Distributions and Descriptives

Summarizing data

We use summary statistics because it is difficult (sometimes impossible) to look at every value in the data.

Summarizing data

We use summary statistics because it is difficult (sometimes impossible) to look at every value in the data.

Means are useful, but they only give us part of the picture.

Summarizing data

We use summary statistics because it is difficult (sometimes impossible) to look at every value in the data.

Means are useful, but they only give us part of the picture.

Consider the grades for these two quizzes:

Summarizing data

We use summary statistics because it is difficult (sometimes impossible) to look at every value in the data.

Means are useful, but they only give us part of the picture.

Consider the grades for these two quizzes:

- Quiz 1: 0, 0, 100, 0, 100, 0, 100, 100

Summarizing data

We use summary statistics because it is difficult (sometimes impossible) to look at every value in the data.

Means are useful, but they only give us part of the picture.

Consider the grades for these two quizzes:

- Quiz 1: 0, 0, 100, 0, 100, 0, 100, 100
- Quiz 2: 50, 60, 55, 45, 40, 40, 40, 70

Summarizing data

We use summary statistics because it is difficult (sometimes impossible) to look at every value in the data.

Means are useful, but they only give us part of the picture.

Consider the grades for these two quizzes:

- Quiz 1: 0, 0, 100, 0, 100, 0, 100, 100
- Quiz 2: 50, 60, 55, 45, 40, 40, 40, 70

Mean: 50

Summarizing data

We use summary statistics because it is difficult (sometimes impossible) to look at every value in the data.

Means are useful, but they only give us part of the picture.

Consider the grades for these two quizzes:

- Quiz 1: 0, 0, 100, 0, 100, 0, 100, 100
- Quiz 2: 50, 60, 55, 45, 40, 40, 40, 70

Mean: 50

We need to consider more than just a mean

Distributions

Every continuous variable may be viewed as a *distribution*

Distributions

Every continuous variable may be viewed as a *distribution*

We may be interested in how observations fall along a scale:

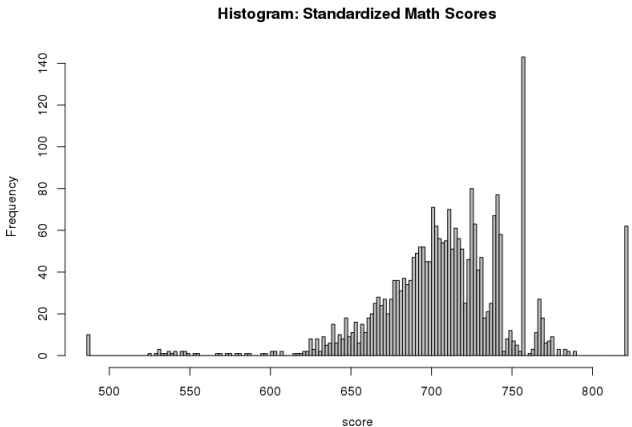
Distributions

Every continuous variable may be viewed as a *distribution*

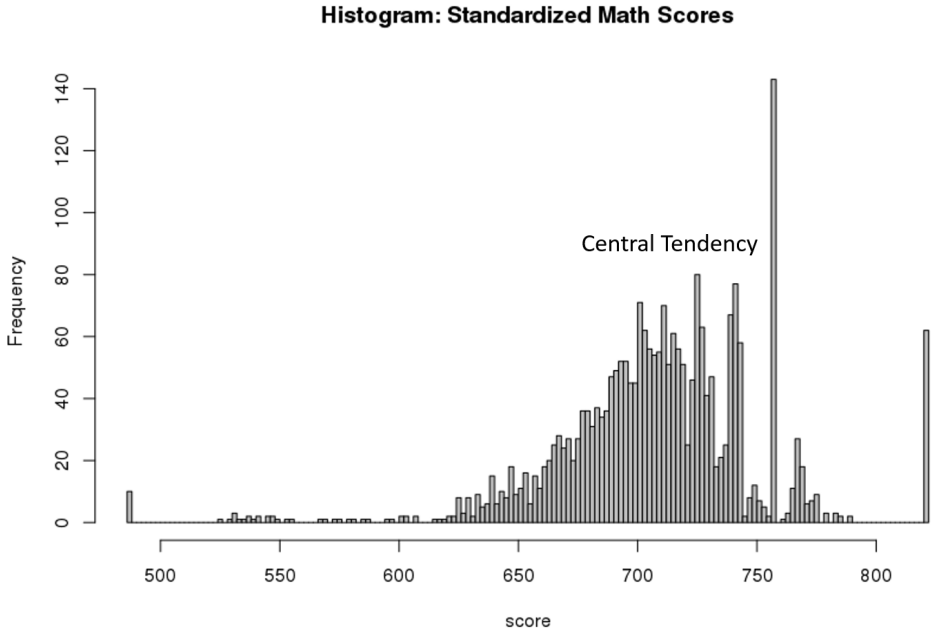
We may be interested in how observations fall along a scale:

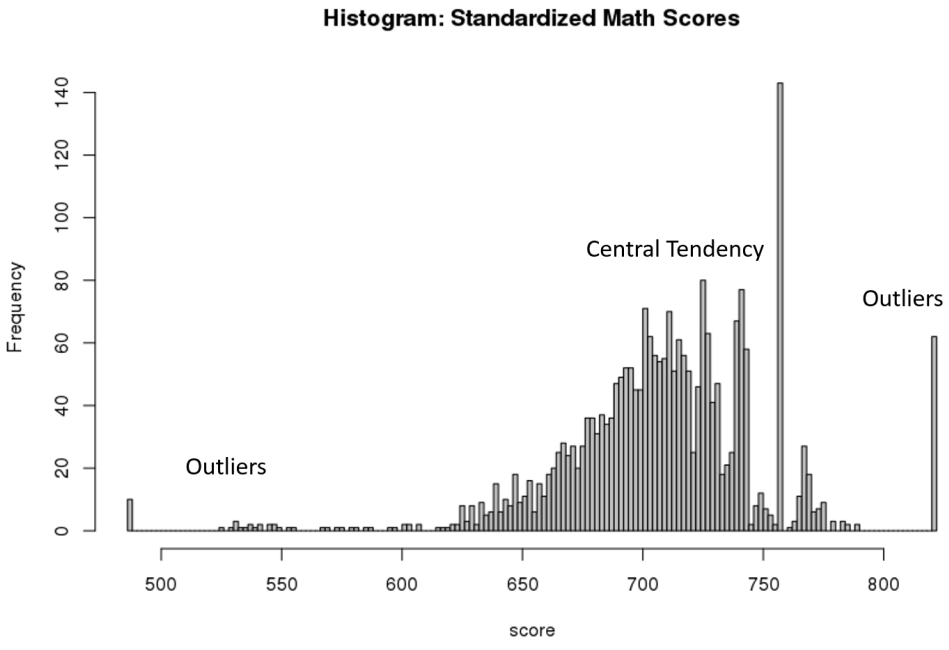
- How varied were student scores on a standardized math exam?

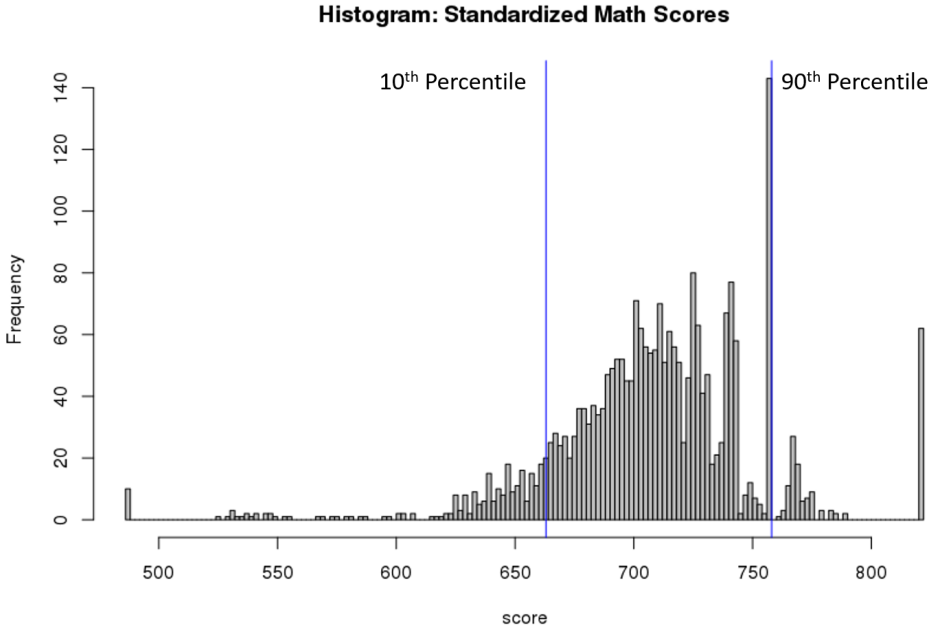
Consider

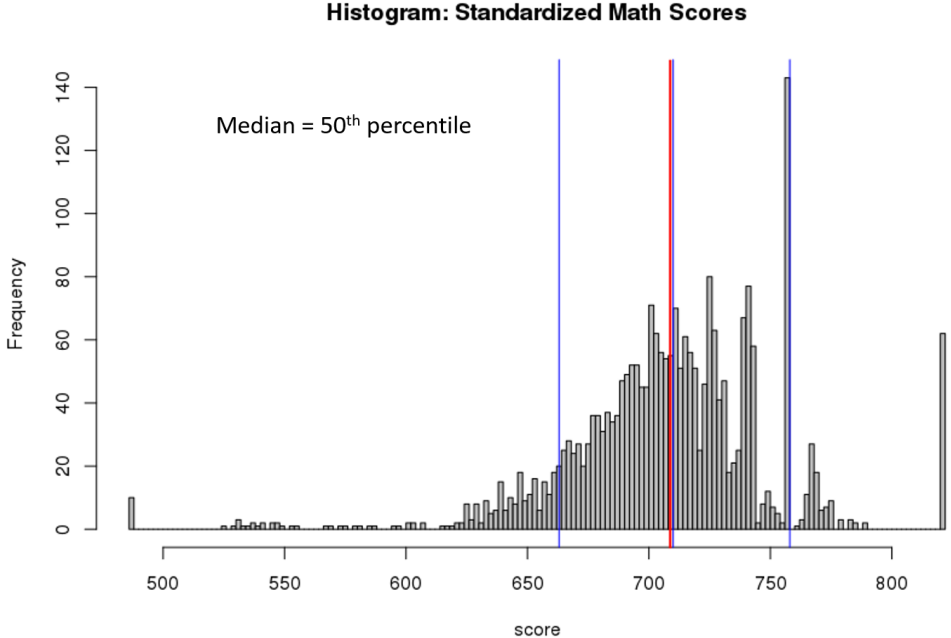


Does the mean reflect the data?


Histogram: Standardized Math Scores
Central Tendency
Frequency
0
20
40
60
80
100
120
140
500
550
600
650
700
750
800
score


Histogram: Standardized Math Scores
Frequency
0
20
40
60
80
100
120
140
500
550
600
650
700
750
800
score
Central Tendency
Outliers
Outliers


Histogram: Standardized Math Scores
10th Percentile
90th Percentile
Frequency
0
20
40
60
80
100
120
140
500
550
600
650
700
750
800
score


Histogram: Standardized Math Scores
Median = 50th percentile
Frequency
0
20
40
60
80
100
120
140
500
550
600
650
700
750
800
score

Descriptive Statistics

Skew:

When Mean $<$ Median

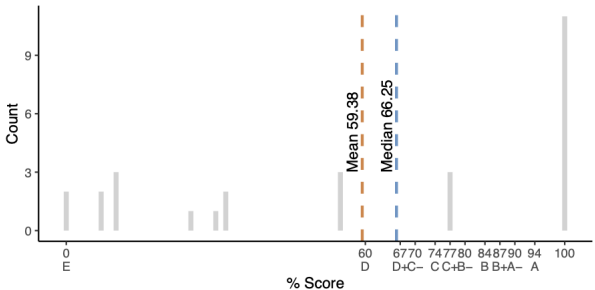
Skewed to the **left** (*long left tail*)

Descriptive Statistics

Skew:

When Mean $<$ Median

Skewed to the **left** (*long left tail*)



Descriptive Statistics

Skew:

When Mean $>$ Median

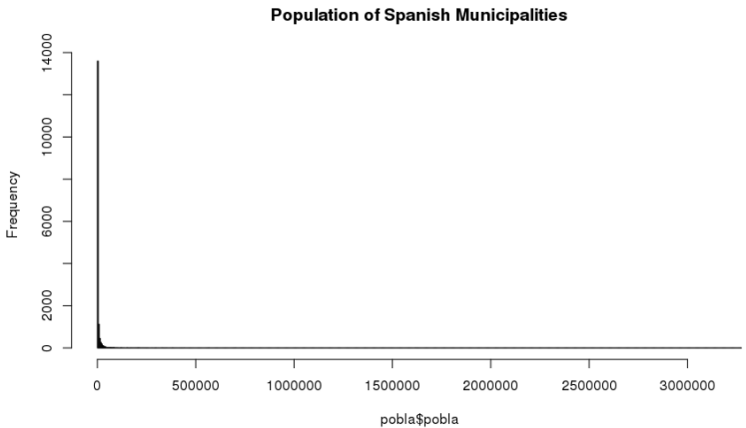
Skewed to the **right** (*long right tail*)

Descriptive Statistics

Skew:

When Mean $>$ Median

Skewed to the **right** (*long right tail*)



Descriptive Statistics

Skew:

When Mean $>$ Median

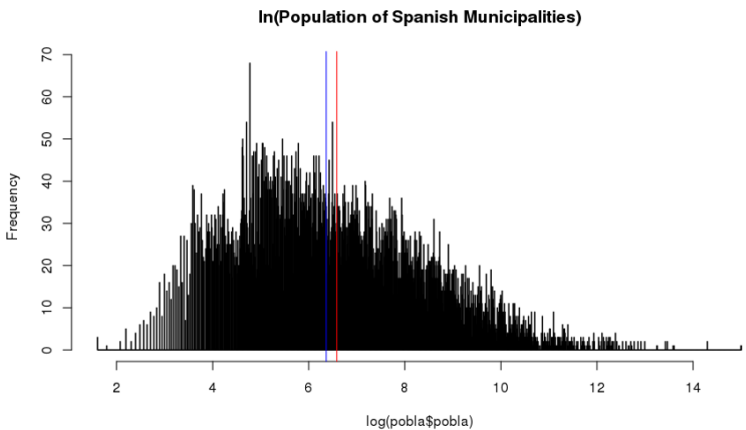
Skewed to the **right** (*long right tail*)

Descriptive Statistics

Skew:

When Mean $>$ Median

Skewed to the **right** (*long right tail*)



Descriptive statistics

Practice in R