Measures and indexes of dispersion

Data Science 101 Team
Measures of spread/dispersion/variability

- A measure of center needs to be complemented by a measure of spread around this center.
- The definition of averages that we explore naturally lead themselves to measures of variability
- **Variance**: average square distance from the mean

\[
V(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

- **Standard Deviation**: 

\[
\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

- Note that R actually divides by \(n - 1\) rather than \(n\). This is because when \(x_1, \ldots, x_n\) are a sample from a larger population of possible values, dividing by \(n - 1\) one has a “better” estimator for the population quantity.
A note: data with frequencies

Often data is summarized so that we have counts of occurrences of the same values: we have a set $v_1, \ldots, v_m$ of possible values, with their frequencies $f_i$

- Calculating averages and standard deviations has to adapt to this different set-up

$$
\bar{v} = \frac{1}{\sum_{i=1}^{m} f_i} \sum_{i=1}^{m} v_i f_i
$$

$$
\text{Variance} = \frac{1}{\sum_{i=1}^{m} f_i} \sum_{i=1}^{m} (v_i - \bar{v})^2 f_i
$$
A note: the maximal variance of \(x_1, \ldots, x_n\)

- Generally speaking, the variance of a dataset can be arbitrarily large
- Let’s consider some restrictions that make the statement meaningful
  - \(x_i \geq 0 \ \forall i\)
  - fix the total sum of values \(\sum_{i=1}^{n} x_i = n\bar{x}\)

\[
\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 + \bar{x}^2 - 2x_i\bar{x})
\]

\[
= \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \bar{x}^2 - 2\bar{x} \sum_{i=1}^{n} x_i
\]

\[
= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}(n\bar{x})
\]

\[
= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\]
A note: the maximal variance of $x_1, \ldots, x_n$

So, $V(x_1, \ldots, x_n) = \left( \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)/n$. Now,

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \leq (\sum_{i=1}^{n} x_i)^2 - n\bar{x}^2 = n^2\bar{x}^2 - n\bar{x}^2 = \bar{x}^2 n(n - 1)$$

Which means that

$$V(x_1, \ldots, x_n) \leq \bar{x}^2(n - 1)$$

- Can we imagine a set of values of $x_1, \ldots, x_n$ for which the variance is actually equal to this max?
Index of concentration

The opposite of spread-out is “concentrated.”

Let’s consider variables like the one we just talked about, that is with only positive values. One such variable might be the income of households in a nation.

It is interesting to study how “concentrated” or not such income is. One can imagine that the total income of a nation is the total amount of a resource that one could distribute.
Income inequality in the media
Income inequality in politics

Income and Wealth Inequality

Today, we live in the richest country in the history of the world, but that reality means little because much of that wealth is controlled by a tiny handful of individuals.

The issue of wealth and income inequality is the great moral issue of our time, it is the great economic issue of our time, and it is the great political issue of our time.
How can we measure “income inequality”? 

- Let’s think we have a population with \( n \) individuals, each with income \( x_1, \ldots, x_n \).
- \( n \bar{x} \) is the total income in the population (with \( \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \))
- What would be the values of \( x_1, \ldots, x_n \) in the case of maximal “income equality”?
- What would be the values of \( x_1, \ldots, x_n \) in the case of maximal “income inequality”?
- How are we going to judge cases in the middle?
- Any known measure?
- Any measure we can come up with given what we already know?
A graphical display for income distribution

We take the values \( x_1, \ldots, x_n \) and order them

\[
x(1) \leq x(2) \leq \cdots \leq x(n)
\]

For simplicity, we are going to drop the parentheses from the index notation, and just remember that \( x_1 \) is the smallest index.

We now calculate two quantities:

\[
F_i = \frac{i}{n} \quad Q_i = \frac{\sum_{j=1}^{i} x_j}{\sum_{j=1}^{n} x_j}
\]

- \( F_1 \) is the fraction of the population that correspond to the bottom earner; \( F_2 \) is the fraction of the population that correspond to the two bottom earners etc.
- \( Q_1 \) is the fraction of the national income earned by the bottom earner; \( Q_2 \) is the fraction of the national income earned by the two bottom earners etc.
A graphical display for income distribution

- Let’s think about the relation between $F_i$ and $Q_i$ in the case of perfect income equality.
- In general, $Q_i \leq F_i$. To see this, let’s look at their definition and multiply by $\sum_{j=1}^n x_j$ and divide by $i$

\[
\frac{Q_i}{\sum_{j=1}^n x_j} \leq \frac{F_i}{\sum_{j=1}^n x_j} \leq \frac{i}{n} \leq \frac{\sum_{j=1}^i x_j}{i} \leq \frac{\sum_{j=1}^n x_j}{n}
\]

and remember that the $x_i$ are increasing.
A graphical display for income distribution

Income values = 1,2,3,10,15,15,30,50
A graphical display for income distribution

How could we use this to construct an Index?
An idea for the index
From area to index

- Index varies between 0 and 1
- Area in between curves \( A = \frac{1}{2} \) - area under bottom curve
- Area under bottom curve: sum of areas of trapezoids. Thus

\[
A = \frac{1}{2} - \sum_{i=1}^{n} \frac{(F_i - F_{i-1})(Q_i + Q_{i-1})}{2}
\]

- Gini’s index \( G = \frac{A}{1/2} = 1 - \sum_{i=1}^{n} (F_i - F_{i-1})(Q_i + Q_{i-1}) \)
How do things change if we have repetition?

- data in the form
  \[ x_1 \leq x_2 \leq \cdots \leq x_k \]
  \[ n_1 \quad n_2 \quad \cdots \quad n_k \]

  with \( \sum_j n_j = n \)

- Define
  \[ F_i = \frac{\sum_{j=1}^i n_j}{n} \quad Q_i = \frac{\sum_{j=1}^i n_j x_j}{\sum_{j=1}^k n_j x_j} \]

- Everything else stays the same.
Revisiting the Income data

Gini Index

Race

- White
- Hispanic
- Black
- Asian
- All

0.47 0.48 0.49 0.50
Gini index for other data
Gini index for other data

Gini Index - Income Disparity since World War II
where 0 is perfect equality, and 100 is perfect inequality (i.e., one person has all the income)

Graph showing the Gini index over time for various countries.
We can calculate the following summary of "mutual variability"

\[
\Delta = \sum_{i=1}^{k} \sum_{j=1}^{k} |x_i - x_j| \frac{n_i}{n} \frac{n_j}{n}
\]

And one can show that

\[
G = \frac{\Delta}{2\bar{x}}
\]