澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Exploring Data

CISC7204: DATA SCIENCE & VISUALIZATION

Derek F. Wong

NLP²CT – *Natural Language Processing & Portuguese-Chinese Machine Translation Research Group*

derekfw@um.edu.mo

E11-4010 (Ext: 4478)

Office Hours: Thu – 16:00~17:30, Fri 11:00~12:30

Natural Language Processing & Portuguese – Chinese Machine Translation Laboratory
自然語言處理與中葡機器翻譯實驗室

Content

- *Tendency* and *Dispersion*
  - *Mode, Median, Mean*
  - *Range, Variance, Standard Deviation*
- *Exploring Data* with Simple Charts

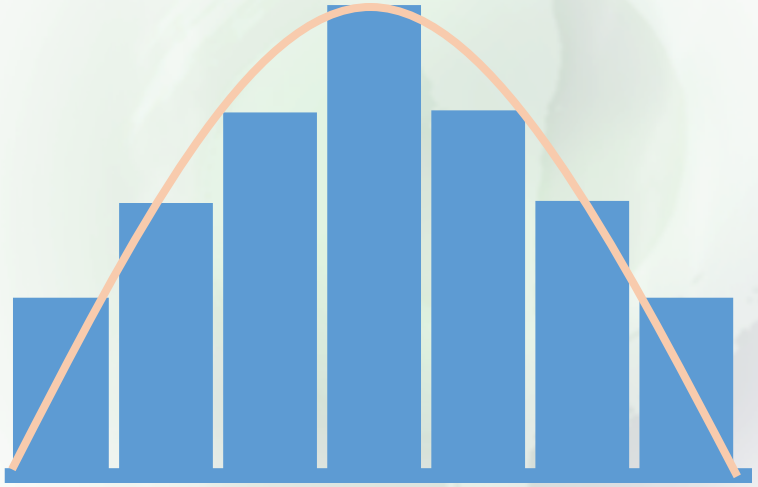
Measures of Central Tendency

Measures of Central Tendency

Descriptive Statistic

There are at least *three characteristics* you look for in a *descriptive statistic* to *represent* a set of data

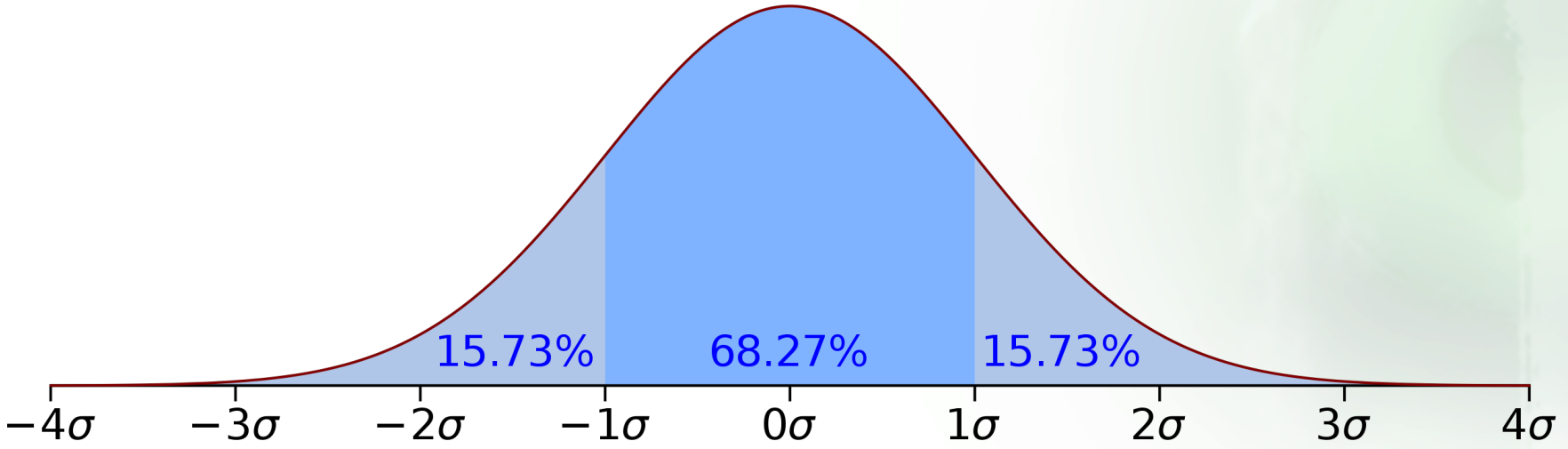
- *Represented*: A *good descriptive statistic* should be similar to many scores in a distribution (*High Frequency*)
- *Well balanced*: neither *greater-than* or *less-than* scores are overrepresented
- *Inclusive*: Should take individual values from the distribution into account so *no value is left out*



Measures of Central Tendency

Descriptive Statistic

- In a *normal distribution* the *most frequent scores* cluster near the *center* and *less frequent scores* fall into the *tails*
- *Central tendency* means *most scores* (68%) in a normally distributed set of data tend to cluster in the *central tendency area*



The Mode

The *mode* (Mo) is the *most frequently* occurring score in a *distribution*

- Example with a set of quiz scores (X)

$$10\ 10\ 10\ 9\ 9\ 9\ 9\ 8\ 8\ 8\ 8\ 7\ 7\ 7\ 7\ 7\ 6\ 6\ 6\ 6\ 6\ 6\ 5\ 5\ 5\ 5\ 5\ 5\ 5$$
$$5\ 4\ 4\ 4\ 4\ 4\ 4\ 4\ 4\ 3\ 3\ 3\ 3\ 2\ 2\ 1$$

What is the Mo?

▶ $X = 5$ has the *greatest frequency* (9), hence, the mode is 5

The Mode

Limitations

- **Multi-modal**: There can be *more than one mode*
- **Lack of Representativeness**: It may *not* be a *good representative* of all values

The Mode

Limitations

- **Multi-modal**: There can be *more than one mode*

Examples

$$10\ 10\ 9\ 9\ 9\ 9\ 9\ 9\ 8\ 8\ 7\ 7\ 6\ 6\ 5\ 5\ 4\ 4\ 4\ 4\ 4\ 4\ 3\ 3\ 2\ 2\ 1\ 1$$

Two most frequent scores, $X = 9$ and $X = 4$

▶️ Or in case of a *rectangular distribution*

$$10\ 9\ 8\ 7\ 6\ 5\ 4\ 3\ 2\ 1\ 0$$

The Mode

Limitations

- **Lack of Representativeness**: It may *not* be a *good representative* of all values

  Examples

10 10 10 10 10 10 10 9 8 7 6 5 4 3 2 1 0

The *mode might* be at *one end* of a distribution, *not at the center*

The Mode

Limitations

- **Lack of Representativeness**: It may *not* be a *good representative* of all values

$$10\ 10\ 10\ 9\ 9\ 9\ 9\ 8\ 8\ 8\ 8\ 7\ 7\ 7\ 7\ 7\ 6\ 6\ 6\ 6\ 6\ 6\ \underbrace{5\ 5\ 5\ 5\ 5\ 5\ 5\ 5}_{9}$$

$$5\ \underbrace{4\ 4\ 4\ 4\ 4\ 4\ 4\ 4}_{8}\ 3\ 3\ 3\ 3\ 3\ 2\ 2\ 1$$

It is possible that the most frequent score has a frequency that is only one or two counts greater than the second most frequent score. In the example above the mode was $X = 5$. However, $X = 4$ had a frequency of eight, which is only one less than the frequency of nine for $X = 5$. *Why is 5 a better mode than 4?* There is no answer!

The Median

The *median* (Md) is the *middle score of a distribution*

- Half on the *left half* on *the right* (the 50th percentile)
  - *Better* measure of *central tendency* than the *mode* since it *balances* perfectly distribution

How to find it?

The Median

How to find it?

Two simple steps

1. Determine the *median's location*
2. Find the *value* at that location
   - It differs whether you have an *even* or an *odd* number of scores

With ODD number of observations: For example: $n = 11$

10 10 9 7 7 6 5 4 3 2 2

Position of median

Use the equation: $Md = \frac{n+1}{2}$ $Md = \frac{11+1}{2} = 6^{th}$

The Median

How to find it?

Two simple steps

1. Determine the *median's location*
2. **Find the *value* at that location**
   - It differs whether you have an *even* or an *odd* number of scores

Order the data: rank-ordered from *smallest* to *largest*

Count off *six positions* starting with the smallest value

X	2	2	3	4	5	6	7	7	9	10	10
Position	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}

The Median

How to find it?

Two simple steps

1. **Determine the *median's location***
2. Find the *value* at that location
   - It differs whether you have an *even* or an *odd* number of scores

With Even number of observations: For example: $n = 12$

20 19 18 16 15 14 12 11 11 11 10 9

Determine the *two positions* around the *median*:

$$Md = \frac{n}{2}, \frac{n+2}{2} \qquad Md = \frac{12}{2}, \frac{12+2}{2} = 6^{th}, 7^{th}$$

The Median

How to find it?

Two simple steps

1. Determine the *median's location*
2. **Find the *value* at that location**
   - It differs whether you have an *even* or an *odd* number of scores

Order the data: rank-ordered from *smallest* to *largest*

The median is between *sixth* and *seventh* positions

X	9	10	11	11	11	12	14	15	16	18	19	20
Position	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}

The average $(12 + 14)/2 = 13$ is the median value

The Median

Limitations

The median *does not take into account* the *actual values* of the scores in a set of data

Example: take the following *two sets of data*, each with $n = 5$ scores

Dataset #1	1	4	5	7	10
Dataset #2	4	4	5	6	6

Median is equal to 5, which is unaffected by the values greater and less than its value

The Mean

Is the *arithmetic average* of all the *scores in a distribution*

- The *mean* is the most-often used *measure of central tendency*
  - It *evenly balances* a distribution so both the *large* and *small values* are equally represented
  - Takes into account all *individual values*

To estimate it in two steps

1. *Add* together all the scores in a distribution $\sum X$
2. *Divide* that sum $\sum X$ by the *total number of scores* N in the distribution

Sample: $\bar{X} = M = \frac{\sum X}{n}$ Population: $\mu = \frac{\sum X}{N}$

The Mean

Example

Calculate the mean from the sample of $n = 11$:

$$2\ 2\ 3\ 4\ 5\ 6\ 7\ 7\ 9\ 10\ 10$$

Mean: $$\bar{X} = M = \frac{\sum X}{n} = \frac{65}{11} = 5.909$$

The median $= 6$. The difference between the *mean* and the *median* reflects the mean's taking into account the *individual values*

Mean vs Median

Just to *emphasize* the fact that the mean takes into account all values in a set of data. *Recall this example*

Dataset #1	1	4	5	7	10
Dataset #2	4	4	5	6	6

The mean for each values would be:

Dataset #1: $\bar{X} = \frac{27}{5} = 5.4$ Dataset #2: $\bar{X} = \frac{25}{5} = 5$

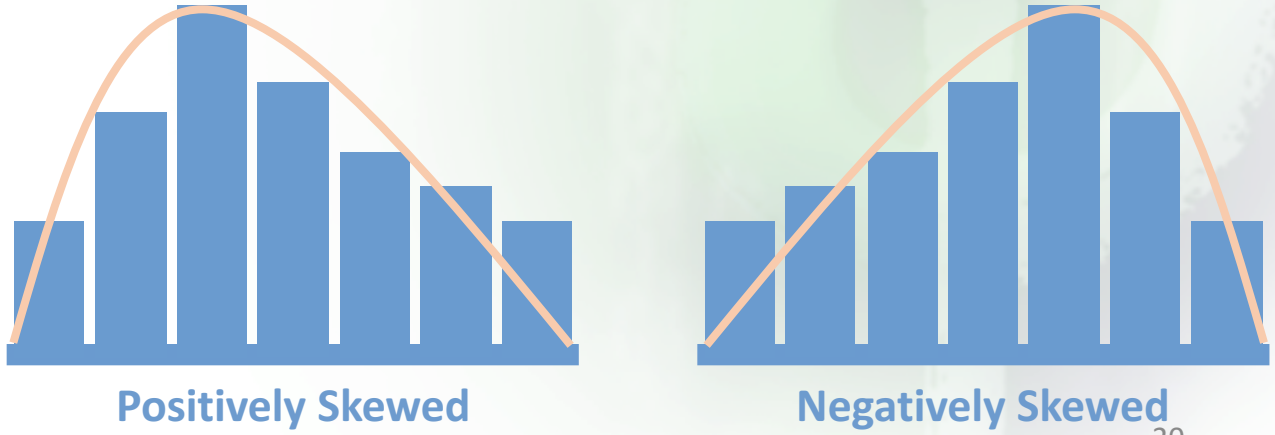
The mean is a more accurate measure of central tendency as it takes the individual scores into account; the *median does not*

The Mean

Limitations

The mean has the *following problems*:

- By *taking individual values into account* the mean can be *influenced* by *extremely large* or *extremely small values* (outliers)
- Specifically, extremely *small values* pull the *mean down* and extremely *large values* pull the *mean up*
- This only occurs in *skewed distributions*: it is good to use the *median* as a *measure of central tendency*



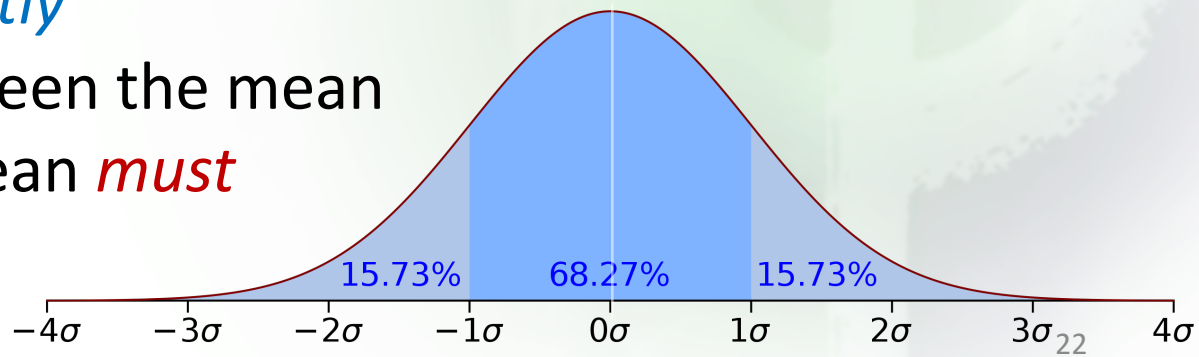
What to Use, When?

- **Nominal Data** ⇒ the *mode* is the best measure of central tendency (*i.e. Gender*)
  - Data are categorical in nature and values can *only fit into one class*
- **Ordinal Scale** ⇒ the *median* may be more appropriate. The *individual values* on an ordinal scale are *meaningless*
  - *Extreme scores* and you *do not want to distort the average*
- **Interval** or **Ratio Scale** ⇒ the *mean* is generally preferred (*Counts for all observations*)
  - *No extreme scores* and are not categorical

Normal and Skewed Distributions

The *positions* of the *mean*, *median*, and *mode* are affected by whether a distribution is *normally distributed* or *skewed*

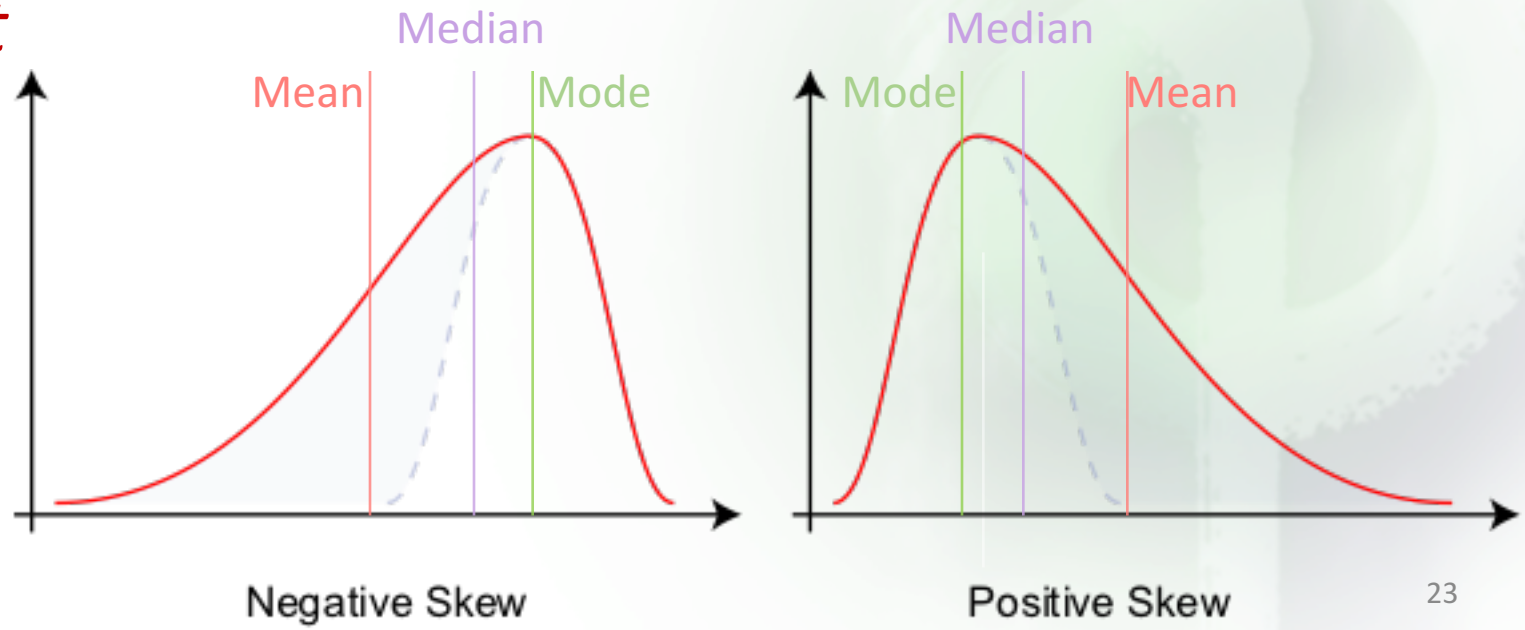
- In data *normally distributed*: *mean = median = mode*
  - *50% of the scores* must lie above the *center point*, which is also the *mode*, and the other *50%* of the scores *must lie below*
  - Because the distribution is *perfectly symmetrical* the *differences* between the mean and the values larger than the mean *must cancel out*



Normal and Skewed Distributions

Negatively skewed: The *tail* of the distribution is to the *left* and the *hump* is to the *right*

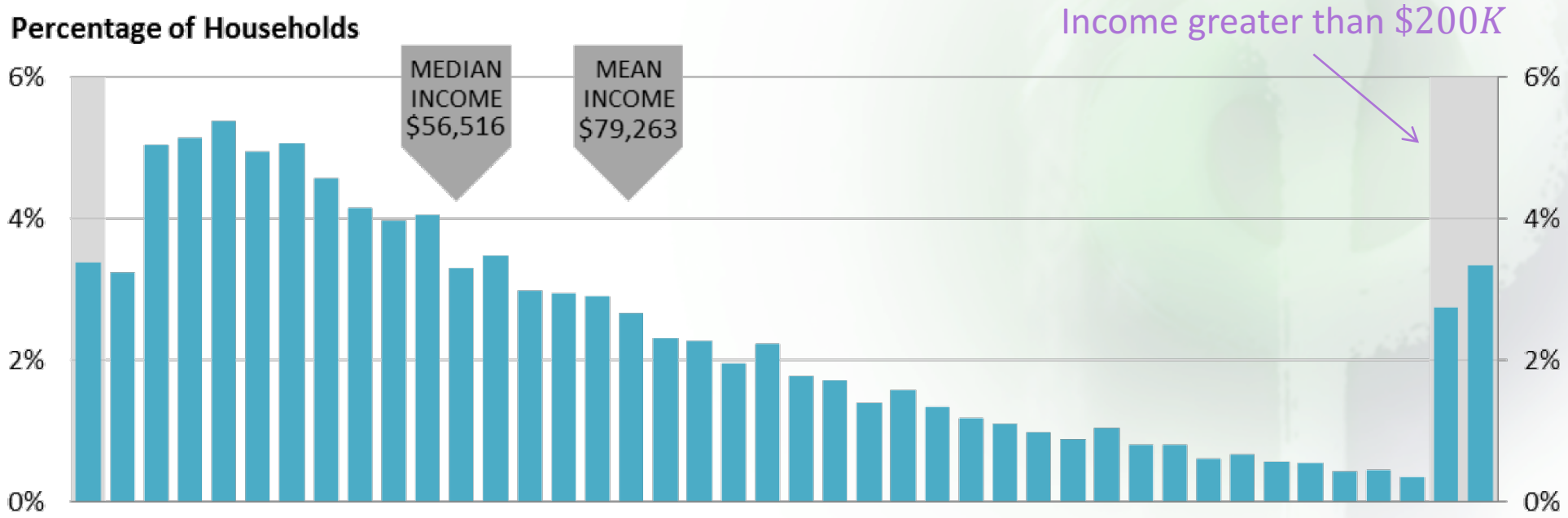
Positively skewed: The *tail* of the distribution is to the *right* and the *hump* is to the *left*



Normal and Skewed Distributions

So if the *mean differs from the median/mode*, the distribution is *skewed*

- The *median* is better as the *measure of central tendency* when the distribution is *positively* or *negatively skewed*



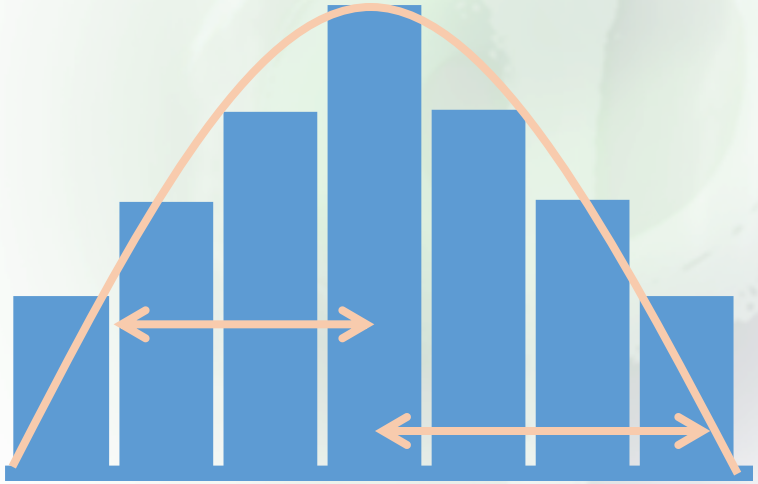
Variability

Describe how much data sets vary

What is Variability?

Variability is simply the *differences among items*, which could be differences in *eye color*, *height*, *weight*, *gender*, *intelligence*, etc.

- If *measures of central tendency* (mean, median, and mode) estimate *where a distribution falls*
- *Variability* measures the *dispersion* or *similarity* among scores and tell us about the *variety of the scores in a distribution*



What is Variability?

The *mean just tells* you where a *distribution of data tends to fall*

- *For example, if the mean for a set of ten scores on a Data Visualization quiz is 7, those ten scores could be:*

$$7 + 7 + 7 + 7 + 7 + 7 + 7 + 7 + 7 + 7 = 70/10 = 7$$

or

$$6 + 6 + 6 + 6 + 6 + 8 + 8 + 8 + 8 + 8 = 70/10 = 7$$

or

$$1 + 1 + 5 + 5 + 9 + 9 + 10 + 10 + 10 + 10 = 70/10 = 7$$

Measures of Variability

There are *several measures of variability*. We will go from the easiest to more complex ones

- The *range*
- *Sum of squares*
- The *variance*
- The *standard deviation*

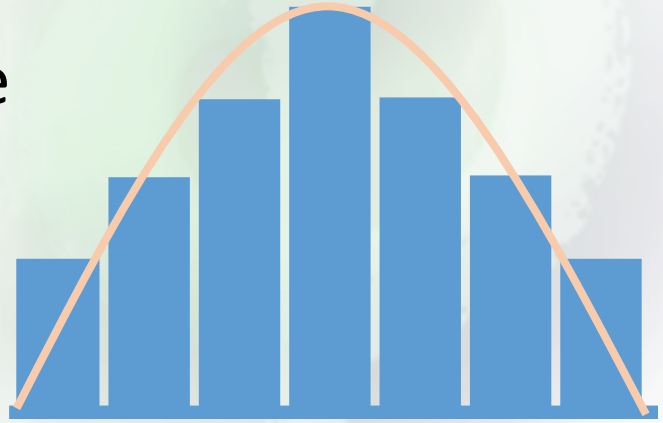
The Range

For the following *two sets of* $n = 10$ *scores*, in both sets the *sum of scores* and the *mean* are *identical*

Set #1	2	2	2	4	5	5	6	6	8	10	$\sum X = 50$	$\bar{X} = 5$
Set #2	4	4	5	5	5	5	5	5	5	6	$\sum X = 50$	$\bar{X} = 5$

The *range* is the *largest* value *minus* the *smallest* value

- Set #1 the range is $10 - 2 = 8$
- Set #2 is $6 - 4 = 2$



It provides information about the *area a distribution covers*, BUT does *not say anything* about *individual values*

The Range

These two sets have the *same range*

Set #3	1	1	1	1	10
Set #4	1	2	4	5	10

Clearly there are *more differences* among scores in Set #4 than Set #3, but the *range does not account for this variability*

- What we need is some way to *measure the variability* among the scores

Sum of Squares (SS)

The sum of the squared *deviation scores* from the *mean* and measures the summed or *total variation* in a set of data

$$SS = \sum(X - \bar{X})^2$$

- Could be *equal to zero* if *there is no variability* and all the scores are equal
- *Cannot be negative* (mathematically impossible)

Sum of Squares (SS)

Estimate the SS for the sample:

Set #1	2	2	2	4	5	5	6	6	8	10	$\sum X = 50$	$\bar{X} = 5$
Set #2	4	4	5	5	5	5	5	5	5	6	$\sum X = 50$	$\bar{X} = 5$

Set #1

X	$(X - \bar{X})$	$(X - \bar{X})^2$
10	5	25
8	3	9
6	1	1
6	1	1
5	0	0
5	0	0
4	-1	1
2	-3	9
2	-3	9
2	-3	9

$SS = 64$

Larger sum of squares indicates there is more variability among the scores

Set #2

X	$(X - \bar{X})$	$(X - \bar{X})^2$
6	1	1
6	1	1
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
4	-1	1
4	-1	1

$SS = 4$

Variance

Issues with SS

- Sum of squares measures the *total variation* among scores in a distribution
- Sum of squares *does not measure average variability*
- We want a measure of variability that takes into account *both the variation* of the scores and *number of scores* in a distribution

Sample Variance S^2

- The *average* sum of the squared deviation scores from a mean
- Measures the *average variability* among scores in a distribution

$$S^2 = \frac{\sum(X - \bar{X})^2}{n} = \frac{SS}{n}$$

Variance

$$S^2 = \frac{\sum(X - \bar{X})^2}{n} = \frac{SS}{n}$$

In our examples?

- *Set #1*

$$S^2 = \frac{64}{10} = 6.4$$

- *Set #2*

$$S^2 = \frac{4}{10} = 0.4$$

Set #1	Set #2
10	6
8	6
6	5
6	5
5	5
5	5
4	5
2	5
2	4
2	4
$SS = 64$	$SS = 4$
$S^2 = 6.4$	$S^2 = 0.4$

Larger measurements of variance indicate *greater variability*

Variance

Issues

The *one issue* with *sample variance* is it is in *squared units*, not the *original unit* of measurement

- If the *original measurements* in Sets #1 and #2 were of *length* in *centimeters*, when the values were squared the measurement units are now cm^2

$$S^2 = \frac{\sum(X - \bar{X})^2}{n} = \frac{SS}{n}$$

What is the solution to this issue?

Standard Deviation

Once *sample variance* has been calculated, calculating the sample *standard deviation* (S) is as simple as taking the *square root* of the variance

$$S = \sqrt{\frac{\sum(X - \bar{X})^2}{n}} = \sqrt{\frac{SS}{n}} = \sqrt{S^2}$$

So, in our examples

- *Set #1* $S = \sqrt{6.4} = 2.53$
- *Set #2* $S = \sqrt{0.4} = 0.63$

The standard deviation measures the average deviation between a score and the mean of a distribution

Summary

	Nominal	Ordinal	Interval or Ratio
Central Tendency (Representative)	Mode	Median	Mode, Median, Mean
Variability (Dispersion)	Proportion	Range	Range, Variance, Standard Deviation

Exploring Data

Find patterns and trends lurking in the data and then observe the deviations from those patterns

Case Study

Ideb: Quality in Basic Education

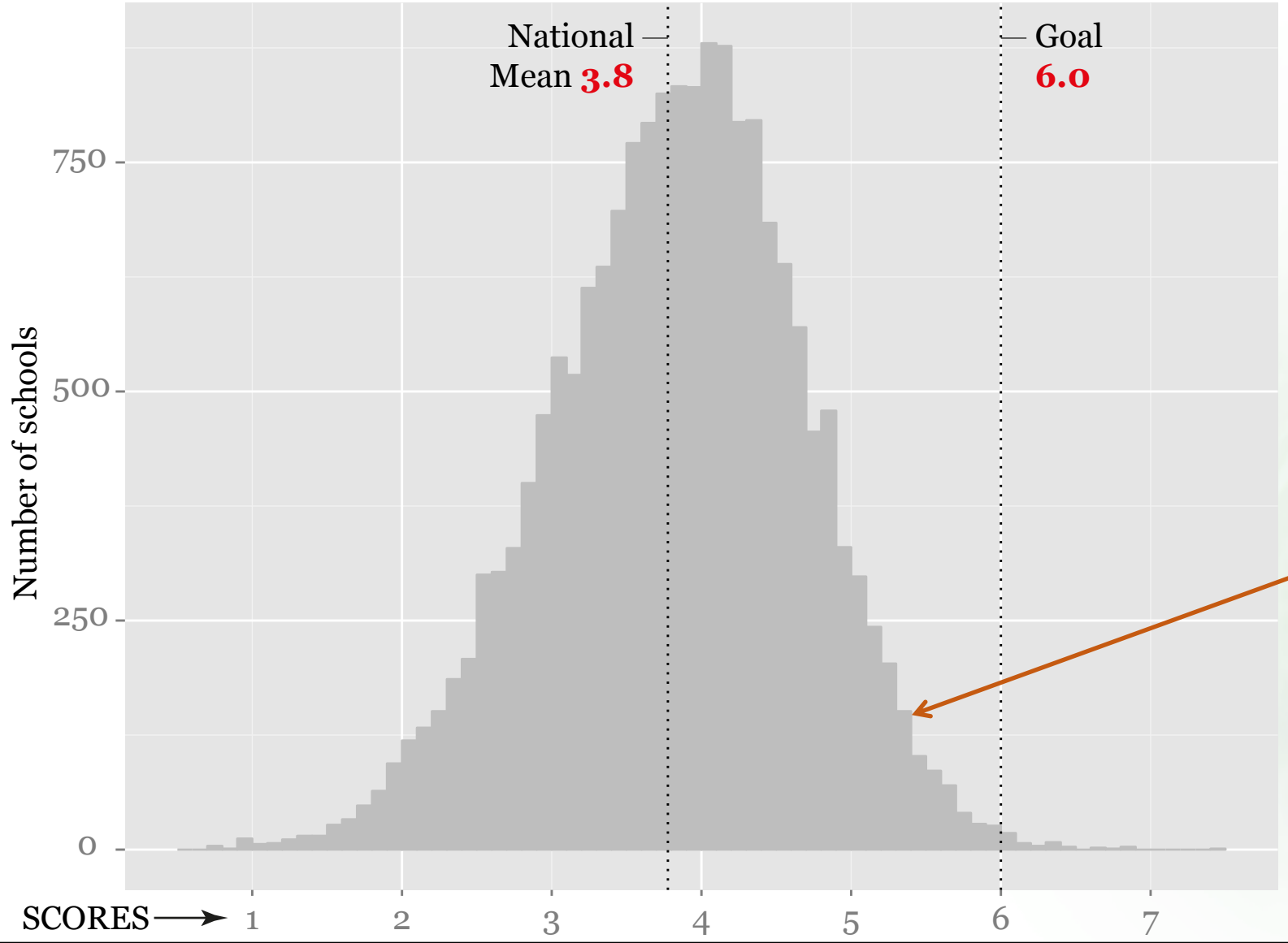
- *The Ideb*: an index that *measures quality in basic education* in Brazil, released by *Brazilian Ministry of Education*
- *19,386* records

Attributes	Examples	
State	RO	AM
City	ALTA FLORESTA D'OESTE	MANAUS
School	EEEF EURIDICE LOPES PEDROSO	ALDEIA DO CONHECIMENTO PROF. RUTH PRESTES GONCALVES
Score	3.70	4.2

This materials of this section are from the Chapter 6 of "The Truthful Art – by Alberto Cairo, 2016"

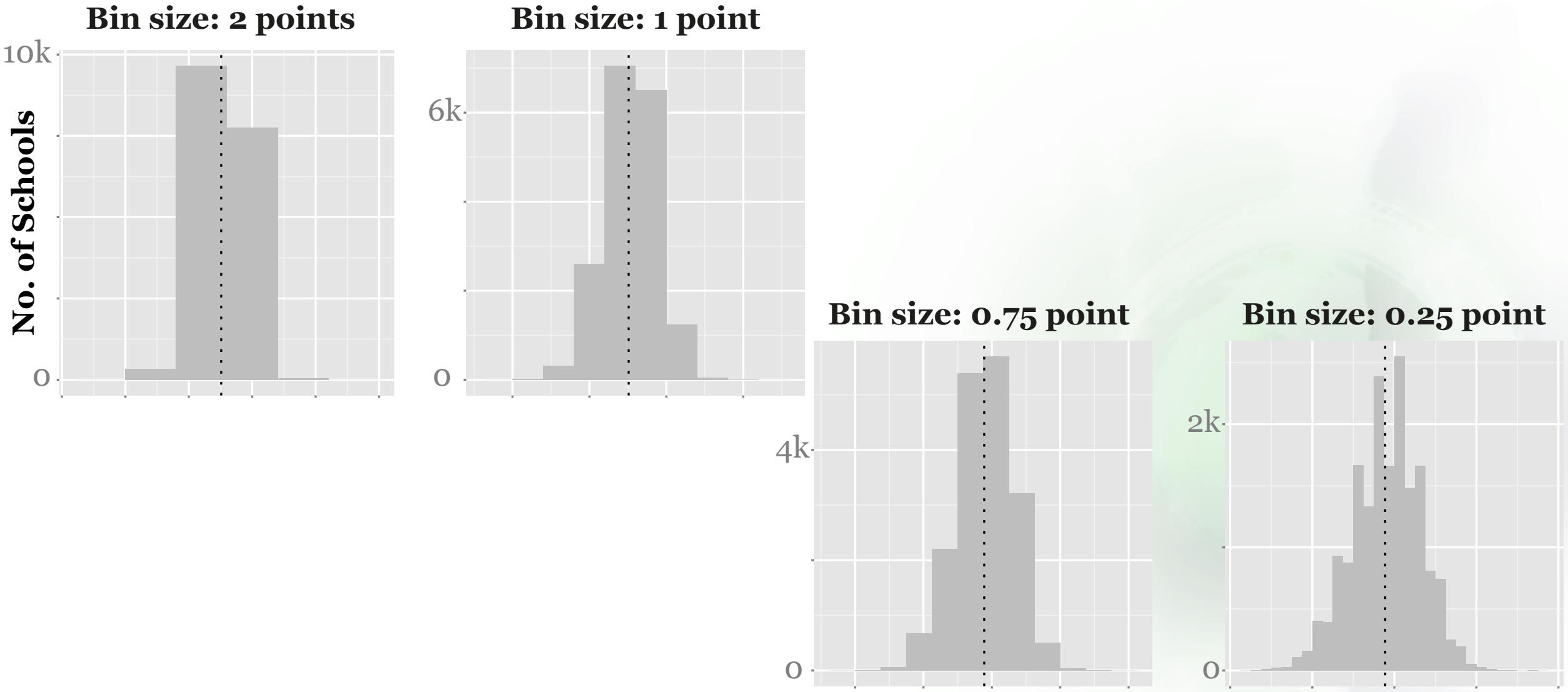
Descriptive Statistics

Descriptive Statistics



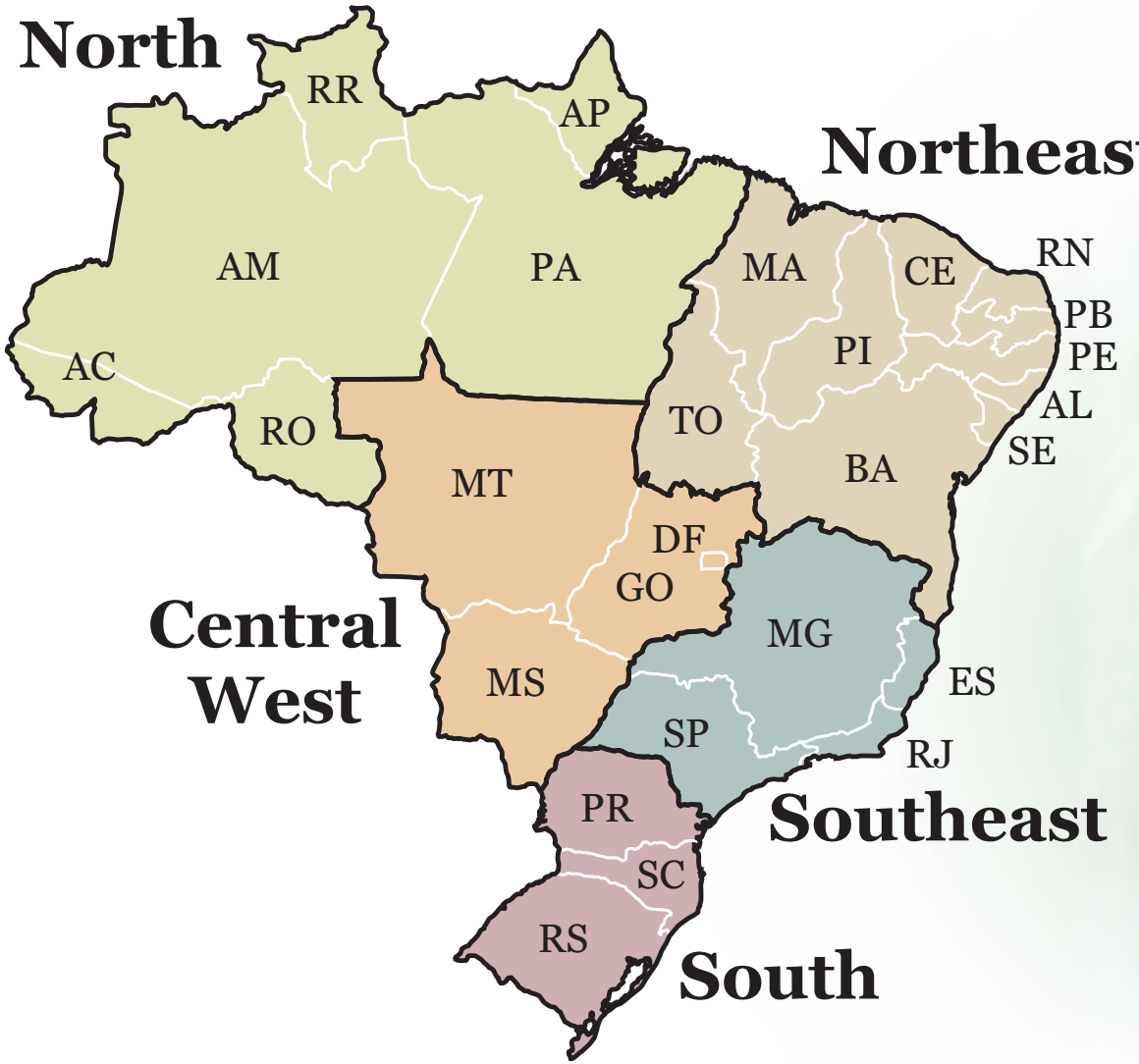
The range of bins is 0.1 score point

Descriptive Statistics

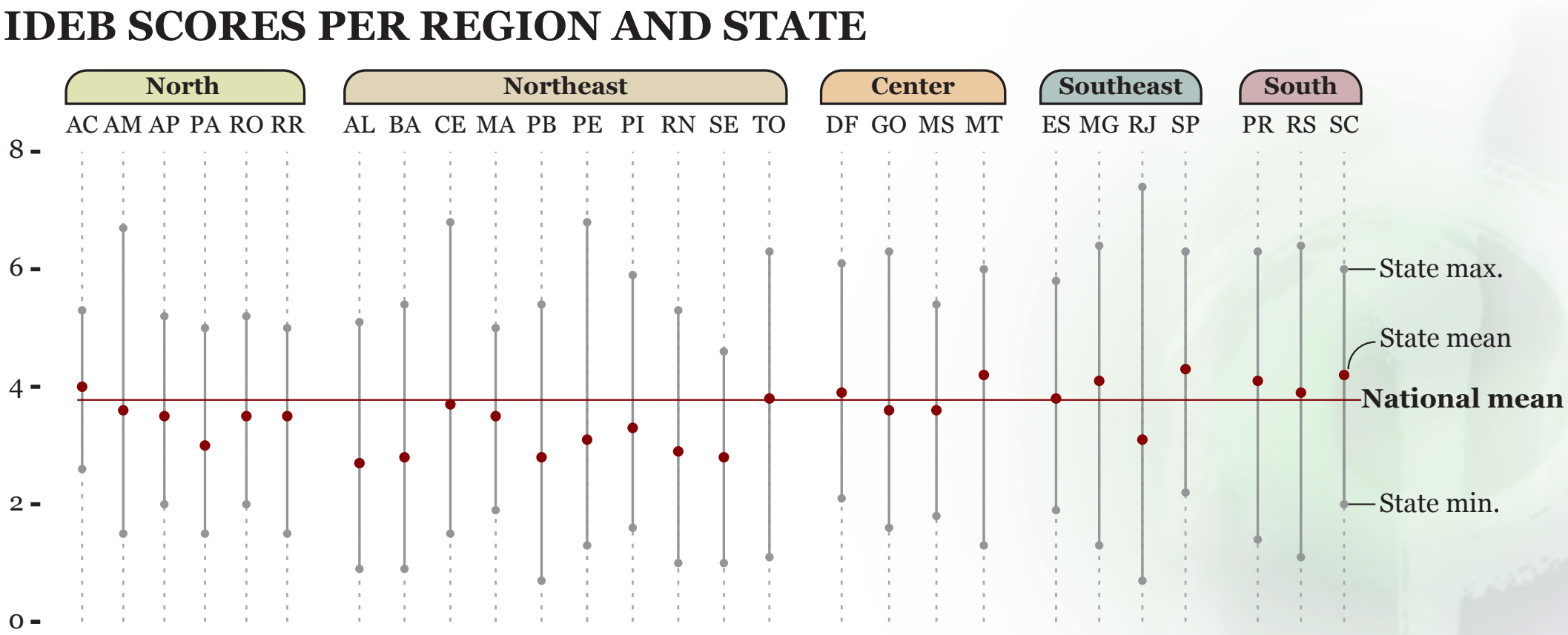


Map of Brazil

27 States

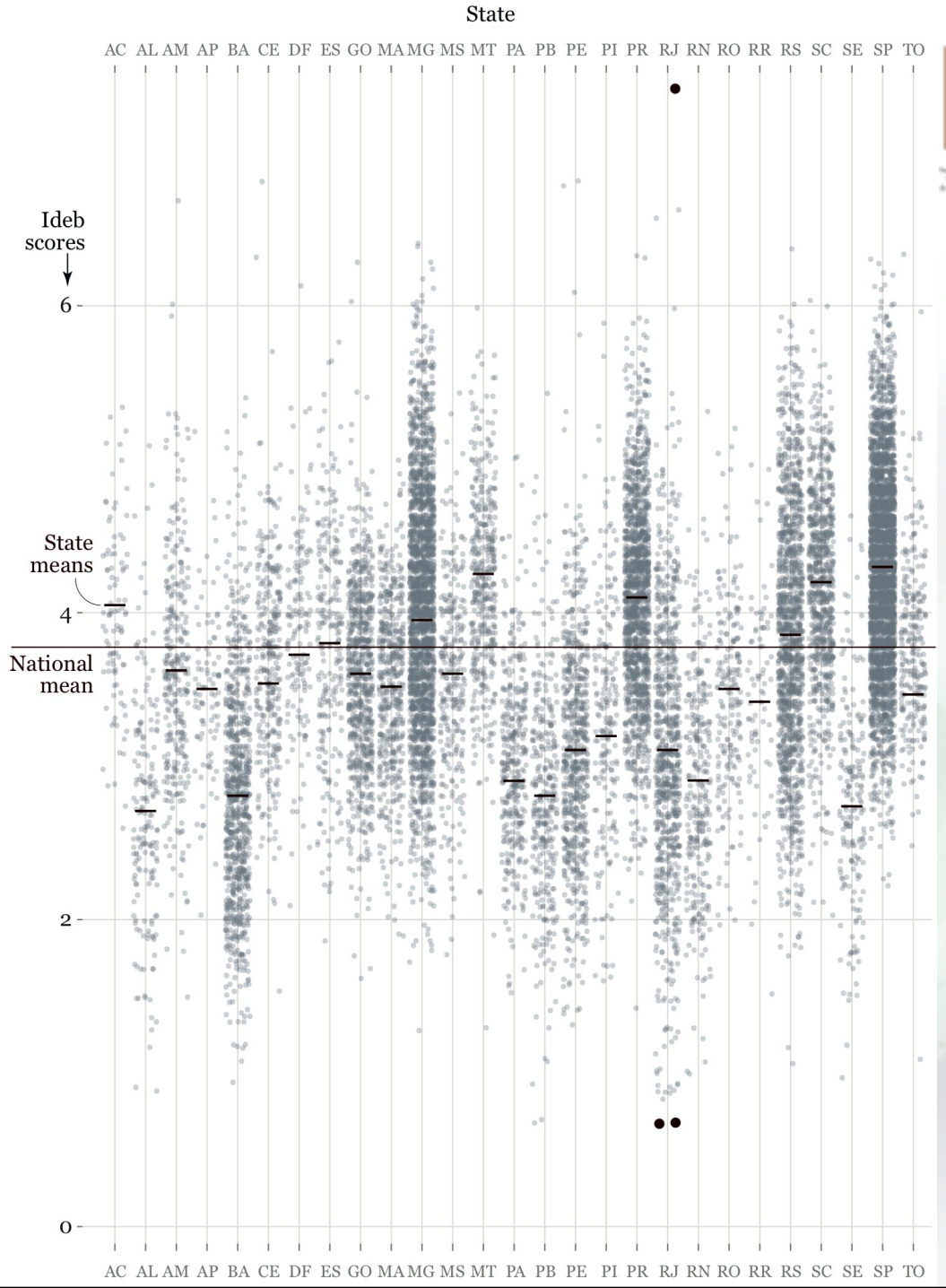


Descriptive Statistics



Descriptive Statistics

- RJ (Rio de Janeiro) has the widest range
- The best and the worst schools

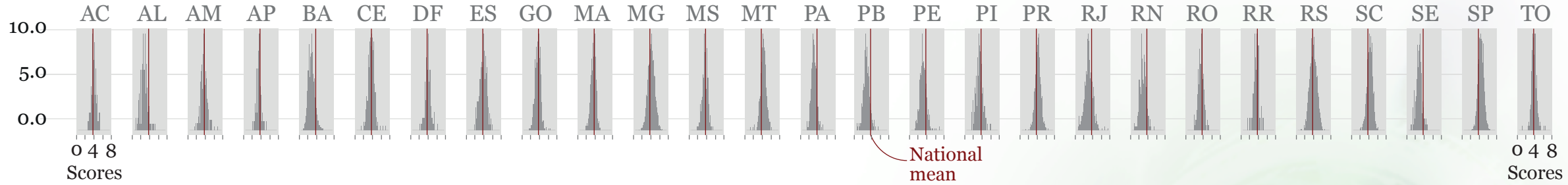


Doing Exploratory Work

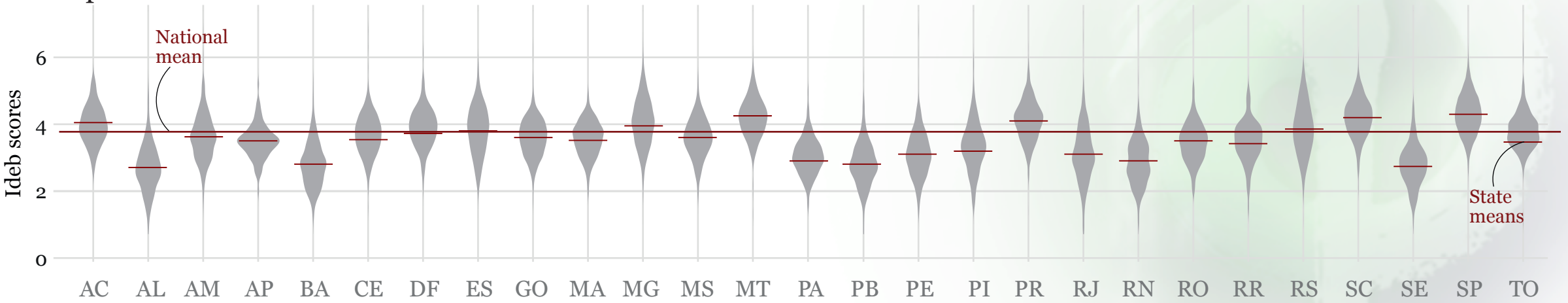
A visualization designer should never rely on a single statistic or a single chart or map

Levels of Details

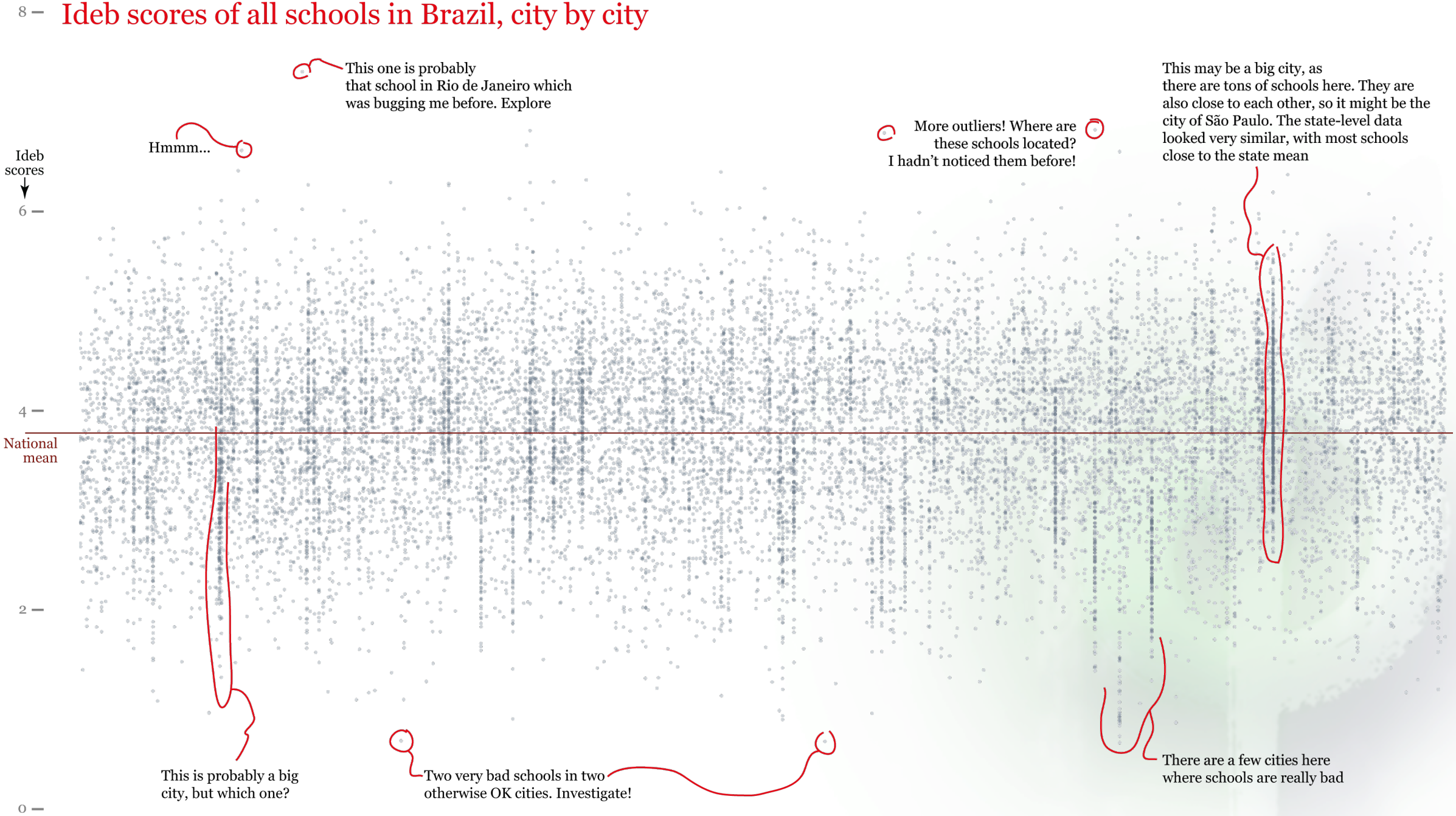
Histograms



Violin plot



Ideb scores of all schools in Brazil, city by city



References

- GravyAnecdote (2014). *Control Your Message with Title, Color and Orientation*.
- Cairo, A. (2016). *The Truthful Art: Data, Charts, and Maps for Communication*. New Riders.

Acknowledgements

Some of the materials are adapted from:

- Oscar Barrera, 2018
- Alberto Cairo, 2016