Section 1: VISUALIZING DATA

One-way tables

Individuals: The set of elements (whether people or otherwise) that are surveyed to form a set of data about those individuals

Variables: Each property we collect in our data about the individuals

Data: The collection of individuals and variables

Data table: A table that organizes the data, including the individuals and their variables

Categorical Variables: Non-numerical variables, also called "qualitative" variables. Their values aren't represented with numbers.

Quantitative variables: Numerical variables. Their values are numbers.

Discrete variables: Variables we can obtain by by counting. Therefore, they can take on only certain numerical values.

Continuous variables: Variables that can include data such as decimals, fractions, or irrational numbers.

Nominal scale of measurement: Things like favourite food, colors, names, and 'yes' or 'no' responses have a nominal scale of measurement. Only categorical data can be measured with a nominal scale.

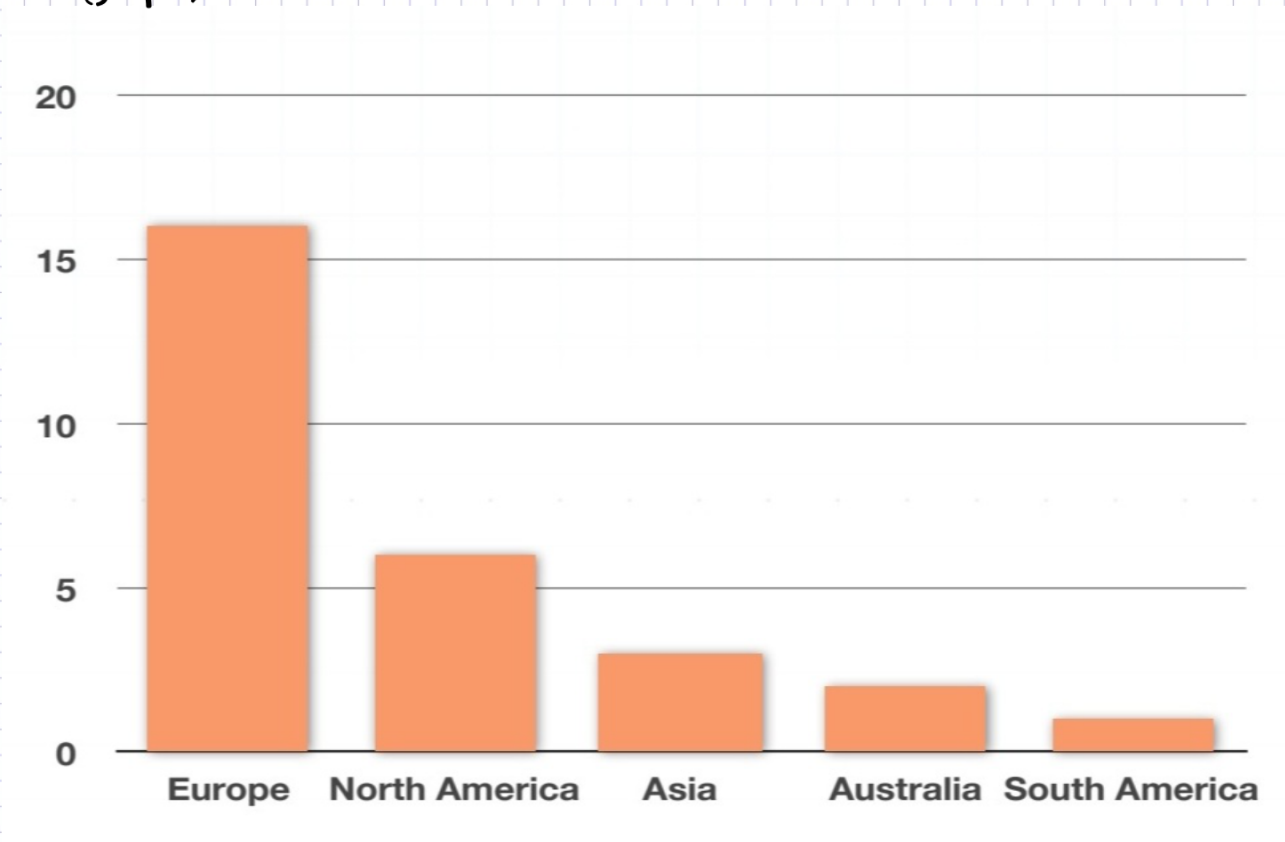
Ordinal scale of measurement: Categorical data can also be ordinal. This type of data can be ordered.

Interval scale of measurement: Data scaled using an interval scale can be ordered like ordinal data. But interval data also gives us a known interval between measurements.

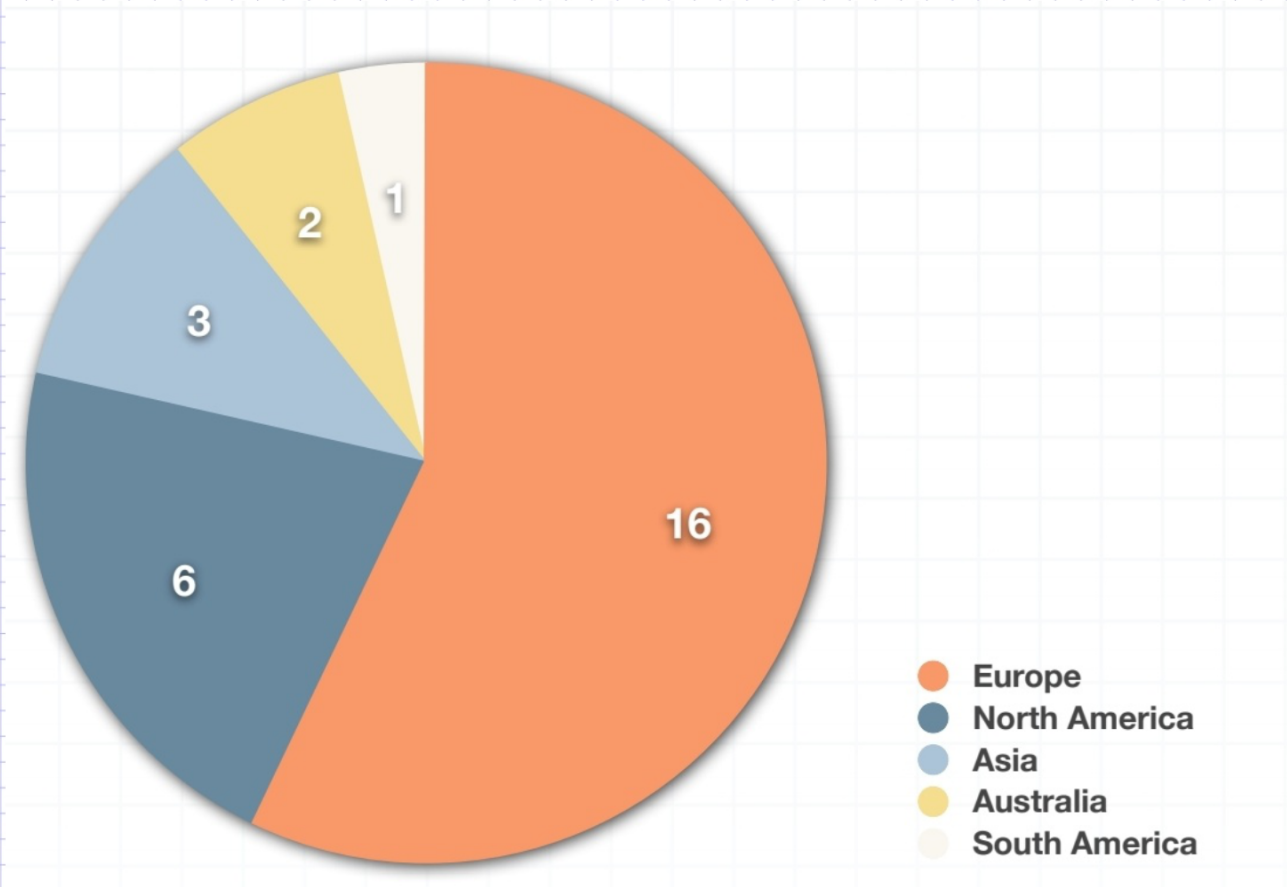
Ratio scale of measurement: Data measured using a ratio scale is just like interval scale data, except that ratio scale data has a starting point, or absolute zero.

Bar graphs and pie charts

Bar graph, bar chart:



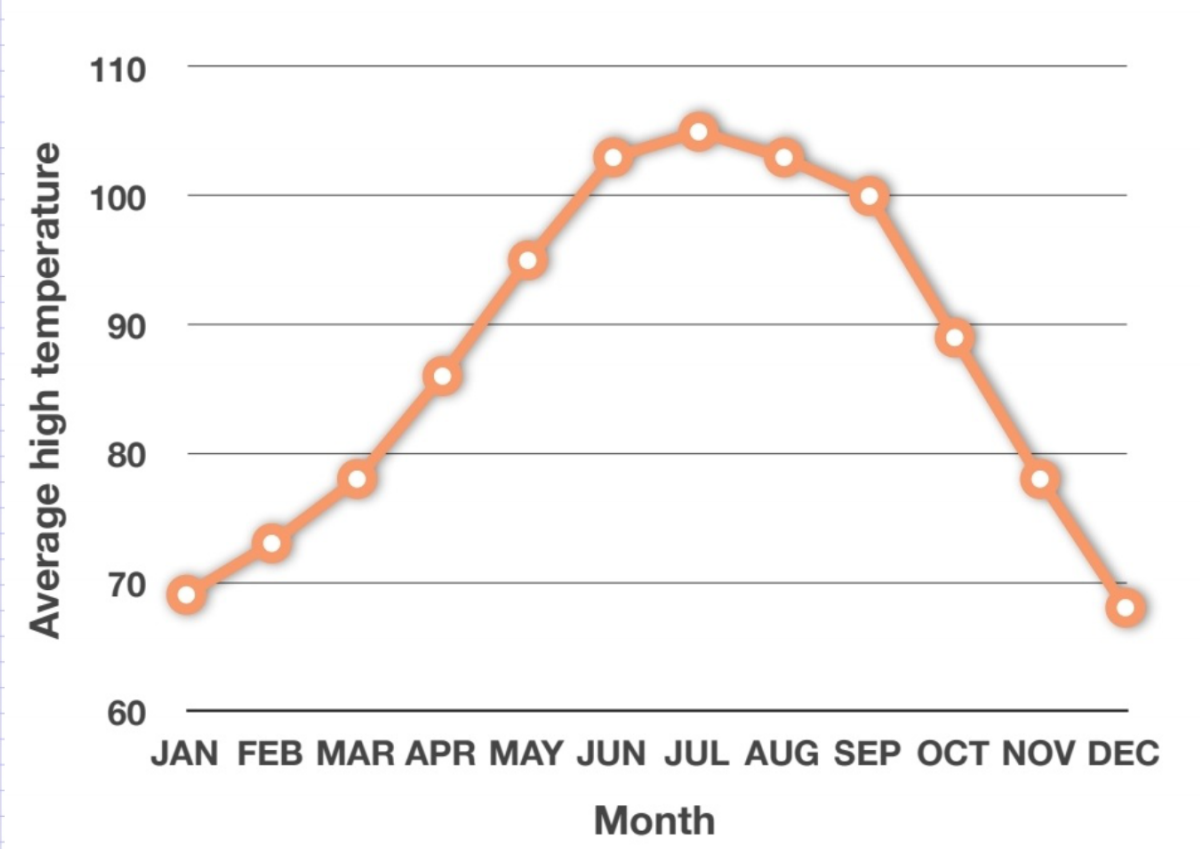
Pie chart



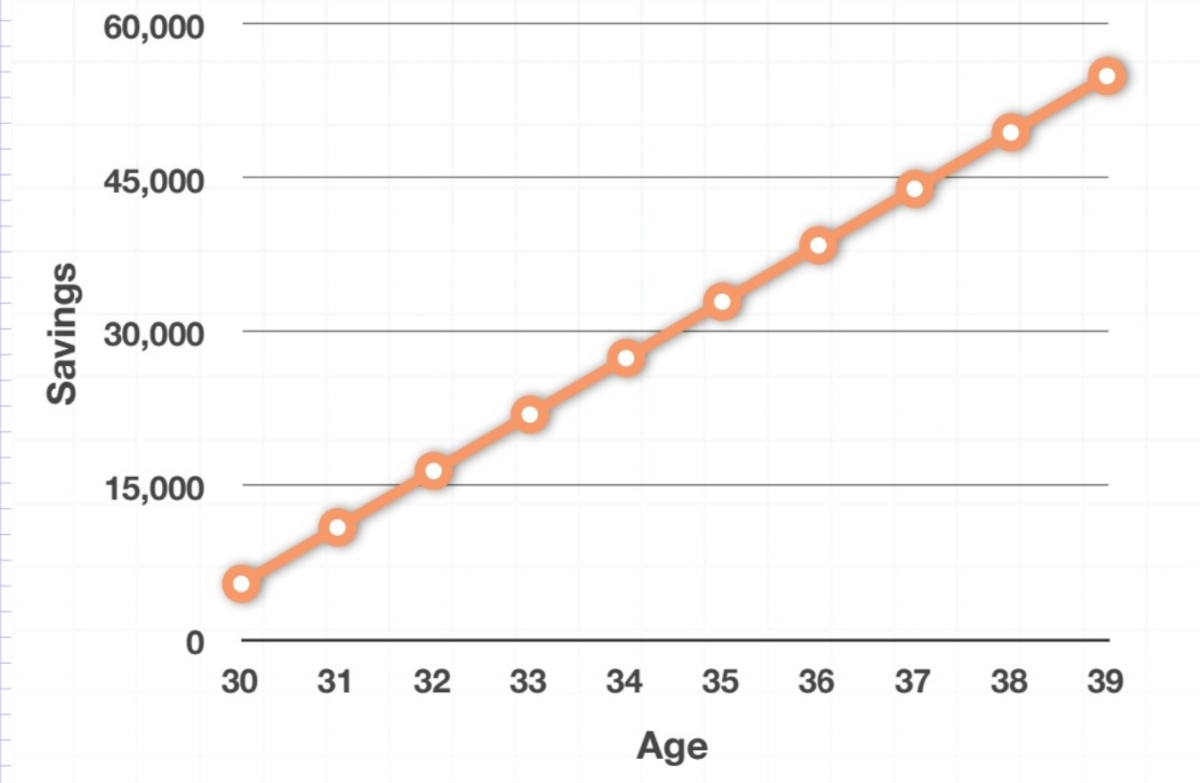
Frequency table: A summary table that shows the frequency or count of each categorical variable.

Line graphs and ogives

Line graph:

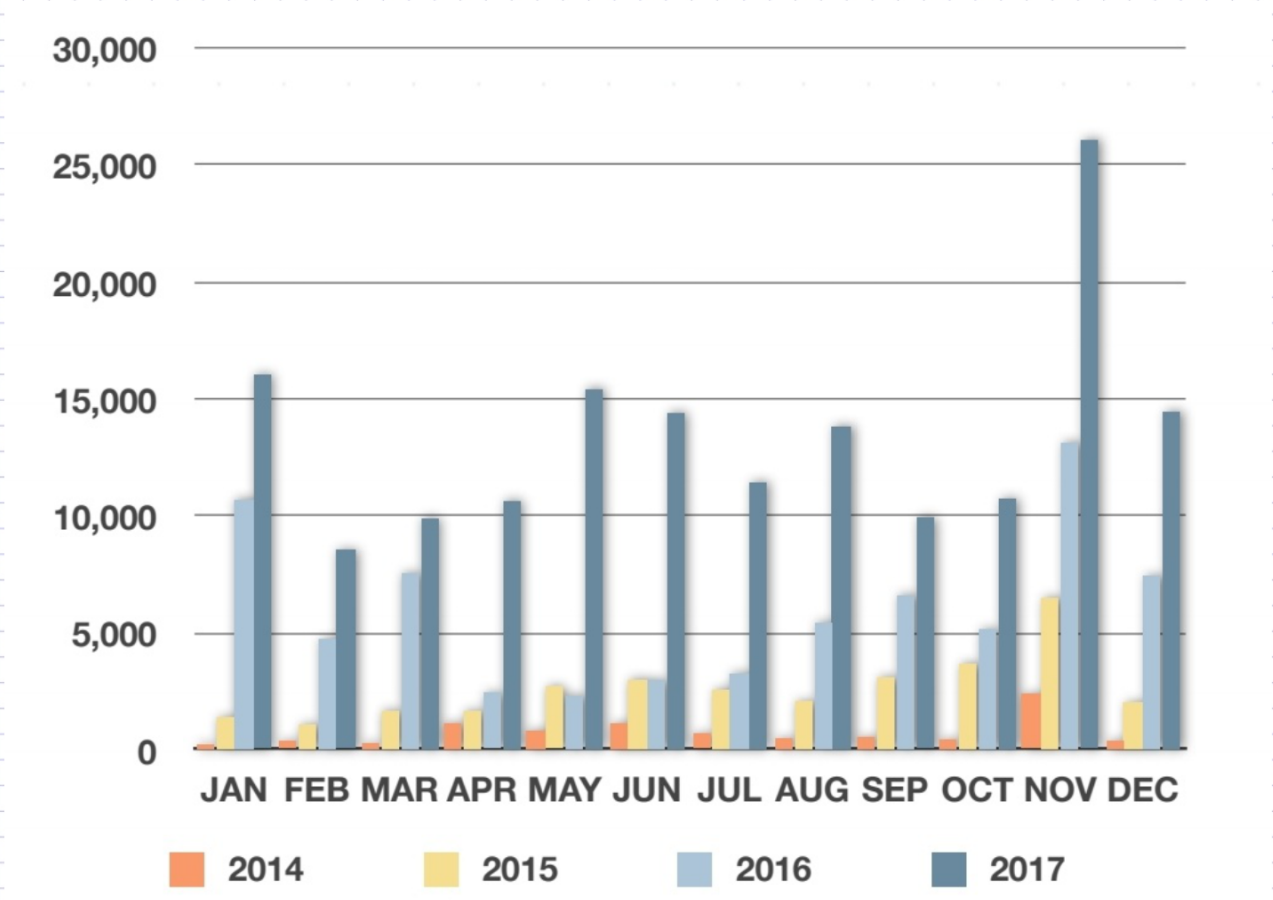


Ogive:

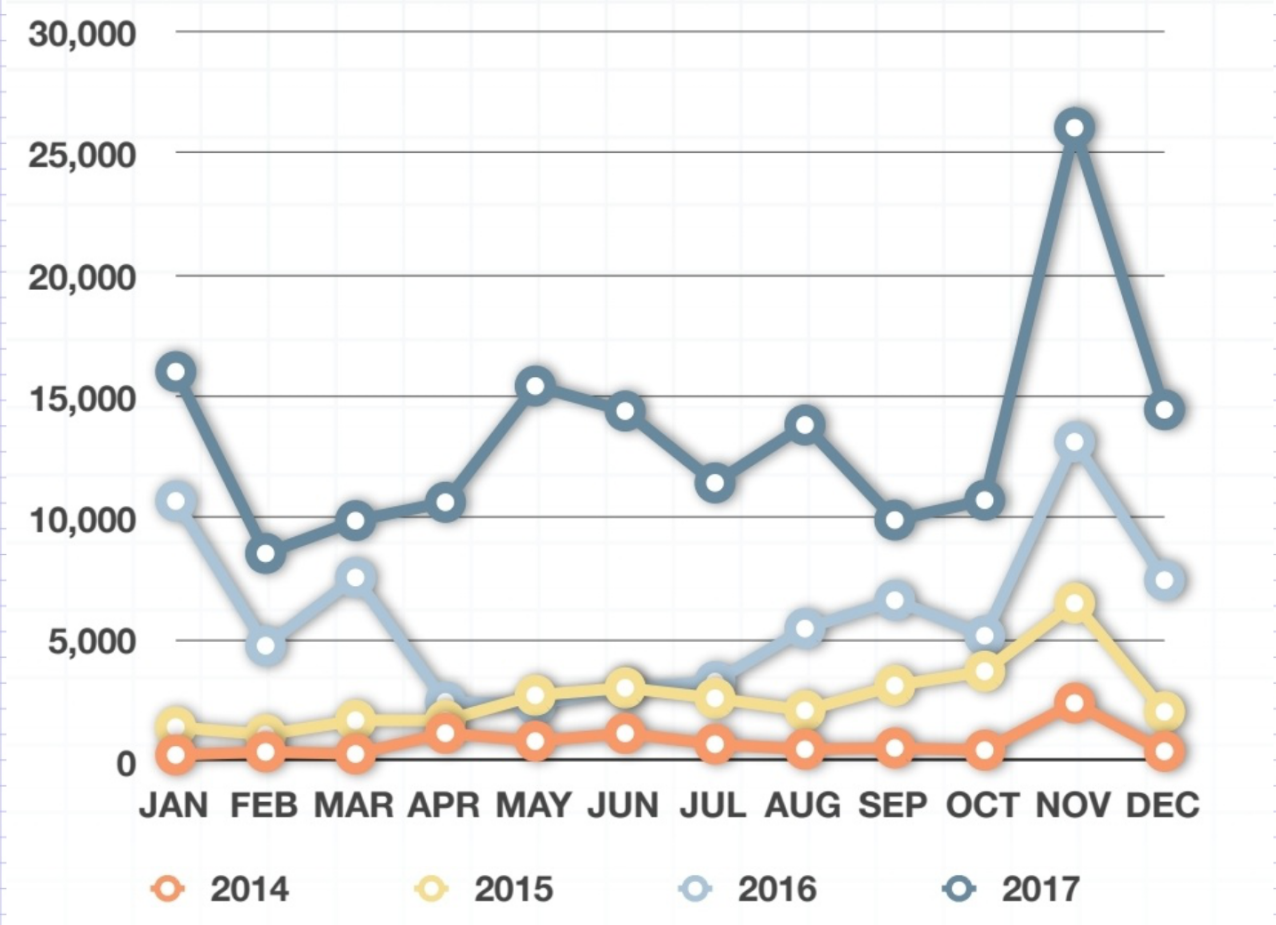


Two-way tables

Comparison bar graph:

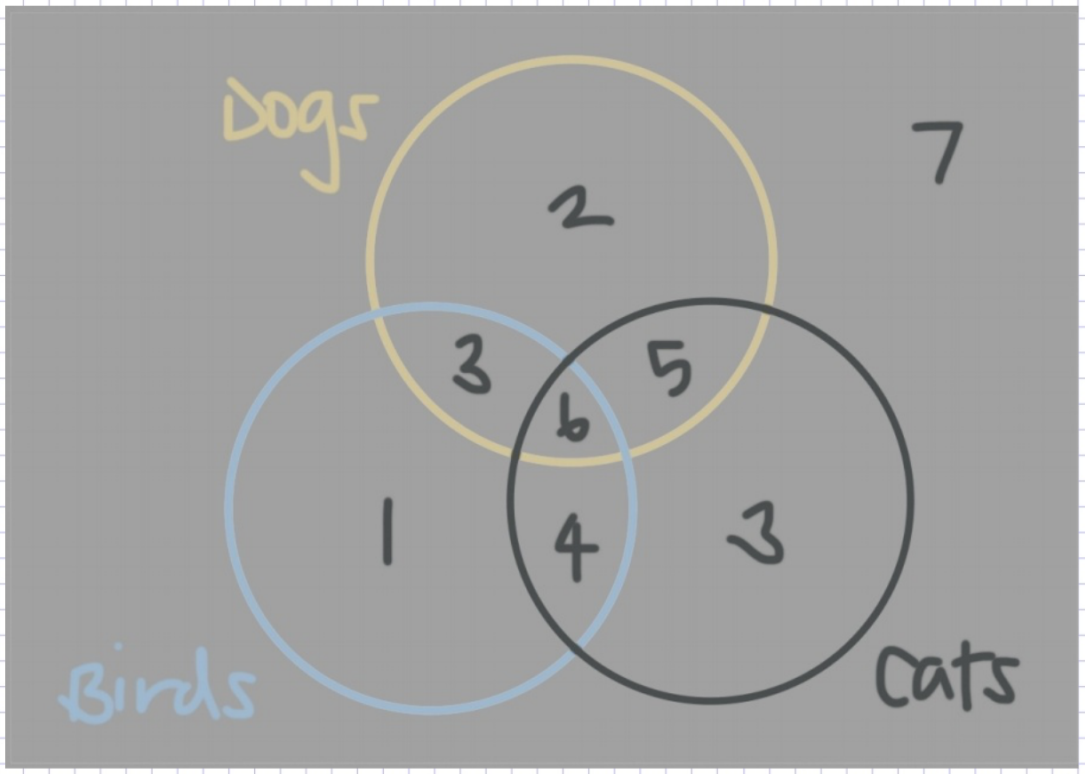


Comparison line graph:



Venn diagrams

Venn diagram:



Frequency tables and dot plots

Frequency table: A table that displays how frequently or infrequently something occurs

Dot plot: A plot that can be used to show the frequency of small data sets.

Relative frequency tables

Relative frequency table: A table that shows percentages instead of actual counts

Column-relative frequency table: A relative frequency table in which the columns sum to 1.00 but the rows do not

Row-relative frequency table: A relative frequency table in which the rows sum to 1.00 but the columns do not.

Total-relative frequency table: A relative frequency table in which both the rows and columns sum to 1.00. A grand total cell is also included, which is always 1.00.

Joint distributions

Joint distribution: A table of percentages similar to a relative frequency table. The difference is that, in a joint distribution we show the distribution of one set of data against the distribution of another set of data.

Marginal distribution: The total column or the total row in a joint distribution.

Conditional distribution: The distribution of one variable given a particular value of the other variable.

Histograms and stem-and-leaf plots

Histogram: Also called a frequency histogram, a histogram is like a bar graph, except that we collect the data into equally-sized buckets or bins, and then sketch a bar for each bucket. Histograms have no gaps between the bars, because they represent a continuous dataset.

Stem- and leaf plots: Also called a stem plot, a stem-and-leaf plot groups data together based on the first digit(s) in each number. Stems are the numbers on the left, and leaves are the numbers on the right.

Building histograms from data sets

Class interval, class, bin: An individual grouping bucket within a histogram

Class width: The interval over which the class is defined

Class midpoint: The value halfway between the lower and upper edges of the class

How to build a histogram:

1. Put the data set in ascending order, then find the range as the difference between the largest and the smallest values.
2. Determine the number of bins, or classes, that we want to have in our histogram. As a rule of thumb, it's best to use 5-6 classes for most of the data. However, we might want to use up to 20 classes when we deal with larger data sets. It all depends on how large our dataset is and the number of classes would best represent data.
3. Divide the range by the number of classes, the round up the result to get the class width.
4. Build a table, putting each class in a separate row.
5. Find the frequency for each class by counting the data points that fall into each one. It is worth mentioning that we can choose either overlapping intervals or non-overlapping intervals in the previous step. However, for overlapping intervals like 0-4 and 4-8 the data point 4 should be included in the second interval. In other words, the interval 0-4 actually contains data from 0 to $3.\overline{9}$, and the interval 4-8 contains data from 4 to $7.\overline{9}$. Overlapping intervals are particularly useful when the data set contains decimals.
6. Graph the histogram by placing the classes along the horizontal axis and their frequencies along the vertical axis, such that the height of each bar is the frequency of each class.