Data summaries

Data Science Team

April 16, 2018
Admin Items

- Midterm moved from Monday April 30 to Friday May 4 (still in class, no laptop)
- p-set solutions have been posted to Canvas
- labs will Wednesday and Friday this week
Goals of the module

- Data has many variables and many observations
- In order to understand what is going on, it is often useful to boil down this complexity to one or two numbers
- We will lose some complexity, but it helps us with a “gestalt” of the data and with comparisons
Example questions

- How do different nations compare in terms of their income distribution?
- How diverse is the genetic make-up of different populations?
Interest rates for US treasury security, constant maturity

Federal Reserve Data

The graph shows the yield for various US treasury securities with constant maturity over the years 1960 to 2000.
Figure 1: Income distribution in USA 2015. 
Genotypes of individuals from 52 human populations

**Human Genome Diversity Panel**

---

We will focus on *univariate* data: our analysis will be of one variable at a time.

When summarizing we may choose *different aspects*: center, spread, asymmetry.

We talk about an *index* when we can standardize a summary so that its range is between fixed values (ex. [0,1], [-1,1]).
Summaries of Center

- Imagine the data as a cloud and you have to decide where to place it
- Focus on quantitative variables
- From K-12 education: Mean, Mode, Median
- Some further options: weighted average, trimmed mean, etc..
- Let’s re-think the objectives we have in mind when we calculate a summary of center and see if this gives us new ideas.
Properties for an *average*

*average*: a number $\bar{x}$ that substitutes the entire data $x_1, x_2, \ldots, x_n$ for one variable

1. **Internality:** $\min_i x_i \leq \bar{x} \leq \max_i x_i$

2. **Monotonicity:** if $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are such that $x_i \leq y_i \ \forall i \rightarrow \bar{x} \leq \bar{y}$.

3. **Symmetry:** $\text{average}(x_1, x_2, \ldots, x_n) = \text{average}(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})$, with $\pi$ a permutation of the indexes

4. **Associativity:** Let $\bar{x}_{kl} = \text{average}(x_k, x_{k+1}, \ldots, x_l)$, then $\text{average}(x_1, x_2, \ldots, x_k, x_{k+1}, \ldots, x_l, x_{l+1}, \ldots, x_n) = \text{average}(x_1, x_2, \ldots, \bar{x}_{kl}, \bar{x}_{kl}, \ldots, \bar{x}_{kl}, x_{l+1}, \ldots, x_n)$
One definition of *average*

Let $x_1, x_2, \ldots, x_n$ be the data, and $f(x_1, x_2, \ldots, x_n)$ a real-valued function of it that we are interested in. Then, it makes sense to consider the average $\bar{x}$ such that

$$f(x_1, x_2, \ldots, x_n) = f(\bar{x}, \bar{x}, \ldots, \bar{x})$$

- We can substitute the one number $\bar{x}$ in place of all the values for the variable and get the same result.
- Different functions $f$ are desirable based on various purpose and lead to different *averages* (arithmetic mean, geometric mean).
The average interest rate

- Let $r_1, r_2, \ldots, r_{12}$ the monthly interest rates that a product earned during the months 1, \ldots, 12
- What do we want from an average interest rate?
- We would like one value $\bar{r}$ so that if we substitute it to the 12 different values, we obtain the same income
- $\text{Income}(r_1, r_2, \ldots, r_{12}) = \text{Income}(\bar{r}, \bar{r}, \ldots, \bar{r})$

\[
(1 + r_1)(1 + r_2)\cdots(1 + r_{12}) = (1 + \bar{r})^{12}
\]

\[
\prod_{i=1}^{12}(1 + r_i) = (1 + \bar{r})^{12}
\]

\[
\bar{r} = \left(\prod_{i=1}^{12}(1 + r_i)\right)^{1/12} - 1
\]

- We find that for interest rates the geometric mean of $(1 + r_i)$ is a meaningful summary.
Average speed

Suppose that you have a car and you drive a distance \( d_i \) at speed \( v_i \) for \( i = 1, \ldots, n \).

In looking for an average speed, it might make sense to look for a speed such that travelling at that constant speed, you would cover the same total distance in the same total time.

\[
\text{Time} = \sum_{i=1}^{n} \frac{d_i}{v_i} = \sum_{i=1}^{n} \frac{d_i}{\bar{v}}
\]

Solving for \( \bar{v} \),

\[
\bar{v} = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} \frac{d_i}{v_i}}
\]

We find that the harmonic mean (here with weights \( d_i \)) is a meaningful summary.
Another approach to defining an average

- Maybe it is not possible to find one number $\bar{x}$ such that $f(x_1, x_2, \ldots, x_n) = f(\bar{x}, \bar{x}, \ldots, \bar{x})$
- Maybe we want something a bit more general, that it is suitable for multiple purposes

Let $g(z, x_1, x_2, \ldots, x_n)$ a function that describes the loss we incur when substituting $x_1, x_2, \ldots, x_n$ with $z$, then

$$\text{average}(x_1, x_2, \ldots, x_n) = \bar{x} = \arg\min_z g(z, x_1, x_2, \ldots, x_n)$$

- Ex. 1 $g(z, x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} (x_i - z)^2$
- Ex. 2 $g(z, x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} |x_i - z|$
- Ex. 3 $g(z, x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} 1(x_i \neq z)$
Different loss functions

Let's look at the value of the loss functions for one datapoint $x_i = 0$ as a function of $z$. 

![Graph showing different loss functions for a single datapoint.](image)
Different loss functions

- This approach to defining average says that the average is our “best guess” for a value in the data, and the different loss functions specify how we evaluate the goodness of a guess.
- The square loss penalizes more large discrepancies and downweights small discrepancies.
- The absolute loss considers all discrepancies at their face values.
- The 0-1 loss considers all discrepancies to be the same, with the exception of no error.
- What are the implications of this in terms of robustness to outliers?
Which average does the 0-1 loss give rise to?

Let’s calculate

$$\sum_{i=1}^{n} 1(x_i \neq z) = \#\{x_i \neq z\} = n - \#\{x_i = z\}$$

So, to find the $z$ that minimizes this loss we need to look for the $z$ such that the number of $x_i$ equal to $z$ is maximal.

This is called the **mode** and it is useful as a measure of the “center” for qualitative variables.
Other averages

- The square error loss gives rise to the **arithmetic mean**, which is often what we mean when we say “average”

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

- The absolute loss gives rise to the **median**.

\[
\text{sort}(x) = (x_{(1)}, x_{(2)}, \ldots, x_{(11)})
\]

\[
\text{median} = x_{(6)}
\]