Description of Samples and Populations

Introduction

Variable: a characteristic of a person or a thing that can be assigned a number or a category.

- *Categorical variable*: a variable that records which of several categories a person or thing is in, either nominal or ordinal.
  - Gender (male, female), eye color (blue, brown, green), or country of origin (USA, Canada, UK).
  - Educational attainment (elementary, high school, college, postgraduate) or Likert scale responses (strongly disagree, disagree, neutral, agree, strongly agree).
- *Numeric variable*: a variable that records the amount of something, either continuous or discrete.
  - Weight of a baby.
  - Cholesterol concentration in a blood specimen.
  - Number of bacteria colonies in a petri dish.
  - Length of a DNA segment in basepairs.

Frequency Distributions

Frequency distribution: a display of the frequency, or number of occurrences, of each value in the data set.

Example: color of poinsettias

Poinsettias can be red, pink, or white. In one investigation of the hereditary mechanism controlling the color, 182 progeny of a certain parental cross were categorized by color.

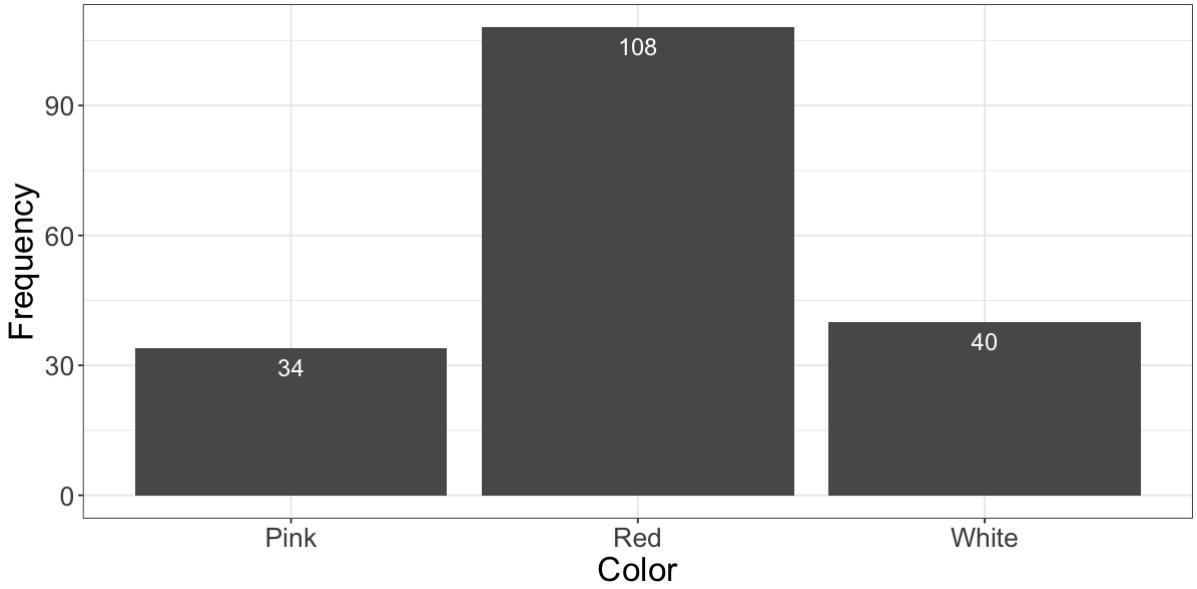
- Table:

Color	Frequency (number of plants)
Pink	34
Red	108
White	40
Total	182

- Bar chart:

In [2]:
```
library(ggplot2)
g <- ggplot(data = data.frame(Frequency = c(108, 34, 40),
```

```
                                Color = c('Red', 'Pink', 'White')),
            aes(x = Color, y = Frequency)) +
    geom_bar(stat="identity") +
    geom_text(aes(label = Frequency), vjust = 1.6, color = "white", size = 5)+
    theme_bw() +
    theme(text = element_text(size = 20))
options(repr.plot.width=10, repr.plot.height=5)
g
```

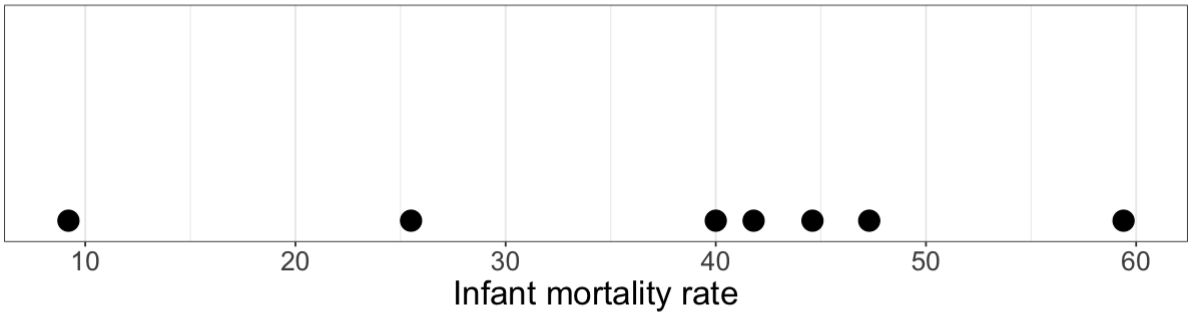


Example: infant mortality

The following table shows the infant mortality rate (infant deaths per 1,000 live births) in each of seven countries in South Asia, as of 2013.

Country	Infant mortality rate (deaths per 1,000 live births)
Bangladesh	47.3
Bhutan	40.0
India	44.6
Maldives	25.5
Nepal	41.8
Pakistan	59.4
Sri Lanka	9.2

In [7]:
```
g <- ggplot(data = data.frame(x = c(47.3, 40.0, 44.6, 25.5, 41.8, 59.4, 9.2)),
            aes(x = x)) +
    geom_dotplot(binwidth = 1) +
    scale_x_continuous(name = 'Infant mortality rate',
                       breaks = seq(10, 60, 10)) +
    scale_y_continuous(NULL, breaks = NULL) +
    theme_bw() +
    theme(text = element_text(size = 20), aspect.ratio=1/5)
options(repr.plot.width=10, repr.plot.height=5)
g
```



Relative frequency

The frequency scale is often replaced by a relative frequency scale:

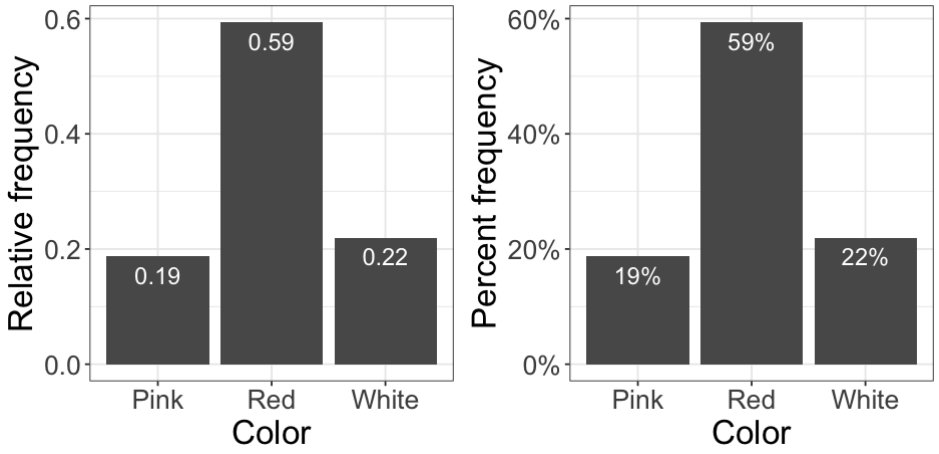
$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

As another option, a relative frequency can be expressed as a percentage frequency.

Color	Frequency (number of plants)	Relative frequency	Percent frequency
Pink	34	.19	19
Red	108	.59	59
White	40	.22	22
Total	182	1.00	100

In [90]:
```
g1 <- ggplot(data = data.frame(Frequency = c(108, 34, 40),
                               Color = c('Red', 'Pink', 'White')),
             aes(x = Color, y = Frequency / sum(Frequency))) +
    geom_bar(stat="identity") +
    geom_text(aes(label = round(Frequency / sum(Frequency), 2)),
              vjust = 1.6, color = "white", size = 5)+
    labs(y = 'Relative frequency') +
    theme_bw() +
    theme(text = element_text(size = 20))
g2 <- ggplot(data = data.frame(Frequency = c(108, 34, 40),
                               Color = c('Red', 'Pink', 'White')),
             aes(x = Color, y = Frequency / sum(Frequency))) +
    geom_bar(stat="identity") +
    geom_text(aes(label = paste0(100 * round(Frequency / sum(Frequency), 2),
                                 '%')), vjust = 1.6, color = "white",
              size = 5)+
    scale_y_continuous(name = 'Percent frequency', labels=scales::percent) +
    theme_bw() +
    theme(text = element_text(size = 20))
library(patchwork)
```

```
options(repr.plot.width=8, repr.plot.height=4)
g1 + g2
```



Grouped frequency distributions and histograms

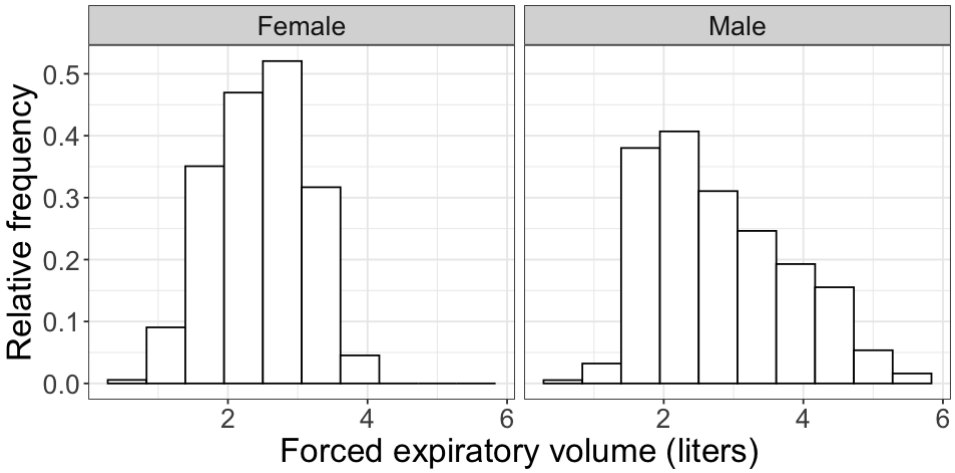
For many data sets, it is necessary to group the data in order to condense the information adequately. (This is usually the case with continuous variables.)

Example: forced expiratory volume in children

A total of 654 children, comprising 336 boys and 318 girls, underwent examination to measure their forced expiratory volume in liters.

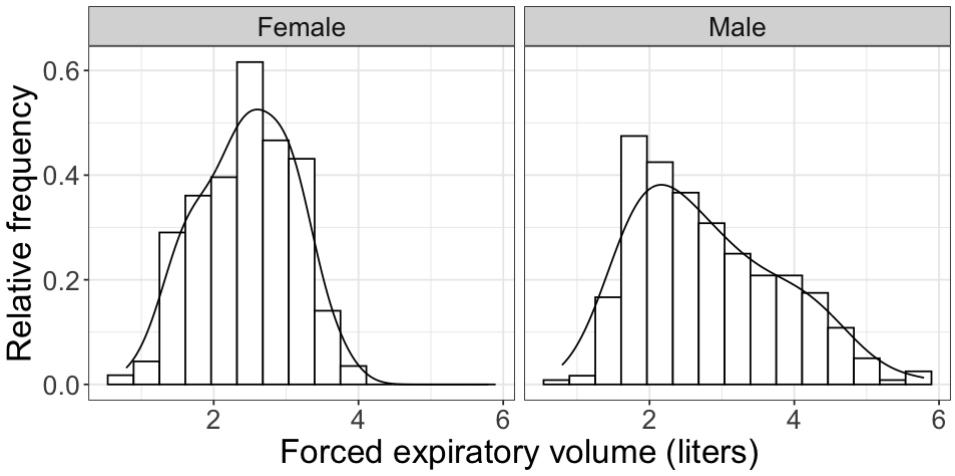
In [91]:
```
library(isdals)
data(fev)
fev$Gender <- ifelse(fev$Gender == 0, 'Female', 'Male')
table(fev$Gender)
g <- ggplot(data = fev, aes(x = FEV, y = ..density..)) +
    geom_histogram(color="black", fill = 'white', bins = 10) +
# can also change bins to obtain finer or coarser histograms
    labs(x = "Forced expiratory volume (liters)", y = "Relative frequency") +
    facet_wrap(~Gender) +
    theme_bw() +
    theme(text = element_text(size = 20))
options(repr.plot.width=8, repr.plot.height=4)
g
```

```
Female   Male
   318    336
```



When discussing a set of data, we want to describe the shape, center, and spread of the distribution. The shape of a distribution can be indicated by a smooth curve that approximates the histogram.

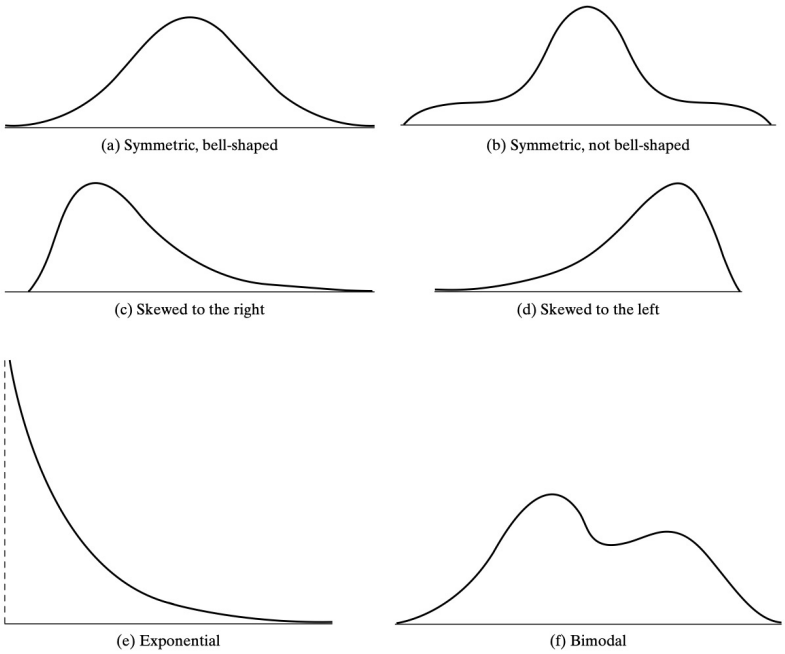
In [92]:
```
g <- ggplot(data = fev, aes(x = FEV, y = ..density..)) +
    geom_histogram(color="black", fill = 'white', bins = 15) +
# can also change bins to obtain finer or coarser histograms
    geom_density(adjust = 1.5) +
    #geom_vline(aes(xintercept = mean(FEV)), col = 'orange')+
    #geom_vline(aes(xintercept = median(FEV)), col = 'skyblue')+
    labs(x = "Forced expiratory volume (liters)", y = "Relative frequency") +
    facet_wrap(~Gender) +
    theme_bw() +
    theme(text = element_text(size = 20))
options(repr.plot.width=8, repr.plot.height=4)
g
```



Shapes of distributions

A common shape for biological data is unimodal (has one mode) and is somewhat skewed to the right, as in (c). Approximately bell-shaped distributions, as in (a), also occur. Sometimes a distribution is symmetric but differs from a bell in having long tails; an exaggerated version is shown in (b). Left-skewed (d) and exponential (e) shapes are less

common. Bimodality (two modes), as in (f), can indicate the existence of two distinct subgroups of observational units.



How to tell if a distribution is left skewed or right skewed

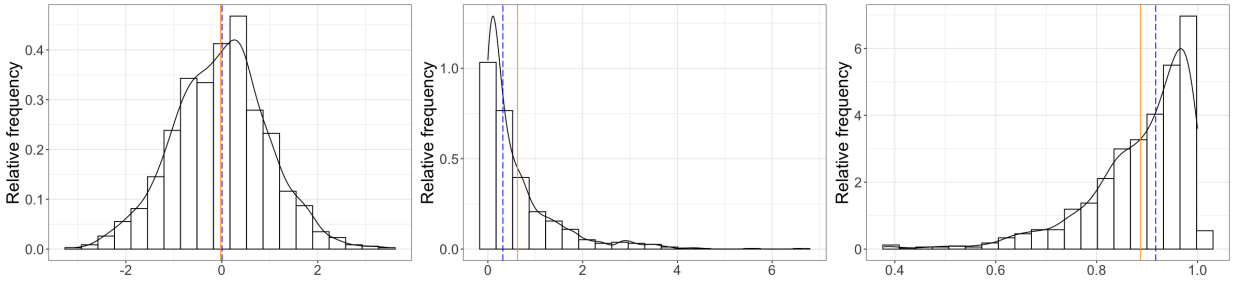
A skewed distribution occurs when one tail is longer than the other. Skewness defines the asymmetry of a distribution.

- Skewed to the right: The mean is **greater** than the median.
- Skewed to the left: The mean is **less** than the median.
    - The **mean** is the average of a data set.
    - The **mode** is the most common number in a data set.
    - The **median** is the middle of the set of numbers.

In [9]:
```
# Normal distribution
g1 <- ggplot(data = data.frame(x = rnorm(1000)),
             aes(x = x, y = after_stat(density))) +
    geom_histogram(color="black", fill = 'white', bins = 20) +
    geom_density() +
    geom_vline(aes(xintercept = mean(x)), col = 'orange') +
    geom_vline(aes(xintercept = median(x)), col = 'blue',
               linetype = 'longdash') +
    labs(x = "", y = "Relative frequency") +
    theme_bw() +
    theme(text = element_text(size = 20))
# Gamma distribution
g2 <- ggplot(data = data.frame(x = rgamma(1000, shape = 0.6)),
             aes(x = x, y = after_stat(density))) +
    geom_histogram(color="black", fill = 'white', bins = 20) +
    geom_density() +
    geom_vline(aes(xintercept = mean(x)), col = 'orange') +
    geom_vline(aes(xintercept = median(x)), col = 'blue',
               linetype = 'longdash') +
```

```
    labs(x = "", y = "Relative frequency") +
    theme_bw() +
    theme(text = element_text(size = 20))
# Beta distribution
g3 <- ggplot(data = data.frame(x = rbeta(1000, shape1 = 8, shape2 = 1)),
             aes(x = x, y = after_stat(density))) +
    geom_histogram(color="black", fill = 'white', bins = 20) +
    geom_density() +
    geom_vline(aes(xintercept = mean(x)), col = 'orange') +
    geom_vline(aes(xintercept = median(x)), col = 'blue',
               linetype = 'longdash') +
    labs(x = "", y = "Relative frequency") +
    theme_bw() +
    theme(text = element_text(size = 20))
```

In [10]:
```
# orange solid: mean
# blue dashed: median
library(patchwork)
options(repr.plot.width=20, repr.plot.height=5)
g1 + g2 + g3
```



Descriptive Statistics: Measures of Center

- A numerical measure calculated from sample data is called a **statistic**.
- **Descriptive statistics** are statistics that describe a set of data.
- Usually the descriptive statistics for a sample are calculated in order to provide information about a population of interest.
- Two most widely used measures of center: the **median** and the **mean**.

Median $\tilde{y}$

- The sample median is the value that most nearly lies in the middle of the sample, i.e., the data value that splits the ordered data into two equal halves.
- To find the median, first arrange the observations in increasing order. In the array of ordered observations, the median is the middle value (if n is odd) or midway between the two middle values (if n is even).

Example: weight gain of lambs

- The following are the 2-week weight gains (lb) of six young lambs of the same breed that had been raised on the same diet:

    - 11 13 19 2 10 1
- Suppose the sample contained one more lamb, with 2-week weight gains (lb) being 10.

A more formal way to define the median is in terms of rank position in the ordered array (counting the smallest observation as rank 1, the next as 2, and so on). The rank position of the median is equal to $(0.5)(n+1)$. Note that the formula $(0.5)(n+1)$ does not give the median, it gives the location of the median within the ordered list of the data.

Mean $\bar{y}$

- The most familiar measure of center is the ordinary average or mean (sometimes called the arithmetic mean).
- The mean of a sample (or "the sample mean") is the sum of the observations divided by the number of observations.
- The general definition of the sample mean is $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, where the y_i are the observations in the sample and n is the sample size.

Example: weight gain of lambs

- The following are the 2-week weight gains (lb) of six young lambs of the same breed that had been raised on the same diet:
    - 11 13 19 2 10 1

Robustness

A statistic is said to be robust if the value of the statistic is relatively unaffected by changes in a small portion of the data, even if the changes are dramatic ones.

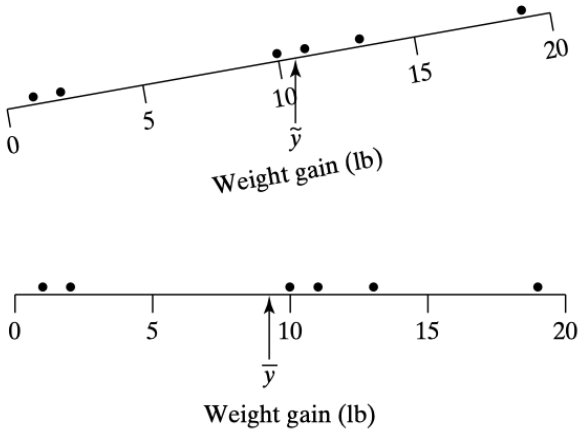
Recall that for the lamb weight-gain data,

- if the observation 19 is changed 14, the mean becomes 8.5 and the median does not change;
- if the observation 19 is changed 29, the mean becomes 11 and the median does not change.

Median v.s mean

- While the median divides the data into two equal pieces (i.e., the same number of observations above and below), the mean is the "point of balance" of the data.
- The median is more robust than the mean.
- An advantage of the mean is that in some circumstances it is more efficient than the median. Partly because of its efficiency, the mean has played a major role in classical methods in statistics.

- If the frequency distribution is symmetric, the mean and the median are equal and fall in the center of the distribution. If the frequency distribution is skewed, both measures are pulled toward the longer tail, but the mean is usually pulled farther than the median.



Boxplots

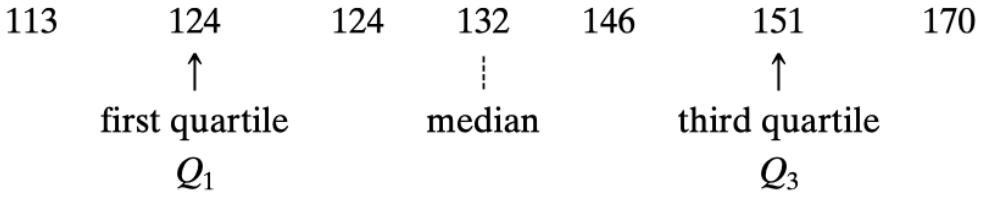
One of the most efficient graphics, both for examining a single distribution and for making comparisons between distributions, is known as a boxplot.

Quartile and the interquartile range

- The median of a distribution splits the distribution into two parts, a lower part and an upper part. The quartiles of a distribution divide each of these parts in half, thereby dividing the distribution into four quarters.
- The minimum, the maximum, the median, and the quartiles, taken together, are referred to as the **five-number summary** of the data.

Example: blood pressure

- The systolic blood pressures (mm Hg) of seven middle-aged men were as follows:
  - 151 124 132 170 146 124 113



- Suppose one more observation 130 is added in the sample.

Interquartile range

The interquartile range is the difference between the first and third quartiles and is abbreviated as IQR, which measures the spread of the middle 50\% of the distribution.

$$\text{IQR} = Q_3 - Q_1$$

Recall that for the blood pressure data, $Q_1 = 124$ and $Q_3 = 151$. It follows that $\text{IQR} = 151 - 124 = 27$.

Outliers

- Sometimes a data point differs so much from the rest of the data that it doesn't seem to belong with the other data. Such a point is called an outlier.
- An outlier might occur because of a recording error or typographical error when the data are recorded, because of an equipment failure during an experiment, or for many other reasons.

To given a definition of outlier, we first discuss what are known as fences.

- The **lower fence** of a distribution is

$$\text{lower fence} = Q_1 - 1.5 \times \text{IQR}$$

- The **upper fence** of a distribution is

$$\text{upper fence} = Q_3 + 1.5 \times \text{IQR}$$

An outlier is a data point that falls outside of the fences. That is, if

$$\text{data point} < Q_1 - 1.5 \times \text{IQR}$$

or

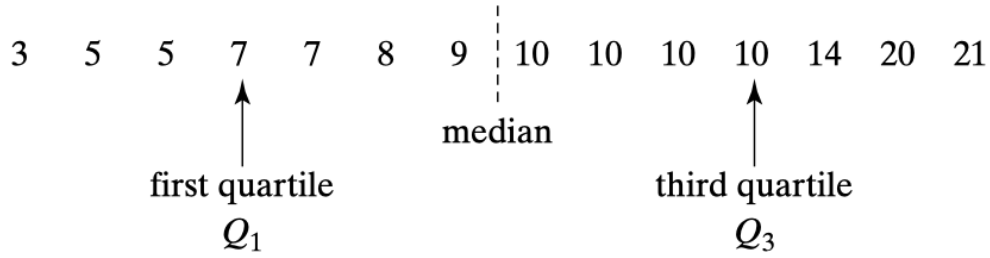
$$\text{data point} > Q_3 + 1.5 \times \text{IQR}$$

then we call the point an outlier.

Recall that for the blood pressure data, $Q_1 = 124$, $Q_3 = 151$, and $\text{IQR} = 27$. It follows that the lower fence is $124 - 1.5 \times 27 = 83.5$ and the upper fence is $151 + 1.5 \times 27 = 191.5$. Any point less than 83.5 or greater than 191.5 would be an outlier. There is thus no outliers in this data set.

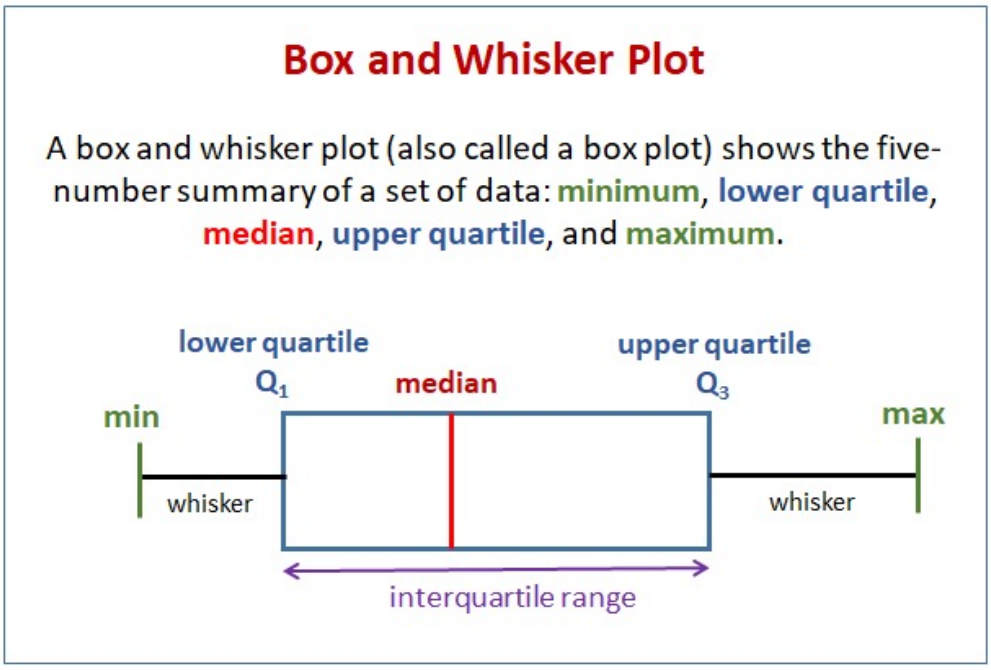
Example: radish growth in light

A common biology experiment involves growing radish seedlings under various conditions. In one experiment students grew 14 radish seedlings in constant light. The observations, in order, are



Boxplots for data with no outliers

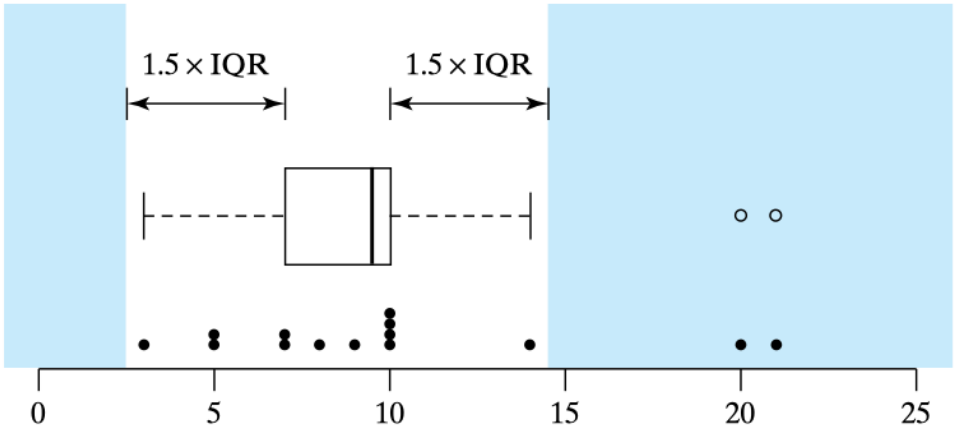
A boxplot is a visual representation of the five-number summary.



The boxplot of blood pressure data is

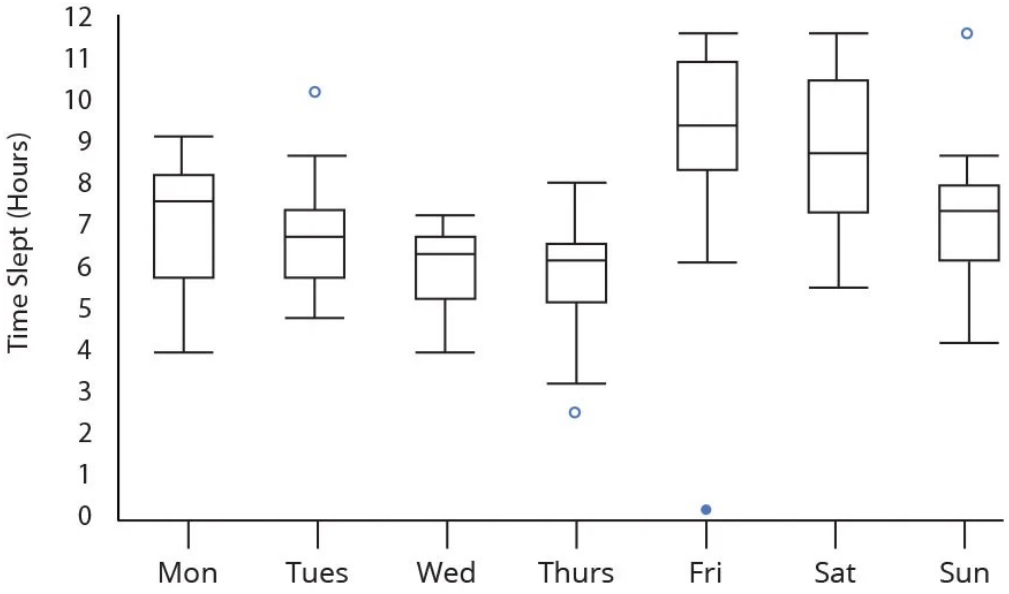
Boxplots for data with outliers

If there are outliers in the lower or upper part of the distribution, we identify them with dots and extend a whisker from Q_1 down to the smallest observation that is not an outlier or from Q_3 up to the largest data point that is not an outlier.

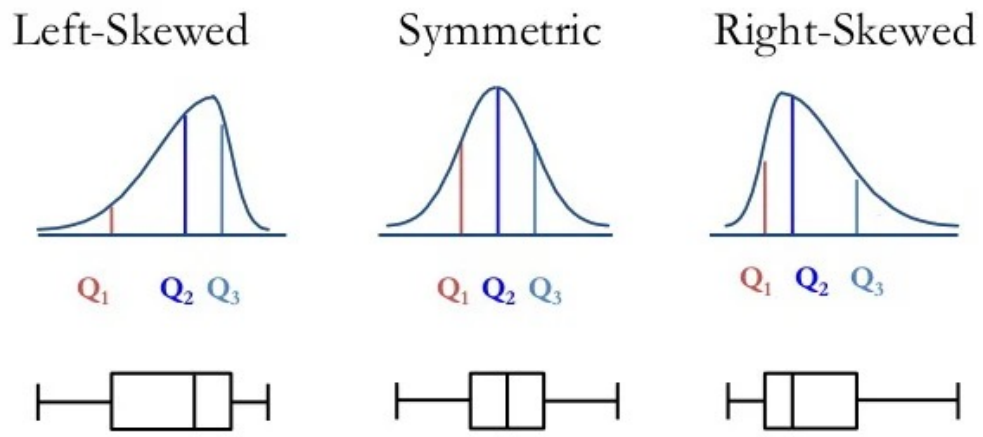


How to read boxplots?

- **Spread**: The length of the box and the whiskers provides information about the spread or variability of the data. A longer box and longer whiskers indicate greater variability, while a shorter box and shorter whiskers indicate less variability.
- **Outliers**: Any data points plotted beyond the whiskers are considered outliers and may be worth further investigation as they deviate significantly from the rest of the dataset.



- **Skewness**: The position of the median within the box can indicate the skewness of the distribution. If the median is closer to the upper quartile, the distribution is left-skewed, while if it is closer to the lower quartile, the distribution is right-skewed.



Percentiles and quantiles

- The **percentile** is a data value where a certain percentage of observations fall below that data value. The qth percentile of a sample is the value below which q percent of the individuals lie.
- The same information in a percentile is sometimes represented as a **quantile**. This only means that the proportion less than or equal to the given value is represented as a *decimal* rather than as a percentage.
  - Median: 50th percentile
  - First quartile: 25th percentile
  - Third quartile: 75th percentile
  - Minimum: 0th percentile
  - Maximum: 100th percentile
  - 10th percentile: 0.10 quantile

Relationship between Variables

Categorical-categorical relationships

Suppose we are studying the relationship between the diet (plant-based or animal-based) and the occurrence of a specific health condition (e.g., high blood pressure) among a group of individuals.

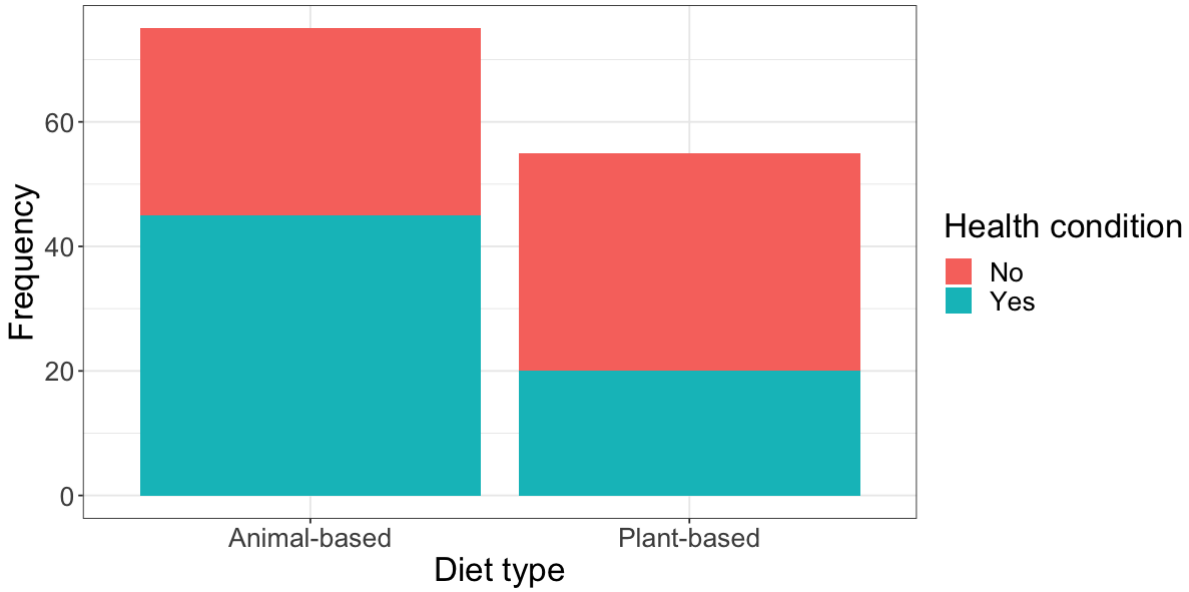
- bivariate frequency table

Diet Type	Health Condition: Yes	Health Condition: No
Plant-based	20	35
Animal-based	45	30

- stacked bar charts
- stacked relative frequency (or percentage) bar charts

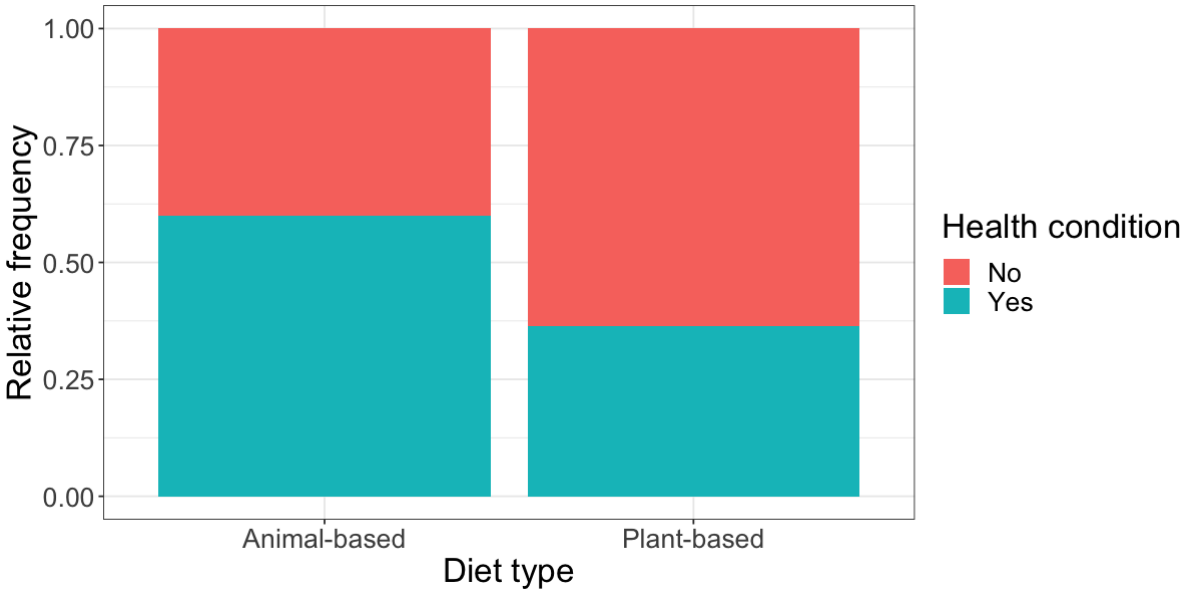
In [3]:
```r
# Create a data frame with the bivariate frequency table data
data <- data.frame(
    Diet_Type = c("Plant-based", "Plant-based", "Animal-based",
                  "Animal-based"),
    Health_Condition = c("Yes", "No", "Yes", "No"),
    Frequency = c(20, 35, 45, 30)
)

# Create the stacked bar chart
g <- ggplot(data, aes(x = Diet_Type, y = Frequency,
                      fill = Health_Condition)) +
    geom_bar(stat = "identity") +
    labs(x = "Diet type", y = "Frequency", fill = "Health condition") +
    theme_bw() +
    theme(text = element_text(size = 20))
options(repr.plot.width=10, repr.plot.height=5)
g
```



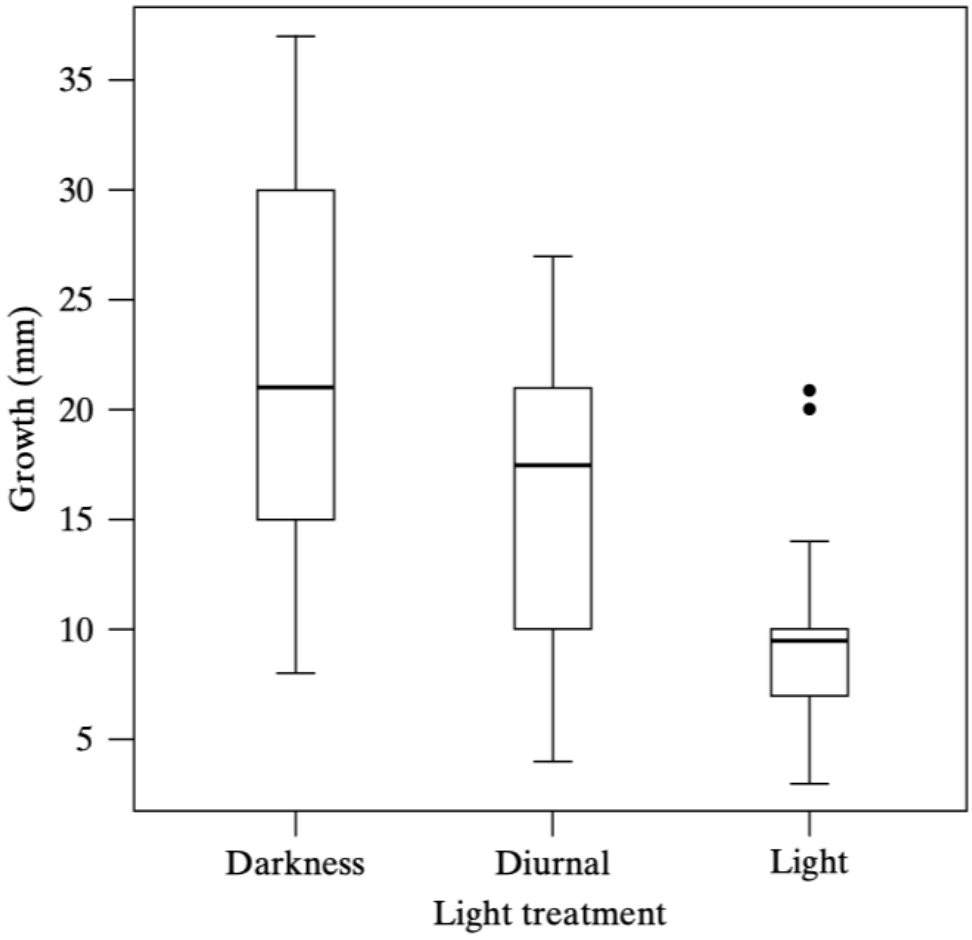
In [4]:
```r
# Calculate relative frequencies within each diet type
data <- transform(data, Relative_Frequency = Frequency /
                  tapply(Frequency, Diet_Type, sum)[Diet_Type])

# Create the stacked relative frequency bar chart
g <- ggplot(data, aes(x = Diet_Type, y = Relative_Frequency,
                      fill = Health_Condition)) +
    geom_bar(stat = "identity") +
    labs(x = "Diet type", y = "Relative frequency",
         fill = "Health condition") +
    theme_bw() +
    theme(text = element_text(size = 20))
options(repr.plot.width=10, repr.plot.height=5)
g
```



Numeric-categorical relationships

- Side-by-side boxplots of radish growth under three conditions: constant darkness, half light–half darkness, and constant light.

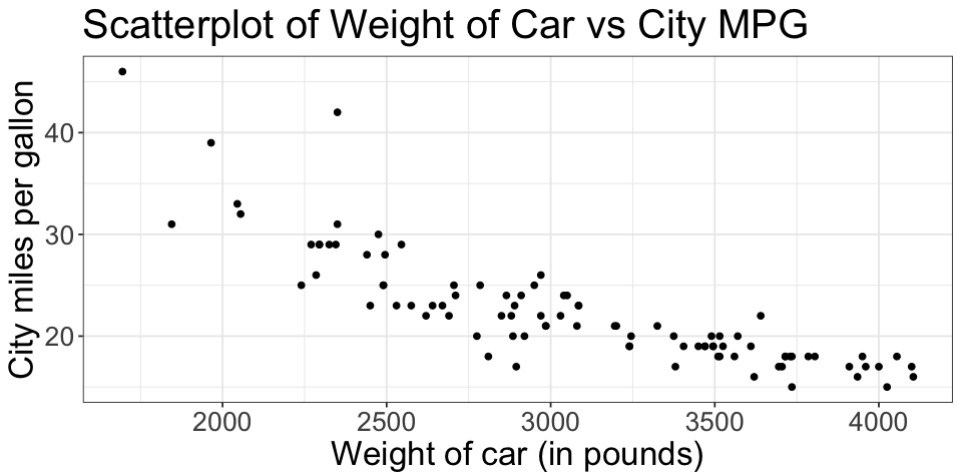


Numeric-numeric relationships

- scatterplot

In [16]:
```
library(MASS)

g <- ggplot(data = Cars93, aes(x = Weight, y = MPG.city)) +
    geom_point() +
    labs(title = "Scatterplot of Weight of Car vs City MPG",
             x = "Weight of car (in pounds)",
             y = "City miles per gallon")+
    theme_bw() +
    theme(text = element_text(size = 20))
options(repr.plot.width=8, repr.plot.height=4)
g
```



Measures of Dispersion

- We have considered the shapes and centers of distributions, but a good description of a distribution should also characterize how spread out the distribution is; are the observations in the sample all nearly equal, or do they differ substantially?
- We defined the interquartile range (IQR) in previous section, which is one measure of dispersion. Here we consider other measures of dispersion: the range and the standard deviation.

The range

The sample **range** is the difference between the largest and smallest observations in a sample.

Recall the blood pressure data: The systolic blood pressures (mm Hg) of seven middle-aged men were as follows:

151 124 132 170 146 124 113

- The range is easy to calculate, but it is very sensitive to extreme values; that is, it is not robust.
- Unlike the range, the IQR is robust.

The standard deviation

The standard deviation is the classical and most widely used measure of dispersion. The sample standard deviation is denoted by s and is defined by the following formula:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Here $y_i - \bar{y}$ is called the deviation between observation y_i, and the sample mean and $\sum_{i=1}^{n}(y_i - \bar{y})^2$ denotes the sum of the squared deviations.

The sample variance, denoted by s^2, is simply the standard deviation squared:

$$\text{variance} = s^2 \text{ or } s = \text{variance}.$$

We will frequently abbreviate "standard deviation" as "SD"; the symbol "s" will be used in formulas.

Example: chrysanthemum growth

In an experiment on chrysanthemums, a botanist measured the stem elongation (mm in 7 days) of five plants grown on the same greenhouse bench. The results were as follows:

76 72 65 70 82

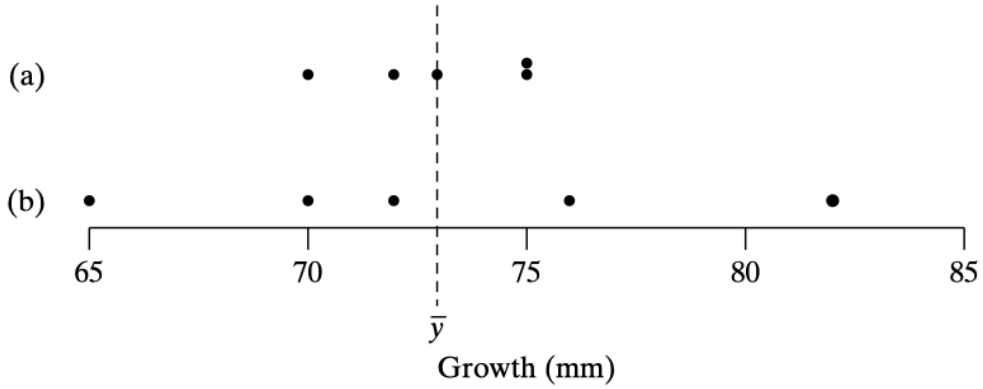
Observation	Deviation	Squared deviation
76		
72		
65		
70		
82		
Sum		

Interpretation of the definition of s

If the chrysanthemum growth data are

75 72 73 75 70

then the mean is the same (y = 73 mm), but the SD is smaller (s = 2.1 mm), because the observations lie closer to the mean.



Why $n-1$?

Note that the sum of the deviations $y_i - y$ is always zero. Thus, once the first $n-1$ deviations have been calculated, the last deviation is constrained. This means that in a sample with n observations, there are only $n-1$ units of information concerning deviation from the average. The quantity $n-1$ is called the **degrees of freedom** of the standard deviation or variance.

Consider the extreme case when $n=1$ and $n=2$ with $y_1 = y_2$.

The Empirical Rule

For "nicely shaped" distributions; that is, unimodal distributions that are not too skewed and whose tails are not overly long or short, we usually expect to find

- about 68% of the observations within ± 1 SD of the mean.
- about 95% of the observations within ± 2 SDs of the mean.
- >99% of the observations within ± 3 SDs of the mean.

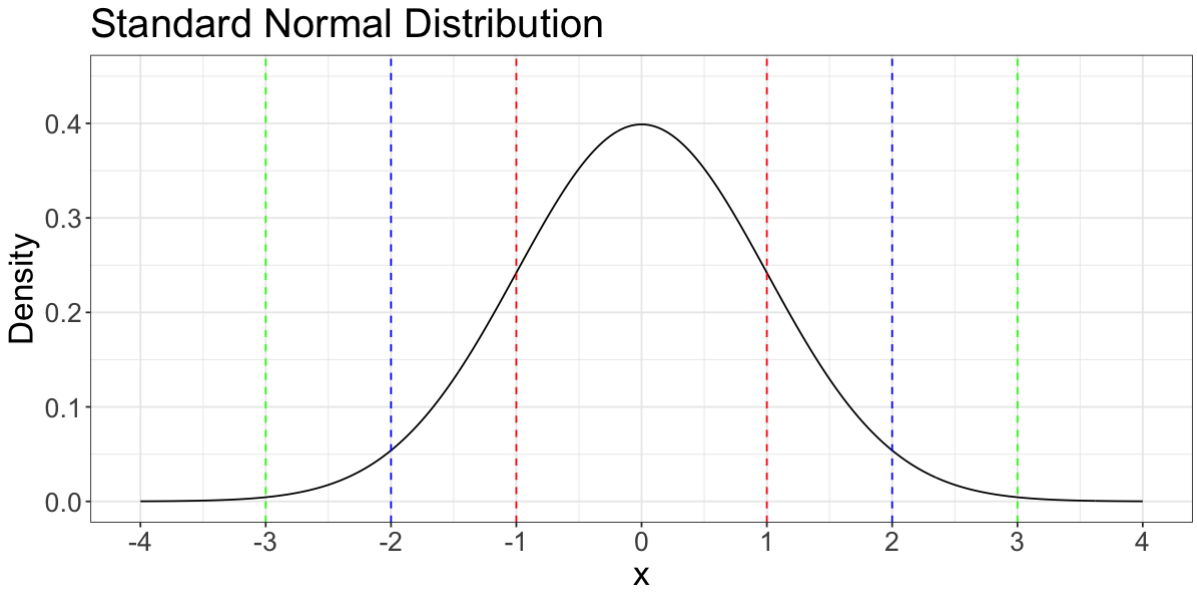
In [13]:
```r
# Generate x values for the density plot
x <- seq(-4, 4, by = 0.01)

# Calculate the density values for the standard normal distribution
density <- dnorm(x)

# Create a data frame with x and density values
data <- data.frame(x = x, density = density)

# Create the plot using ggplot2
g <- ggplot(data, aes(x = x, y = density)) +
    geom_line() +
    labs(x = "x", y = "Density", title = "Standard Normal Distribution") +
    geom_vline(xintercept = c(-1, 1), col = "red", linetype = "dashed") +
    geom_vline(xintercept = c(-2, 2), col = "blue", linetype = "dashed") +
    geom_vline(xintercept = c(-3, 3), col = "green", linetype = "dashed") +
    scale_x_continuous(breaks = seq(-4, 4, by = 1)) +
    scale_y_continuous(limits = c(0, 0.45)) +
    theme_bw() +
    theme(text = element_text(size = 20))
```

In [14]:
```
options(repr.plot.width=10, repr.plot.height=5)
g
```



Robustness: IQR > SD > range

In this course, we will rely primarily on the mean and SD rather than other descriptive measures.

Effect of Transformation of Variables

For example, we might convert from inches to centimeters or from °F to °C. Transformation, or reexpression, of a variable Y means replacing Y by a new variable, say Y'.

Linear transformations

For linear transformations, a graph of Y against Y' would be a straight line. A familiar reason for linear transformation is a change in the scale of measurement.

- Multiplicative transformations: Suppose Y represents the weight of an animal in kg, and we decide to reexpress the weight in lb. Then

$$Y = \text{Weight in kg}$$

$$Y' = \text{Weight in lb}$$

so

$$Y' = 2.2Y$$

- Additive and multiplicative transformations:

$$Y = \text{Temperature in } ^\circ\text{C}$$

$$Y' = \text{Temperature in } ^\circ\text{F}$$

then

$$Y' = 1.8Y + 32$$

A linear transformation consists of (1) multiplying all the observations by a constant, or (2) adding a constant to all the observations, or (3) both.

Under a linear transformation $Y' = aY + b$,

- $\bar{y}' = a\bar{y} + b$
- $\tilde{y}' = a\tilde{y} + b$
- $s' = as$
- $\mathrm{IQR}' = a\mathrm{IQR}$

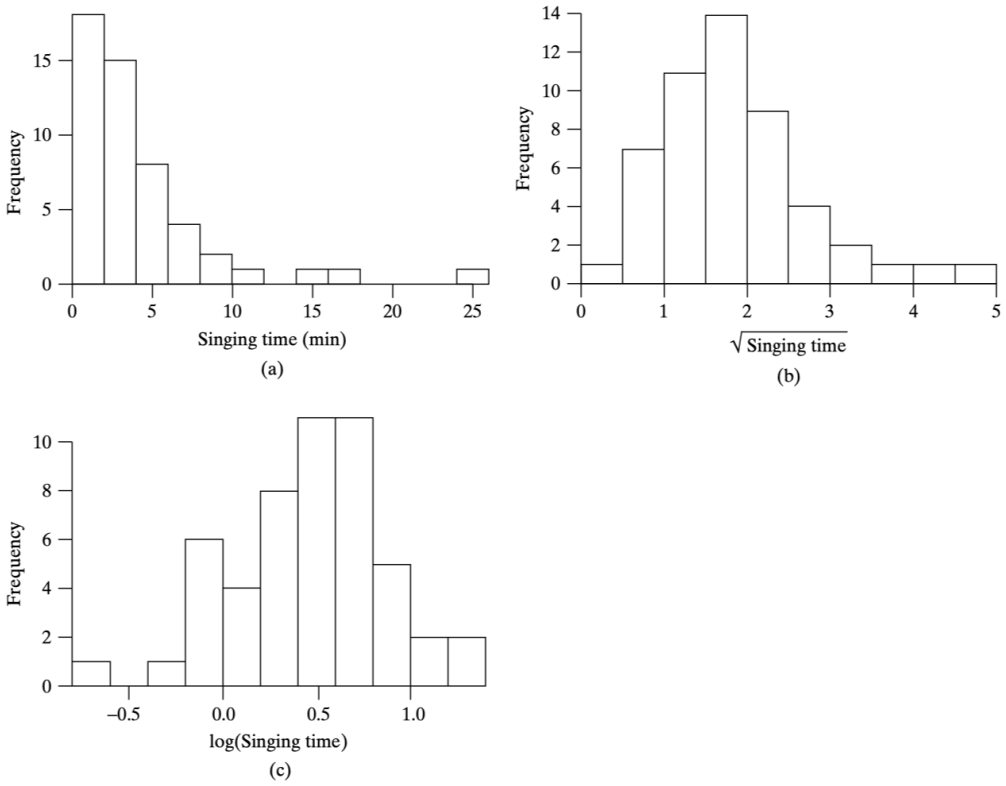
Nonlinear transformations

Data are sometimes reexpressed in a nonlinear way. Examples of nonlinear transformations are

- $Y' = \sqrt{Y}$
- $Y' = \log Y$
- $Y' = \frac{1}{Y}$
- $Y' = Y^2$

- The logarithmic transformation is especially common in biology because many important relationships can be simply expressed in terms of logs. For instance, there is a phase in the growth of a bacterial colony when log(colony size) increases at a constant rate with time.

- If a distribution is skewed to the right, we may wish to apply a transformation that makes the distribution more symmetric, by pulling in the right-hand tail. Using $Y' = \sqrt{Y}$ will pull in the right-hand tail of a distribution and push out the left-hand tail. The transformation $Y' = \log Y$ is more severe than $\sqrt{Y}$ in this regard.



Statistical Inference

The process of drawing conclusions about a population, based on observations in a sample from that population, is called statistical inference.

Example: blood types

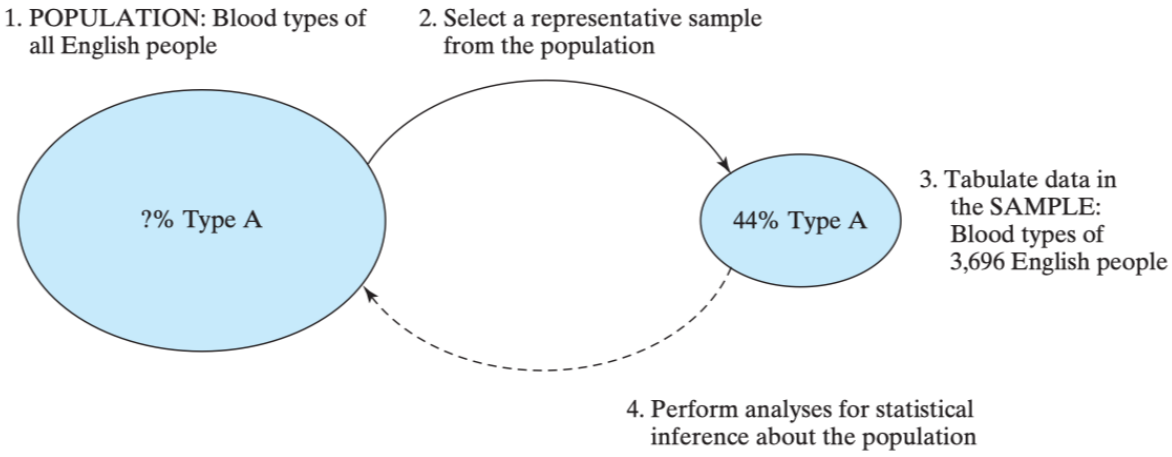
In an early study of the ABO blood-typing system, researchers determined blood types of 3,696 persons in England.

Blood type	Frequency
A	1,634
B	327
AB	119
O	1,616
Total	3,696

These data were not collected for the purpose of learning about the blood types of those particular 3,696 people. Rather, they were collected for their scientific value as a source of information about the distribution of blood types in a larger population. For instance, one might presume that the blood type distribution of all English people should resemble the distribution for these 3,696 people. In particular, the observed relative frequency of type A blood was

$$\frac{1634}{3696} \text{ or } 44\% \text{ type A}$$

One might conclude from this that approximately 44% of the people in England have type A blood.



In making a statistical inference, we hope that

- the sample is representative of the population.
- the sample size cannot be too small.

Describing a population

- Just as each sample has a distribution, a mean, and an SD, so also we can envision a population distribution, a population mean, and a population SD.
- In statistical language, we say that the sample characteristic is an estimate of the corresponding population characteristic.
- A sample characteristic is called a **statistic**; a population characteristic is called a **parameter**.

Proportions

For a categorical variable, we can describe a population by simply stating the proportion, or relative frequency, of the population in each category. The sample proportion of a category is an estimate of the corresponding population proportion.

$$p = \text{ Population proportion}$$

$$\hat{p} = \text{ Sample proportion}$$

The symbol "^" can be interpreted as "estimate of". Thus,

$$\hat{p} \text{ is an estimate of } p$$

Mean and SD

If the observed variable is quantitative, one can consider descriptive measures such as the mean, the SD, the median, the quartiles and so on. Each of these quantities can be computed for a sample of data, and each is an estimate of its corresponding population analog.

The population mean is denoted by μ (mu), and the population SD is denoted by σ (sigma). We may define these as follows for a quantitative variable Y:

$$\mu = \text{ Population average value of } Y$$

$$\sigma = \sqrt{\text{Population average value of } (Y - \mu)^2}$$

Measure	Sample value (statistics)	Population value (parameter)
Proportion	$\hat{p}$	p
Mean	$\bar{y}$	μ
Standard deviation	s	σ