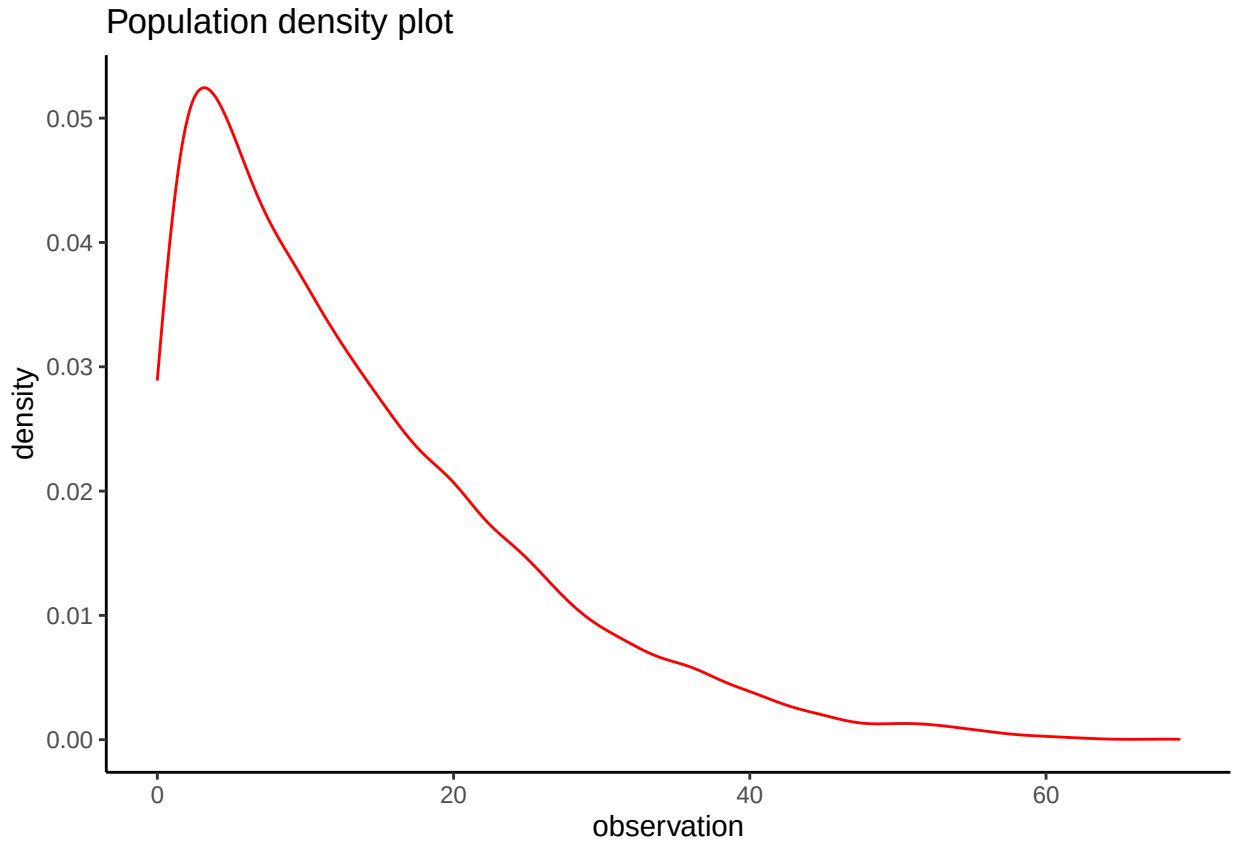
List & explain Statistical Measures of a distribution

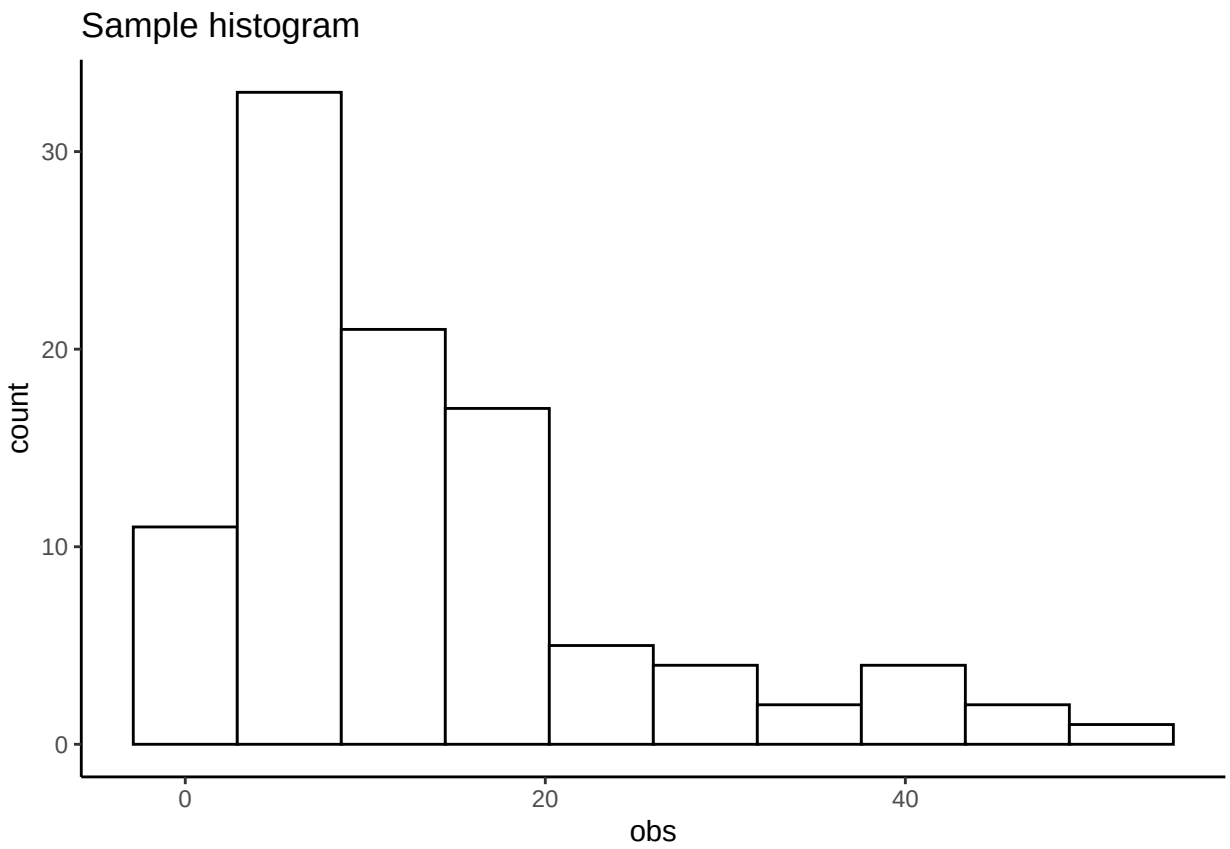
Introduction to Data Science homework - 17/11/2021

Pedro Magalhães - 200202298

Population and sample

All statistical measures present on this report will be calculated using a sample of 100 observations of a population of a continuous variable which closely follows a Beta distribution with $\alpha = 1$ and $\beta = 5$.





The data used on this exercise is therefore right skewed (eg: age distribution of a day center with kindergarden and elderly support) which affects the main calculation and interpretation of the main statistical measures.

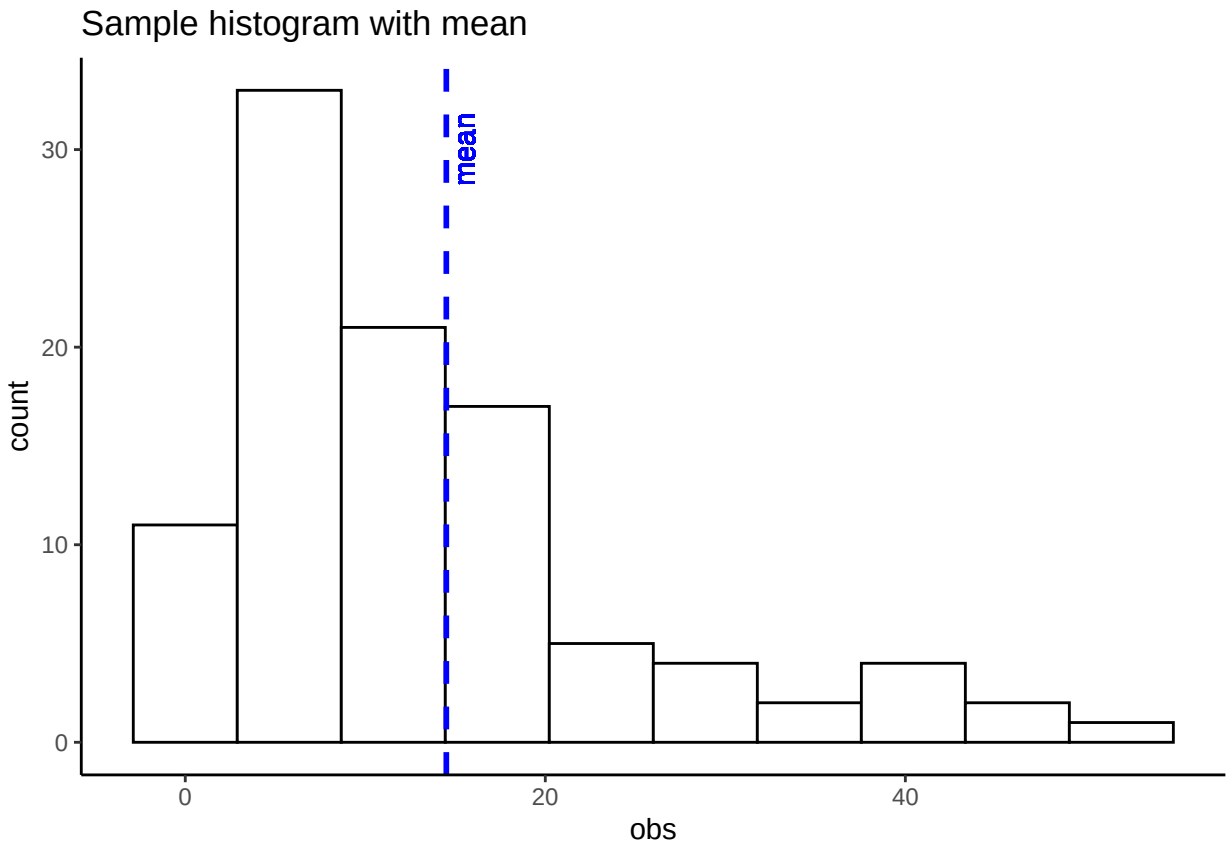
Measures of centrality

Mean

Wikipedia definition: In mathematics and statistics, the arithmetic mean, or simply the mean or the average (when the context is clear), is the sum of a collection of numbers divided by the count of numbers in the collection.

Therefore, given a sample of $\{x_1, x_2, ..., x_n\}$ the the average $\overline{x}$ is defined as follows

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

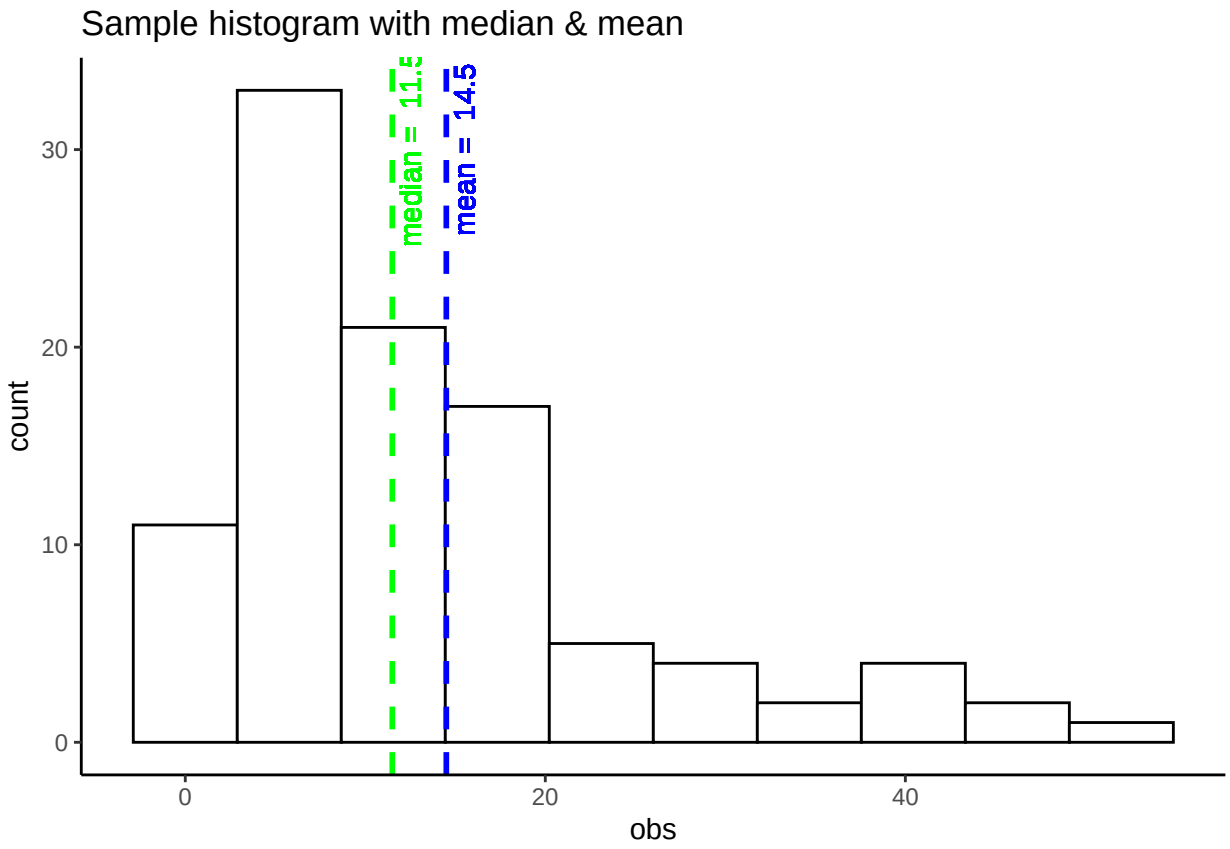


Given the skewed nature of our sample data the interpretation of the mean should be done with caution since its not a suitable measure to describe the above distribution and does not provide a "true" center.

Median

Wikipedia definition: In statistics and probability theory, the median is the value separating the higher half from the lower half of a data sample, a population, or a probability distribution. For a data set, it may be thought of as "the middle" value.

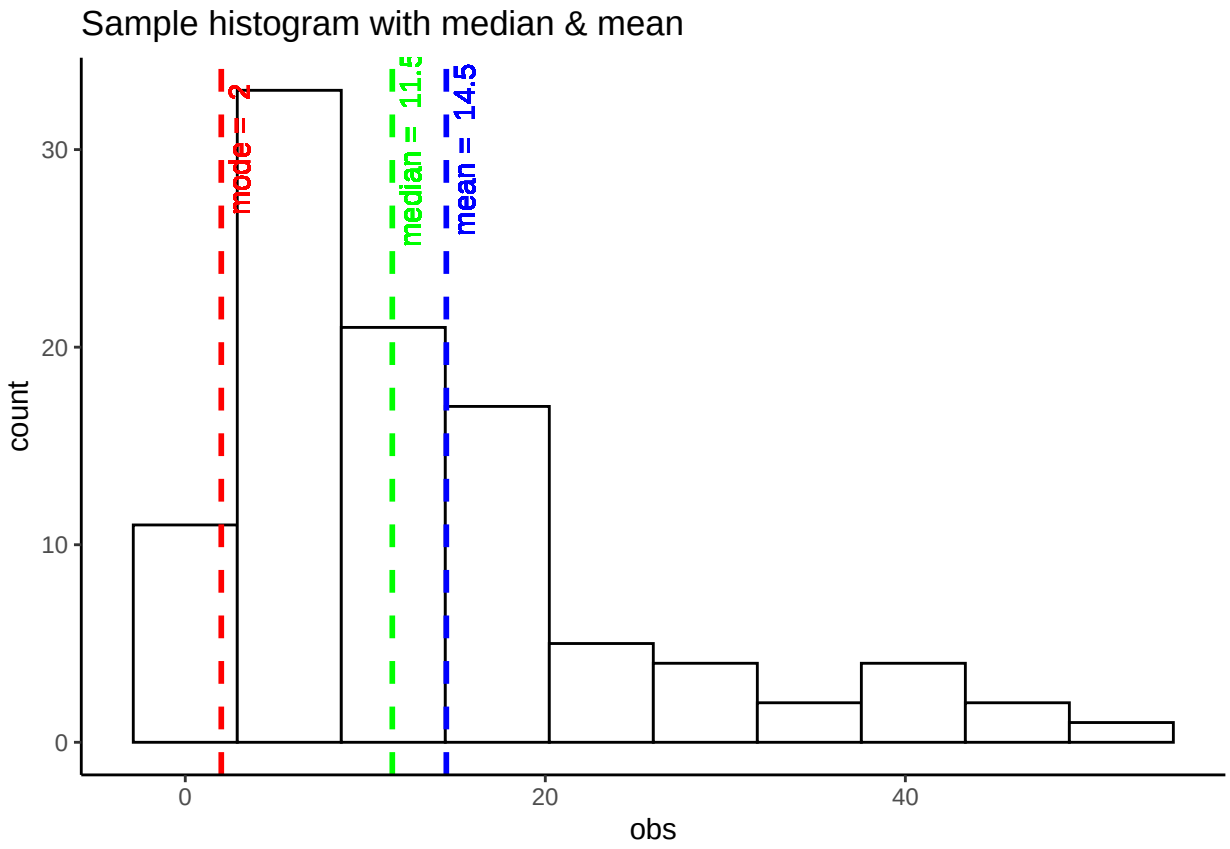
The median represents the percentile 50 of a distribution



Mode

Wikipedia definition: The mode is the value that appears most often in a set of data values.

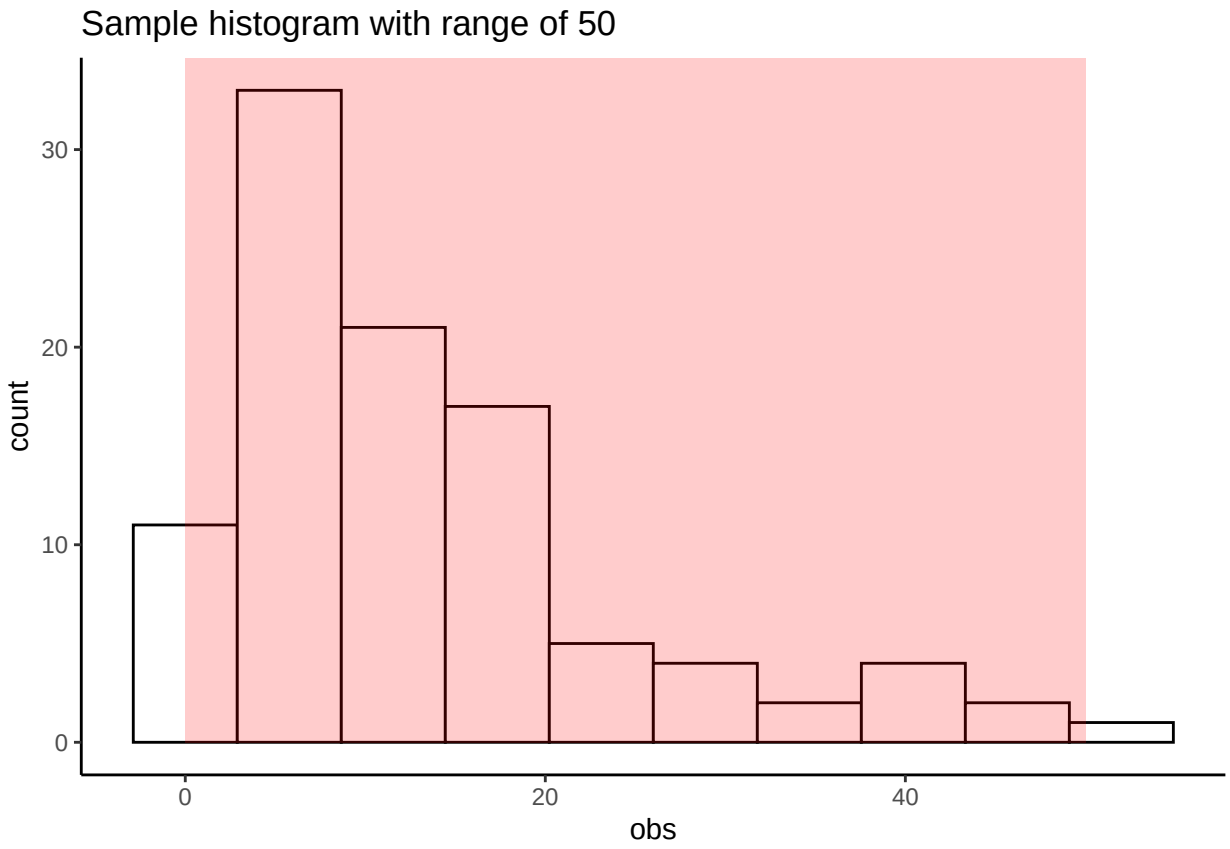
In our scenario it shows us that most observations are of the value 5.



Measures of dispersion

Range

Wikipedia definition: In statistics, the range of a set of data is the difference between the largest and smallest values.[1] Difference here is specific, the range of a set of data is the result of subtracting the sample maximum and minimum.



Variance & Standard deviation

Wikipedia definition: In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values.[1] A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range.

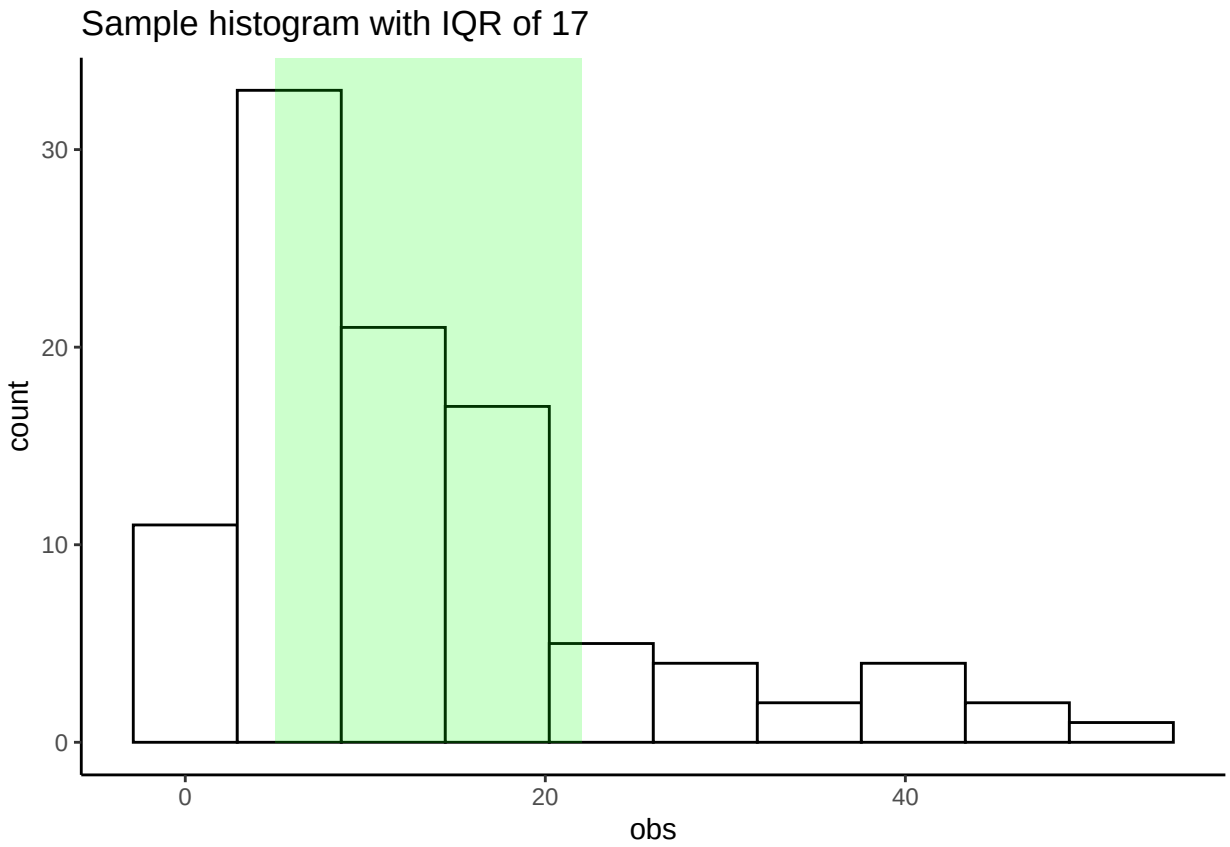
Therefore, given a sample of $\{x_1, x_2, ..., x_n\}$ the the standard deviation $\overline{x}$ is defined the squared root of the variation:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Given the the skewed nature of our sample the varianace and standard deviation does not provide a "true" image on the dispersion around the mean. In order to understand the dispersion of this distibution it is advisable to use the Quartiles and inter quartiles information.

Quartiles and InterQuartile Range

Wikipedia definition: In statistics, a quartile is a type of quantile which divides the number of data points into four parts, or quarters, of more-or-less equal size. The data must be ordered from smallest to largest to compute quartiles; as such, quartiles are a form of order statistic.


Sample histogram with IQR of 17
count
30
20
10
0
0
20
40
obs