{cases} y_{\left(\frac{N+1}{2}\right)}, & \text{if } N \text{ is odd} \\ \frac{y_{\left(\frac{N}{2}\right)} + y_{\left(\frac{N}{2}+1\right)}}{2}, & \text{if } N \text{ is even} \end{cases}$$

- the population quartiles:
  - Q_1 is 25^{th} percentile, or the first quartile,
  - Q_2 is 50^{th} percentile, or the median, and
  - Q_3 is 75^{th} percentile, or the third quartile.

Variability Attributes based on Order Statistics

- The population range:
$$a(\mathcal{P}) = \max_{u \in \mathcal{P}} y_u - \min_{u \in \mathcal{P}} y_u = y_{(N)} - y_{(1)}$$

- The population inter-quartile range IQR:
$$a(\mathcal{P}) = Q_3 - Q_1$$
where Q_1 and Q_3 are 25^{th} and 75^{th} percentiles or the first and third quartiles, as above.

- The Median Absolute Deviation (MAD) is the median of the absolute differences between each y_u and the median:
$$a(\mathcal{P}) = \text{median}_{u \in \mathcal{P}} \left| y_u - \text{median}_{u \in \mathcal{P}} y_u \right|$$

Skewness Attributes

These are measures of asymmetry in a population. A symmetric distribution of population values should result in a skewness attribute of zero.

- Pearson's moment coefficient of Skewness:

$$a(\mathcal{P}) = \frac{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \bar{y})^3}{[S_{D_\mathcal{P}}(y)]^3}$$

- Pearson's second skewness coefficient (median skewness) given by:

$$a(\mathcal{P}) = \frac{3 \times (\bar{y} - \text{median}_{u\in\mathcal{P}} y_u)}{S_{D_\mathcal{P}}(y)}$$

- Bowley's measure of skewness based on the quartiles:

$$a(\mathcal{P}) = \frac{(Q_3 + Q_1)/2 - Q_2}{(Q_3 - Q_1)/2}$$

The above content is for Lecture 1 on Jan 9, 2024

```r
# R code goes here
summary(cars)  # for example, to summarize the 'cars' dataset
```