Data summaries and measures of spread/dispersion/variability

Data Science Team

Goals of the module

- Typically, data come with many observations on multiple variables
- Therefore it is helpful to summarize the data with a few numbers
- Such summaries obviously don't contain as much information as the whole data set, but they are useful for answering some questions and to get a first overview of the data

Summaries & Indexes

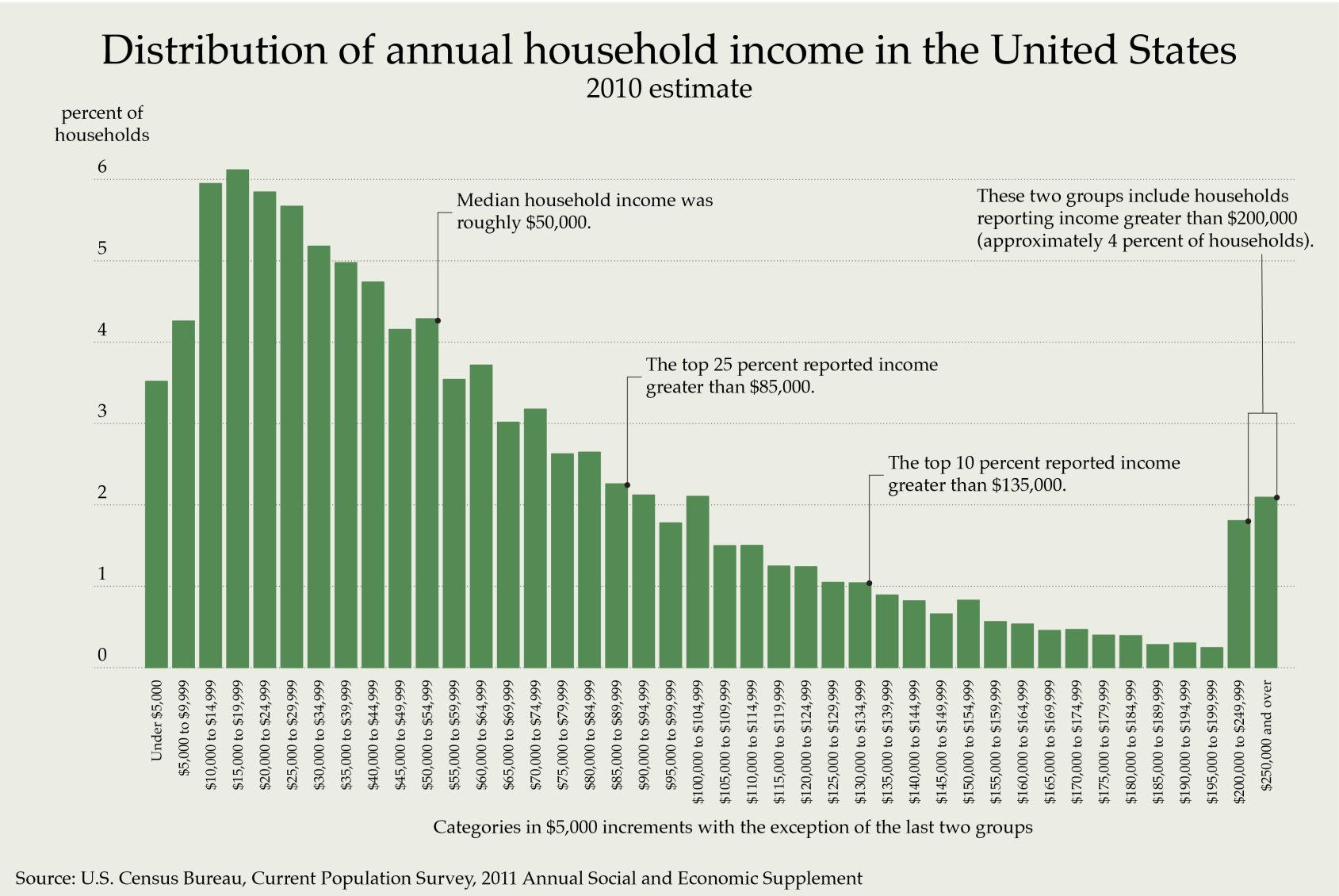
- We will focus on *univariate* data: our analysis will be of one variable at a time
- When summarizing we may choose *different aspects* of the data: center, spread, asymmetry
- We talk about an *index* when we can standardize a summary so that its range is between fixed values (ex. [0,1], [-1,1])

Summaries of Center

- We now focus on quantitative variables, i.e. they can be put on a number line
- The idea is to find a "representative value" or the "middle" of the data set
- Popular summaries of center are: Mean (average), Mode (most frequent value), Median (the number that is larger than half the data and smaller than the other half)
- How can you locate the median in a histogram? (Recall the basic principle of the histogram)

Mean vs. median

When the histogram is skewed to the right, then the mean can be much larger than the median:



Source: U.S. Census Bureau, Current Population Survey, 2011 Annual Social and Economic Supplement

When the histogram is very skewed (e.g. incomes, house prices), then it is better to use the median:

- If the median sales price of 10 home is $ 1 million, then we know that 5 homes sold for $ 1 million or more.
- What conclusions can we draw if instead we are told that the average sales price is $ 1 million?

There are different versions of an *average*

Which version of a summary measure we want to use depends on the question we interested in.

Say we have data $x_1, x_2, \ldots, x_n$ and we are interested in a function $f(x_1, x_2, \ldots, x_n)$ of the data.

Then, we might require that the average $\bar{x}$ is a good summary of the data in the sense that we can substitute the one number $\bar{x}$ in place of all the data values and get the same result for f:

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

Depending on the function f that we are interested in, this results in different versions of averages: arithmetic mean, geometric mean, harmonic mean, etc.

The average interest rate

- Let $r_1, r_2, \ldots, r_{12}$ the monthly interest rates that a financial product earned during the months 1, ..., 12. So after the first month, an investment of \$ 1 has grown to \$ $1 + r_1$, after two months to \$ $(1 + r_1)(1 + r_2)$, ...
- What would be an appropriate definition of an average interest rate?

- We would like **one value** $\bar{r}$ so that if we substitute it for the 12 different monthly interest rates, **we obtain the same compound interest**:
- Comp. interest$(r_1, r_2, \ldots, r_{12})$=Comp. interest$(\bar{r}, \bar{r}, \ldots, \bar{r})$

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

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

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

- 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 traveling at that constant speed, you would cover the same total distance in the same total time.

$$\text{Total travel time} = \sum_{i=1}^{n} \frac{d_i}{v_i} = \sum_{i=1}^{n} \frac{d_i}{\bar{v}}$$

Solving for $\bar{v}$,

$$\bar{v} = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} \frac{d_i}{v_i}} = \frac{1}{\sum_{i=1}^{n} w_i \frac{1}{v_i}}, \text{ where } w_i = \frac{d_i}{\sum_{i=1}^{n} d_i}$$

We find that the **harmonic mean** (here with weights proportional to d_i) is a meaningful summary.

Another approach to defining a *center*

- Maybe it is not possible to find one number $\bar{x}$ such that $f(x_1, x_2, \ldots, x_n) = f(\bar{x}, \bar{x}, \ldots, \bar{x})$
- Instead we could look for a single number such that the data are as close as possible to this number "on average." This seems like a reasonable notion of the "center"
- What does it mean for two points to be **close**? It means the **distance** between the two points is small.
- Can you give an example of a mathematical notion of distance? That is, a **function** $d(x, y)$ that assigns to each pair of points x, y a **number** that expresses how "close" they are?

Distance functions

You probably thought of the **Euclidean distance**, which we usually denote by $\|x - y\|$. For two points x, y on the real line, the Euclidean distance is

$$d(x, y) = \|x - y\| = \sqrt{(x - y)^2}$$

Today, we'll actually work with the **squared** Euclidean distance, which is just $\|x - y\|^2 = (x - y)^2$

Other distances

- However, there are other reasonable ways to define a distance. For example, the "city block" distance $|x - y|$, the **absolute value** of the distance between x and y
- Another way to define the distance between points is

$$d(x, y) = \mathbf{1}\{x \neq y\} = \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}$$

How do we know these are "real" distances?

For a function d to be a **distance function** (or **metric**), it needs to have certain properties. Specifically

1. nonegativity $d(x, y) \geq 0$
2. $d(x, y) = 0 \Leftrightarrow x = y$ (the distance is only zero when the two points are the same)
3. symmetry $d(x, y) = d(y, x)$ (the distance from x to y is the same as distance from y to x)
4. The **triangle inequality** $d(x, z) \leq d(x, y) + d(y, z)$

Although the two "other" distances I suggested before may be unfamiliar to you (especially $d(x, y) = \mathbf{1}\{x \neq y\}$), they all satisfy these properties and are therefore valid **distance functions**

Minimizing the average distance

We said earlier that another way to define the "center" of the data might be to pick the point that minimizes the **average distance** between that point and all the data points $x_1, \ldots, x_n$

So we are looking for a point z which minimizes $\frac{1}{n}\sum_{i=1}^{n} d(z, x_i)$

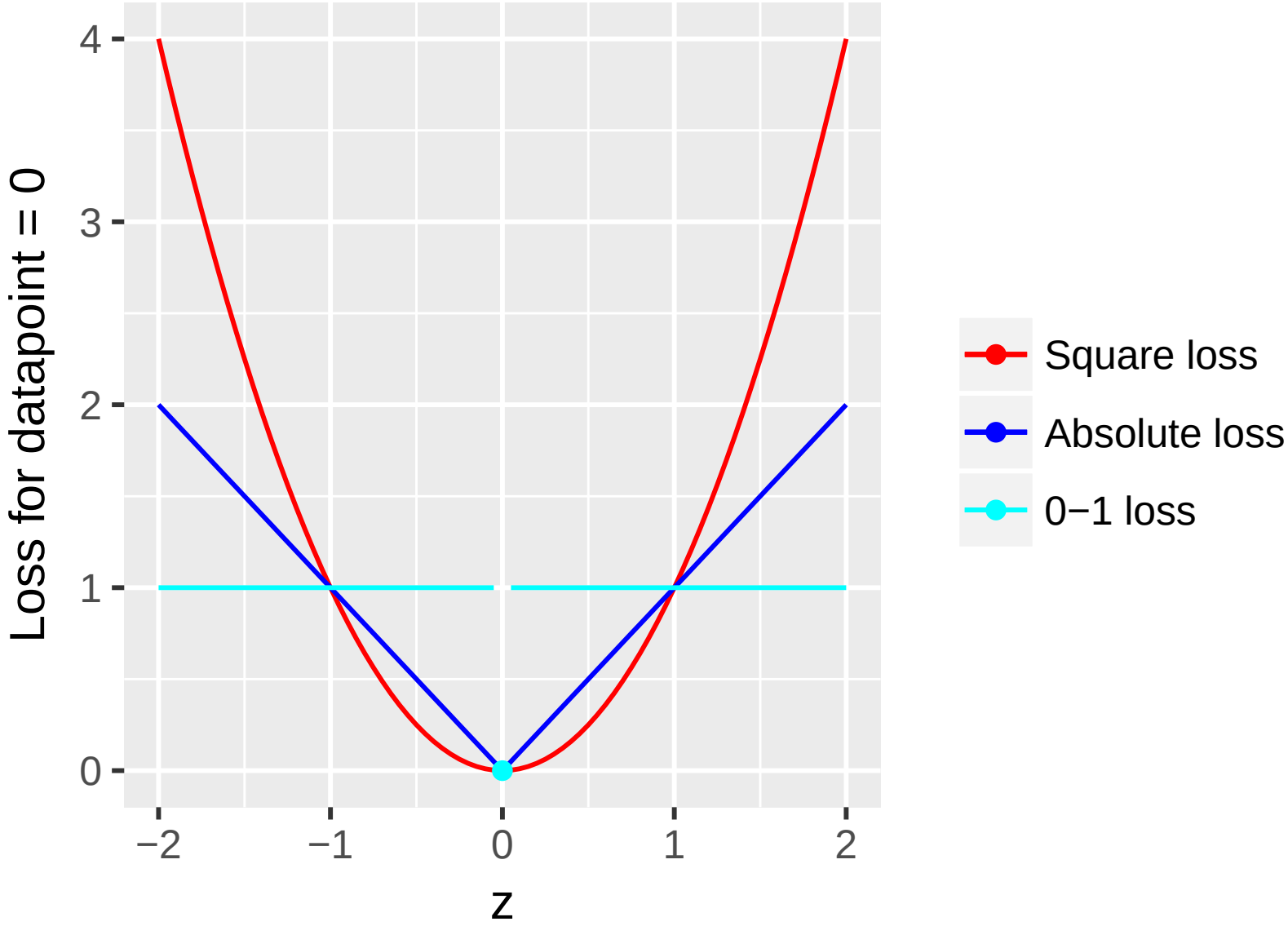
- Ex. 1: use the squared Euclidean distance $n^{-1}\sum_{i=1}^{n}(x_i - z)^2$
- Ex. 2: use the city block distance $n^{-1}\sum_{i=1}^{n}|x_i - z|$
- Ex. 3: use the 0-1 distance $n^{-1}\sum_{i=1}^{n}\mathbf{1}\{x_i \neq z\}$

Loss functions

- The average distance is an example of a **loss function**. What's that?
- Imagine playing a game where the objective is to guess the "center" of the data. How would we know who wins?
- Finding the center means finding the point that minimizes the average distance to the data. So we would take everyone's guesses and compute the average distance for each guess. The winner would be the one whose guess gave the **smallest** average distance.
- A bad guess results in a large average distance. So the average distance tells us how much we lose with respect to the best guess. That's why it's called a **loss function**.

Different loss functions

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



Different loss functions

- The center is the point that **minimizes the loss function**
- Question: How sensitive is each loss function is to "extreme" data points?

Minimizers of these average distances/loss functions

First consider the 0-1 loss

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

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/average squared Euclidean distance is minimized by 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/average city block distance is minimized by the **median**

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

$$\text{median} = x_{(6)}$$

Measures of spread/dispersion/variability

- A notion of the "center" needs to be complemented by a measure of the "spread" around this center.
- The **basic idea** we used was to define the "center" as the point that minimized the average distance to the data points
- The **magnitude** of this average distance is a measure of how much the data vary around the center
- If we use the squared Euclidean distance to define the center, then the measure of spread we obtain is called **Variance**

$$\text{Var}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Standard deviation

- The **Standard Deviation** is just the square root of variance

$$\mathrm{SD}(x_1, \ldots, x_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Typically one uses the standard deviation rather than the variance as a measure of spread: It gives a more useful interpretation of spread (details later), and it is on the correct scale: If the measurements are in (say) pounds, then the SD is also in pounds, whereas Var has the unit pounds^2.
- Sometimes one divides by $n-1$ rather than n in calculating variance and standard deviation, and software typically uses $n-1$. This distinction is not crucial for most problems.

A note: data with frequencies

- Sometimes data are grouped or rounded so that we have counts for occurrences of various values
- Suppose we have a set $v_1, \ldots, v_m$ of possible unique values, with their frequencies $f_i, i = 1, \ldots, m$, giving us a data table of the form

$$\begin{array}{cccc} v_1 & v_2 & \cdots & v_m \\ f_1 & f_2 & \cdots & f_m \end{array}$$

- **Question**: why would we want to store the data this way by grouping/rounding, instead of storing each observation?

How to compute averages with frequencies

- Let's see how to compute the average (arithmetic mean) when we have data in this form:
- If we "expanded" the data to just be f_i many copies of v_i, we would get

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

but that is just

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

since $\sum_{j=1}^{f_i} 1 = f_i$.

Weighted average

- Also, since $n = \sum_{i=1}^{m} f_i$ we have

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

where $w_i = \frac{f_i}{\sum_{j=1}^{m} f_j}$ are the **proportions** of the data in each "bin". This is a **weighted average**, where the weights are given by **relative frequency**.

- How about variance? Recall the formula

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

- But this is itself just an average! We average the quantities $(x_i - \bar{x})^2$ instead of x_i. So when the data are provided via the v_i and their frequencies f_i, then the variance is

$$\frac{1}{\sum_{i=1}^{m} f_i} \sum_{i=1}^{m} (v_i - \bar{v})^2 f_i = \sum_{i=1}^{m} w_i (v_i - \bar{v})^2$$

An alternative to measure spread

- We saw that one very outlying observation can change the mean a lot. This is also true for the SD and variance.
- The median is more "robust" to such outliers. A corresponding robust measure of spread is the **interquartile range**, see below.

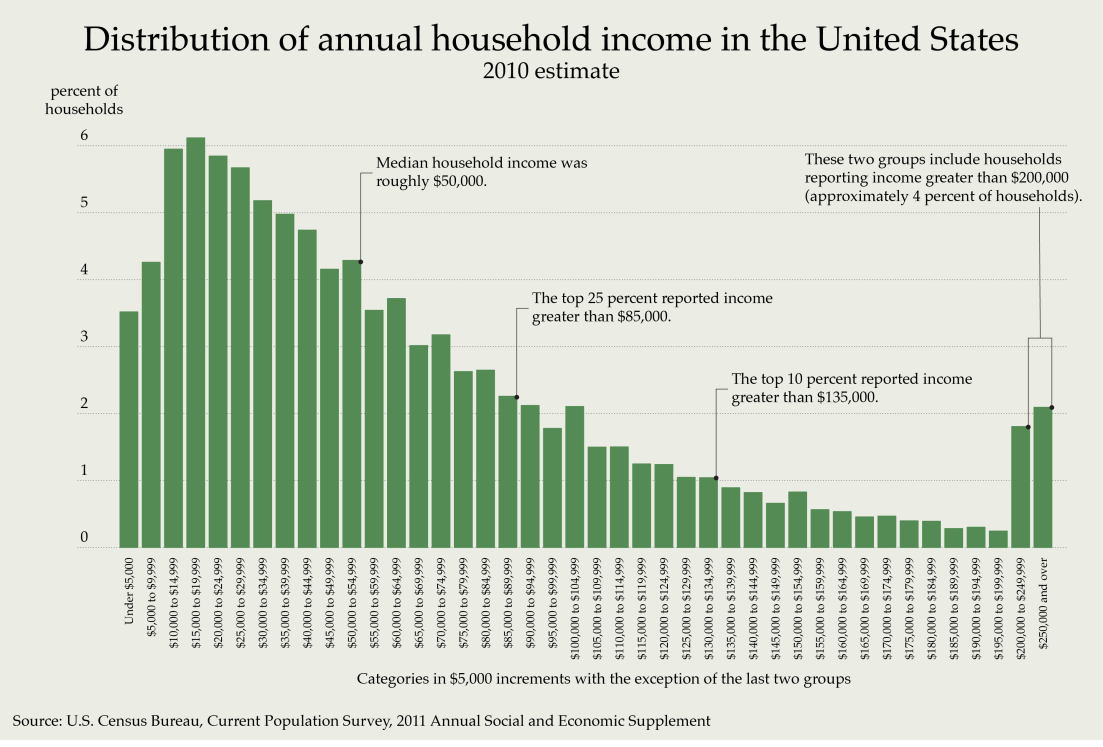
Five-number summary

Recall that the boxplot gives a **five-number summary** of the data:

smallest number, 1st quartile, median, 3rd quartile, largest number

The **interquartile range** = 3rd quartile - 1st quartile. It is an alternative measure of spread.

Percentiles



The 75th percentile is called **3rd quartile**: $ 85,000

The 50th percentile is the **median**: $ 50,000

The 25th percentile is called **1st quartile**

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 income is. One can imagine that the total income of a nation is the **total amount of a resource that one could distribute**.

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 $\bar{x} = \sum_{i=1}^{n} x_i/n$)
- 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"?
- Any measure we can come up with given what we already know?

Order statistics

We take the values $x_1, \ldots x_n$ and **order** them (aside: these are called the "order statistics")

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

For simplicity, we are going to drop the parentheses from the index notation, and just remember that x_1 is the smallest income and x_n the largest.

Computing some statistics

Suppose initially that we have no "repeats" in the data (everyone has a different income)

We now calculate two quantities:

$$F_i = \frac{i}{n} \qquad Q_i = \frac{\sum_{j=1}^{i} x_j}{\sum_{j=1}^{n} x_j}$$

- 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

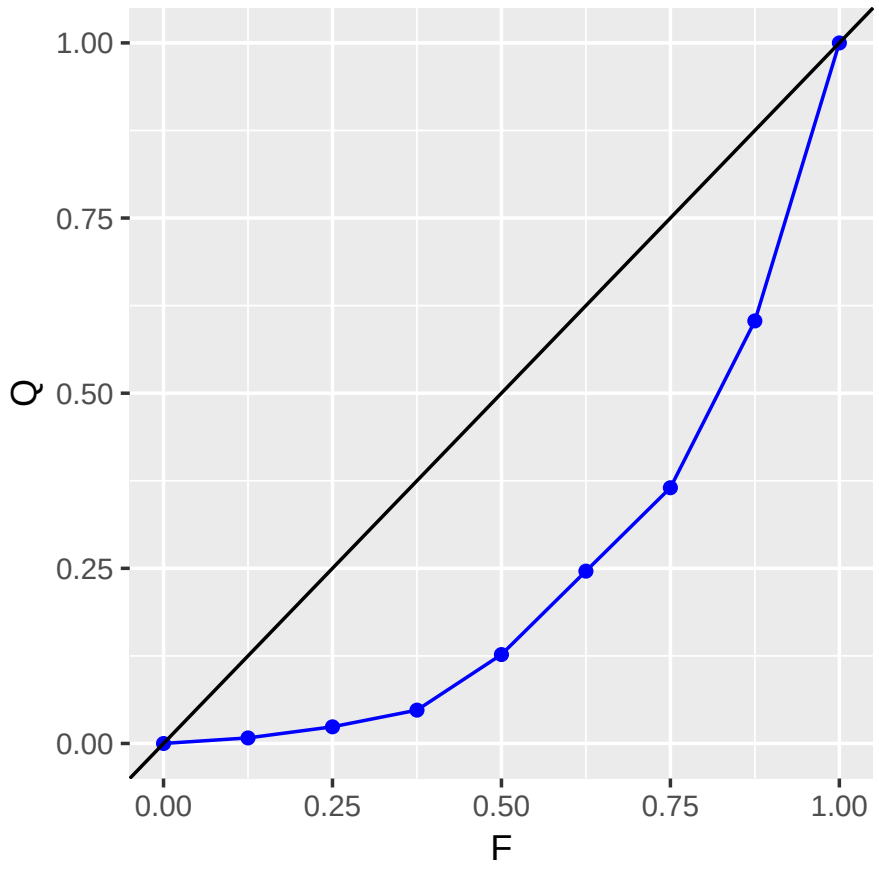
- Let's think about the relation between F_i and Q_i in the case of perfect income equality
- In general, $Q_i \leq F_i$. To see this, recall that the x_i are **increasing** and so

$$
\begin{aligned}
Q_i &= \frac{\sum_{j=1}^{i} x_j}{\sum_{j=1}^{n} x_j} \\
\frac{Q_i}{F_i} &= \frac{\sum_{j=1}^{i} x_j}{\sum_{j=1}^{n} x_j} \frac{n}{i} = \bar{x}_{(i)} \frac{1}{\bar{x}} \\
&= \frac{\bar{x}_{(i)}}{\bar{x}} \leq 1
\end{aligned}
$$

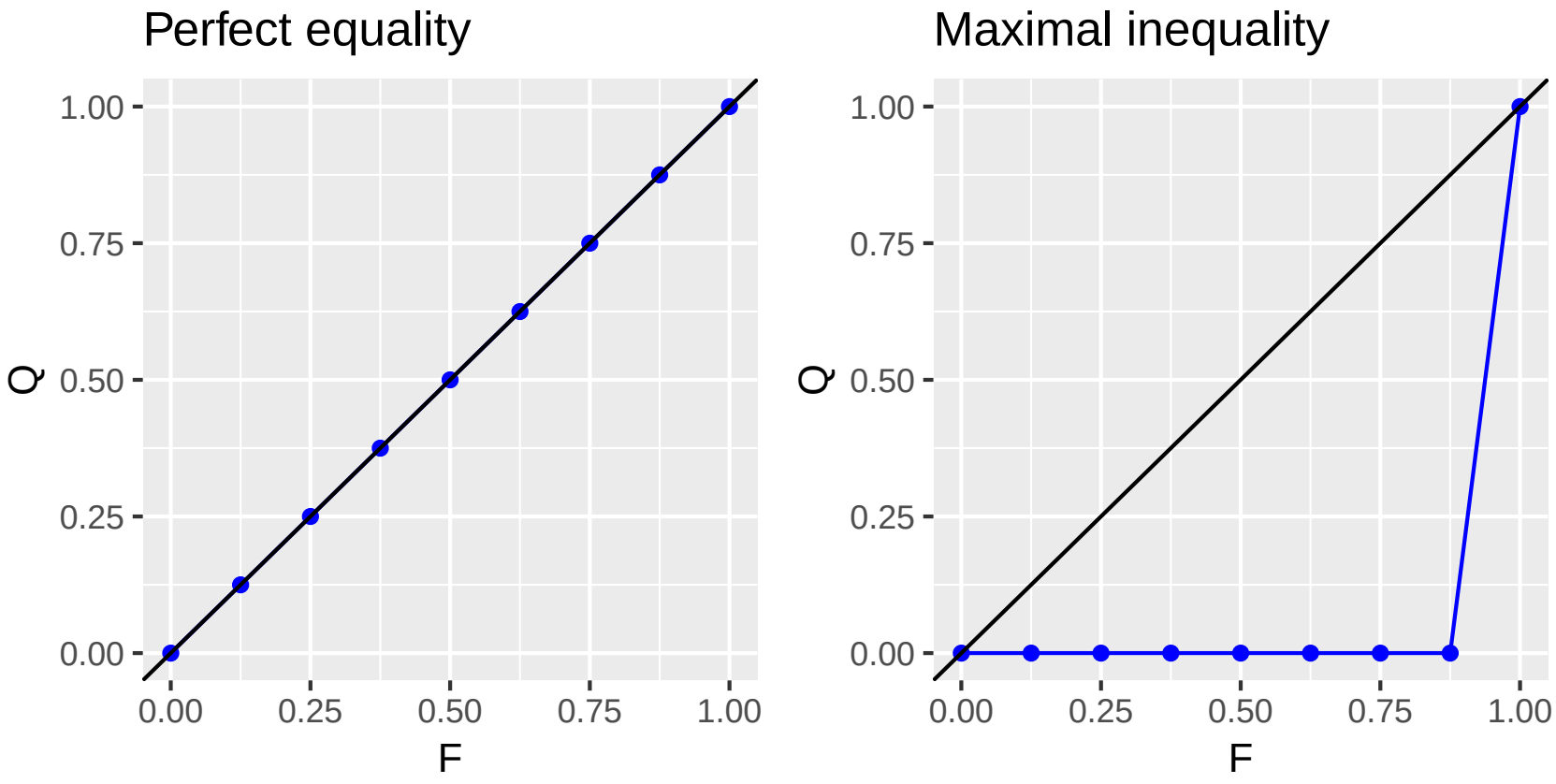
where $\bar{x}_{(i)} = \frac{1}{i} \sum_{j=1}^{i} x_i$

A graphical display for income distribution

Income values = 1,2,3,10,15,15,30,50

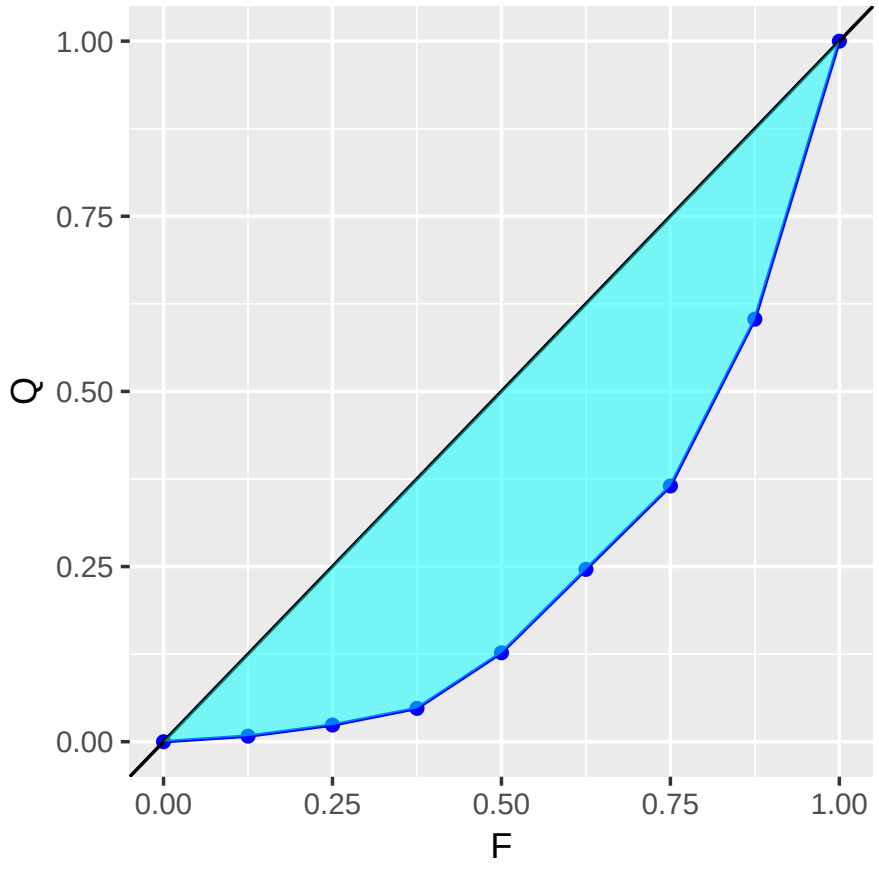


A graphical display for income distribution



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 A between the curves $= 1/2$- area under bottom curve
- The area under the bottom curve equals the sum of areas of trapezoids. So

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

- Gini's index$= G = \dfrac{A}{1/2} = 1 - \sum_{i=1}^{n}(F_i - F_{i-1})(Q_i + Q_{i-1})$
- Larger Gini index means more inequality

How do things change if we have repetition?

- Data in the form:

$$\begin{array}{cccc} x_1 & \leq x_2 & \leq \cdots & \leq x_k \\ n_1 & n_2 & \cdots & n_k \end{array}$$

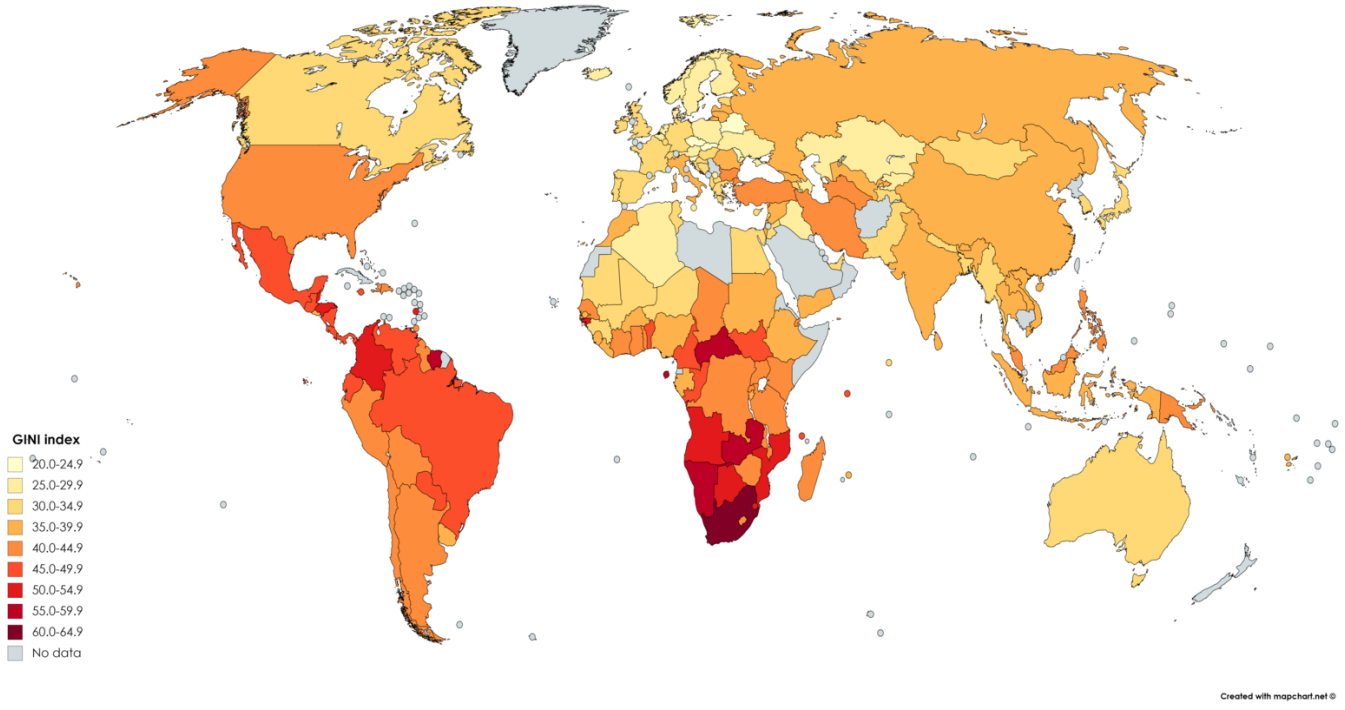
with $\sum_j n_j = n$

- 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.

GINI index Map



Source: World map of Gini coefficients by country. Based on World Bank data ranging from 1992 to 2018.