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
Days & Times	TTh 2:30PM - 3:50PM
Section	001
Date Created	January 05, 2024
Last Modified	January 16, 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While I have made diligent efforts to ensure the accuracy of the content, there is a potential for errors, outdated information, or inaccuracies, especially in sections sourced from ChatGPT.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

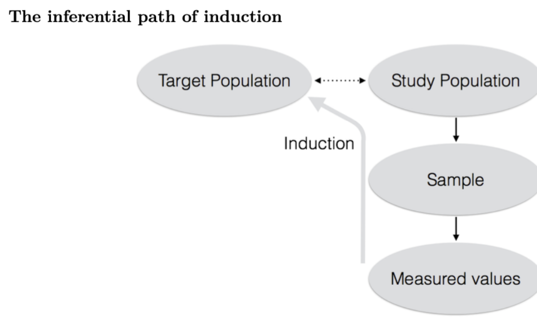

Figure 1: The inferential path of induction

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

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.

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

- 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

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

In general, attributes can be numerical or graphical – as long as they summarize the whole population.

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. Notice these are functions of entire population.
- 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:

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

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

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

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

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

NAs in R: Note that many programs in R accommodate missing data (represenated as `NA`s) and do something appropriate (typically they omit them).

- For your own code and analyses, you either need to decide what to do with `NA`s or ensure that the data do not have any `NA`s.
- If you choose to simply omit `NA`s, for example, the function `na.omit(...)` may be helpful (it will remove rows which contain an `NA` from a data set). For other possibilities see `help("na.omit")` in R.

2.2.2 Attribute Properties

Definition 2.4
A population attribute is a function of measured variates y_u:

$$a(\mathcal{P}) = f_y, y_2, \ldots, y_N$$

and the variates y_u are typically associated with some measurement units.

Definition 2.5
Location Invariance and Equivariance

For an attribute $a(\mathcal{P}) = a(y_1, \ldots, y_N)$ we say that for any $m > 0$ and $b \in \mathbb{R}$, that the attribute is

- location invariant if

$$a(y_1 + b, \ldots, y_N + b) = a(y_1, \ldots, y_N)$$

- location equivariant if

$$a(y_1 + b, \ldots, y_N + b) = a(y_1, \ldots, y_N) + b$$

Example 2.1

The population average is location equivariant:

$$\begin{aligned} a(\mathcal{P}) = a(y_1, y_2, \ldots, y_N) &= \frac{1}{N} \sum_{i=1}^{N} y_i \\ a(y_1 + b, y_2 + b, \ldots, y_N + b) &= \frac{1}{N} \sum_{i=1}^{N} (y_i + b) \\ &= \frac{1}{N} \sum_{i=1}^{N} y_i + \frac{Nb}{N} = a(\mathcal{P}) + b \end{aligned}$$

But is the population variance location equivariant? No!

Definition 2.6

Scale Invariance and Equivariance

For an attribute $a(\mathcal{P}) = a(y_1, \ldots, y_N)$ we say that for any $m > 0$ and $b \in \mathbb{R}$, that the attribute is

- scale invariant if

$$a(m \times y_1, \ldots, m \times y_N) = a(y_1, \ldots, y_N)$$

- scale equivariant if

$$a(m \times y_1, \ldots, m \times y_N) = m \times a(y_1, \ldots, y_N)$$

- location-scale invariant if it is both location invariant and scale invariant, i.e.

$$a(m \times y_1 + b, \ldots, m \times y_N + b) = a(y_1, \ldots, y_N)$$

- location-scale equivariant if it is both location equivariant and scale equivariant, i.e.

$$a(m \times y_1 + b, \ldots, m \times y_N + b) = m \times a(y_1, \ldots, y_N) + b$$

Example 2.2

The population average is location-scale equivariant

$$\begin{aligned} a(my_1+b, my_2+b, \ldots, my_N+b) &= \frac{1}{N}\sum_{i=1}^{N}(my_i+b) \\ &= \frac{m}{N}\sum_{i=1}^{N} y_i + \frac{Nb}{N} \\ &= ma(\mathcal{P})+b \end{aligned}$$

Definition 2.7
Replication

Another invariance/equivariance property of interest for population attributes is replication invariance and replication equivariance.

If a population $\mathcal{P}$ is duplicated $k-1$ times (so that there are k copies of it), how does the attribute change on this new population denoted by $\mathcal{P}^k$?

$$\mathcal{P}^k = \{y_1, y_2, \ldots, y_N, y_1, y_2, \ldots, y_N, \ldots, y_1, y_2, \ldots, y_N\} = \{\underbrace{x_1, x_2, \ldots, x_{kN}}_{kN \text{ elements}}\}$$

The attribute $a(\mathcal{P})$ is

- replication invariant whenever $a(\mathcal{P}^k) = a(\mathcal{P})$
- replication equivariant whenever $a(\mathcal{P}^k) = k \times a(\mathcal{P})$

Example 2.3
The population average is replication invariant.

$$a(\mathcal{P}^k) = \frac{1}{kN}\sum_{j=1}^{kN} y_j = \frac{1}{kN}\sum_{i=1}^{N} ky_i = \frac{1}{N}\sum_{i=1}^{N} y_i = a(\mathcal{P})$$

2.2.3 Influence, Sensitivity Curves, and Breakdown Points

Definition 2.8
Influence(outlier detection)

- If we remove variate y_u (i.e. remove unit u) then the influence of that variate on the population attribute is quantified by

$$\Delta(a,u) = a(\underbrace{y_1, \ldots, y_{u-1}, y_u, y_{u+1}, \ldots, y_N}_{\textit{population with the unit } u}) - a(\underbrace{y_1, \ldots, y_{u-1}, y_{u+1}, \ldots, y_N}_{\textit{population without the unit } u})$$

- Ideally, no single unit's value should have greater influence than any other.

- If a unit has larger influence than the rest;
  1. it would require further investigation as it might be in error, or
  2. it might be the most interesting unit in the population.

The population average, $a(y_1, y_2, \ldots, y_n) = \bar{y}$ and the average without unit u can be written as

$$a(y_1, \ldots, y_{u-1}, y_{u+1}, \ldots, y_N) = \frac{1}{N-1} \sum_{\substack{k \in \mathcal{P}, \\ k \neq u}} y_k = \frac{\sum_{k \in \mathcal{P}} y_k - y_u}{N-1} = \frac{N\bar{y} - y_u}{N-1}$$

and $\Delta(a, u)$, the influence for a given u, is:

$$\Delta(a, u) = \bar{y} - \frac{N\bar{y} - y_u}{N-1} = \frac{(N-1)\bar{y} - (N\bar{y} - y_u)}{N-1} = \frac{y_u - \bar{y}}{N-1}$$

The above content is for Lecture 2 on Jan 11, 2024

Definition 2.9
Sensitivity Curve

- We can also examine the effect on an attribute when we add a variate. To examine this effect,
  - suppose we have a population of size $N - 1$ and
  - add a variate with the value y.
  - Then our new population with N elements is $\{y_1, \ldots, y_{N-1}, y\}$.
- We define the *sensitivity curve* of an attribute as

$$\begin{aligned} SC(y; a(\mathcal{P})) &= \frac{a(y_1, \ldots, y_{N-1}, y) - a(y_1, \ldots, y_{N-1})}{\frac{1}{N}} \\ &= N\left[a(y_1, \ldots, y_{N-1}, y) - a(y_1, \ldots, y_{N-1})\right] \end{aligned}$$

- We can then plot the *sensitivity curve* as a function of the new variate value y.
  - the sensitivity curve gives a scaled measure of the effect that a single variate value y has on the value of a population attribute $a(\mathcal{P})$.
- We can explore the sensitivity curve for any attribute. These can be determined *mathematically* in general, but can also be determined *computationally* for any particular population and any particular attribute.

The following is a general-purpose sensitivity curve function in R which accommodates any population and any attribute:

```
sc = function(y.pop, y, attr, ...) {
    N = length(y.pop) + 1
```

```
    sapply(y, function(y.new) { N * (attr(c(y.new, y.pop), ...) - attr(y.pop,
    ...)) })
}
# ... means "carry through any additional arguments".
```

Example 2.4

Derive the sensitivity curve for Arithmetic Mean

$$a(y_1, \ldots, y_N) = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$$

$$\begin{aligned}
P &= \{y_1, \ldots, y_{N-1}\} \\
P^* &= \{y_1, \ldots, y_{N-1}, y\} \\
a(P) &= \frac{1}{N-1}\sum_{i=1}^{N-1} y_i = \overline{y}_{N-1} \\
a(P^*) &= \frac{1}{N}\left[\sum_{i=1}^{N-1} y_i + y\right] \\
&= \frac{(N-1)\overline{y}_{N-1} + y}{N} \\
\therefore \mathrm{SC}(y, a) &= N\left[a(P^*) - a(P)\right] \\
&= N\left[\frac{(N-1)\overline{y}_{N-1} + y}{N} - \overline{y}_{N-1}\right] \\
&= (N-1)\overline{y}_{N-1} + y - N\overline{y}_{N-1} \\
&= y - \overline{y}_{N-1}
\end{aligned}$$

Notes:

- A single observation can change the average by a huge (even infinite) amount.
- Averages may not be the best choice for a population attribute representing the location of a population – particularly if extreme values exist in the population.

Example 2.5

Derive the sensitivity curve for maximum

$$a(y_1, \ldots, y_N) = \max\{y_1, \ldots, y_N\} = y_{(N)}$$

$$
\begin{aligned}
&P = \{y_1, \dots, y_{N-1}\} \\
&P^* = \{y_1, \dots, y_{N-1}, y\} \\
&a(P) = y_{(N-1)} \\
&a(P^*) = \begin{cases} y_{(N-1)} & \text{if } y \leq y_{(N-1)} \\ y & \text{if } y > y_{(N-1)} \end{cases} \\
&\therefore \mathrm{SC}(y, \alpha) = N\left[a(P^*) - a(P)\right] \\
&\qquad\qquad = \begin{cases} 0 & \text{if } y \leq y_{(N-1)} \\ N[y - y_{(N-1)}] & \text{if } y > y_{(N-1)} \end{cases}
\end{aligned}
$$

If we draw the sensitivity curve for the maximum, we would find out it is unbounded for large y, the maximum is very sensitive to large outliers.

Example 2.6

Derive the sensitivity curve for 2^{nd} Order Statistic

$$a(y_1, \dots, y_N) = y(2)$$

$$
\begin{aligned}
&P = \{y_1, \dots, y_{N-1}\} \\
&a(P) = y(2) \\
&P^* = \{y_1, \dots, y_{N-1}, y\} \\
&a(P^*) = \begin{cases} y_{(1)} & \text{if } y < y_{(1)} \\ y & \text{if } y_{(1)} \leq y < y_{(2)} \\ y_{(2)} & \text{if } y \geq y_{(2)} \end{cases} \\
&\therefore \mathrm{SC}(y, a) = N\left[a(P^*) - a(P)\right] \\
&\qquad\qquad = \begin{cases} N(y_{(1)} - y_{(2)}) & \text{if } y < y_{(1)} \\ N(y - y_{(2)}) & \text{if } y_{(1)} \leq y < y_{(2)} \\ 0 & \text{if } y \geq y_{(2)} \end{cases}
\end{aligned}
$$

Definition 2.10
Breakdown Points

Another measure of robustness that exists is called the breakdown point.

- It gives an assessment of just how large a proportion of the data must be contaminated before the statistic breaks down (and becomes useless).
- The breakdown point of a statistic is the smallest possible fraction of the observations that can be changed to something very extreme (i.e., plus or minus infinity) to make the error large (infinite)

- e.g. the break-point for
  - the average is $1/N$ (or asymptotically zero), and
  - the median is $1/2$ (i.e., that is half of the data has to go to infinity before the median breaks down).
- Attributes with high breakdown points are called resistant or robust.

2.2.4 Graphical Attributes

Population attributes can also be entirely graphical as in

- histograms of y_u values (univariate graphical summaries)
- bar plots of y_u values (univariate graphical summaries)
- box plots of y_u values (univariate graphical summaries)
- scatter-plots of pairs (x_u, y_u) (bivariate graphical summaries)
- scatter-plots of quantiles and ranks of y_u (bivariate graphical summaries)

Each of these plots summarizes the entire population, and so they're all attributes.

Histograms

Consider the population $\mathcal{P} = \{y_1, y_2, \ldots, y_N\}$.

- Partition the range of the population into k non-overlapping intervals, called bins, $I_j = [a_{j-1}, a_j)$, for $j = 1, 2, \ldots, k$ and then calculate the number (frequency) or proportion (relative frequency) of observations in the jth bin for $j = 1, \ldots, k$.
- Histograms help determine how the values are concentrated.

We can define bins two ways:

- bins of equal size, or (most common)
- bins with equal number of elements but varying size. ("equal area" histogram)
- Below are some examples of histograms with equal-sized bins (top row) and bins of varying sizes (bottom row)

```
x = agpop$farms87
par(mfrow=c(2,3), mar=2.5*c(1,1,1,0.1))
rx = range(x)
hist(x, breaks=seq(rx[1], rx[2], length.out=4), prob=TRUE, main="3 Bins", col
    = "grey")
hist(x, breaks=seq(rx[1], rx[2], length.out=5), prob=TRUE, main="4 Bins", col
    = "grey")
hist(x, breaks=seq(rx[1], rx[2], length.out=16), prob=TRUE, main="15 Bins",
    col = "grey")

# For the histograms in the bottom row, the areas of all rectangles in each
```

```
    panel are the same.
hist(x, breaks=quantile(x, p=seq(0, 1, length.out=4)), prob=TRUE, main="3 Bins
    ", col = "grey")
hist(x, breaks=quantile(x, p=seq(0, 1, length.out=5)), prob=TRUE, main="4 Bins
    " , col = "grey")
hist(x, breaks=quantile(x, p=seq(0, 1, length.out=16)), prob=TRUE, main="15
    Bins", col = "grey")
```

The bins with equal numbers of elements but varying size can help identify asymmetry in the popula- tion.

Rules for the Number of Bins

- Sturges rule:

$$\text{the number of bins should be } = \lceil \log_2(N) + 1 \rceil$$

- Freedman–Diaconis rule:

$$\text{Bin size } = 2\frac{\text{IQR}(x)}{N^{1/3}}$$

- Scott's rule:

$$\text{Bin size } = 3.5\frac{\sigma}{N^{1/3}}$$

Histograms using different rules for bin size selection:

- the first row is `Number of farms` and
- the second row is `log(Number of farms+1)`.

Question: Which scale would you prefer to work with? The original scale or the transformed scale?

Answer: Advantages

- Raw data: data values are easily interpretable
- Transformed data: symmetric data are often easier to work with, statistically speaking

Scatter-plots

- A scatter-plot is a plot of the points (x_u, y_u) for all units in the population.
  - It is used to see whether two variates x and y are related in some way.
- A scatter-plot of the number of farms and total acreage of farming in 1987 by US county is below.

```
par(mfrow=c(1,2))
plot(agpop$farms87, agpop$acres87, pch = 19, cex=0.5, col=adjustcolor("black",
     alpha = 0.3), xlab = "Number of farms", ylab = "Total acreage of farming"
    , main = "US counties 1987")

plot(agpop$acres87, agpop$farms87, pch = 19, cex=0.5, col=adjustcolor("black",
     alpha = 0.3), ylab = "Number of farms", xlab = "Total acreage of farming"
    , main = "US counties 1987")
```

- Sometimes, the scatter-plot of a transformed version of the data provides more insight.

```
par(mfrow=c(1,2))
plot(log(agpop$farms87+1), log(agpop$acres87+1), pch = 19, cex=0.5, col=
    adjustcolor("black", alpha = 0.3), xlab = "log(Number of farms + 1)", ylab
     = "log(Total acreage of farming + 1)", main = "US counties 1987")
plot(log(agpop$acres87+1),log(agpop$farms87+1), pch = 19, cex=0.5, col=
    adjustcolor("black", alpha = 0.3), ylab = "log(Number of farms + 1)", xlab
     = "log(Total acreage of farming + 1)", main = "US counties 1987")
```

The above content is for Lecture 3 on Jan 16, 2024
