Graphical data displays

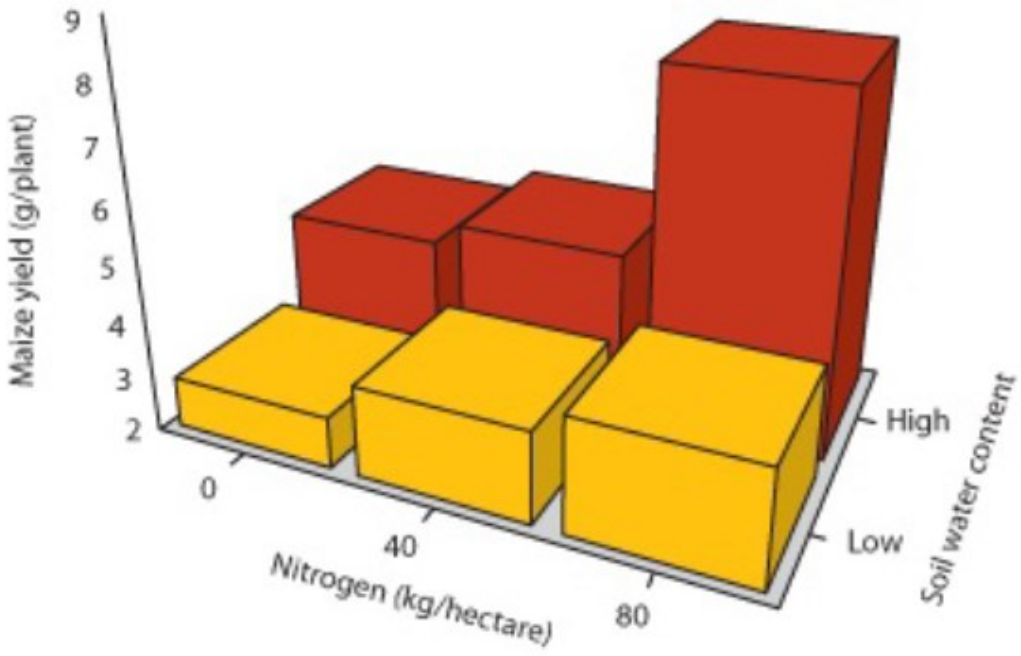
How to draw a bad graph

Mistake #1: Where are the data?

Mistake #2: Patterns in the data are difficult to see.

Mistake #3: Magnitudes are distorted.

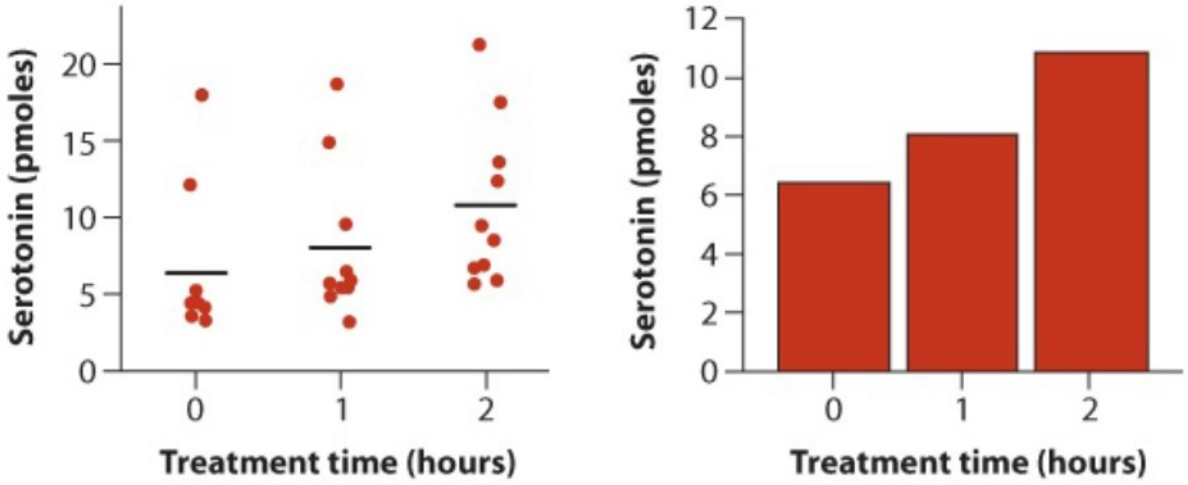
Mistake #4: Graphical elements are unclear.



How to draw a good graph

- Show the data.
- Make patterns in the data easy to see.
- Represent magnitudes honestly.
- Draw graphical elements clearly.

Show the data



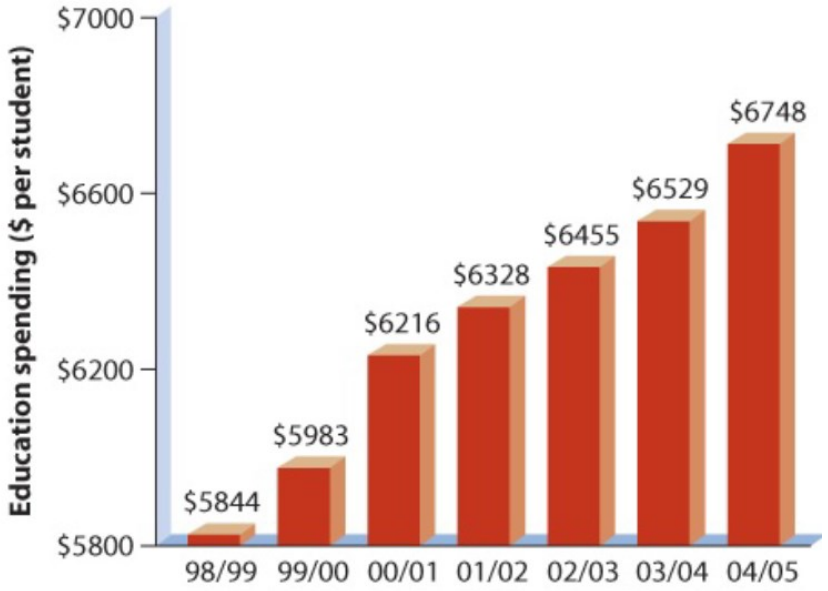
The panel on the left *shows* the data (this type of graph is called a strip chart or dot plot).

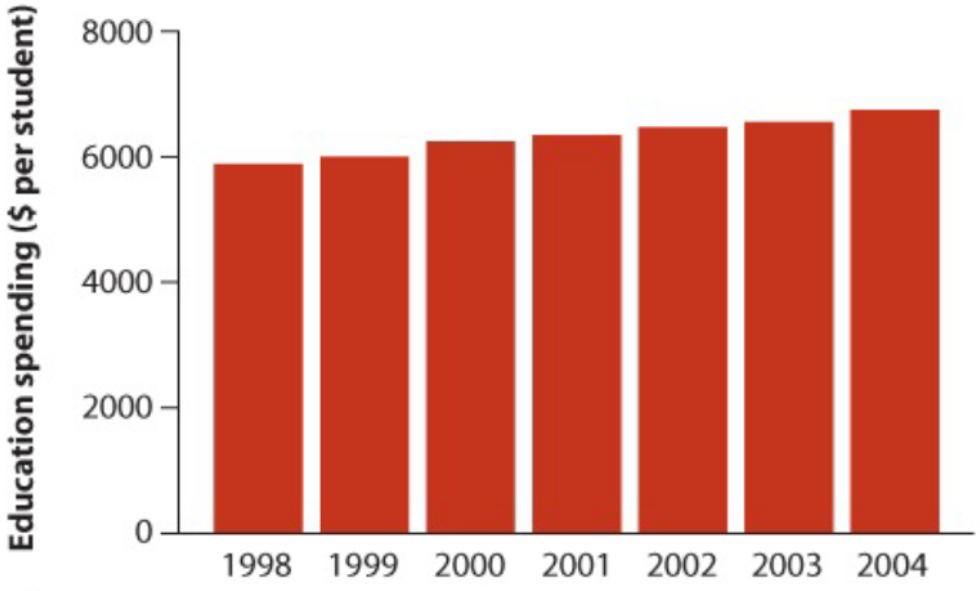
The panel on the right *hides* the data, using bars to show only treatment averages.

How to draw a good graph

Make patterns easy to see.
Try displaying your data in different ways, possibly with different types of graphs, to find the best way to communicate the findings. Avoid putting too much information into one graph.

Represent magnitudes honestly.
One of the most important decisions concerns the smallest value on the vertical axis of a graph (the “baseline”). A bar graph must always have a baseline at zero, because the eye instinctively reads bar height and area as proportional to magnitude. Other types of graphs, such as strip charts, don’t always need a zero baseline if the main goal is to show differences between treatments rather than proportional magnitudes.





How to draw a good graph

Draw graphical elements clearly.
Clearly label the axes and choose unadorned, simple typefaces and colors.

Text should be legible even after the graph is shrunk to fit the final document.

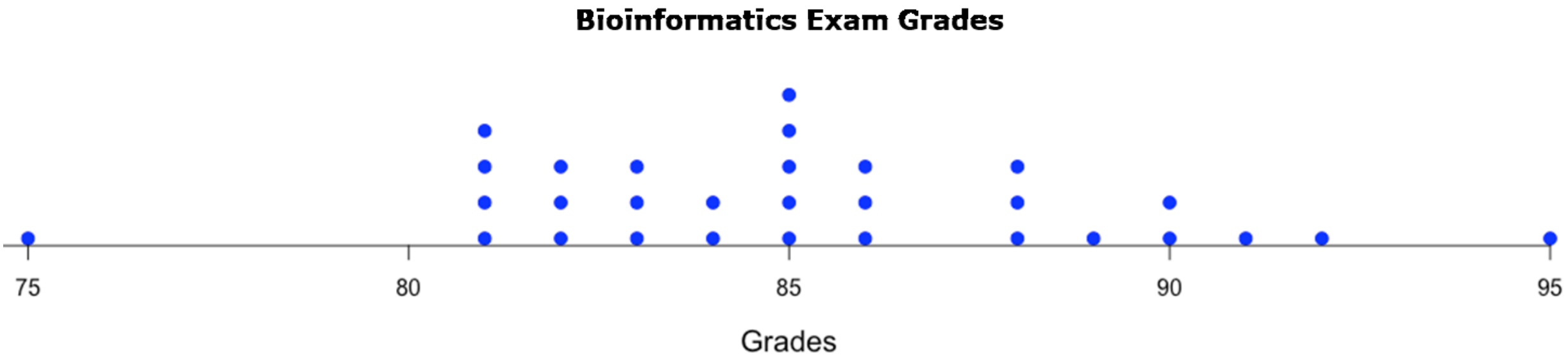
Always provide the units of measurement in the axis label.

Use clearly distinguishable graphical symbols if you plot with more than one kind.

Don’t always accept the default output of statistical or spreadsheet programs.

Graphical data displays

Dot (strip) plots



Graphical data displays

Stem and leaf

A stem and leaf plot, or stem plot, is a technique used to classify either discrete or continuous variables. A stem and leaf plot is used to organize data as they are collected. A stem and leaf plot looks something like a bar graph. Each number in the data is broken down into a stem and a leaf, thus the name. The stem of the number includes all but the last digit. The leaf of the number will always be a single digit. For example, for the number 45, the stem is 4 and leaf is 5.

Elements of a good stem and leaf plot

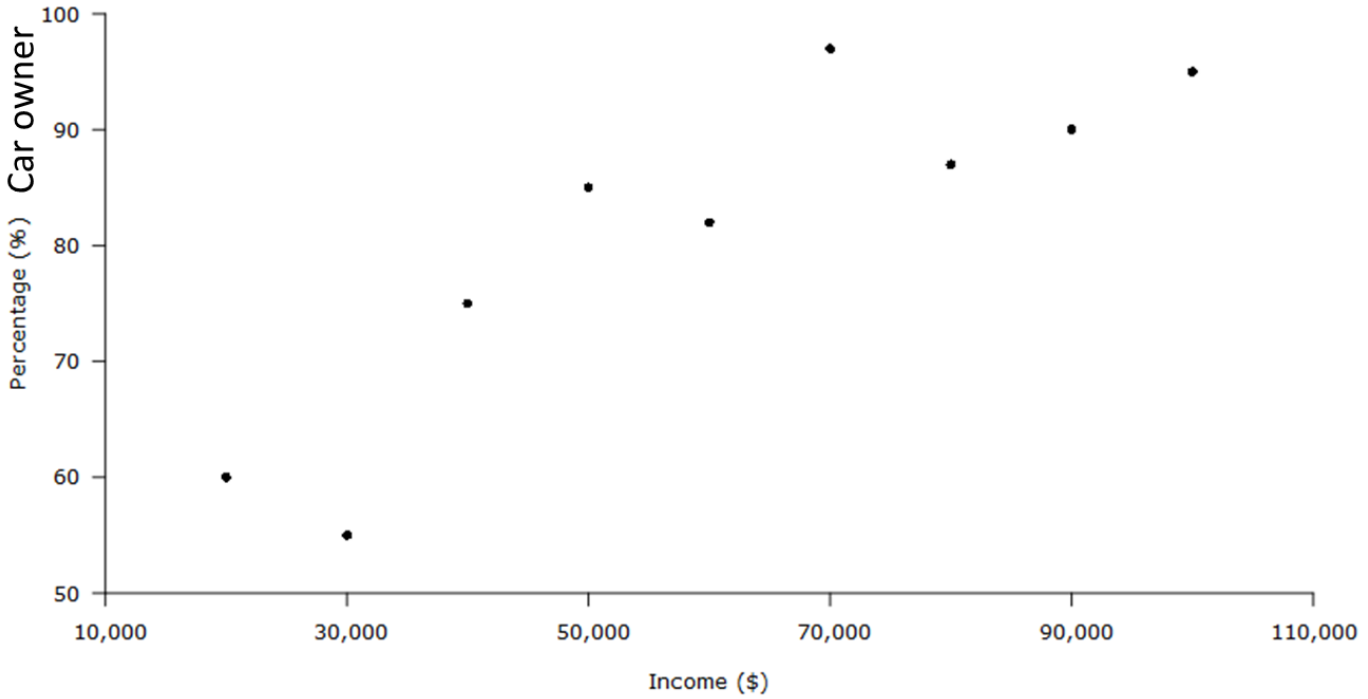
- Anything that has a decimal point is rounded to the nearest whole number.
- Looks like a bar graph when it is turned on its side.
- Shows how the data are spread—that is, highest number, lowest number, most common number and outliers (a number that lies outside the main group of numbers).

Stem	Leaf
4	4, 5, 7, 8, 9, 9
5	7
6	
7	2, 4
8	5, 9
9	2, 3, 3, 6, 7
10	1, 6

Graphical data displays

Scatter plot

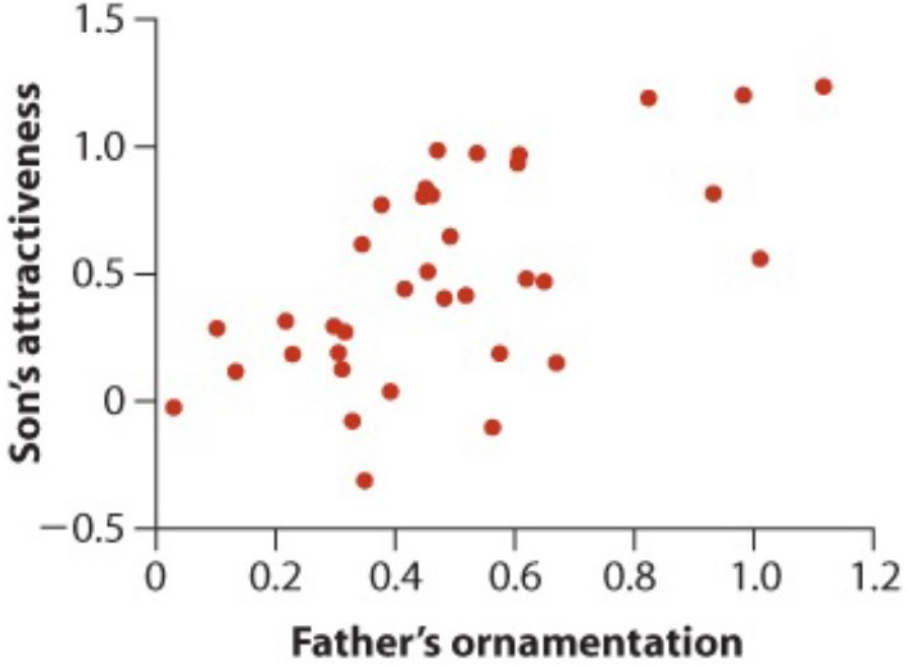
In science, the scatterplot is widely used to present measurements of two or more related variables. It is particularly useful when the values of the variables of the y-axis are thought to be dependent upon the values of the variable of the x-axis. In a scatterplot, the data points are plotted but not joined. The resulting pattern indicates the type and strength of the relationship between two or more variables.



Showing association between two variables

Showing association between numerical variables: scatter plot

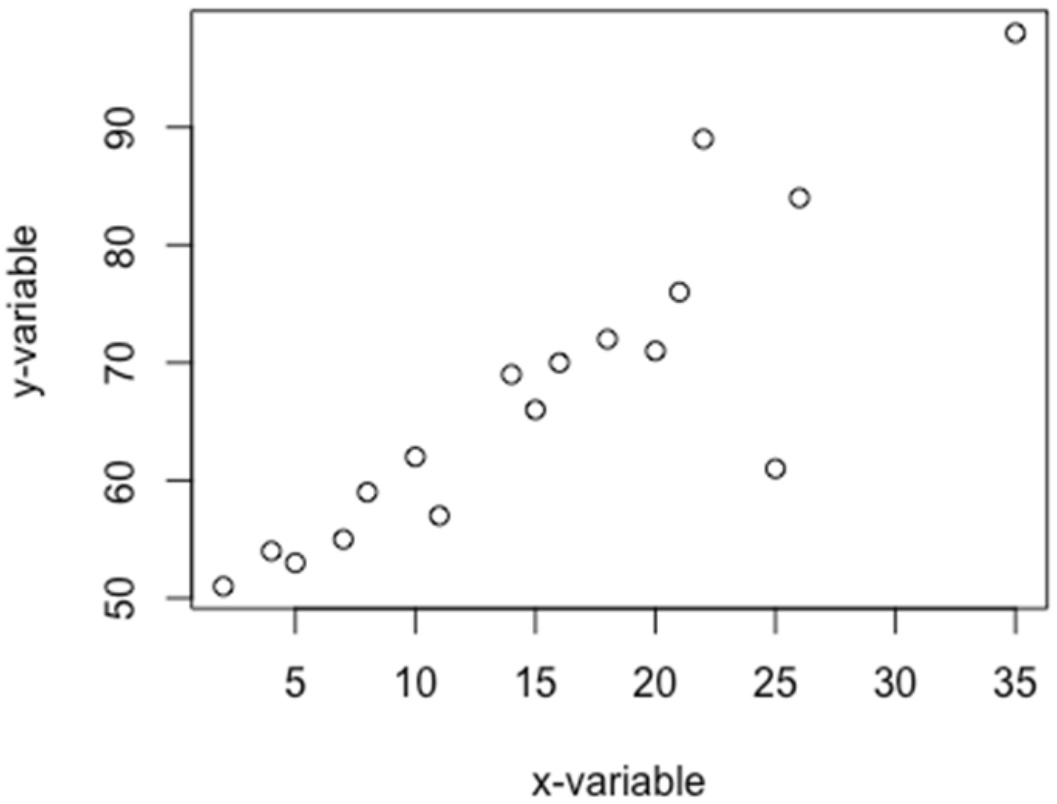
Use a scatter plot to show the association between two numerical variables. Position along the horizontal axis (the *x*-axis) indicates the measurement of the explanatory variable. The position along the vertical axis (the *y*-axis) indicates the measurement of the response variable.



Scatter plot showing the relationship between the ornamentation of male guppies and the average attractiveness of their sons.

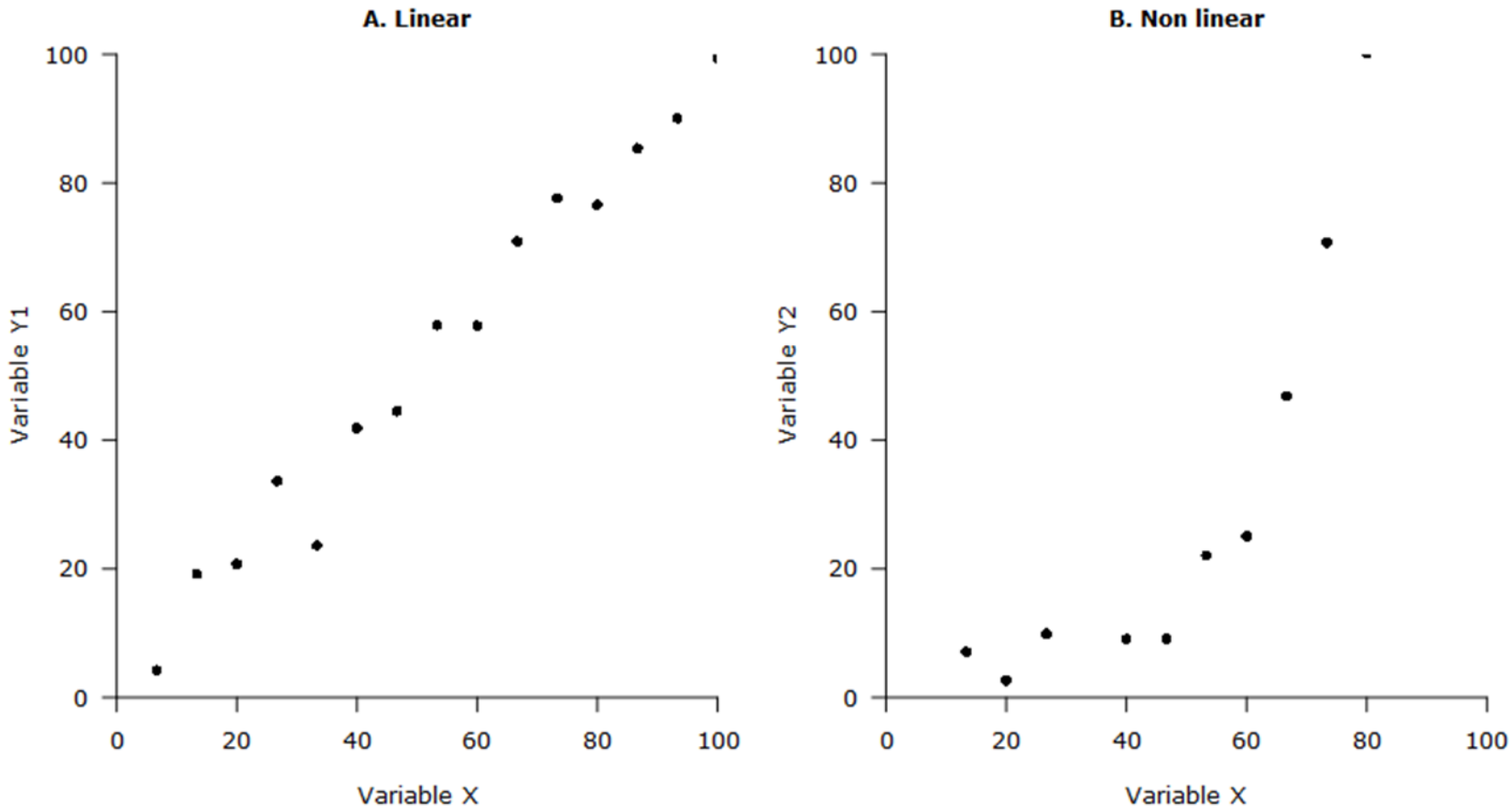
Graphical data displays

Scatter plot: Data are plotted on x-axis and y-axis. The data points are plotted to guess if the data have any association between them.



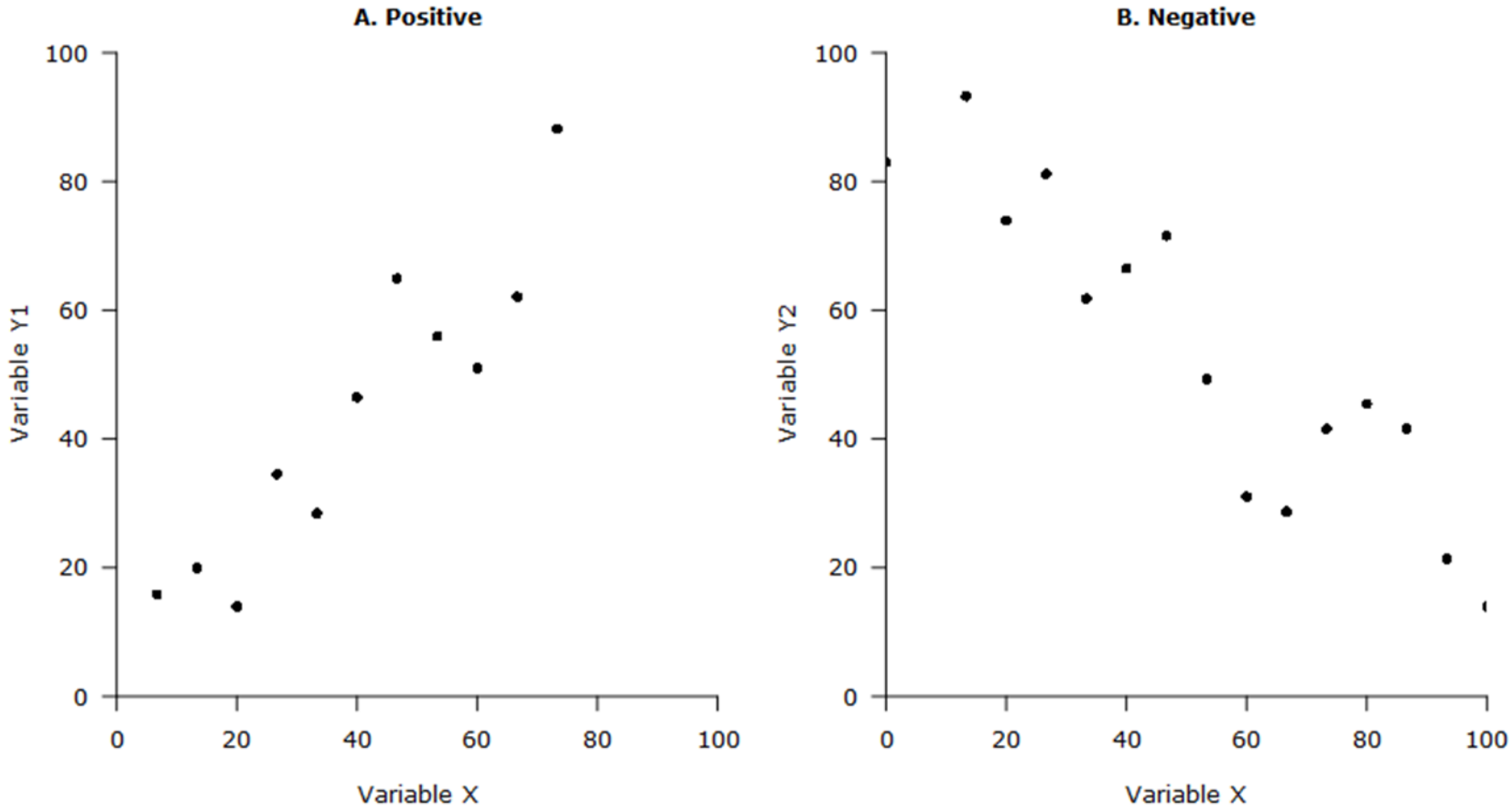
Graphical data displays

Scatter plot: Linear or non-linear relationship



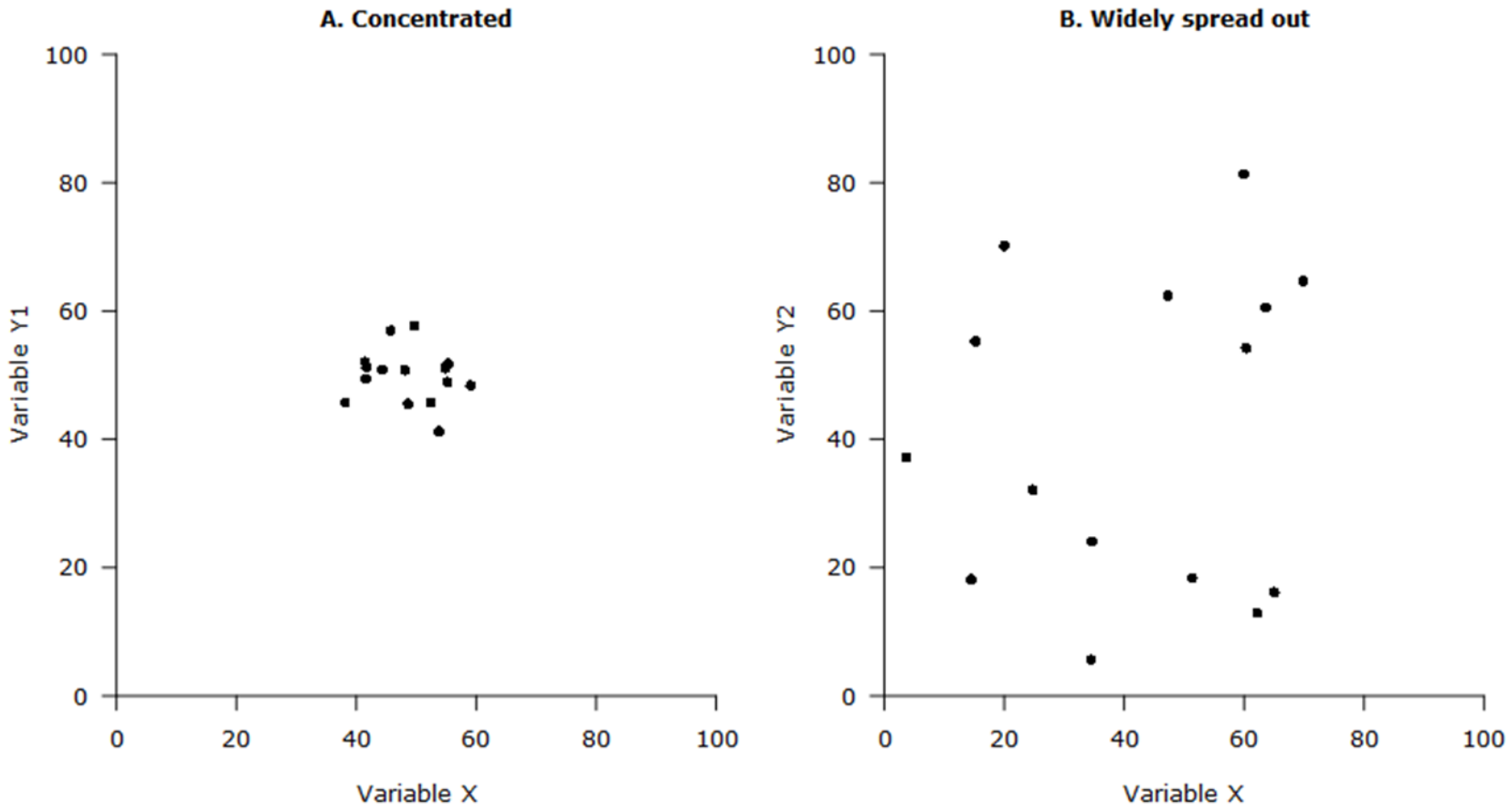
Graphical data displays

Scatter plot: Positive or negative relationship



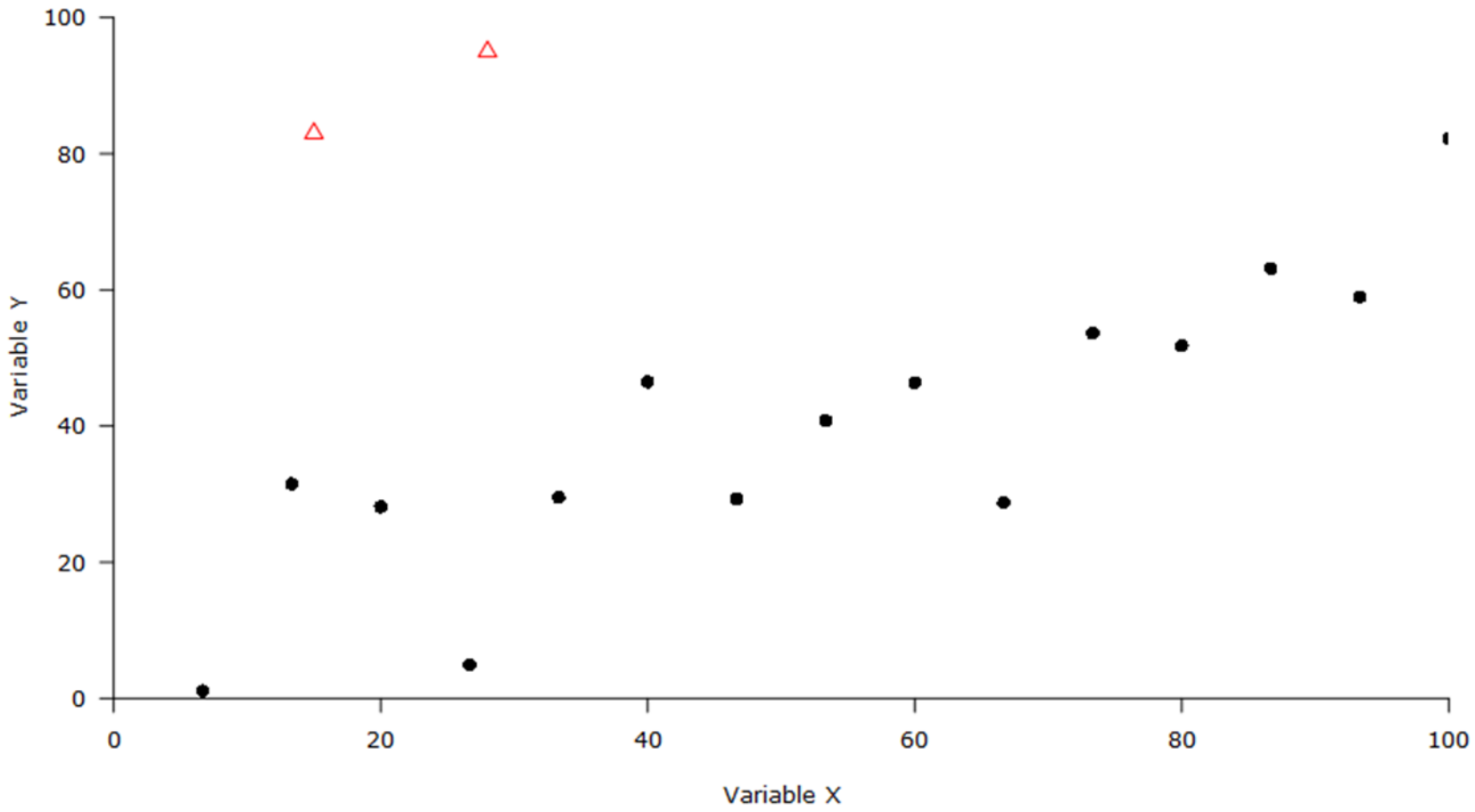
Graphical data displays

Scatter plot: Concentration or spread of data points



Graphical data displays

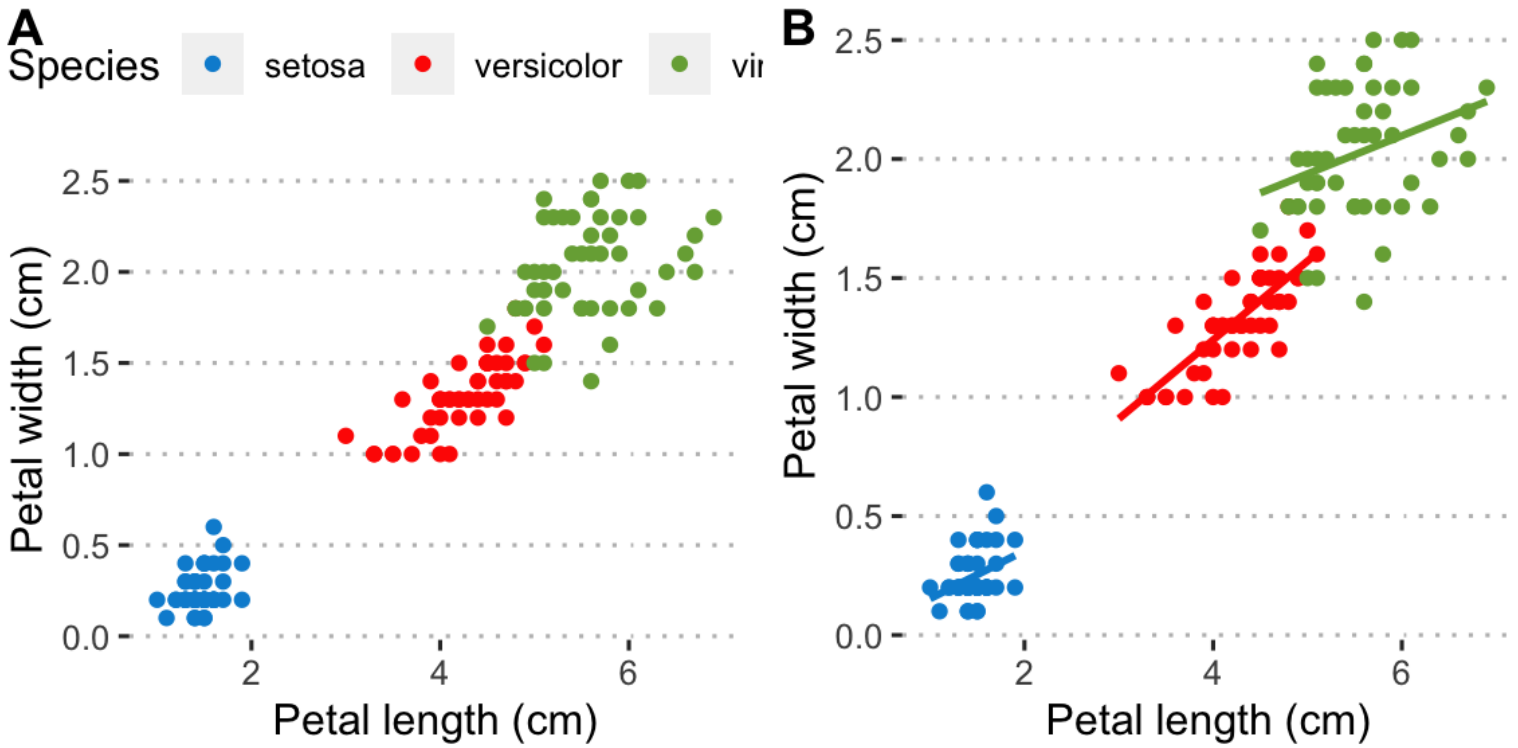
Scatter plot: Presence of outliers



Graphical data displays

Pairwise Scatter Plots

This graph shows the relationship between two (matched) continuous variables. The statistical strength of the relationship can be indicated by a correlation (no causal relationship implied as is the case here) or a regression (when a causal link of x and y is demonstrated).



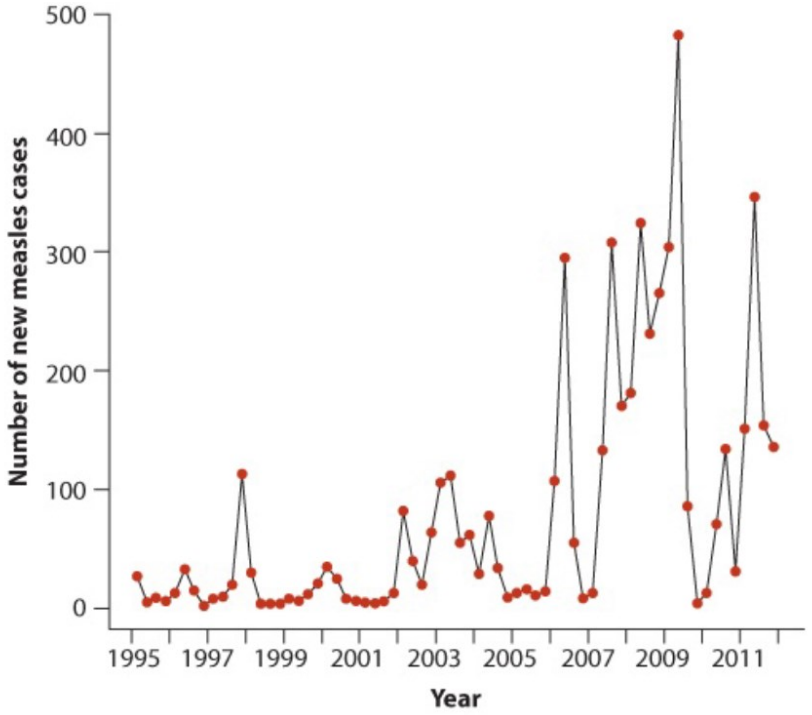
Examples of scatterplots made for the Iris data. (A) A default scatter plot showing the relationship between petal length and width. (B) The same as (A) but with a correlation line added.

Showing trends in time and space

Often a variable of interest represents a summary measurement taken at consecutive points in time or space. In this case, *line graphs* and *maps* are excellent visual tools.

Line graph

A **line graph** is a powerful tool for displaying trends in time or other ordered series. Typically, one *y*-measurement is displayed for each point in time or space, which is displayed along the *x*-axis. Adjacent points along the *x*-axis are connected by a line segment.



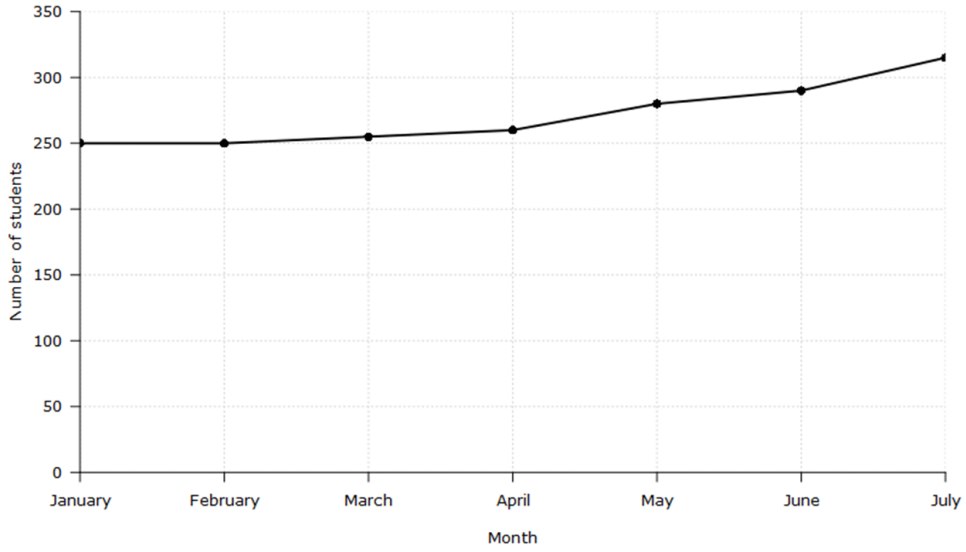
Confirmed cases of measles in England and Wales from 1995 to 2011. The four numbers in each year refer to new cases in each quarter.

Graphical data displays

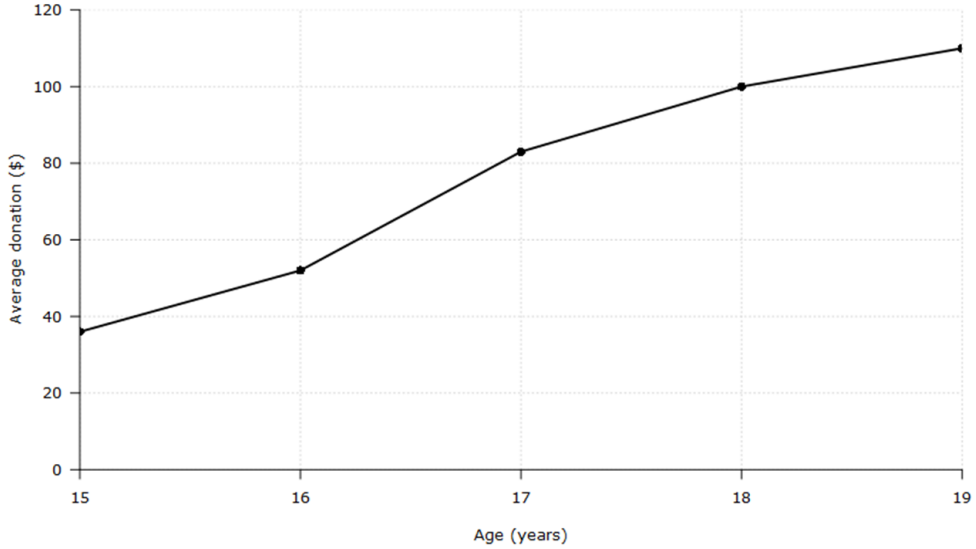
Line chart

Line charts are more popular than all other graphs combined because their visual characteristics reveal data trends clearly and these charts are easy to create.

A line chart is a visual comparison of how two variables—shown on the x- and y-axes—are related or vary with each other. It shows related information by drawing a continuous line between all the points on a grid.



Plotting a trend over time

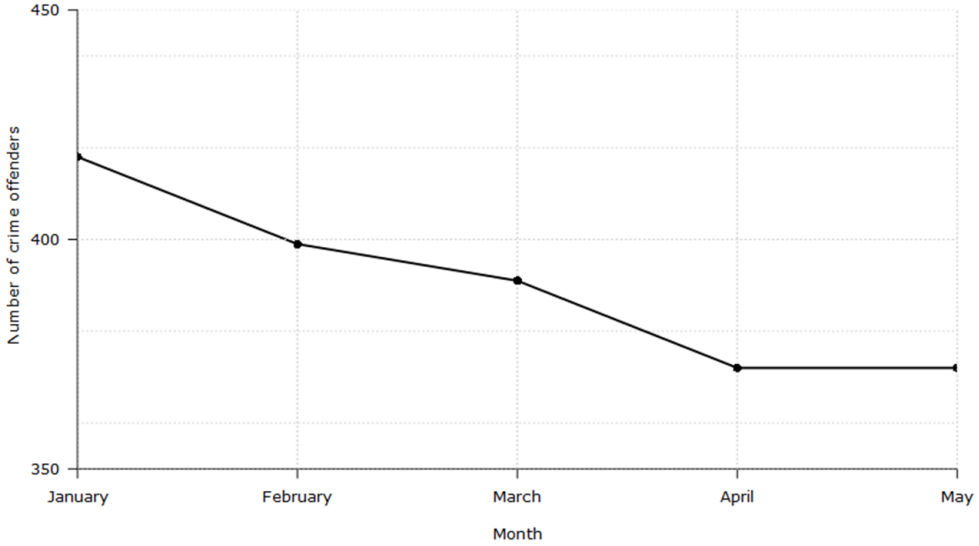


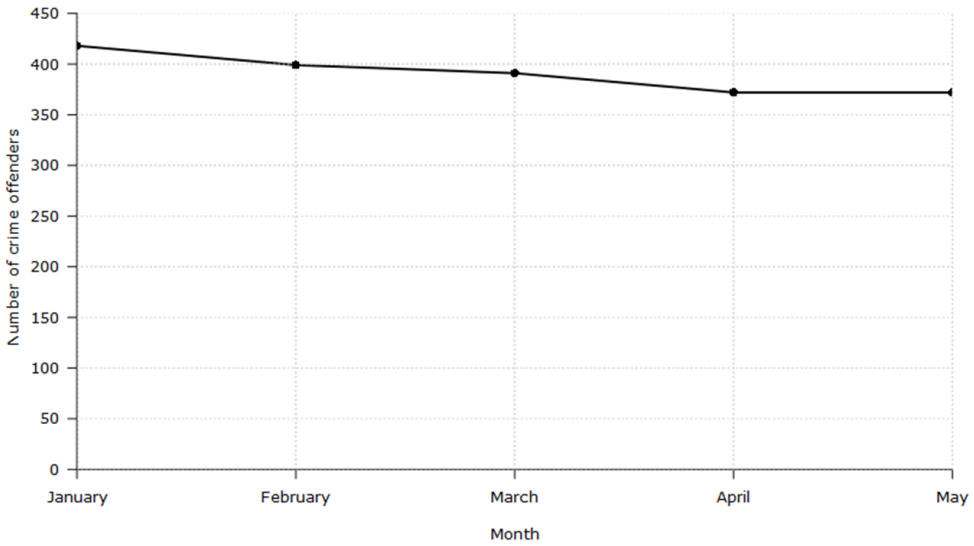
Comparing two related variables

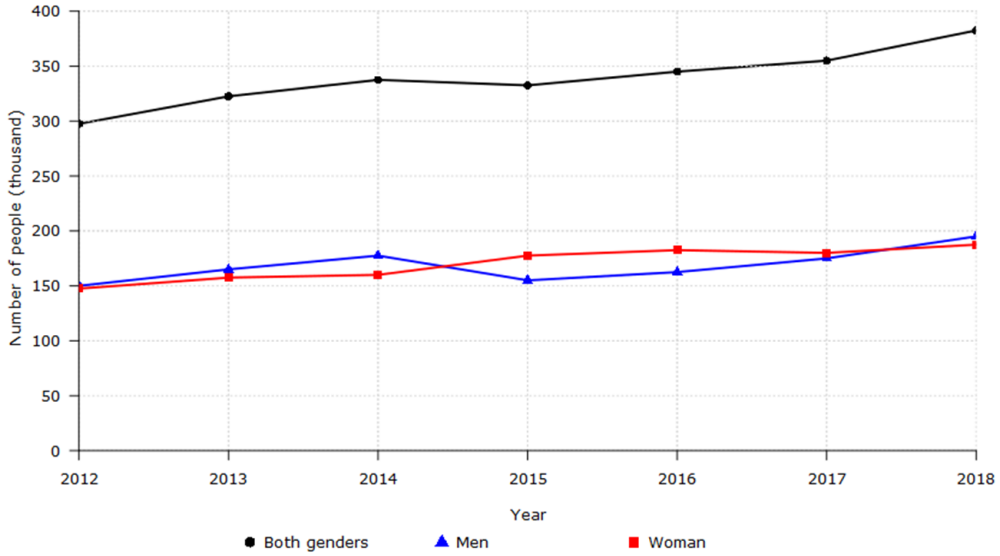
Graphical data displays

Line chart

Using correct scale







Multiple line graphs

Graphical data displays

Ogive graphs

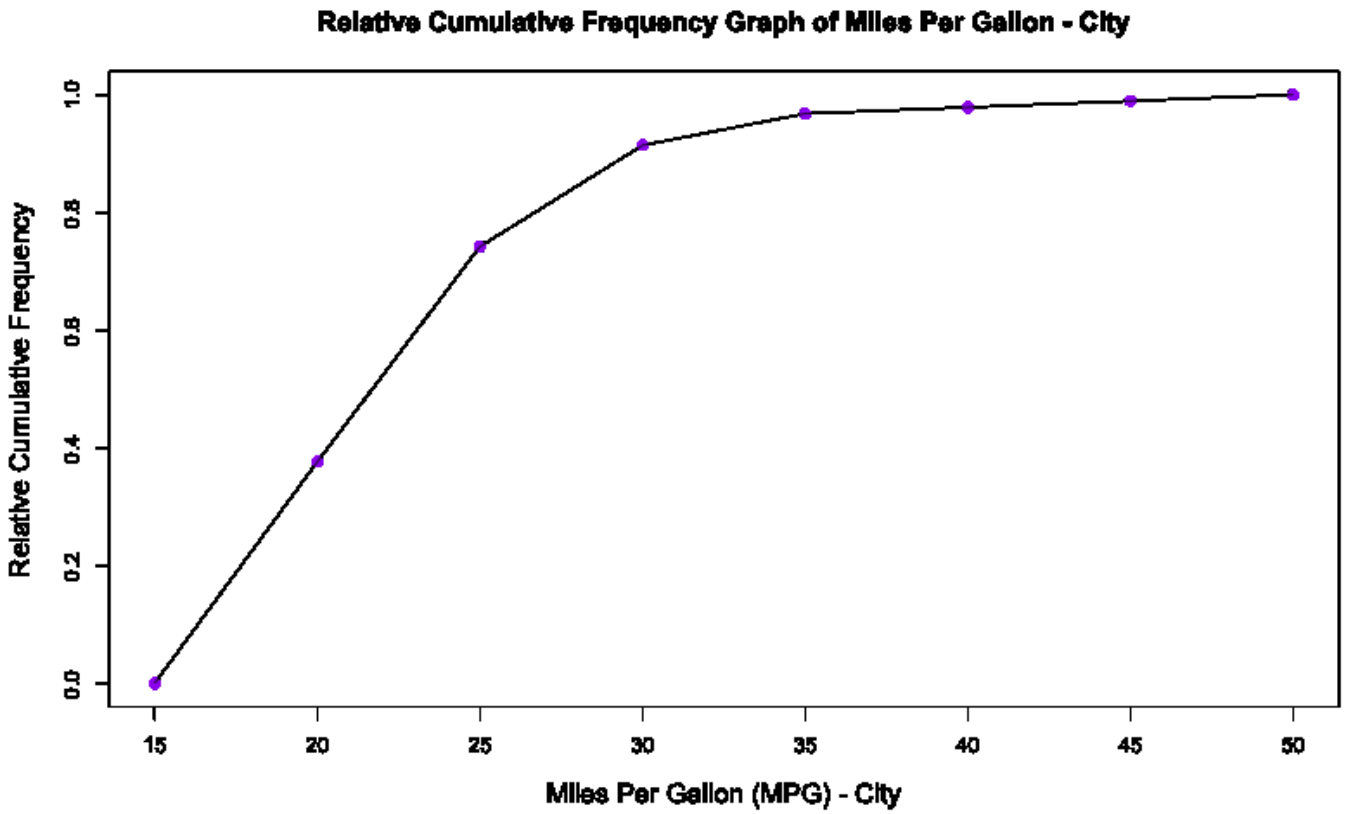
An ogive graph serves as a graphical representation of the cumulative relative frequency distribution for quantitative variables. In other words, these graphs plot the percentile on the y-axis and the quantitative variable on the x-axis. They are interpreted as follows: for example, let's say that the 10th percentile corresponds to an x-value of 20. This would mean that 10% of the data in this given dataset is at or below 20.

Slope: The steeper the slope between two x-values, the higher the frequency of data between those two values. For example, if the slope between x_1 and x_2 is zero, there are no data points between the two x-values. A very high positive slope between x_1 and x_2 indicates that there are a lot of data points between the two x-values.

Point Estimate: Given a percentile of interest, they can be used to estimate the value of your dataset associated with that percentile.

Graphical data displays

Ogive: These graphs plot a variable against its corresponding percentile (the percent of observations at or below that value).



Graphical data displays

Constructing a Cumulative Frequency Graph (Ogive graphs)

Example - The following data represents Miles Per Gallon (city) estimates for cars selected at random from among 1993 passenger car models that were listed in both the Consumer Reports issue and the PACE Buying Guide. The data was obtained from the Cars93 dataset. With the provided dataset, construct a relative cumulative frequency graph.

25	18	20	19	22	22	19	16	19	16	16	25	25	19	21	18	15	17	17	20	23	20	29	23	22	17	21	18	29	20	31
23	22	22	24	15	21	18	46	30	24	42	24	29	22	26	20	17	18	18	17	18	29	28	26	18	17	20	19	23	19	29
18	29	24	17	21	24	23	18	19	23	31	23	19	19	19	20	28	33	25	23	39	32	25	22	18	25	17	21	18	21	20

1. To simplify subsequent steps, begin by sorting the data from least to greatest:

15	15	16	16	16	17	17	17	17	17	17	17	17	18	18	18	18	18	18	18	18	18	18	18	19	19	19	19	19	19	
19	19	19	19	20	20	20	20	20	20	20	20	21	21	21	21	21	21	22	22	22	22	22	22	22	23	23	23	23	23	23
23	23	24	24	24	24	24	25	25	25	25	25	25	26	26	28	28	29	29	29	29	29	29	30	31	31	32	33	39	42	46

Graphical data displays

Constructing a Cumulative Frequency Graph (Ogive graphs)

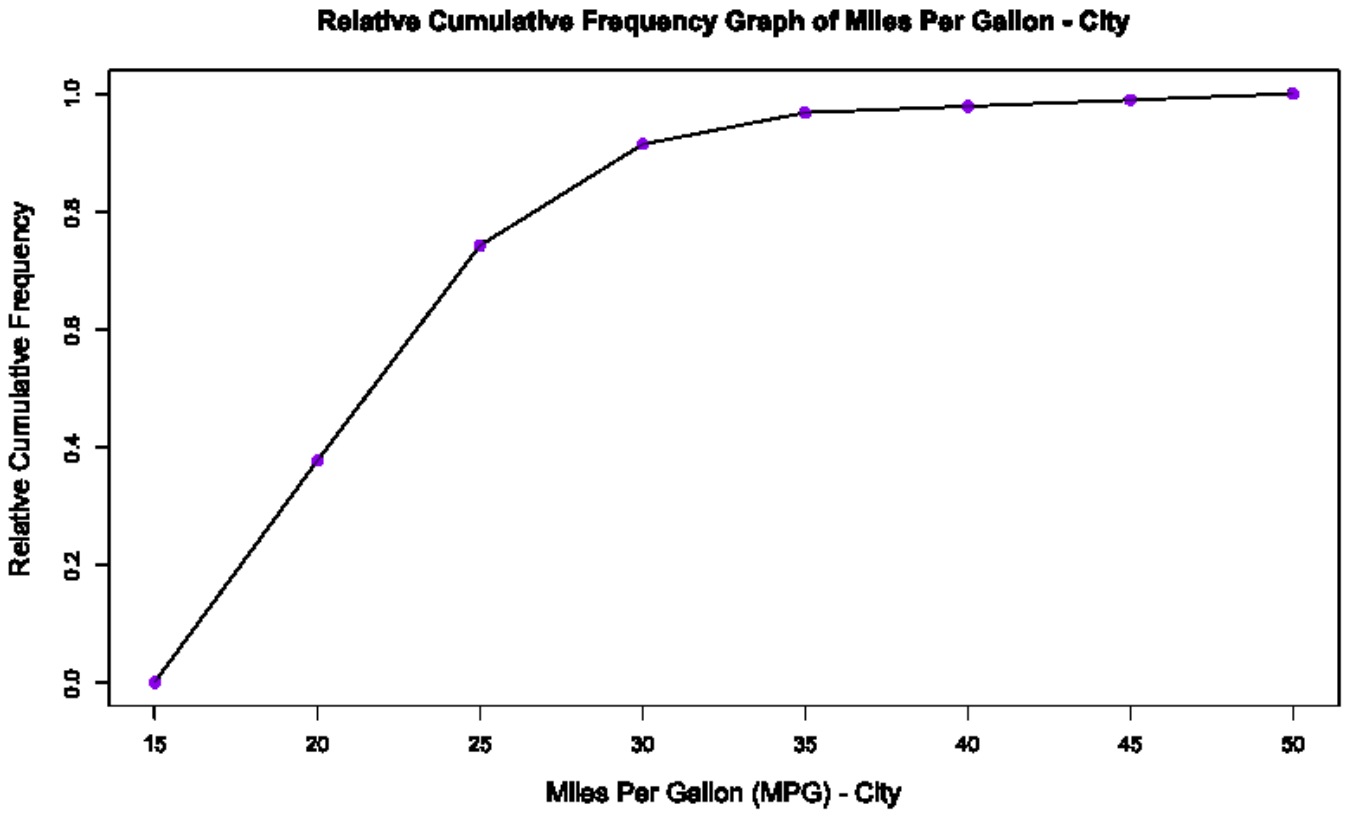
2. Using the sorted data above, construct the following table:

Range	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Percentile
$15 \le x < 20$	35	0.376	35	0.376
$20 \le x < 25$	34	0.366	69	0.742
$25 \le x < 30$	16	0.172	85	0.914
$30 \le x < 35$	5	0.054	90	0.968
$35 \le x < 40$	1	0.011	91	0.978
$40 \le x < 45$	1	0.011	92	0.989
$45 \le x < 50$	1	0.011	93	1.000

Graphical data displays

Constructing a Cumulative Frequency Graph (**Ogive graphs)**

3. Using the range and cumulative percentile data shown above, construct the following graph:



Graphical data displays

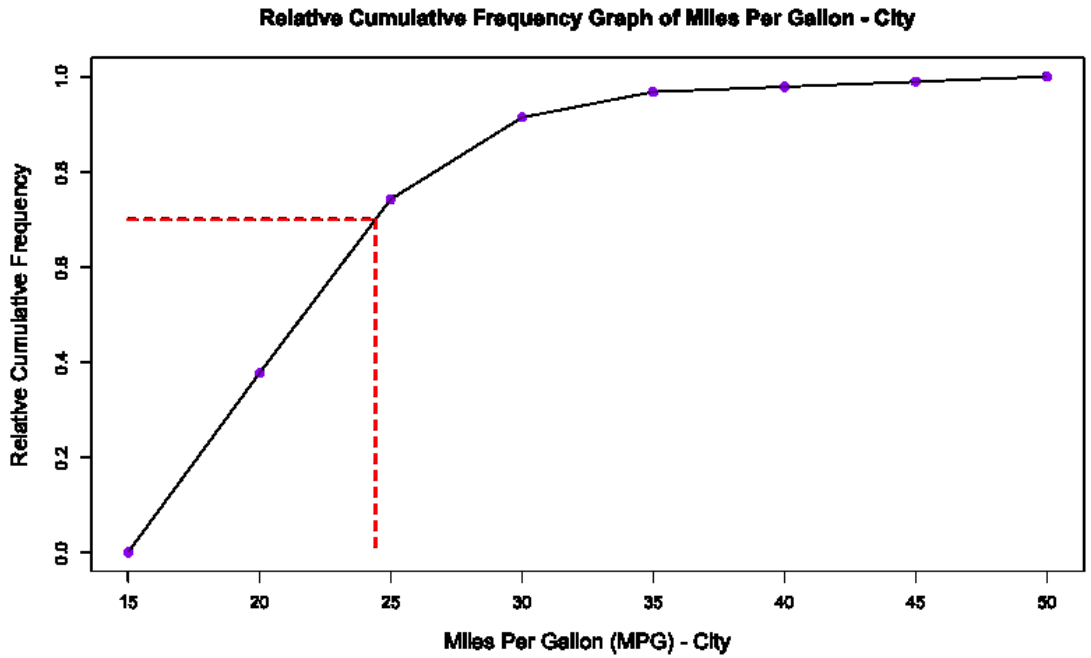
Determining Percentiles Using Cumulative Frequency Graphs

Once a relative cumulative frequency graph has been constructed. One can use the graph to calculate percentile values.

Example: For the Cars93 MPG data set, determine the fuel efficiency that corresponds to the 70th percentile.

Solution:

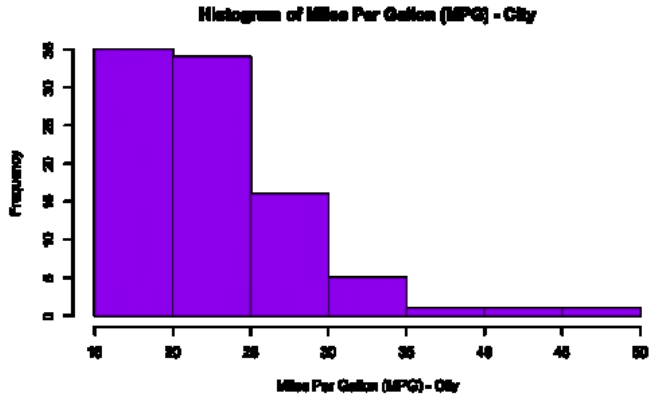
1. On the y-axis, find the value that corresponds to 0.70.
2. As shown on the right, draw a horizontal line to the right of 0.7. Then draw a vertical line at the point where the horizontal hits the ogive's line segment.
3. Use the vertical line to estimate the value on the x-axis.
4. The 70th percentile for fuel efficiency corresponds to 24 MPG.

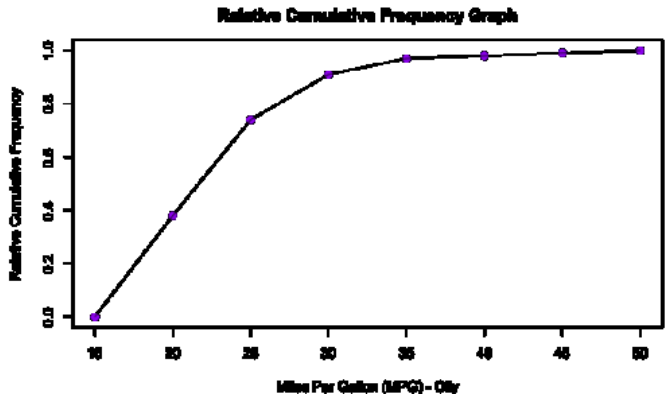


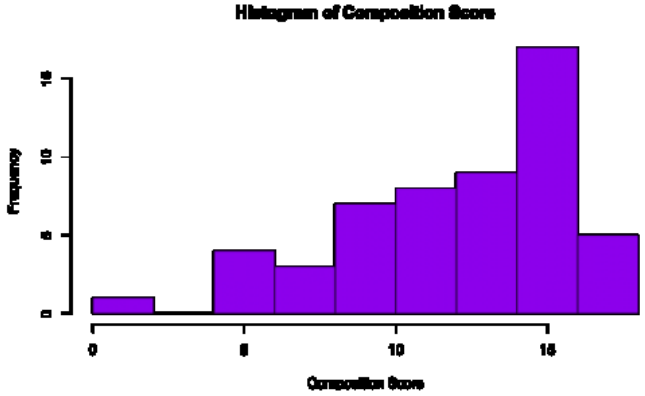
Graphical data displays

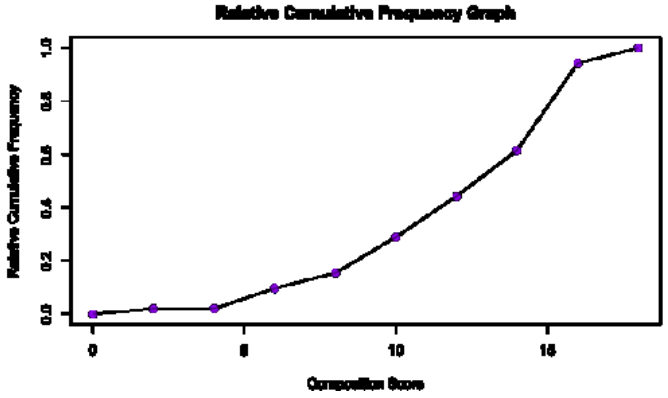
Comparison Between Histograms and Relative Cumulative Frequency Graphs

There is a direct association between the slope of the Cumulative Frequency Graph for certain intervals and the corresponding frequency count in the same interval of the associated histogram. Notice in Example Two (below) that slope between 2 and 4 is 0, and that the frequency count is also 0 for this section of the histogram. Likewise, between interval 14-16, the slope is greatest in the Relative Cumulative Frequency Graph, while the frequency count during this interval is also the greatest among the other bins.







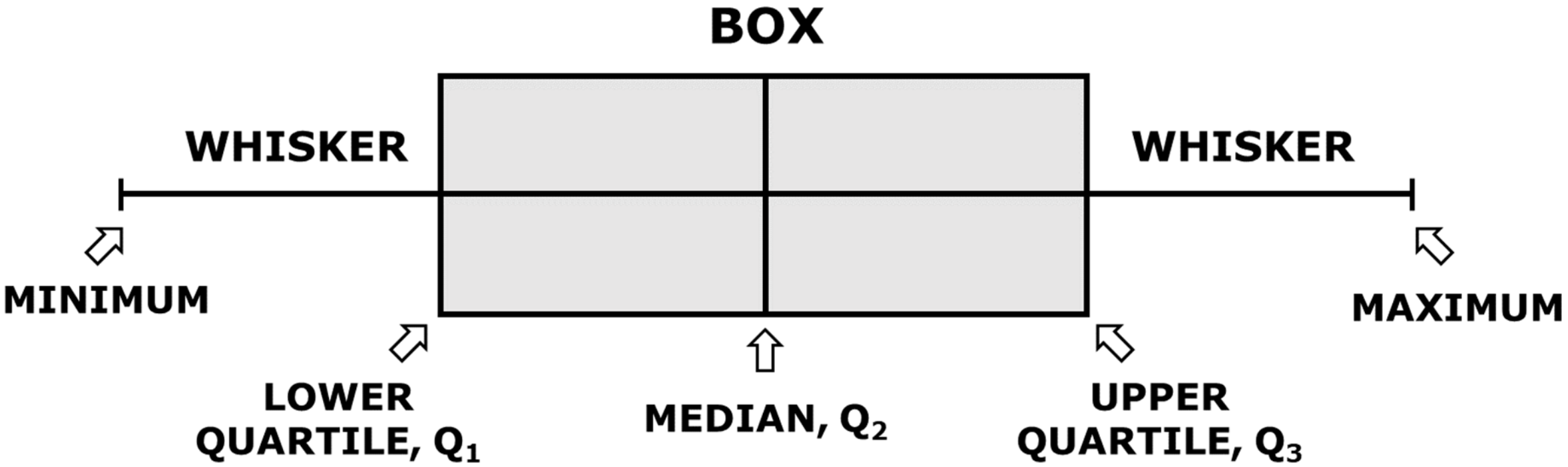


Graphical data displays

Box plots

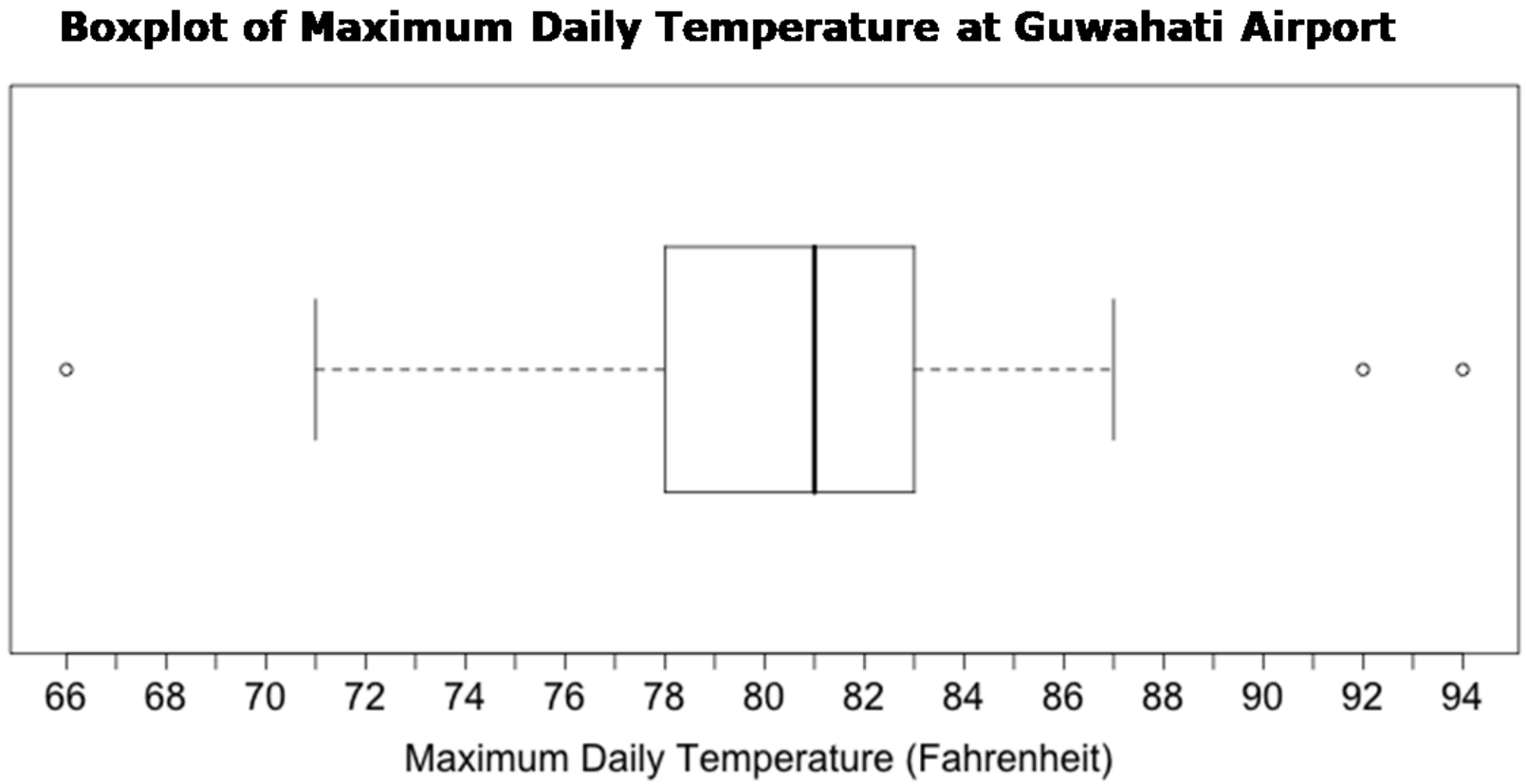
Box plots are sometimes called **box-and-whisker plots**. These graphs are a graphical representation of the data based on its **quartiles** as well as its **smallest (minimum) and largest (maximum) values**.

"The lower and upper hinges correspond to the first (Q1, 25th percentile) and third (Q3, 75th percentile) quartiles."



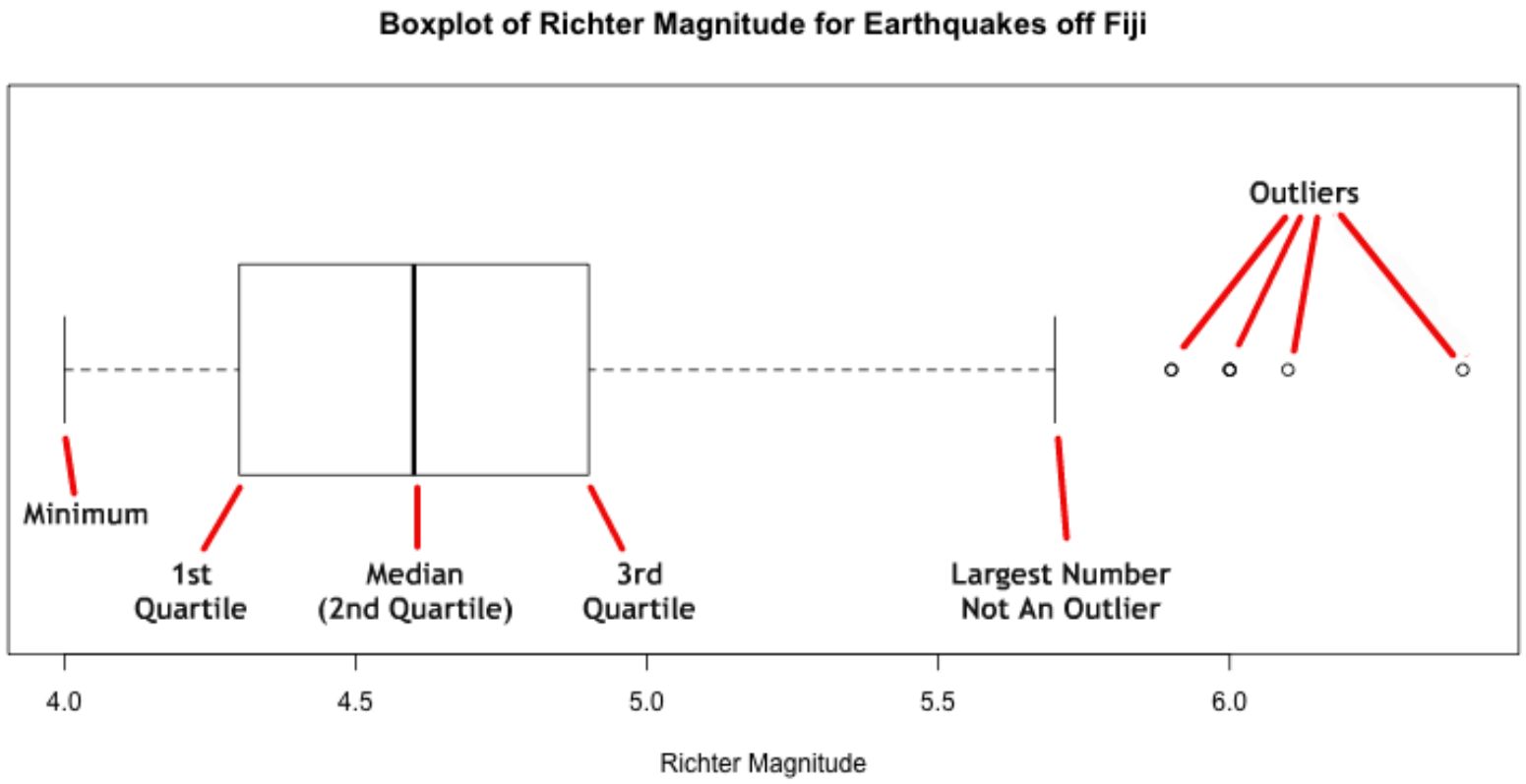
Graphical data displays

Box plot: It represents the **Five Number summary**.



Graphical data displays

Box plot: Data beyond the end of the whiskers are called '**outlying**' points and are plotted individually."



Graphical data displays

Box plots

“The upper whisker extends from the hinge to the largest value no further than 1.5*IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles).

The lower whisker extends from the hinge to the smallest value at most 1.5*IQR of the hinge.

“In a **notched box plot**, the notches extend 1.58*IQR/sqrt(n). This gives a roughly 95% confidence interval for comparing medians.”

One can glance the ‘shape’ of the data distribution; they provide an alternative view to that given by the frequency distribution.

A variation of the basic box-and-whisker plot is to superimpose a jittered scatter plot of the raw data on each bar.

Graphical data displays

Constructing a Boxplot

Construct a modified boxplot using the following dataset of student algebra scores:
95, 79, 68, 93, 86, 87, 83, 84, 85, 88, 82, 90, 80, 86, 84

Use the following steps to construct a boxplot:

STEP 1: Sort the data from least to greatest as shown below:
68, 79, 80, 82, 83, 84, 84, 85, 86, 86, 87, 88, 90, 93, 95

STEP 2: Determine the quartiles:
1^{st} Quartile: 82 Median (2^{nd} Quartile): 85 3^{rd} Quartile: 88

STEP 3: Determine the Interquartile Range:
IQR = 3^{rd} Quartile – 1^{st} Quartile = 88 – 82 = 6
IQR = 6

Graphical data displays

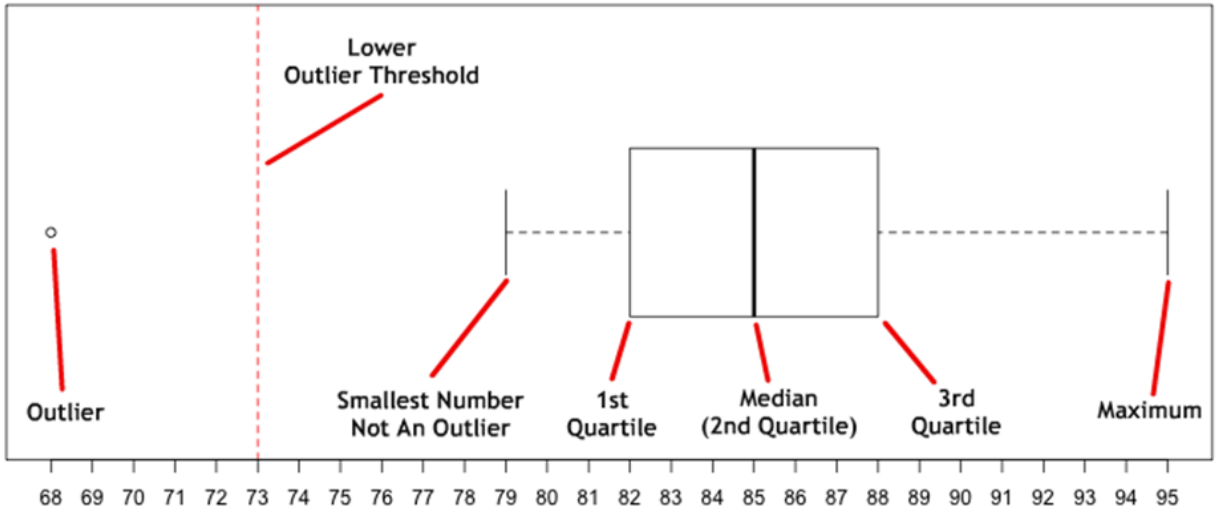
Box Plots

STEP 4: Determine Upper Outlier Threshold & Lower Outlier Thresholds:

Lower Outlier Threshold = Q1 – 1.5(IQR)
Lower Outlier Threshold = 82 – 1.5(6)
Lower Outlier Threshold = 82 – 9 = 73
Lower Outlier Threshold = 73

Upper Outlier Threshold = Q3 + 1.5(IQR)
Upper Outlier Threshold = 88 + 1.5(6)
Upper Outlier Threshold = 88 + 9 = 97
Upper Outlier Threshold = 97

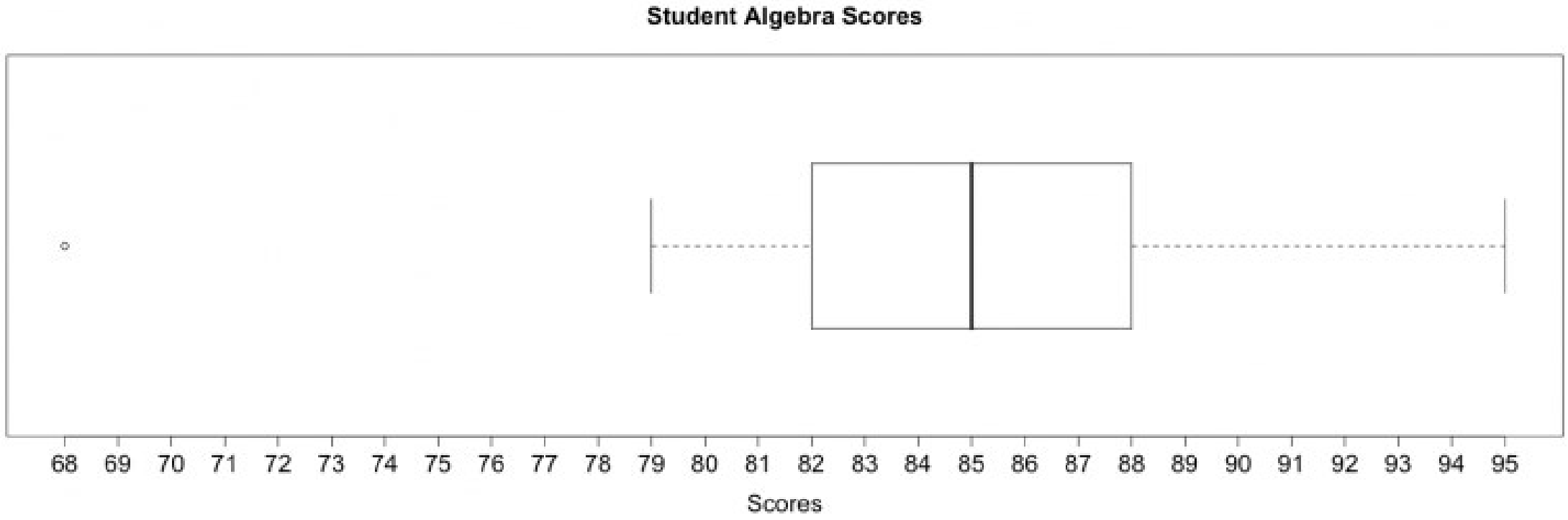
STEP 5: Using the calculated information from above, construct a boxplot as shown below.
NOTE: The red dashed line represents the lower outlier threshold and for demonstration purposes only. Any points that fall below the lower outlier threshold should be considered an outlier. In this particular example, 68 is considered an outlier.



Graphical data displays

Box Plots

STEP 6: Make sure to always label your graphics. For the boxplot below, the main title and x-axis label were added.

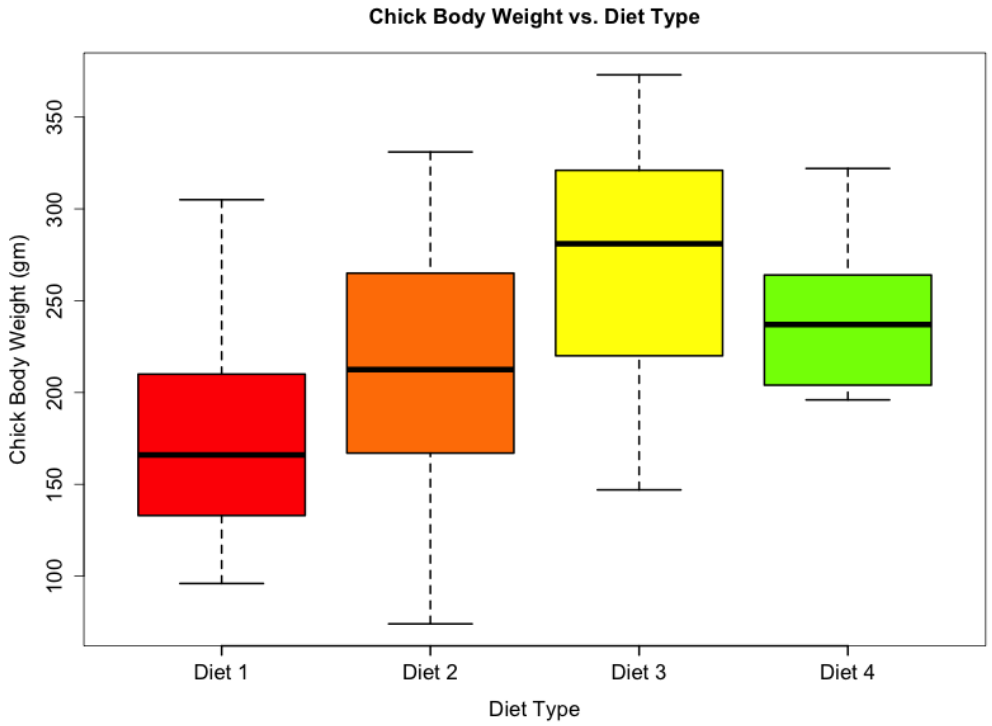


Graphical data displays

Box Plots

Side-By-Side Boxplots

Side-by-side boxplots are useful tools for comparison. The example below shows a side-by-side boxplot of the measured body weights of various chicks on four different protein diets. The weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. The boxplot below is used to compare the final weights, measured on day 21, against the diet type.



From the side-by-side boxplot, one can conclude that...

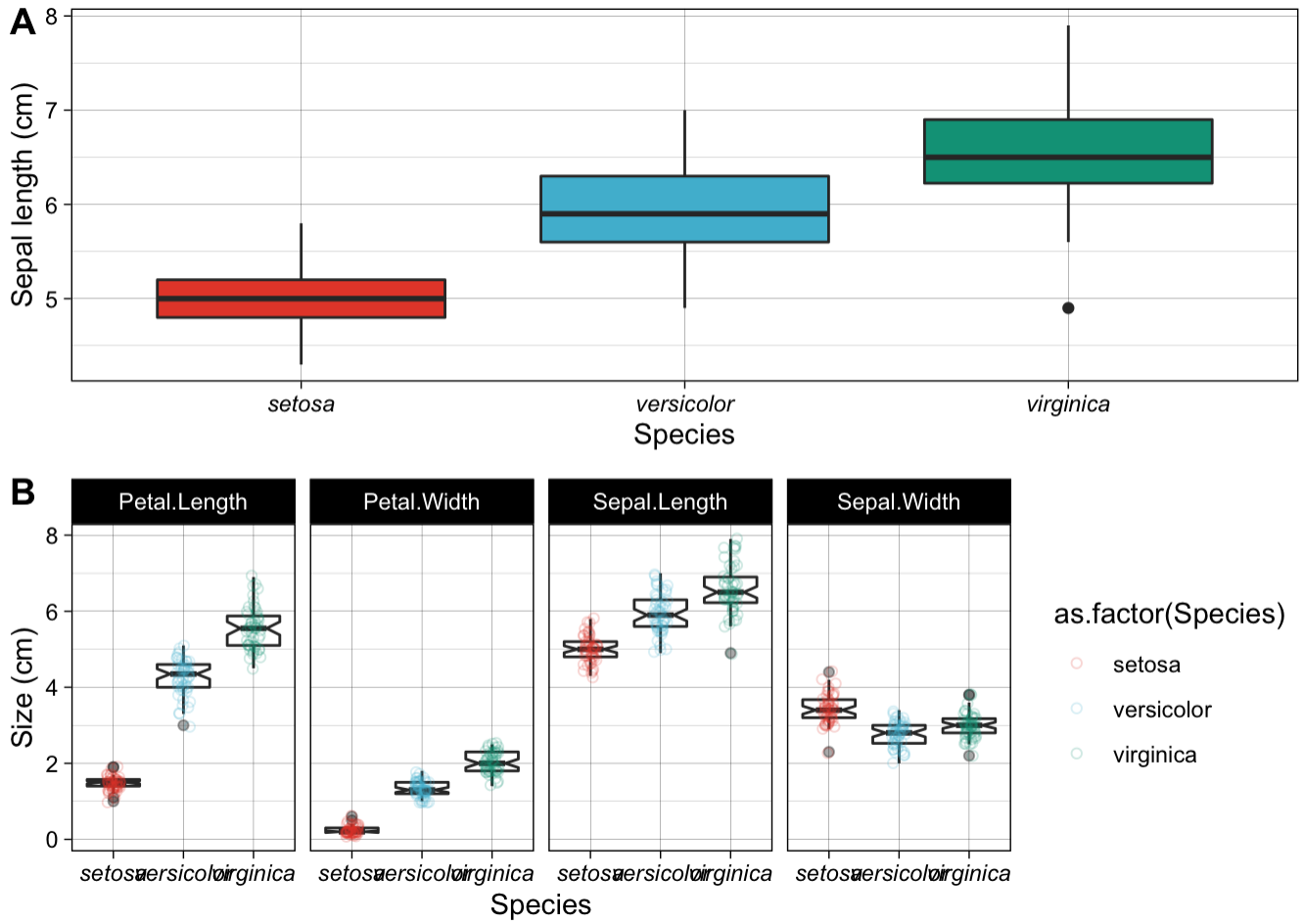
(1) Chicks on Diet 3 had the highest median weight, and chicks on Diet 1 had the lowest median weights.

(2) Chicks on Diet 4 had the smallest range of weights, and chicks on Diet 2 had the largest range of weights.

(3) Although chicks on Diet 2 had a higher median weight than chicks on Diet 1, the minimum weight for chicks on Diet 2 was lower than the minimum weight for chicks on Diet 1.

As you can see, there is a lot of information you can glean from side-by-side boxplots.

Graphical data displays



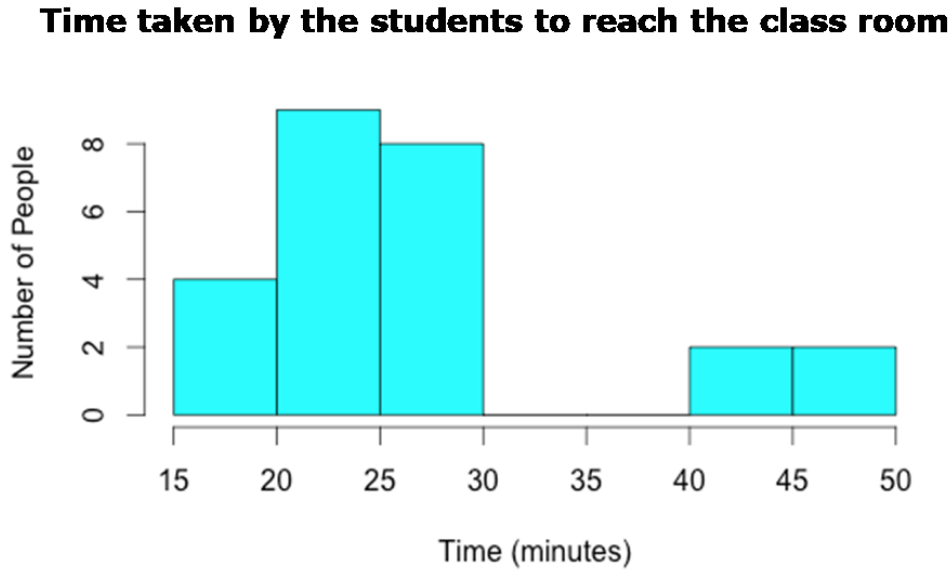
Examples of box plots made for the Iris data. A) A default box plot for one of the variables only. B) A panelled collection of box plots, one for each of the four variables, with a scatterplot to indicate the spread of the actual replicates.

Showing numerical data: frequency table and histogram

A frequency distribution for a numerical variable can be displayed either in a frequency table or in a **histogram**. A histogram uses area of rectangular bars to display frequency. The data values are split into consecutive intervals, or "bins," usually of equal width, and the frequency of observations falling into each bin is displayed.

A ***histogram*** uses the area of rectangular bars to display the frequency distribution (or relative frequency distribution) of a numerical variable.

Histograms: In this type of graph, akin to bar graphs, the heights of the rectangular bars represent the frequency. **Histograms are used for numerical data whereas bar graphs are used for the categorical data**. **Histograms are continuous (no gaps in between the bars) whereas bar graphs have gap between each bar.**



Showing numerical data: frequency table and histogram

Table: Data on the abundance of each species of bird encountered during four surveys in Organ Pipe Cactus National Monument.

Species	Abundance
Greater roadrunner	1
Black-chinned hummingbird	1
Western kingbird	1
Great-tailed grackle	1
Bronzed cowbird	1
Great horned owl	2
Costa's hummingbird	2
Canyon wren	2
Canyon towhee	2
Harris's hawk	3
Loggerhead shrike	3
Hooded oriole	4
Northern mockingbird	5
American kestrel	7

Species	Abundance
Rock dove	7
Bell's vireo	10
Common raven	12
Northern cardinal	13
House sparrow	14
Ladder-backed woodpecker	15
Red-tailed hawk	15
Phainopepla	18
Turkey vulture	23
Violet-green swallow	23
Lesser nighthawk	25
Scott's oriole	28
Purple martin	33
Black-throated sparrow	33

Species	Abundance
Brown-headed cowbird	59
Black vulture	64
Lucy's warbler	67
Gilded flicker	77
Brown-crested flycatcher	128
Mourning dove	135
Gambel's quail	148
Black-tailed gnatcatcher	152
Ash-throated flycatcher	173
Curve-billed thrasher	173
Cactus wren	230
Verdin	282
House finch	297
Gila woodpecker	300
White-winged dove	625

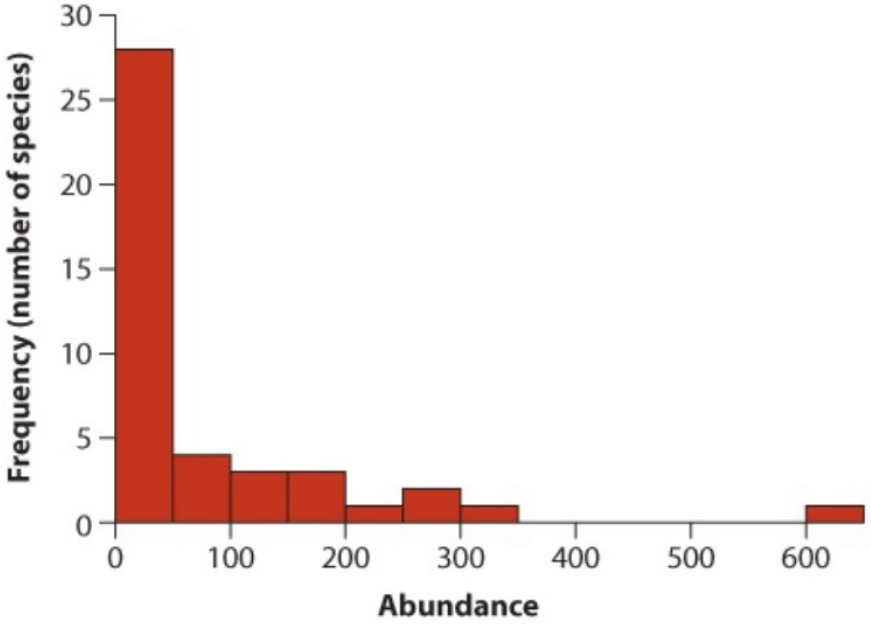
Showing numerical data: frequency table and histogram

The range of abundance values was divided into 13 intervals of equal width (0–50, 50–100, and so on), and the number of species falling into each abundance interval was counted and presented in a frequency table to help see patterns.

Table: Frequency distribution of bird species abundance at Organ Pipe Cactus National Monument.

Abundance	Frequency (Number of species)
0–50	28
50–100	4
100–150	3
150–200	3
200–250	1
250–300	2
300–350	1

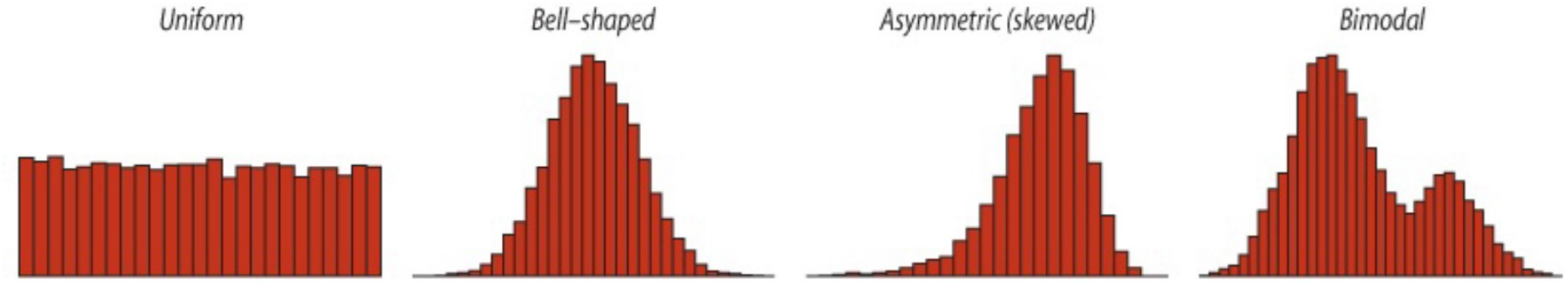
Abundance	Frequency (Number of species)
350–400	0
400–450	0
450–500	0
550–600	0
600–650	1
Total	43



Histogram illustrating the frequency distribution of bird species abundance at Organ Pipe Cactus National Monument. Total number of bird species: $n = 43$.

Describing the shape of a histogram

The histogram reveals the shape of a frequency distribution. Some of the most common shapes are displayed below. Any interval of the frequency distribution that is noticeably more frequent than surrounding intervals is called a peak. The **mode** is the interval corresponding to the highest peak. For example, a bell-shaped frequency distribution has a single peak (the mode) in the center of the range of observations. A frequency distribution having two distinct peaks is said to be **bimodal**.

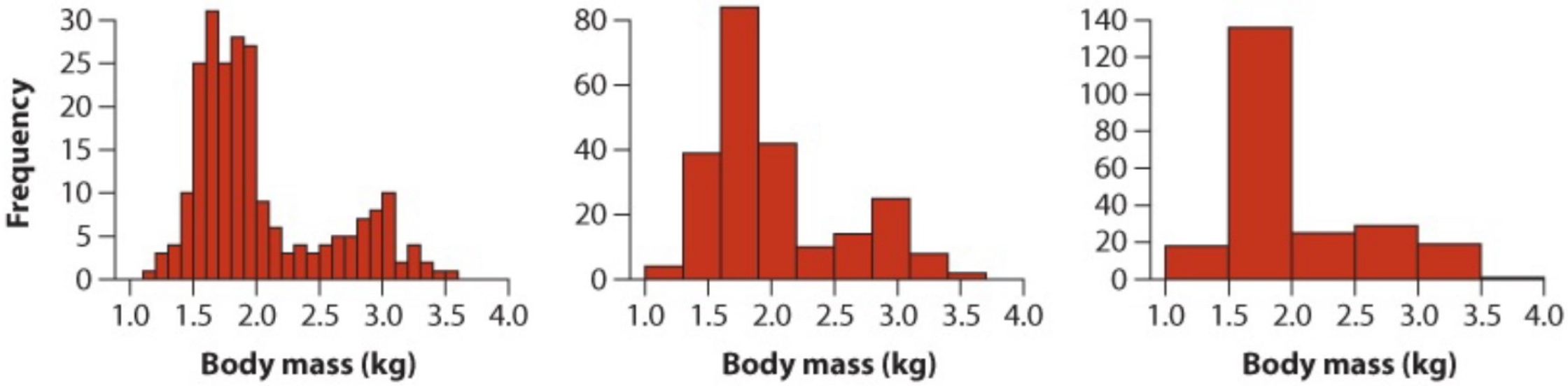


A frequency distribution is **symmetric** if the pattern of frequencies on the left half of the histogram is the mirror image of the pattern on the right half. If a frequency distribution is not symmetric, we say that it is **skewed**. ***Skew*** refers to asymmetry in the shape of a frequency distribution for a numerical variable.

Extreme data points lying well away from the rest of the data are called **outliers**.

How to draw a good histogram

When drawing a histogram, the choice of interval width must be made carefully because it can affect the conclusions.



There are no strict rules about the number of intervals that should be used in frequency tables and histograms. Some computer programs use Sturges's rule of thumb, in which the number of intervals is $(1 + \ln(n)/\ln(2))$, where n is the number of observations and ln is the natural logarithm.

Showing data for one variable

Relative frequency is the proportion of observations having a given measurement, calculated as the frequency divided by the total number of observations. The **relative frequency distribution** is the proportion of occurrences of each value in the data set.

The ***relative frequency distribution*** describes the fraction of occurrences of each value of a variable.

Showing categorical data: frequency table and bar graph

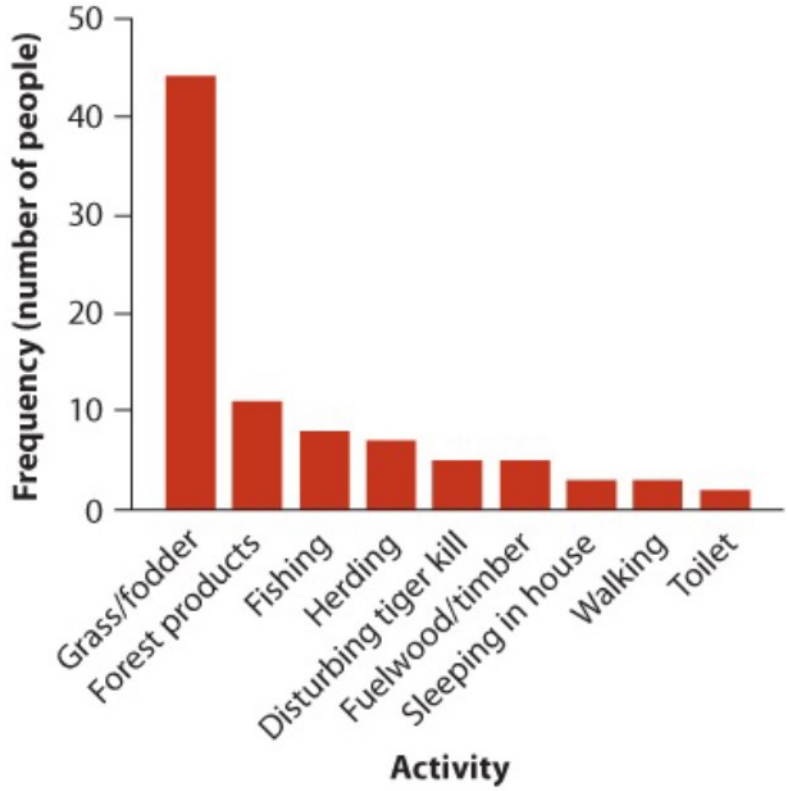
A **frequency table** is a text display of the number of occurrences of each category in the data set. A **bar graph** uses the height of rectangular bars to visualize the frequency (or relative frequency) of occurrence of each category.

A ***bar graph*** uses the height of rectangular bars to display the frequency distribution (or relative frequency distribution) of a categorical variable.

Bar graphs display the mean ± some measure of variation around the mean—typically the standard error or the standard deviation. The mean ± SE and mean ± SD are typically used for normally-distributed data.

Showing categorical data: frequency table and bar graph

Activity	Frequency (number of people)
Collecting grass or fodder for livestock	44
Collecting non-timber forest products	11
Fishing	8
Herding livestock	7
Disturbing tiger at its kill	5
Collecting fuel wood or timber	5
Sleeping in a house	5
Walking in forest	3
Using an outside toilet	2
Total	88



Graphical data displays

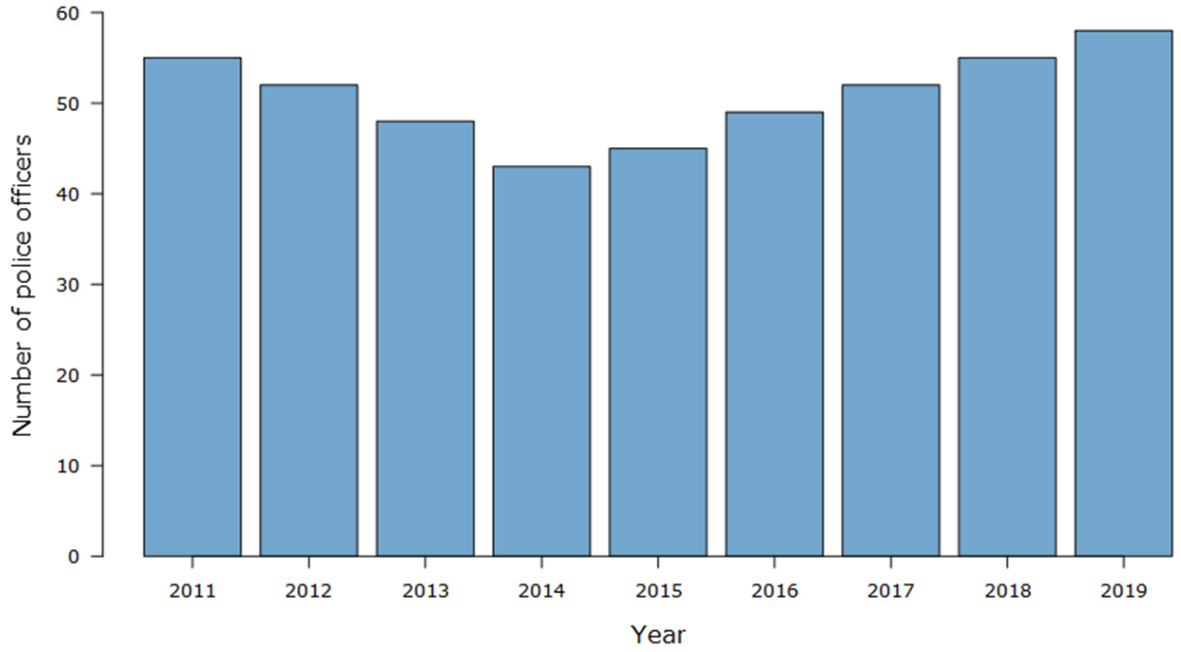
Bar chart

A bar chart may be either horizontal or vertical. The important point to note about bar charts is their bar length or height—the greater their length or height, the greater their value.

Bar charts usually present categorical variables, discrete variables or continuous variables grouped in class intervals. They consist of an axis and a series of labelled horizontal or vertical bars. The bars depict frequencies of different values of a variable or simply the different values themselves. The numbers on the y-axis of a vertical bar chart or the x-axis of a horizontal bar chart are called the scale. Select an arbitrary but consistent width for each bar as well.

Vertical bar charts

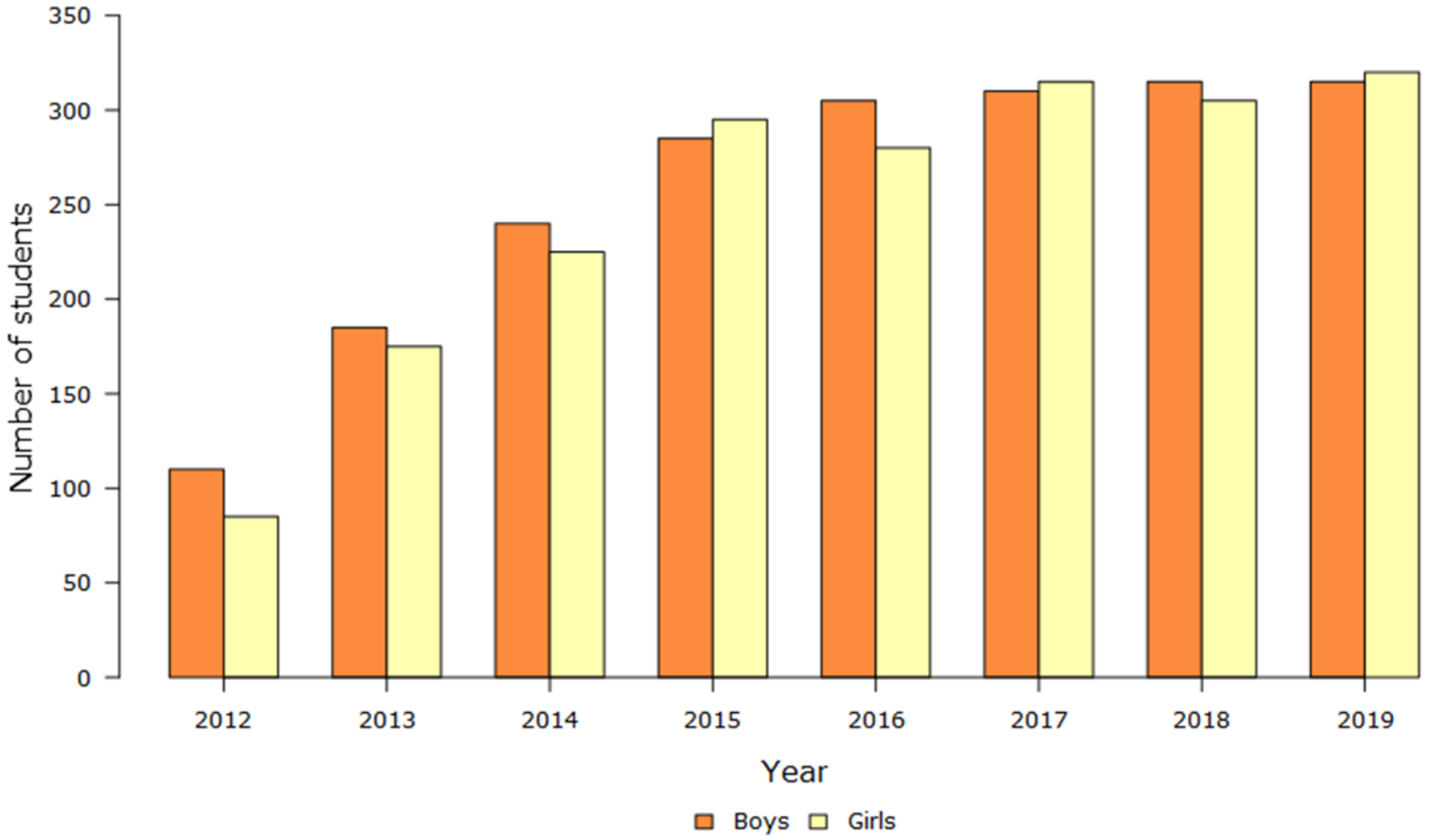
Bar charts should be used when you are showing segments of information. Vertical bar charts are useful to compare different categorical or discrete variables, such as age groups, classes, schools, etc., as long as there are not too many categories to compare. They are also very useful for time series data. The space for labels on the x-axis is small, but ideal for years, minutes, hours or months.



Graphical data displays

Grouped bar charts

The grouped bar chart is another effective means of comparing sets of data about the same places or items.



Showing association between two variables

Showing association between categorical variables

A ***contingency table*** gives the frequency of occurrence of all combinations of two (or more) categorical variables.

Table: Contingency table showing the incidence of malaria in female great tits in relation to experimental treatment.

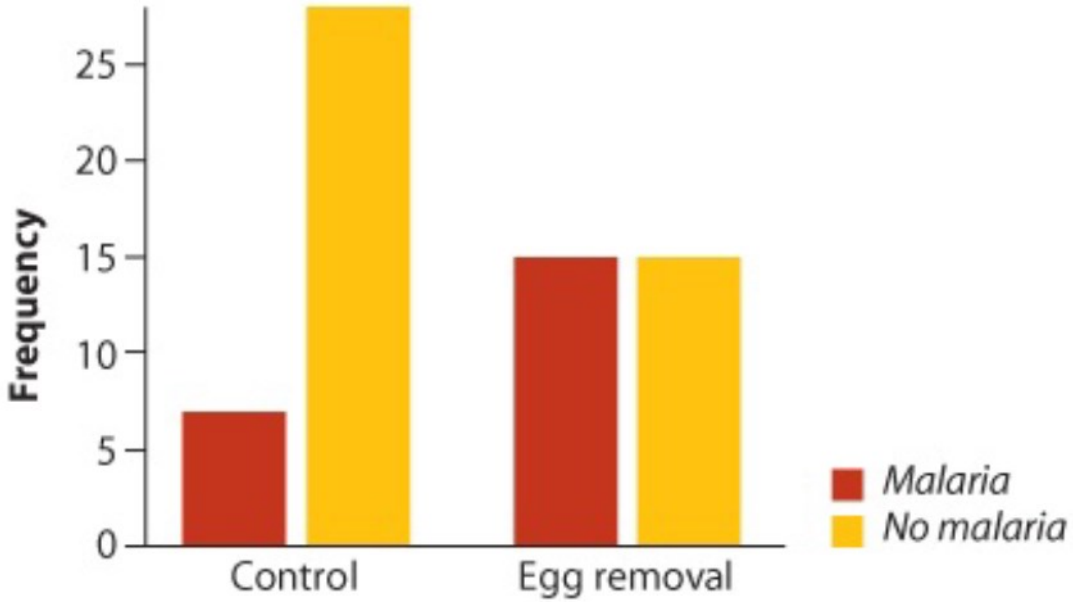
	Experimental treatment group		
	Control group	**Egg-removal group**	**Row total**
Malaria	7	15	22
No malaria	28	15	43
Column total	35	30	65

The explanatory variable (experimental treatment) is displayed in the columns, whereas the response variable, the variable being predicted (incidence of malaria), is displayed in the rows. The frequency of subjects in each treatment group is given in the column totals, and the frequency of subjects with and without malaria is given in the row totals.

Showing association between two variables

Showing association between categorical variables

A ***grouped bar graph*** uses the height of rectangular bars to display the frequency distributions (or relative frequency distributions) of two or more categorical variables.

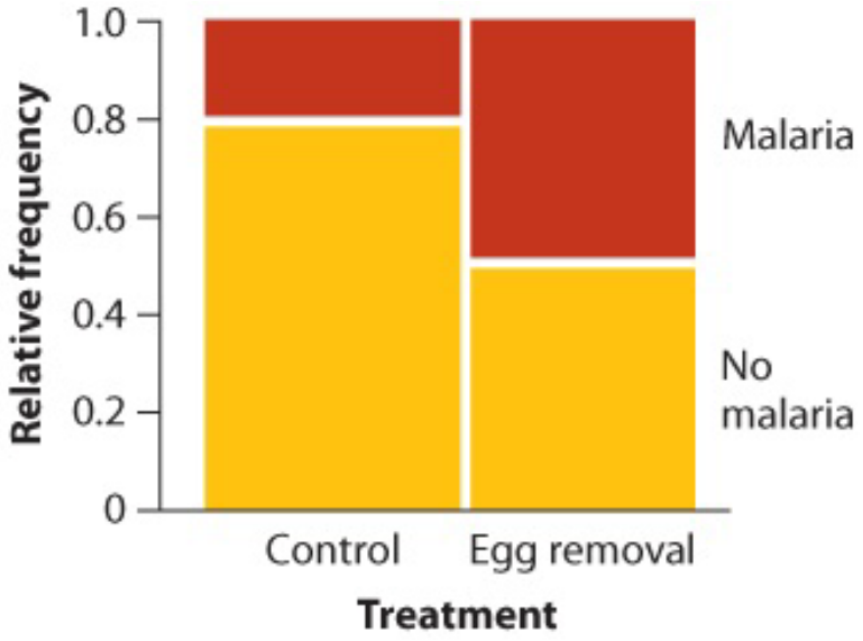


Grouped bar graph for reproductive effort and avian malaria in great tits.

Showing association between two variables

Showing association between categorical variables

A **mosaic plot** is similar to a grouped bar plot except that bars within treatment groups are stacked on top of one another. Within a stack, bar area and height indicate the relative frequencies (i.e., the proportion) of the responses. This makes it easy to see the association between treatment and response variables: if an association is present in the data, then the vertical position at which the colors meet will differ between stacks. If no association is present, then the meeting point between the colors will be at the same vertical position between stacks.

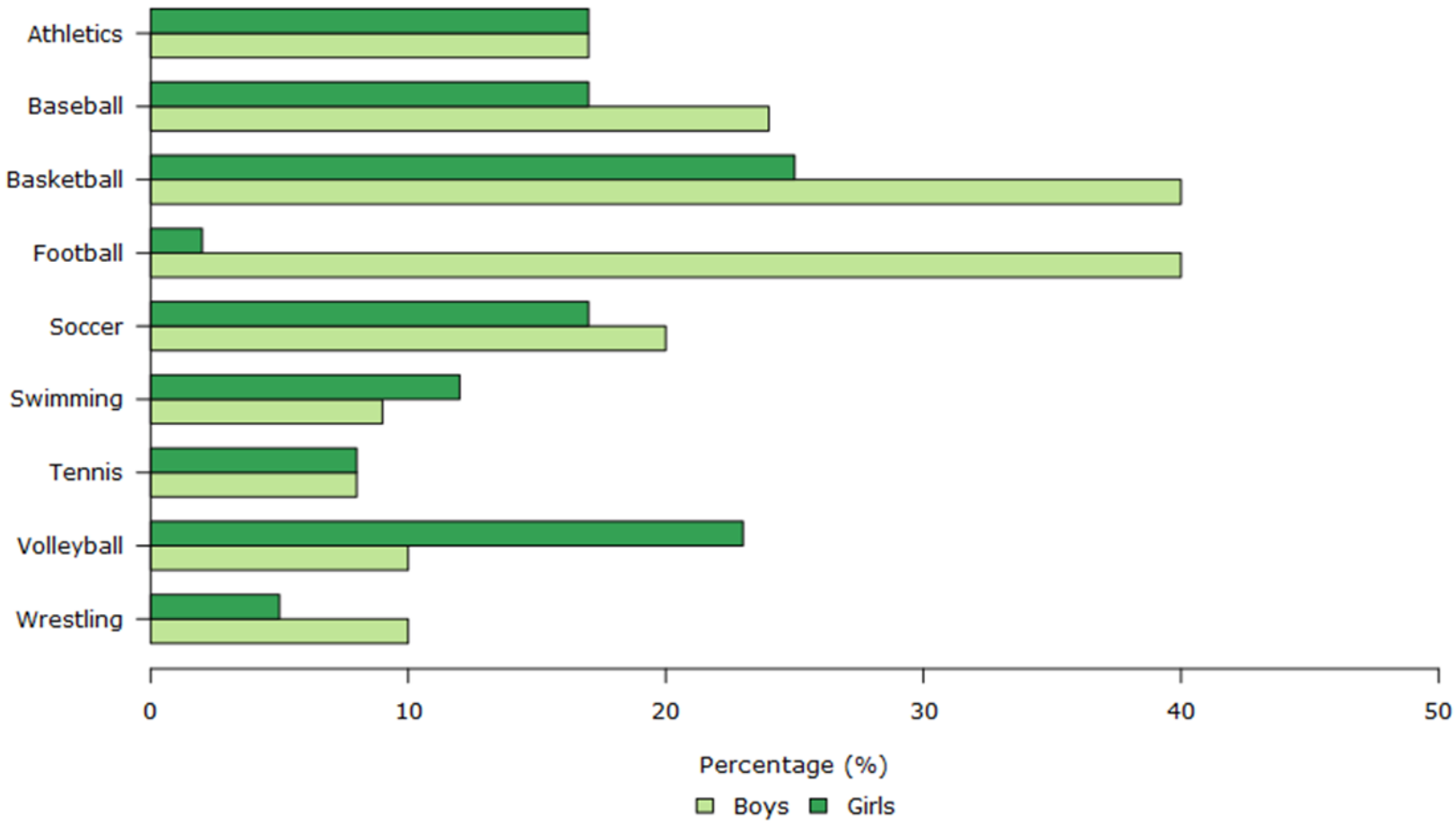


Mosaic plot for reproductive effort and avian malaria in great tits. Red indicates birds with malaria, whereas gold indicates birds free of malaria.

Graphical data displays

Horizontal bar charts

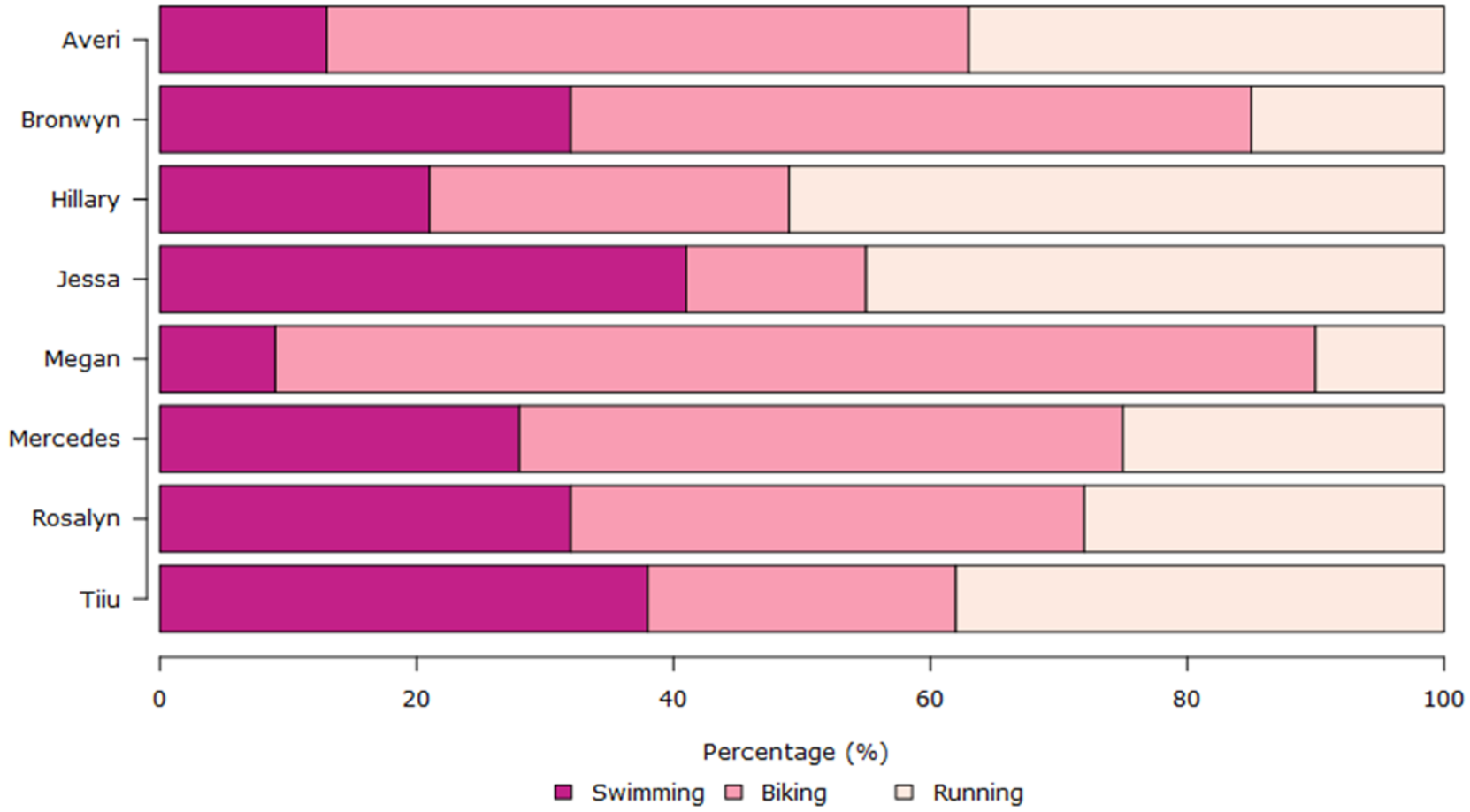
One disadvantage of vertical bar charts, however, is that they lack space for text labelling at the foot of each bar.



Graphical data displays

Stacked bar charts

The stacked bar chart is a preliminary data analysis tool used to show segments of totals.



Graphical data displays

Advices to build bar charts

You should keep the following guidelines in mind when creating bar charts:

- Make bars and columns wider than the space between them.
- Use a single font type on a chart. Try to maintain a consistent font style from chart to chart in a single presentation or document.
- Order your shade pattern from darkest to lightest.
- Avoid using a combination of red and green in the same display.

Graphical data displays

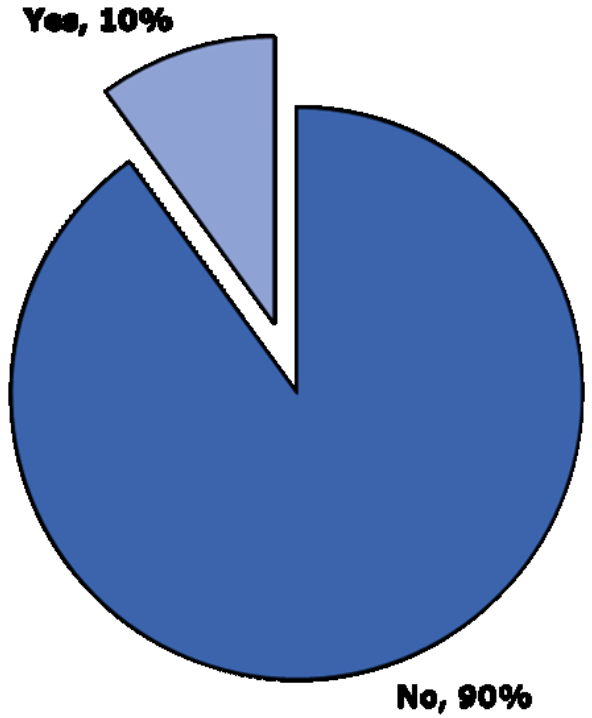
Differences between bar chart and histogram

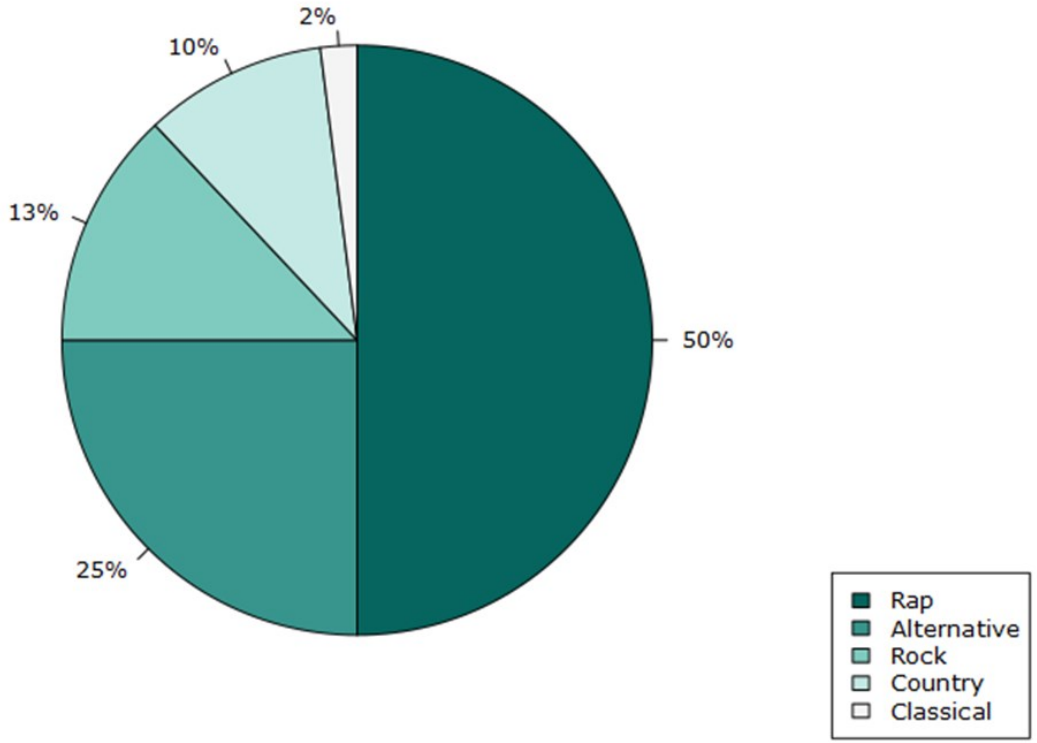
Comparison terms	Bar chart	Histogram
Usage	To compare different categories of data.	To display the distribution of a variable.
Type of variable	Categorical variables	Numeric variables
Rendering	Each data point is rendered as a separate bar.	The data points are grouped and rendered based on the bin value. The entire range of data values is divided into a series of non-overlapping intervals.
Space between bars	Can have space.	No space.
Reordering bars	Can be reordered.	Cannot be reordered.

Graphical data displays

Pie chart

A pie chart, sometimes called a circle chart, is a way of summarizing a set of nominal data or displaying the different values of a given variable (e.g. percentage distribution). This type of chart is a circle divided into a series of segments. Each segment represents a particular category. The area of each segment is the same proportion of a circle as the category is of the total data set.

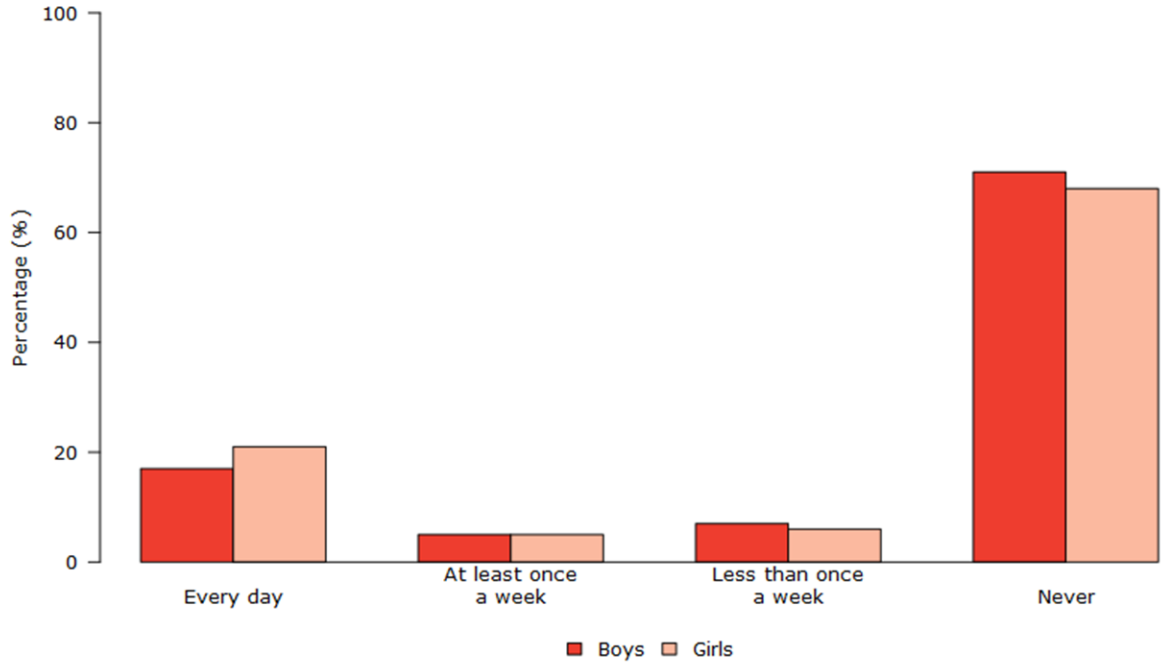




Graphical data displays

Pie charts versus bar charts

When displaying statistical information, refrain from using more than one pie chart for each figure. Below figure shows two pie charts side-by-side, where a grouped bar chart would have shown the information more clearly. A user might find it difficult to compare a segment from one pie chart to the corresponding segment of the other pie chart. However, in a grouped bar chart, these segments become bars which are lined up side by side, making it much easier to make comparisons.



Graphical data displays

A bar graph is usually better than a pie chart

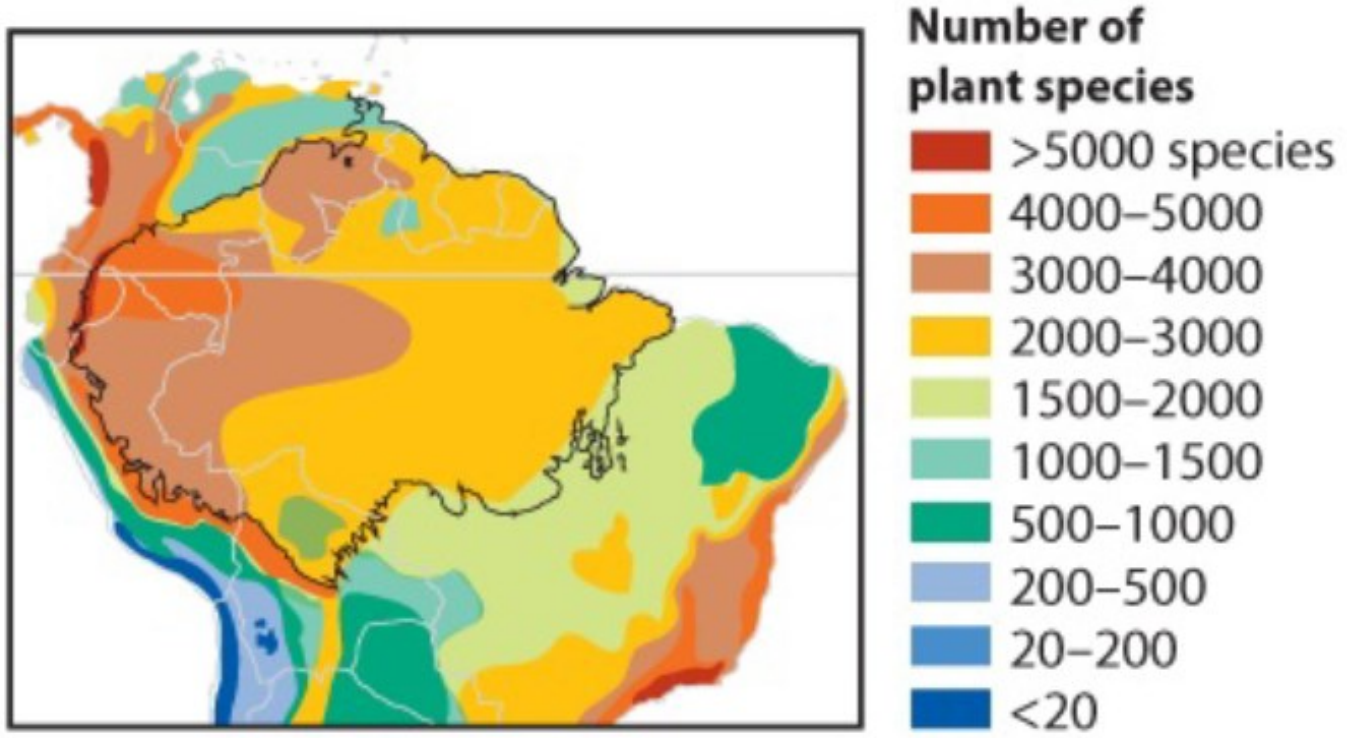
The pie chart is another type of graph often used to display frequencies of a categorical variable. This method uses colored wedges around the circumference of a circle to represent frequency or relative frequency.

One reason is that while it is straightforward to visualize the frequency of deaths in the first and most frequent category (Collecting grass/fodder), it is more difficult to compare frequencies in the remaining categories by eye. This problem worsens as the number of categories increases. Another reason is that it is very difficult to compare frequencies between two or more pie charts side by side, especially when there are many categories. To compensate, pie charts are often drawn with the frequencies added as text around the circle perimeter. The result is not better than a table. The shape of a frequency distribution is more readily perceived in a bar graph than a pie chart, and it is easier to compare frequencies between two or more bar graphs than between pie charts. Use the bar graph instead of the pie chart for showing frequencies in categorical data.

Showing trends in time and space

Maps

A **map** is the spatial equivalent of the line graph, using a color gradient to display a numerical response variable at multiple locations on a surface. The explanatory variable is location in space. One measurement is displayed for each point or interval of the surface.



Map displaying numbers of plant species in northern South America.

Graphical data displays

Density Graphs

Often when we are displaying a distribution of data we are interested in the “shape” of the data more than the actual count of values in a specific category, as shown by a standard histogram.

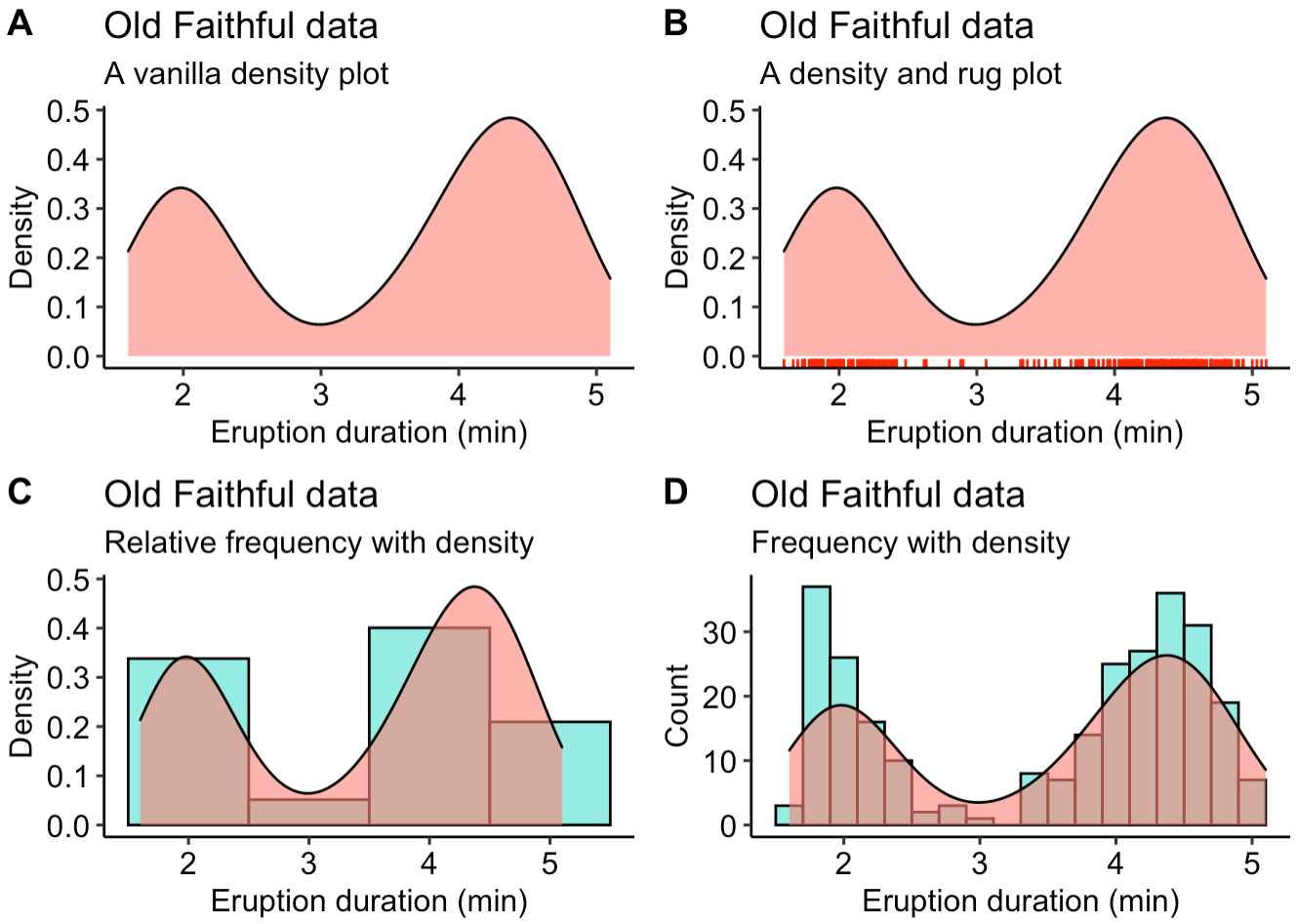
When one wishes to more organically visualize the frequency of values in a sample set a density graphs is used. These may also be thought of as smooth histograms.

These work well with histograms and rug plots, as we may see in the figure below. It is important to note with density plots that they show the relative density of the distribution along the Y axis, and *not* the counts of the data.

This can of course be changed, as seen below, but is not the default setting. Sometimes it can be informative to see how different the count and density distributions appear.

Graphical data displays

Density Graphs

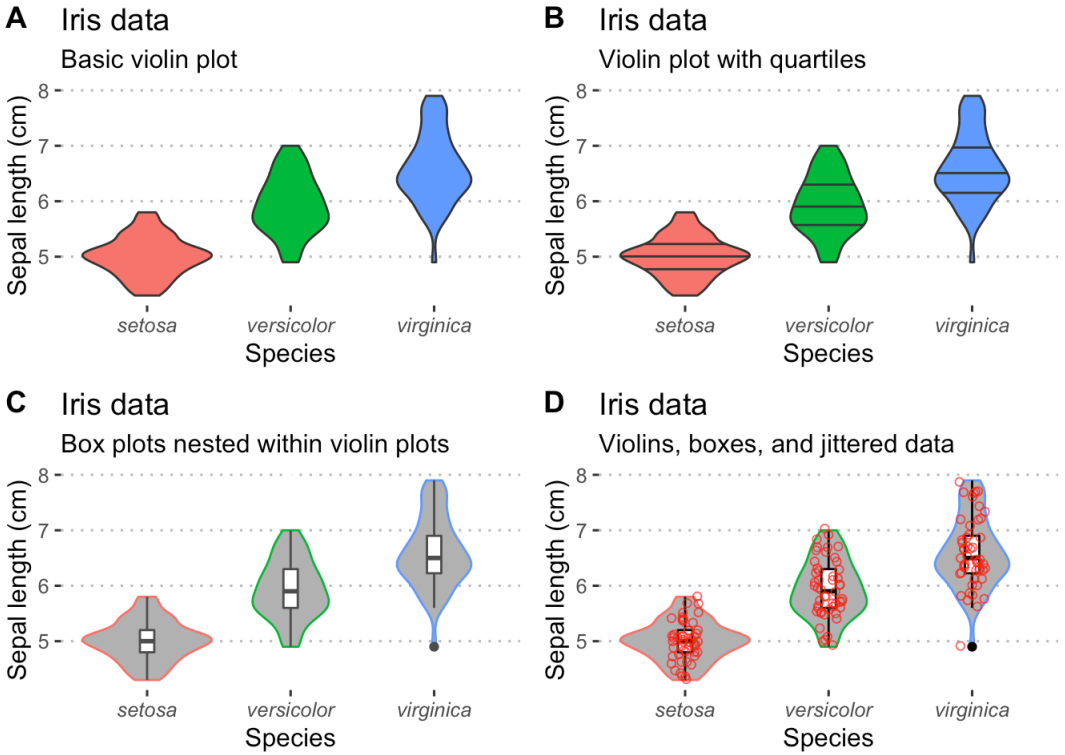


A bevy of density graphs option based on the iris data. (A) A lone density graph. (B) A density graph accompanied by a rug plot. (C) A histogram with a density graph overlay. (D) A ridge plot.

Graphical data displays

Violin Plots

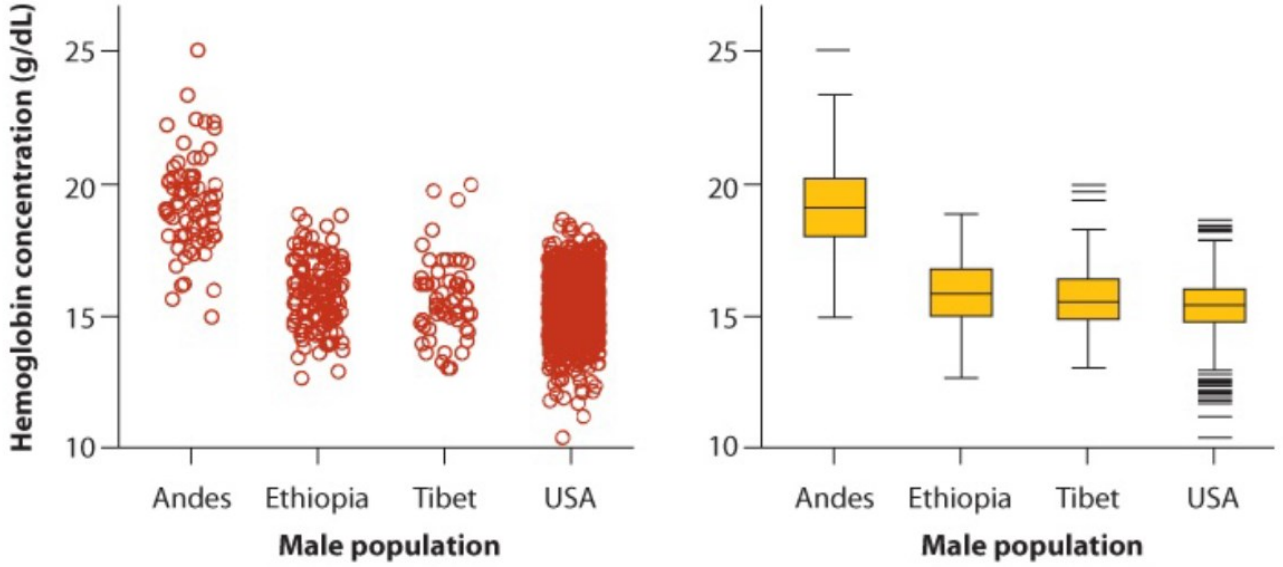
The density graph is not limited to it's use with histograms. We may combine this concept with box plots, too. These are known as violin plots and are very useful when we want to show the distribution of multiple categories of the same variable alongside one another. Violin plots may show the same information as box plots but take things one step further by allowing the shape of the boxplot to also show the distribution of the data within the sample set. We will use the iris data below to highlight the different types of violin plots one may use.



Showing association between two variables

Showing association between a numerical and a categorical variable

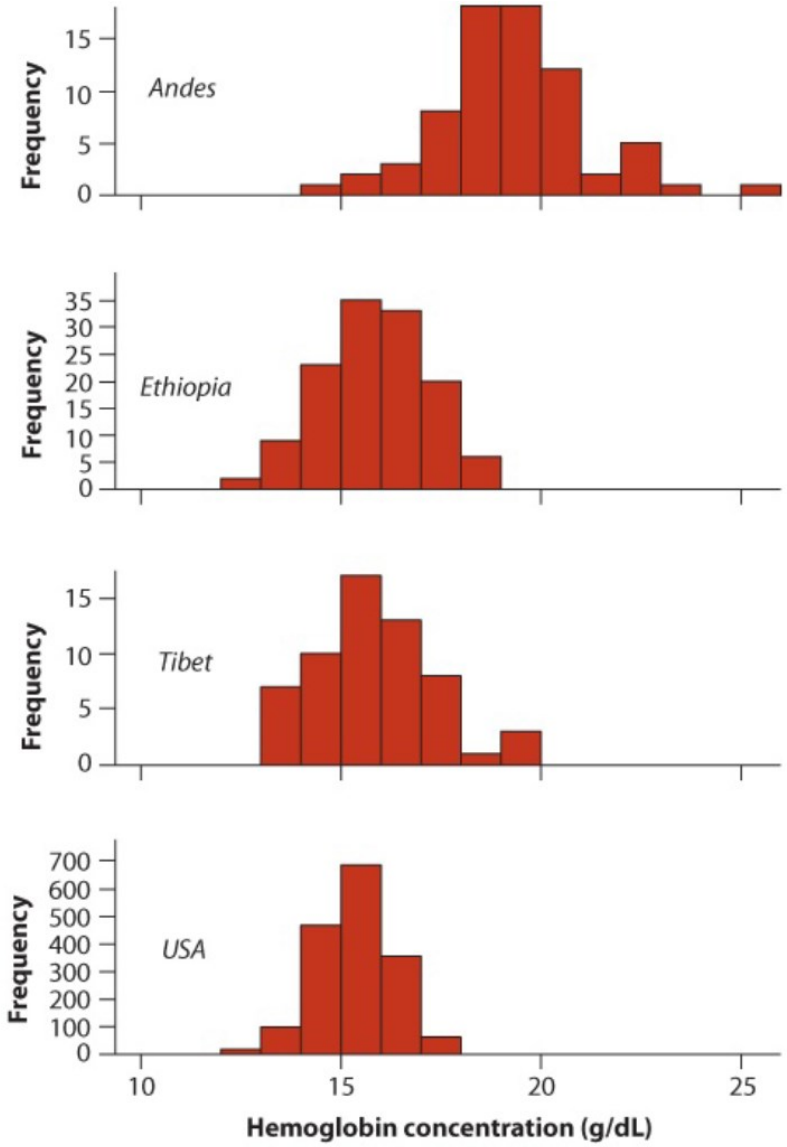
There are several good methods to show an association between a numerical variable and a categorical variable. Three that we recommend are the *strip chart (or dot plot)*, the *box plot,* and the *multiple histograms* method. Here we compare these methods with an example. We recommend against the common practice of using a bar graph because the bars make it difficult to show the data (bar graphs are ideal for frequency data). Showing an association between a numerical and a categorical variable is the same as showing a difference in the numerical variable between groups.



Strip chart (*left*) and box plot (*right*) showing hemoglobin concentration in males living at high altitude in three different parts of the world: the Andes (71), Ethiopia (128), and Tibet (59). A fourth population of 1704 males living at sea level (USA) is included as a control.

Showing association between two variables

Showing association between a numerical and a categorical variable



Multiple histograms showing the hemoglobin concentration in males of the four populations.

Thank You