Quantitative Variables

Ryan Miller

Motivation

- Shown below are the quantitative variables in the "Tips" dataset, but how useful is this information?

total_bill	tip	size
12.69	2.00	2
17.29	2.71	2
7.51	2.00	2
11.35	2.50	2
10.07	1.25	2
14.00	3.00	2
10.33	2.00	2
11.17	1.50	2
24.52	3.48	3
27.05	5.00	6
20.27	2.83	2
12.03	1.50	2
44.30	2.50	3
13.27	2.50	2

Summarization

- **Raw data** is difficult to make sense of
- **Summarization** is way to condense raw data into a more interpretable form
  - Ideally we can summarize a variable using one number, or a small set of numbers, in order to make informed judgements

Summarization

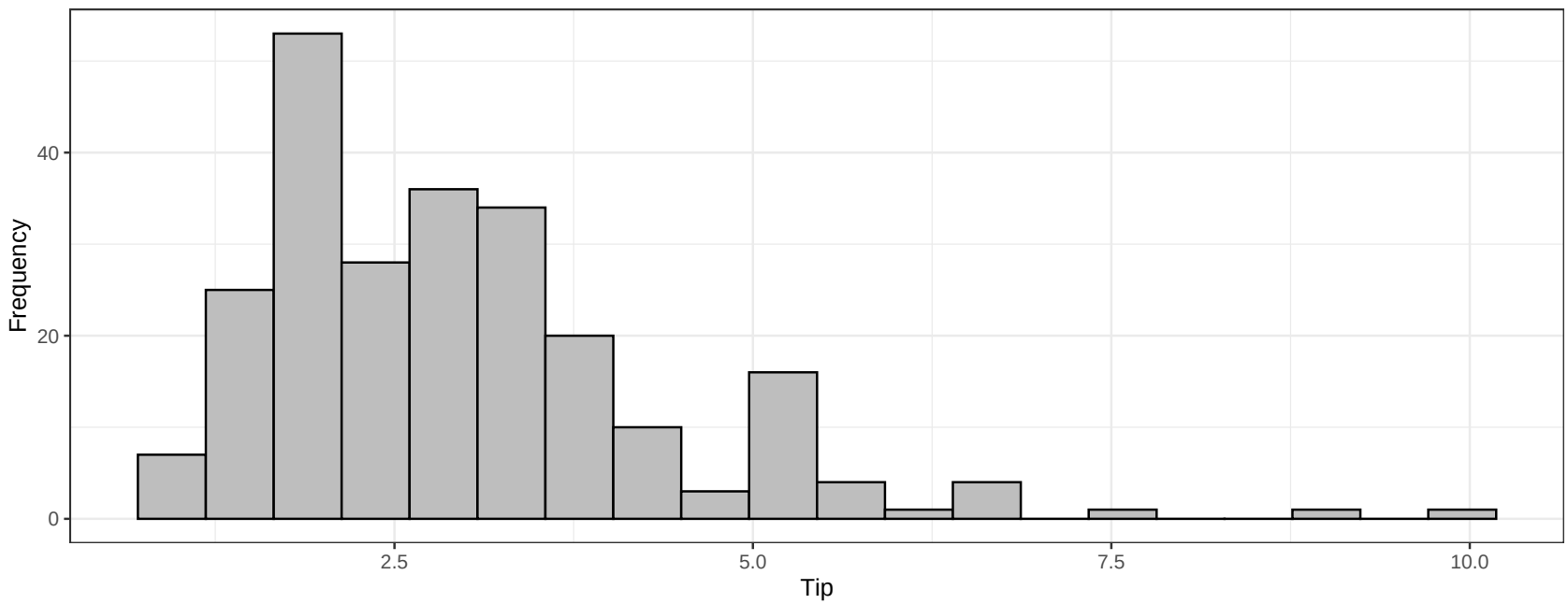
- **Raw data** is difficult to make sense of
- **Summarization** is way to condense raw data into a more interpretable form
  - Ideally we can summarize a variable using one number, or a small set of numbers, in order to make informed judgements
- Today we'll focus on **univariate summaries**, or those involving only a single variable
  - Soon we'll start dealing with more interesting stuff involving multiple variables

Distributions

- Before getting into summarization, we should touch on *distributions*
- A variable's **distribution** describes values that are possible and how frequently they occur

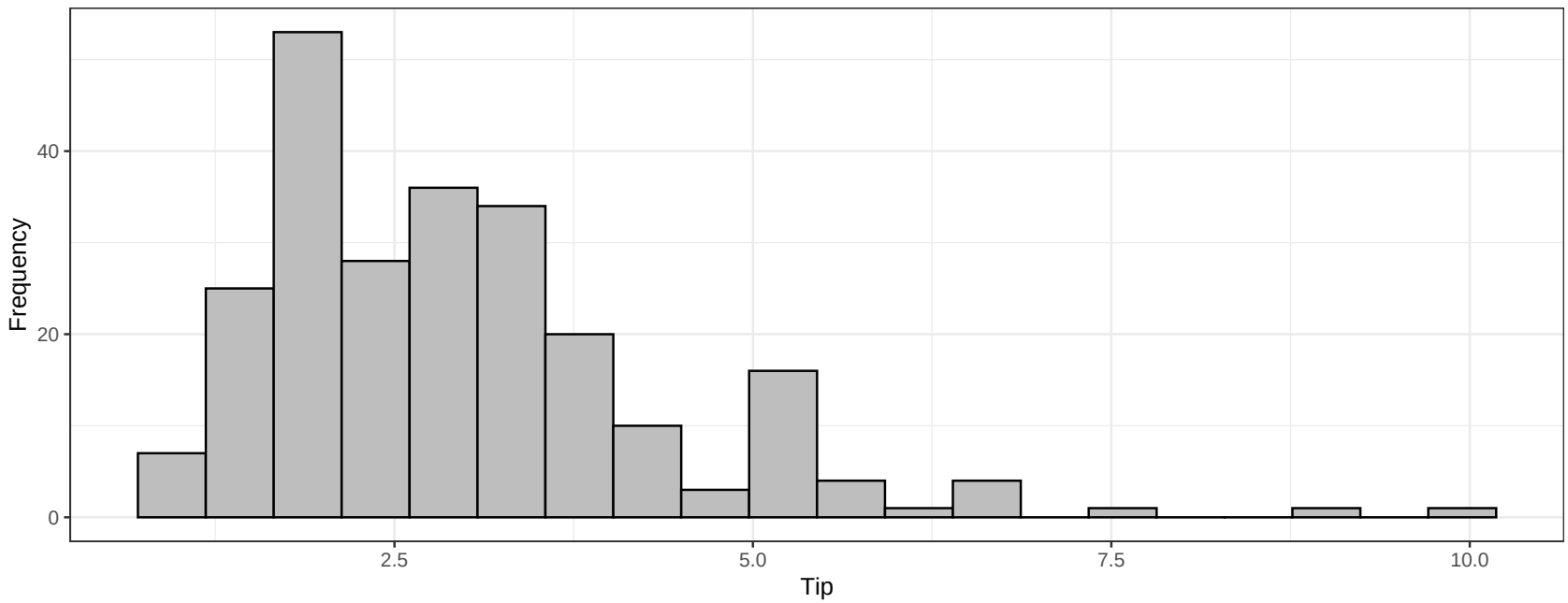
Distributions

- Before getting into summarization, we should touch on *distributions*
- A variable's **distribution** describes values that are possible and how frequently they occur
- Below is a **histogram**, one way of showing a distribution of a quantitative variable



Histograms

- A histogram works by dividing the quantitative variable of interest into **bins**, or equal length intervals
  - The number of cases that belong to each bin are graphed on the y-axis
- Notice how \$2-3 tips are most common, larger tips of \$5+ do occasionally occur, tips over \$10 almost never occur



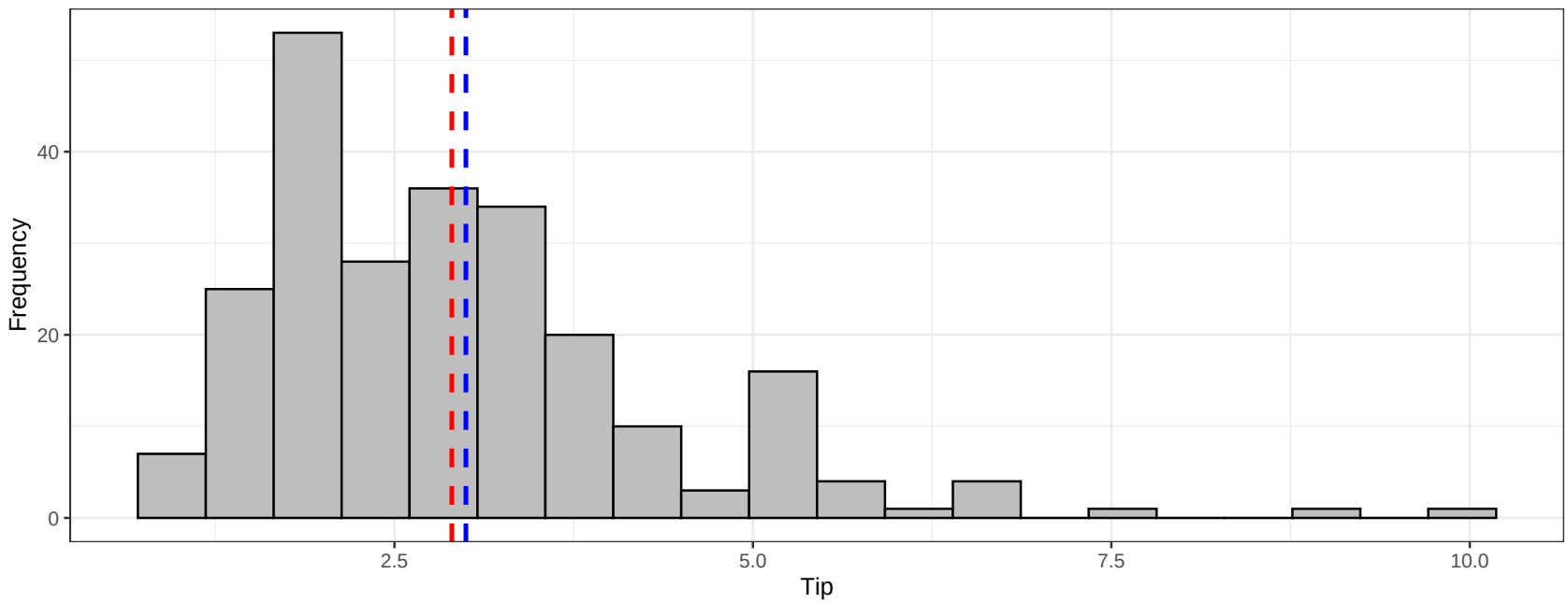
The Mean

- Distributions aren't a summary, but they can help us understand summarization
- The **mean**, or arithmetic average, is way of describing the *center of a distribution*
  - The mean can provide us a sense of what is typical for a quantitative variable

$$\text{Mean} = \frac{\text{Sum across all cases}}{\text{Number of cases}}$$

The Median

- Another way approach to describing the center of a distribution is the **median**, or the midpoint if the variable's values were arranged from smallest to largest
- The histogram below shows the mean tip (blue) and the median tip (red)
  - Why is the mean larger?



Mean vs. Median

- The median is considered a *robust* measure of the center of a distribution because it is not heavily influenced by extreme values
  - The table below shows the impact of adding a 100-dollar tip to our prior data

	Mean	Median
Original	3.00	2.9
With $100 tip	3.39	2.9

Other Summaries

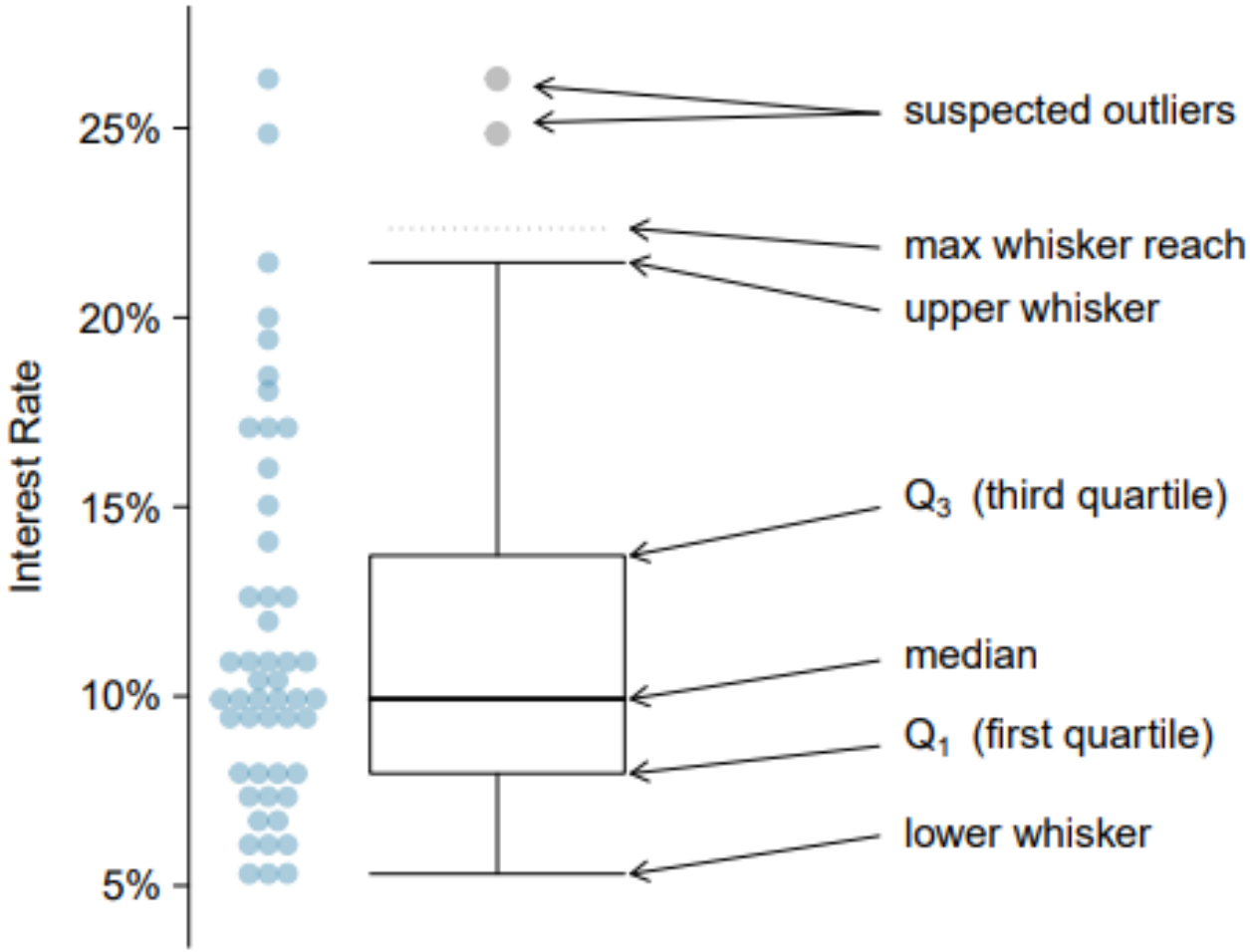
- Sometimes we aren't exclusively interested in the center of a variable's distribution
- The **minimum** and **maximum** are self-explanatory summaries of a variables most extreme values

Other Summaries

- Sometimes we aren't exclusively interested in the center of a variable's distribution
- The **minimum** and **maximum** are self-explanatory summaries of a variables most extreme values
- **Percentiles** describe a cutoff value for which P data falls below
  - The median is the 50^{th} percentile
  - The 25^{th} and 75^{th} percentiles are called the **first quartile**, or Q1, and the **third quartile**, or Q3

Boxplots

- The summary measures presented on the previous slide can be used to construct a visualization known as a **boxplot**



Spread

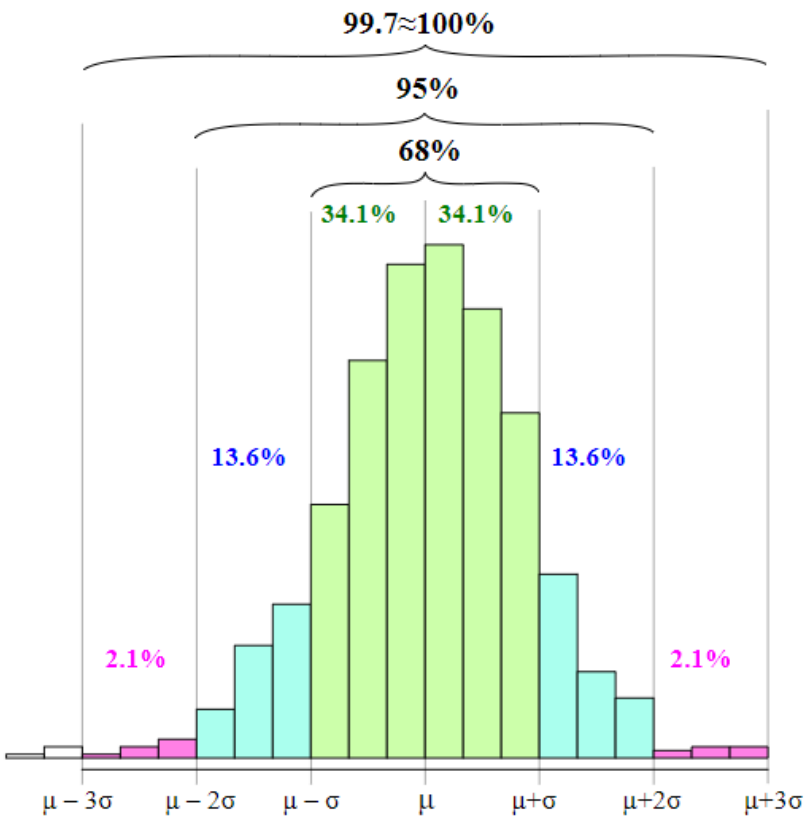
- The mean and median summarize the *center* of a distribution
- It is also useful to summarize the *spread*, or how the data values tend to vary around the center

Spread

- The mean and median summarize the *center* of a distribution
- It is also useful to summarize the *spread*, or how the data values tend to vary around the center
  - The **range** is the difference between the minimum and maximum
  - The **interquartile range**, or **IQR**, is the difference between the third and first quartiles (Q1 and Q3)

Standard Deviation

- The most widely used measure of spread is the **standard deviation**, which roughly corresponds to the *average distance of each data-point from the mean*
- For bell-shaped distributions, the standard deviation is related to the percentage of cases within a certain distance from the mean



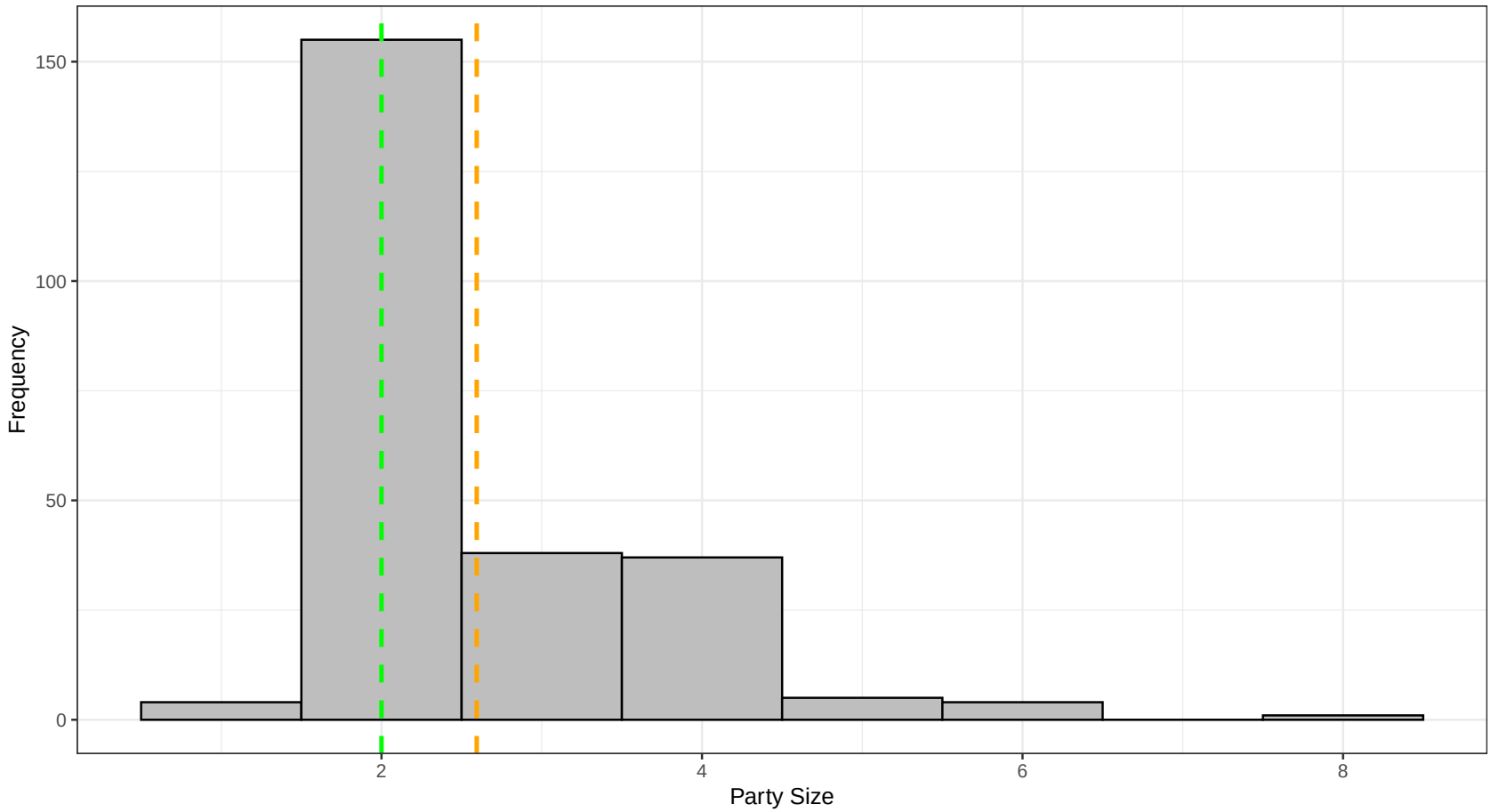
Standard Deviation vs. IQR

- Similar to how the median is more robust to extreme values than the mean, the IQR is more robust than the standard deviation

	Mean	Median	StDev	IQR
Original	3.00	2.9	1.38	1.56
With $100 tip	3.37	2.9	6.35	1.56

Practice

Using the graph below, answer the following: 1) What is the name of this graph? 2) How many bins are displayed? 3) Which color line marks the mean and which marks the median?



Practice (solution)

1) Histogram
2) 8 bins (note that one of them has zero cases in it)
3) green = median, orange/yellow = mean