Biostatistics

David Coffey

2018-07-27

Contents

Chapter 1

General overview

1.1 Introduction

This books provides a consise overview of biostatistics and its applications using the R programming language. The textbook *Fundamentals of Biostatitics* (Rosner, 2016) was used extensivity in the preparation of this book.

1.2 Example dataset

Examples of R functions are performed on a dataset of patients with newly diagnosed multiple myeloma. This dataset contains a variety of categorical and continuous variables. A description of the variables are shown below.

Column	Description
ID	Patient identifier
Sex	Patient sex
Race	Patient race
Age	Patient age in years at the time of diagnosis
Stage	Disease stage according to the international staging system
SurvivalMonths	Duration in months between diagnosis and the last date of contact
Status	Survival status of the patient
DiagnosisYear	Year of diagnosis
Treatment	Initial treatment
TreatmentDurationMonths	Duration of the initial treatment in months
BonyLesions	Number of bony lesions on initial imaging study (MRI or X-ray)
PlasmaCells	Percentage of plasma cells on initial bone marrow biopsy
1q+	FISH result on initial bone marrow biopsy
del13q	FISH result on initial bone marrow biopsy
del17p	FISH result on initial bone marrow biopsy
del1p	FISH result on initial bone marrow biopsy
t(11;14)	FISH result on initial bone marrow biopsy
t(14;16)	FISH result on initial bone marrow biopsy
t(4:14)	FISH result on initial bone marrow biopsy
t(6;14)	FISH result on initial bone marrow biopsy
Albumin	Albumin at the time of diagnosis
B2M	Beta-2 microglobulin at the time of diagnosis
Calcium	Calcium at the time of diagnosis
Creatinine	Creatinine at the time of diagnosis
LightChainRatio	Involved/uninvolved serum free light chain ratio at the time of diagnosis
Hematocrit	Hematocrit at the time of diagnosis
LDH	Lactate dehydrogenase at the time of diagnosis
MProtein	Monoclonal protein at the time of diagnosis

Chapter 2

Descriptive statistics

2.1 Measures of location

2.1.1 Arithmetic mean

The arithmetic mean ($\bar{x}$) is a measure of central location. It is calculated from the sum of all the observations (n) divided by the number of observations:

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

The notation $\sum_{i=1}^{n} x_i$ means the sum of all x_i observations $(x_1 + x_2 + x_n)$. One limitation to the arithmetic mean is that it is overly sensitive to extreme values.

```r
# Import dataset
load("docs/Example-data.Rda")

# Calcuate arithmetic mean
mean(data$Age)
```

```
[1] 59.5
```

2.1.2 Median

If all observations are ordered from smallest to largest, the median is the middle number. More precisely, if n is odd, $\frac{n+1}{2}$, or if n is even, the average of $\frac{n}{2}$ and $\frac{n}{2}+1$. The rationale for using to the median is to ensure an equal number of observations on both sides of the sample median. The main weakness of the sample median is that it is less sensitive to the actual numeric values of the data points. If the sample distribution is symmetric, the arithmetic mean is approximately the same as the median. For positively skewed distributions, the arithmetic mean tends to be larger than the median; for negatively skewed distributions, the arithmetic means tends to be smaller than the median.

```r
# Calcuate arithmetic mean
median(data$Age)
```

```
[1] 62.5
```

2.1.3 Mode

The mode is the most frequently occurring value among all of the observations in a sample. Some distributions have more than one mode. A distribution with one mode is called unimodal; two modes, bimodal; three modes, trimodal.

```
# Calcuate mode
library(DescTools)
Mode(data$Age)
```

```
[1] 64
```

2.1.4 Geometric mean

The geometric mean ($lo\bar{g}x$) is the central number in a geometric progression such as exponential growth. The geometric mean is defined as the nth root of the product of n numbers:

$$lo\bar{g}x = \frac{\sum_{i=1}^{n} logx_i}{n}$$

Any base can be used to compute the logarithms for the geometric mean. It is usually preferable to work in the original scale by taking the antilogarithm of $lo\bar{g}x$ to form the geometric mean.

```
# Calcuate geometric arithmetic mean
library(DescTools)
Gmean(data$Age)
```

```
[1] 58.61499
```

2.2 Measures of spread

2.2.1 Range

The range is the difference between the smallest and largest observations in a sample. Range is very sensitive to extreme observations and depends on the sample size since the large the n, the largest the range tends to be.

```
# Calcuate range
range(data$Age)
```

```
[1] 33 76
```

2.2.2 Percentile

Percentile is the value below which a given percentage of observations in a group of observations fall. The median is the 50th percentile and is a special case of a quantile. Compared to range, percentiles are less sensitive to outliers. The pth percentile is defined by the $(k+1)$th largest sample observation if $np/100$ is not an integer where k is the largest integer less than $np/100$. If $np/100$ is an integer, then the pth percentile is the average of the $(np/100)$th and the $(np/100+1)$th largest observations.

```
# Calcuate the 5th and 95th percentiles
quantile(data$Age, probs = c(0.05, 0.95), type = 2)
```

```
  5%  95%
38.5 74.0
```

2.2.3 Variance

Variance (s^2) is the average of the squared differences from the mean. The reason for squaring the differences is because the sum of the differences are always equal to zero.

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

```r
# Calcuate variance
var(data$Age)
```

```
[1] 95.31579
```

2.2.4 Standard deviation

Standard deviation (s) is the square root of the sample variance. The advantage of using the standard deviation over the variance is that both the mean and standard deviation are in the same units whereas the variance and mean are not.

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

```r
# Calcuate standard deviation
sd(data$Age)
```

```
[1] 9.762981
```

2.2.5 Standard error of the mean

The standard error (se) of the mean is how far the sample mean deviates from the population mean. It is equal to the variance obtained from a set of sample means from repeated samples of size n from a population with underlying variance s. The standard error of the mean is estimated by $\frac{s}{\sqrt{n}}$. The larger the sample size, the more precise the sample mean will estimate the population mean.

```r
# Calcuate standard error
library(DescTools)
MeanSE(data$Age)
```

```
[1] 2.183069
```

2.2.6 Coefficient of varaince

The coefficient of variation (CV) is a measure of relative variability and remains the same regardless of the observations units. CV is defined by $100\% \times s/\bar{x}$. Coefficient of variation is most useful in comparing the variability of several different samples, each with different arithmetic means.

```r
# Calcuate coefficient of variance
library(DescTools)
CoefVar(data$Age)
```

```
[1] 0.1640837
```

2.3 Visualizing descriptive statistics

2.3.1 Single variable

2.3.1.1 Stem and leaf plot

```
# Create a stem and leaf plot
stem(data$Age)
```

```
  The decimal point is 1 digit(s) to the right of the |

  3 | 3
  4 | 4
  5 | 123578
  6 | 0234444459
  7 | 26
```

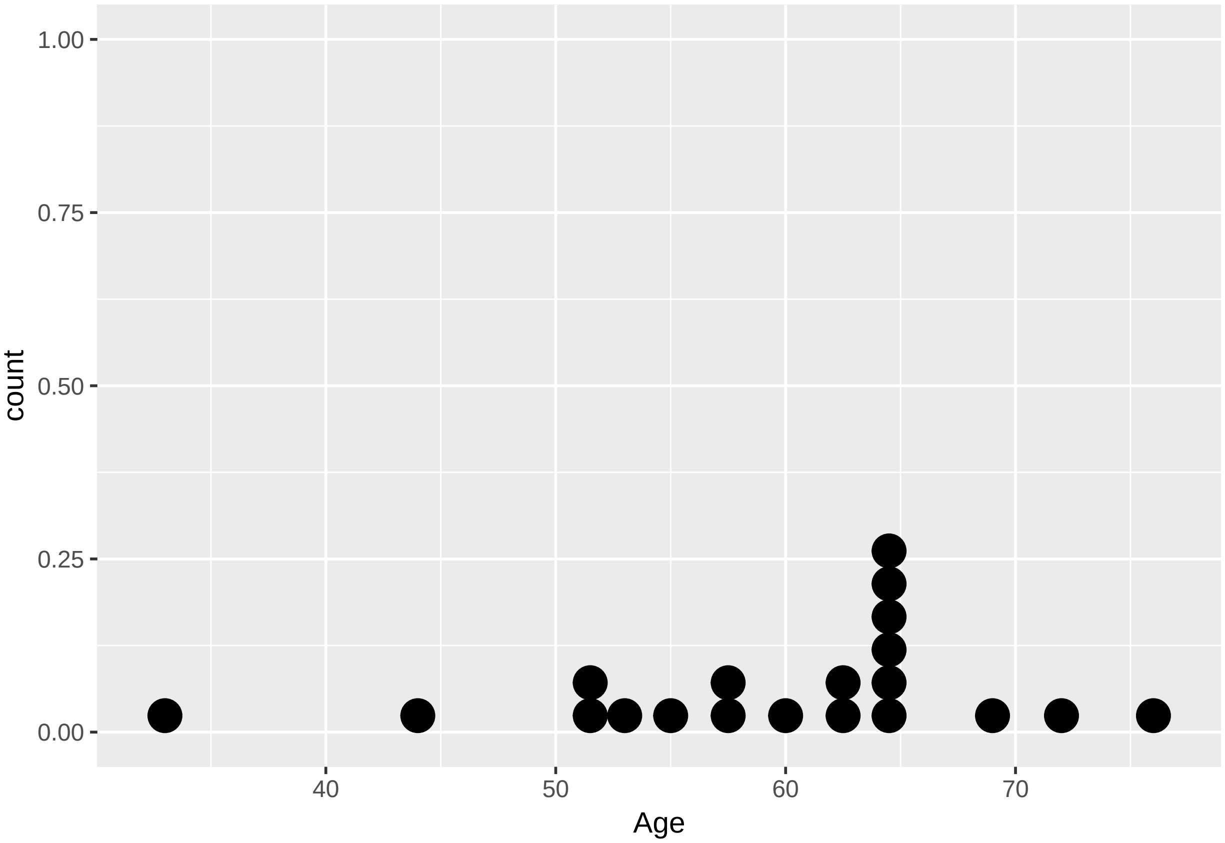
2.3.1.2 Frequency distribution table

```
# Create a frequency distribution table
table(data$Age)
```

```
33 44 51 52 53 55 57 58 60 62 63 64 65 69 72 76
 1  1  1  1  1  1  1  1  1  1  1  5  1  1  1  1
```

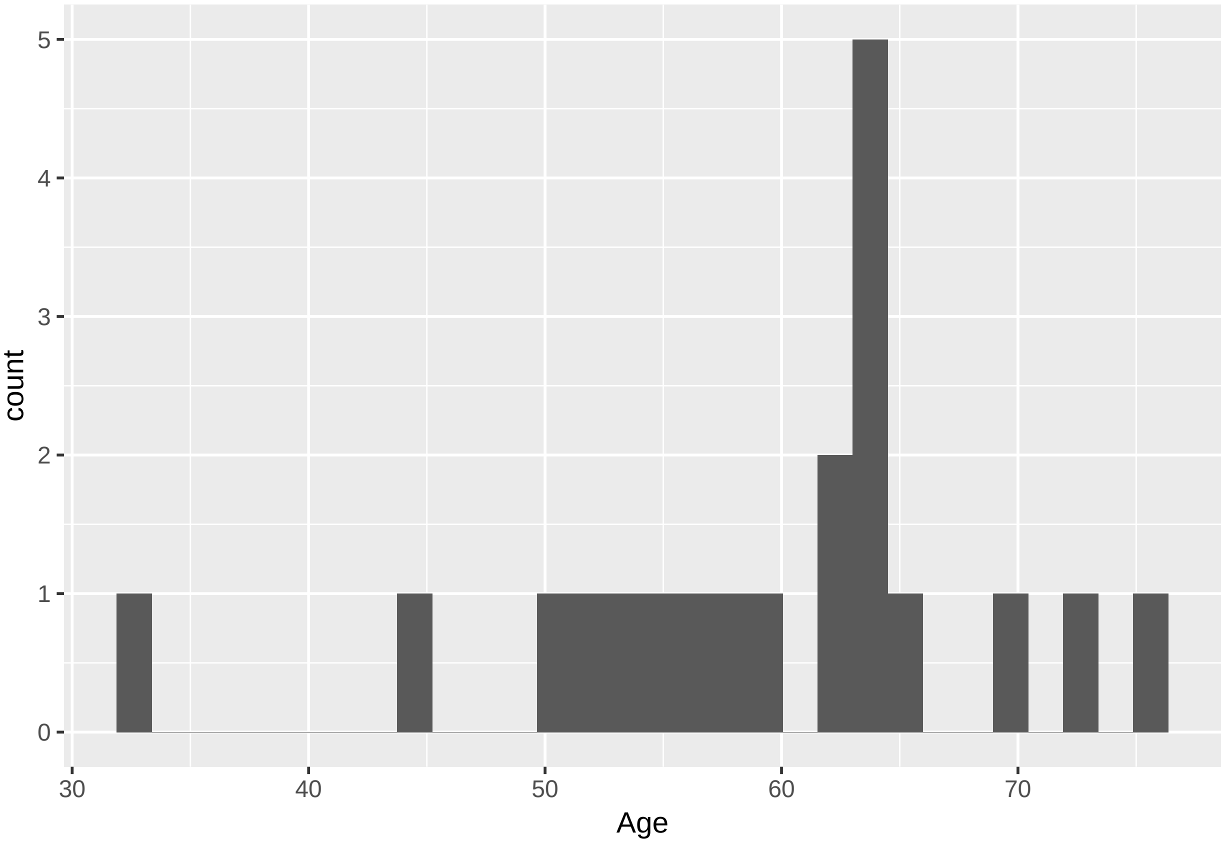
2.3.1.3 Dot plot

```
# Create a dot plot
library(ggplot2)
ggplot(data = data, aes(x = Age)) +
  geom_dotplot()
```



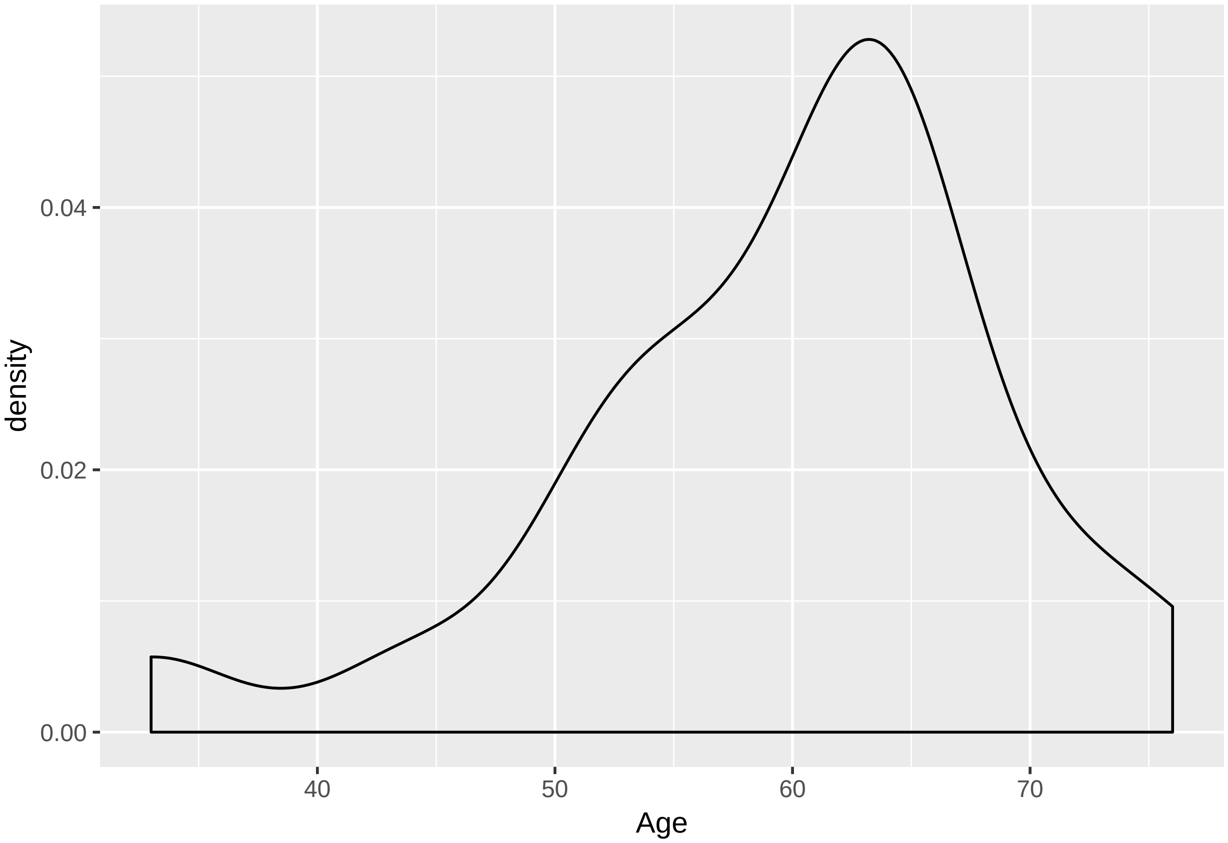
2.3.1.4 Histogram

```
# Create a histogram
library(ggplot2)
ggplot(data = data, aes(x = Age)) +
  geom_histogram()
```



2.3.1.5 Density plot

```
# Create a density plot
library(ggplot2)
ggplot(data = data, aes(x = Age)) +
  geom_density()
```



2.3.2 Muliple variables

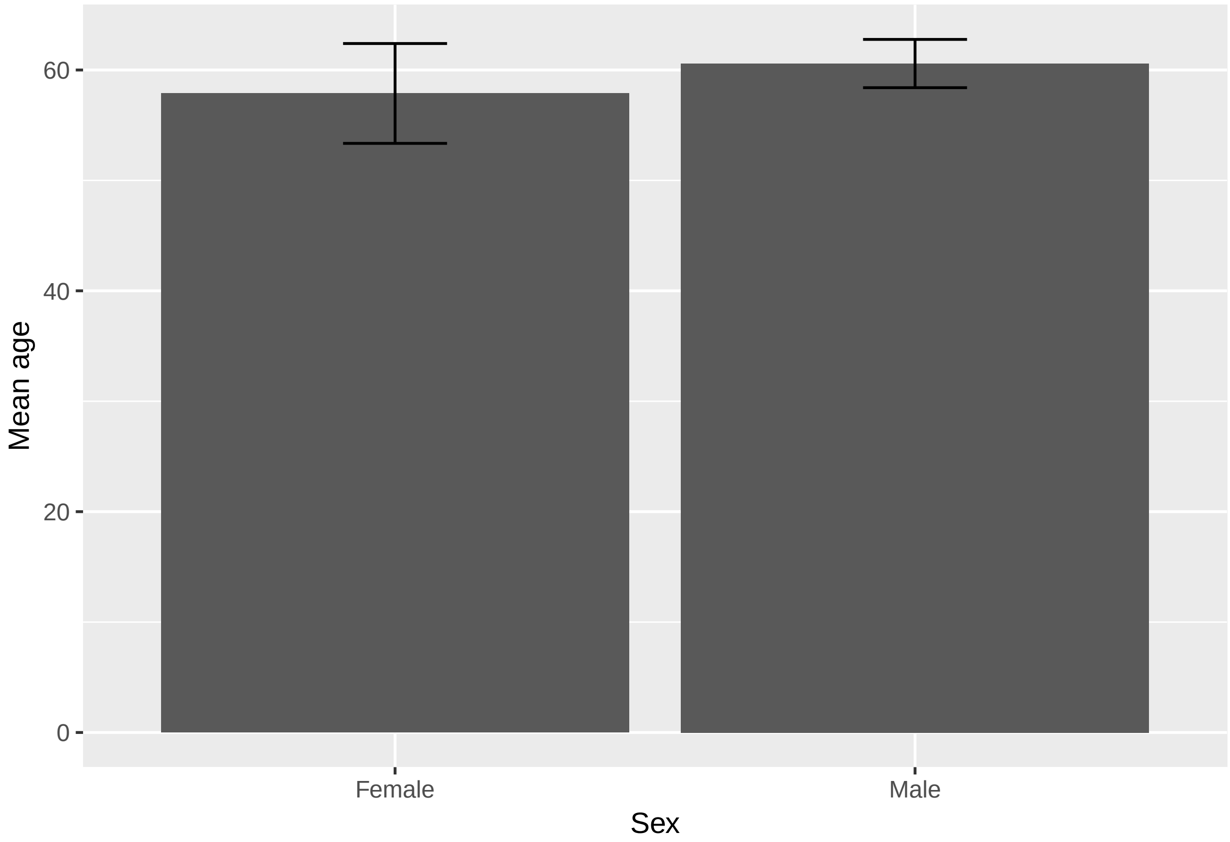
2.3.2.1 Summary table

```
# Create a summary statistics table
library(dplyr)
stats <- group_by(data, Sex) %>%
summarise(Mean = mean(Age), Median = median(Age), SD = sd(Age), SE = MeanSE(Age), N = n())
stats
```

```
# A tibble: 2 x 6
  Sex      Mean Median    SD    SE     N
  <fct>   <dbl>  <dbl> <dbl> <dbl> <int>
1 Female   57.9   63   12.8   4.52     8
2 Male     60.6   60.5  7.57  2.19    12
```

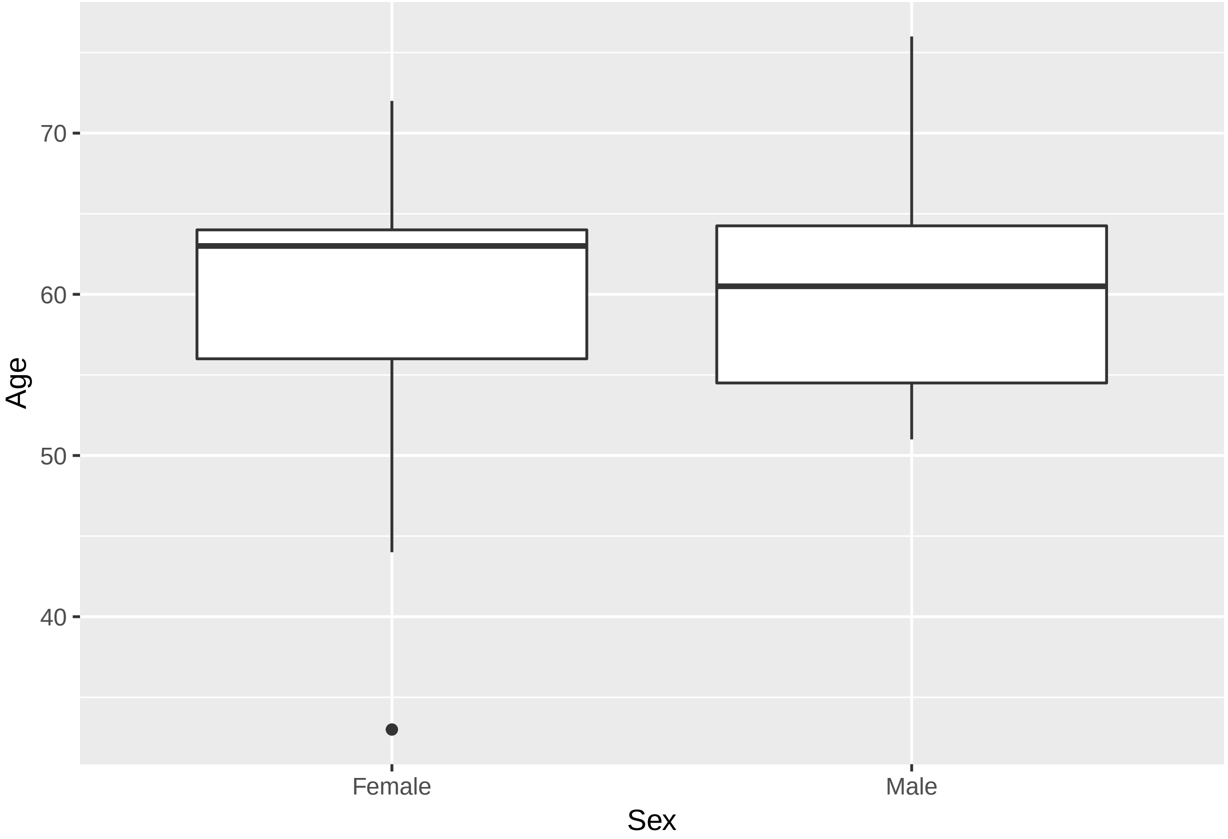
2.3.2.2 Bar graph

```
# Create a bar graph with standard error bars
library(ggplot2)
ggplot(data = stats, aes(x = Sex, y = Mean)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = Mean-SE, ymax = Mean+SE, width = 0.2)) +
  labs(y = "Mean age")
```



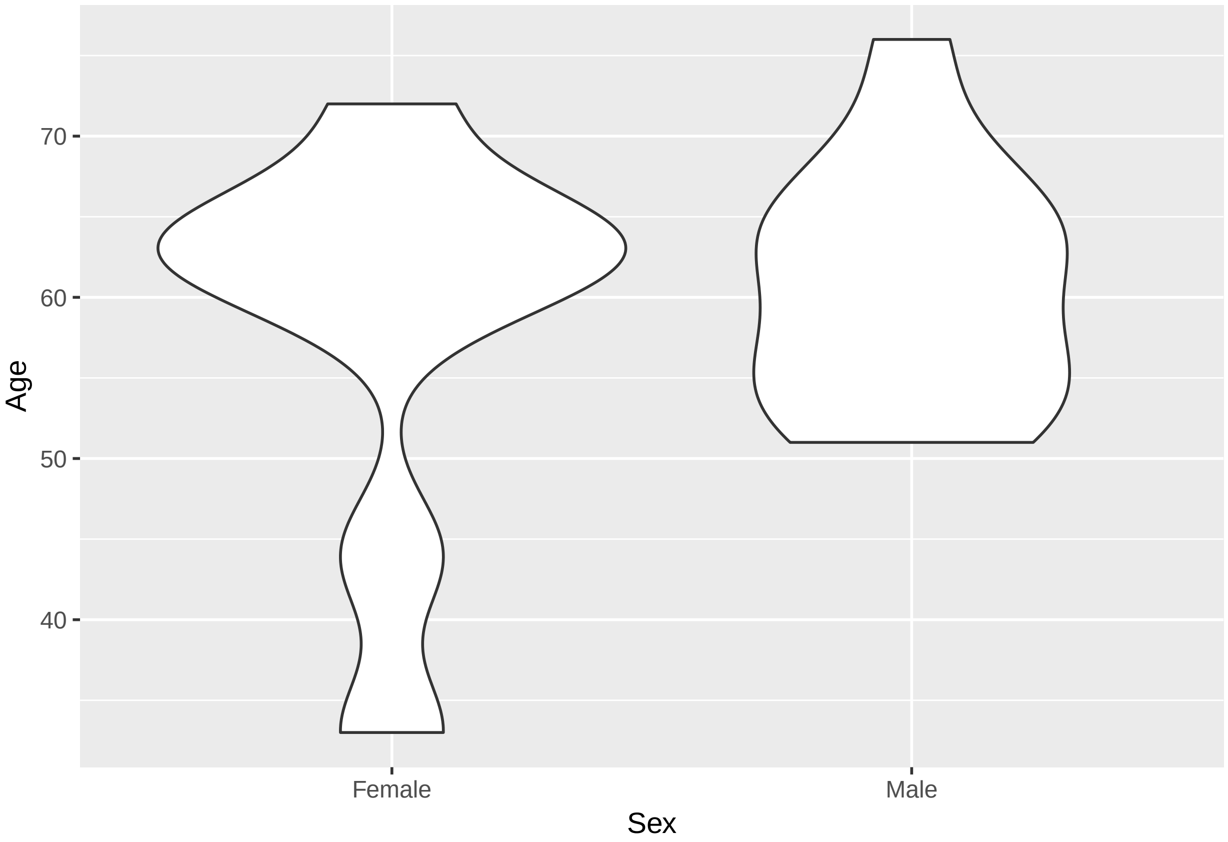
2.3.2.3 Box plot

```
# Create a boxplot
library(ggplot2)
ggplot(data = data, aes(x = Sex, y = Age)) +
  geom_boxplot()
```



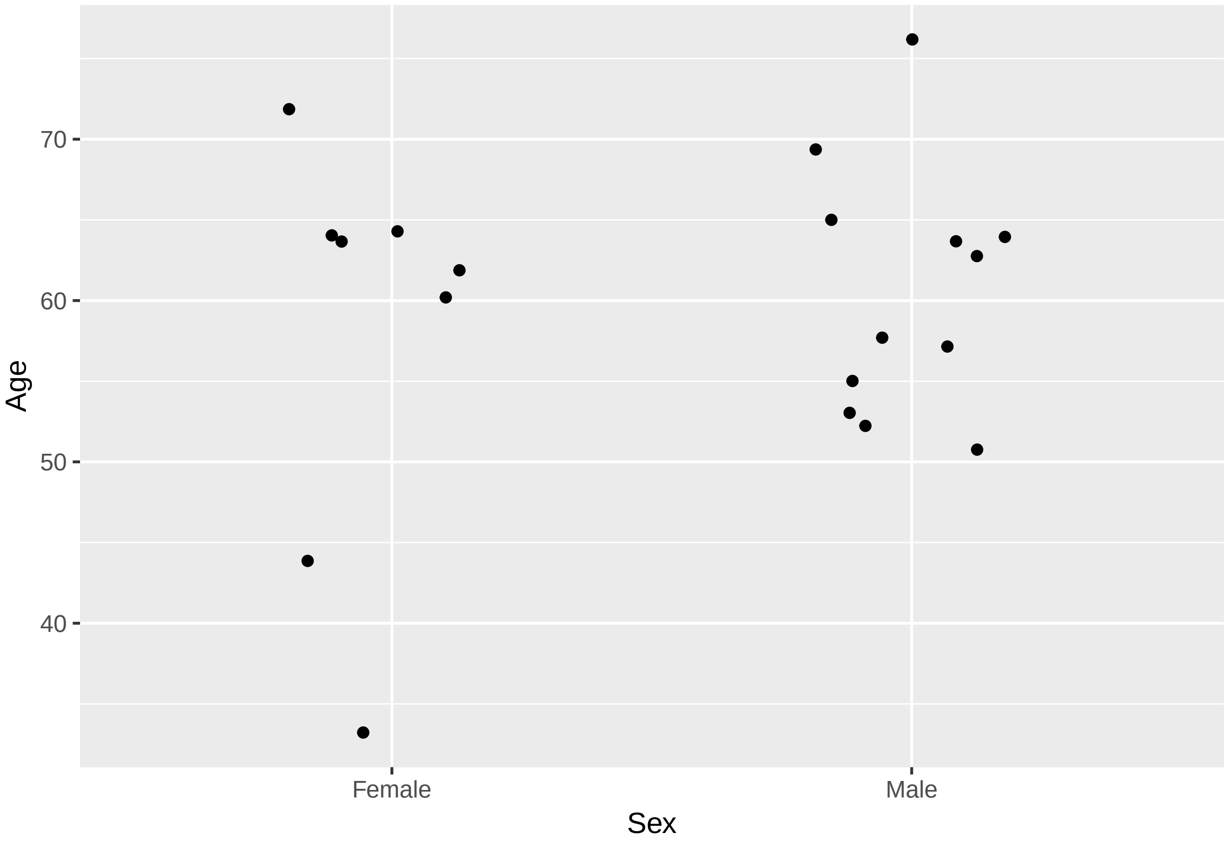
2.3.2.4 Violin plot

```
# Create a violin plot
library(ggplot2)
ggplot(data = data, aes(x = Sex, y = Age)) +
  geom_violin()
```



2.3.2.5 Jitter plot

```
# Create a jitter plot
library(ggplot2)
ggplot(data = data, aes(x = Sex, y = Age)) +
  geom_jitter(width = 0.2)
```


70
60
50
40
Age
Female
Male
Sex

Chapter 3

Probability

3.1 Table of confusion

A table of confusion is a table with 2 rows and 2 columns that reports the number of false positives, false negatives, true positives, and true negatives. Rows represent a prediction outcomes such as a test result and columns represent outcomes of the condition being predicted such as a disease state.

```r
# Import dataset
load("docs/Example-data.Rda")

# Create a table of confusion
BonyLesions <- ifelse(data$BonyLesions != "0" , "Abnormal", "Normal")
MProtein <- ifelse(data$MProtein > 0 , "Abnormal", "Normal")
table(MProtein, BonyLesions)
```

```
          BonyLesions
MProtein   Abnormal Normal
  Abnormal        9      6
  Normal          4      1
```

Sensitivity and specificity to assess the quality of a test. Positive and negative predictive values to interpret the results of the test.

3.2 Sensitivity

Sensitivity is the ability of a test to correctly classify an individual as having a disease. It is defined by the probability of having a positive test when the disease is present:

$$\frac{True\ positives}{True\ positives + False\ negatives} = \frac{True\ positives}{Number\ with\ condition}$$

```r
# Calculate sensitivity
library(caret)
sensitivity(factor(MProtein), factor(BonyLesions)) # sensitivity(prediction,truth)
```

```
[1] 0.6923077
```

3.3 Specificity

Specificity is the ability of a test to correctly classify an individual disease free. It is defined by the probability of having a negative test when the disease absent:

$$\frac{True\ negatives}{True\ negatives + False\ positives} = \frac{True\ negatives}{Number\ without\ condition}$$

```r
# Calculate specificity
library(caret)
specificity(factor(MProtein), factor(BonyLesions)) # specificity(prediction,truth)
```

```
[1] 0.1428571
```

3.4 Postitive predictive value

Positive predictive value is the percentage of patients with a positive test who actually have the disease. It is defined by the probability of having the disease when the test is positive:

$$\frac{True\ positives}{True\ positives + False\ positives} = \frac{True\ positives}{Postive\ predictions}$$

```r
# Calculate postitive predictive value
library(caret)
posPredValue(factor(MProtein), factor(BonyLesions)) # posPredValue(prediction,truth)
```

```
[1] 0.6
```

3.5 Negative predictive value

Negative predictive value is the percentage of patients with a negative test who do not have the disease. It is defined by the probability of not having the disease when the test is negative:

$$\frac{True\ negataives}{True\ negatives + False\ negatives} = \frac{True\ positives}{Negative\ predictions}$$

```r
# Calculate negative predictive value
library(caret)
negPredValue(factor(MProtein), factor(BonyLesions)) # negPredValue(prediction,truth)
```

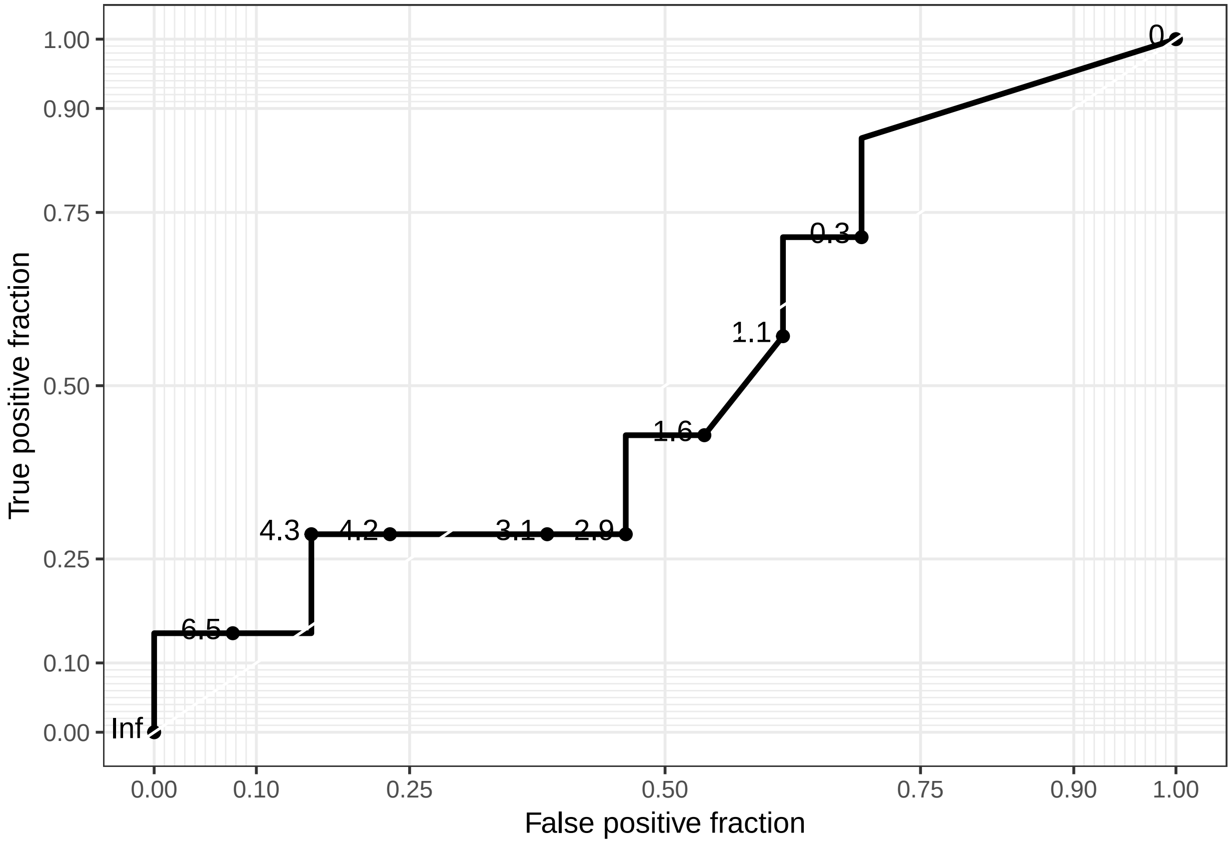
```
[1] 0.2
```

3.6 ROC curve

The receiver operating characteristics (ROC) curve is a plot of the true positive rate versus (sensitivity) versus the false positive rate (specificity) of a screening test. Each point on the curve corresponds to different cutoff points used to designate a positive test. The area under the curve is an estimate of the accuracy of the test.

```r
library(plotROC)
library(ggplot2)

ggplot(data.frame(BonyLesions, MProtein = data$MProtein), aes(d = BonyLesions, m = MProtein)) +
  geom_roc() +
  style_roc()
```



```
# Calculate area under the curve
calc_auc(
  ggplot(data.frame(BonyLesions, MProtein = data$MProtein), aes(d = BonyLesions, m = MProtein)) +
    geom_roc()
  )
```

```
  PANEL group      AUC
1     1    -1 0.521978
```

3.7 Baysian inference

Frequentist inference is a type of statistical inference that draws conclusions from sample data by emphasizing the frequency or proportion of the data. Bayesian inference is an alternative method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. It is given by the following formula:

$$Baye's\ theorem = Pr(A|B) = \frac{Pr(B|A) \times Pr(A)}{Pr(B)}$$

where $Pr(A|B)$ the likelihood of event A occurring given that B is true, $Pr(B|A)$ the likelihood of event B occurring given that A is true, and $Pr(A)$ and $Pr(A)$ are the probabilities of observing A and B independently of each other. Bayesians conceive of two types of probability, a **prior probability** of an event which is the best guess by the observer in the absence of data, and a **posterior probability** which is the likelihood that an event will occur after collecting some empirical data.

Chapter 4

Normal distribution

4.1 Probability distributions

A probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes. Probability distributions are generally divided into two classes. A **discrete probability distribution** applies to discrete random variables such as the probability of being alive or dead as in the binomial distribution or the probability of a given number of deaths occurring in a fixed time interval as in the Poisson distribution. A **continuous probability distribution** applies to continuous random variables such as the measure of hematocrit on any given day. The most common continuous probability distribution is the normal or Gaussian distribution and is defined by the following probability-density function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

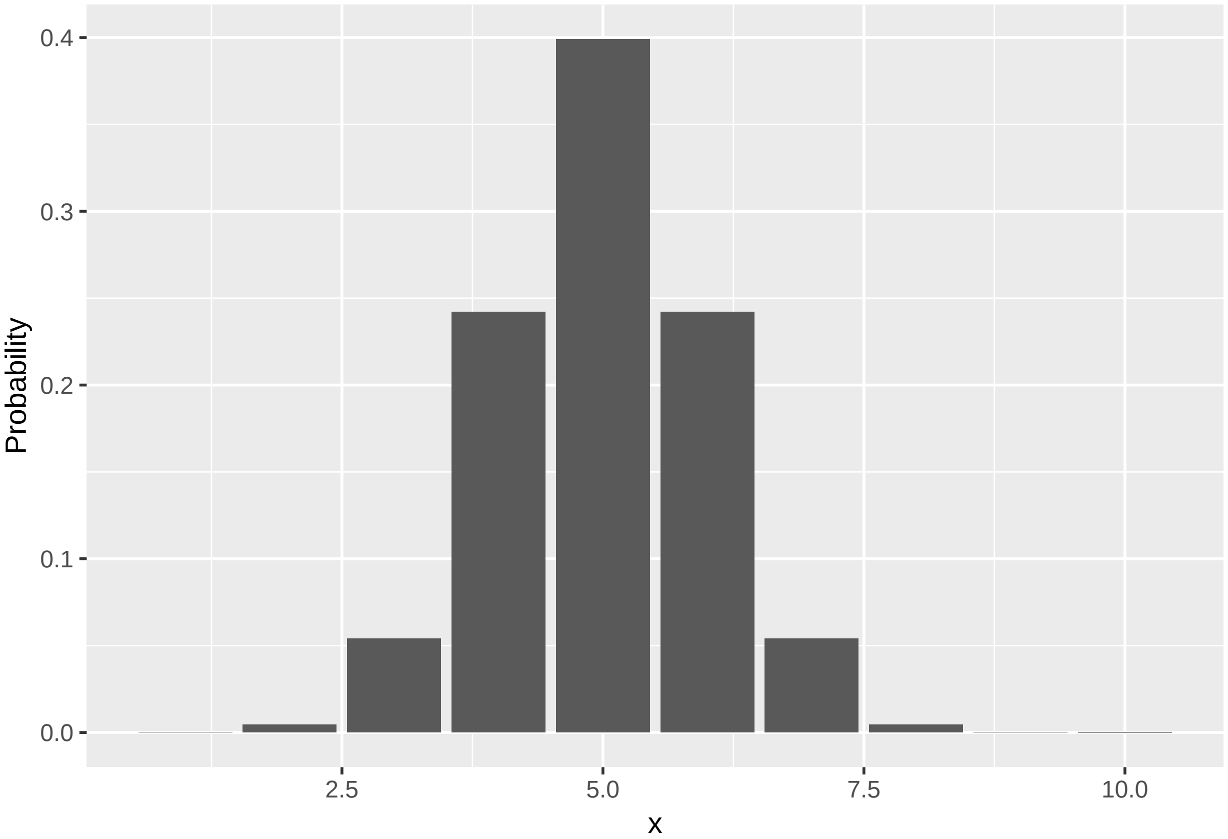
where x is a continuous random variable, σ^2 is the variance, and μ is the mean. Properties of the normal distribution include:

- The curve is symmetric about μ
- The entire shape of the normal distribution is determined by the two parameters μ and σ^2
- The area under the curve between any two points a and b is equal to the probability that the random variable x falls between a and b
- The area under the entire normal density function is always equal to 1
- The probability that the random variable x falls between ±1 standard deviation is approximately 68%, ±2 standard deviations is 95%, and ±2.5 standard deviations is 99%.

```r
# Calculate the probabilities of a normally distributed random variable
library(ggplot2)
p = dnorm(1:10, mean = 5, sd = 1)
p
```

```
 [1] 1.338302e-04 4.431848e-03 5.399097e-02 2.419707e-01 3.989423e-01
 [6] 2.419707e-01 5.399097e-02 4.431848e-03 1.338302e-04 1.486720e-06
```

```r
ggplot(data.frame(Probability = p, x = 1:10), aes(x = x, y = Probability)) +
  geom_bar(stat = "identity")
```



4.2 Central limit theorem

The central limit theorem establishes that the distribution of sample means of an independent random variable tends toward a normal distribution even if the original variables themselves are not normally distributed. For example, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic mean of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the mean will be distributed according to a normal distribution.

4.3 Test for normality

The Shapiro-Wilk test of normality tests if the population is normally distributed. If the p-value is less than a chosen alpha, then the null hypothesis is rejected and there is evidence that the data tested are *not* from a normally distributed population. If the p-value is greater than a chosen alpha, then the data *may* be normally distributed and one cannot reject the hypothesis that the sample comes from a population which has a normal distribution.

```r
# Import dataset
load("docs/Example-data.Rda")

# Shapiro-Wilk test of normality
shapiro.test(data$Age)
```

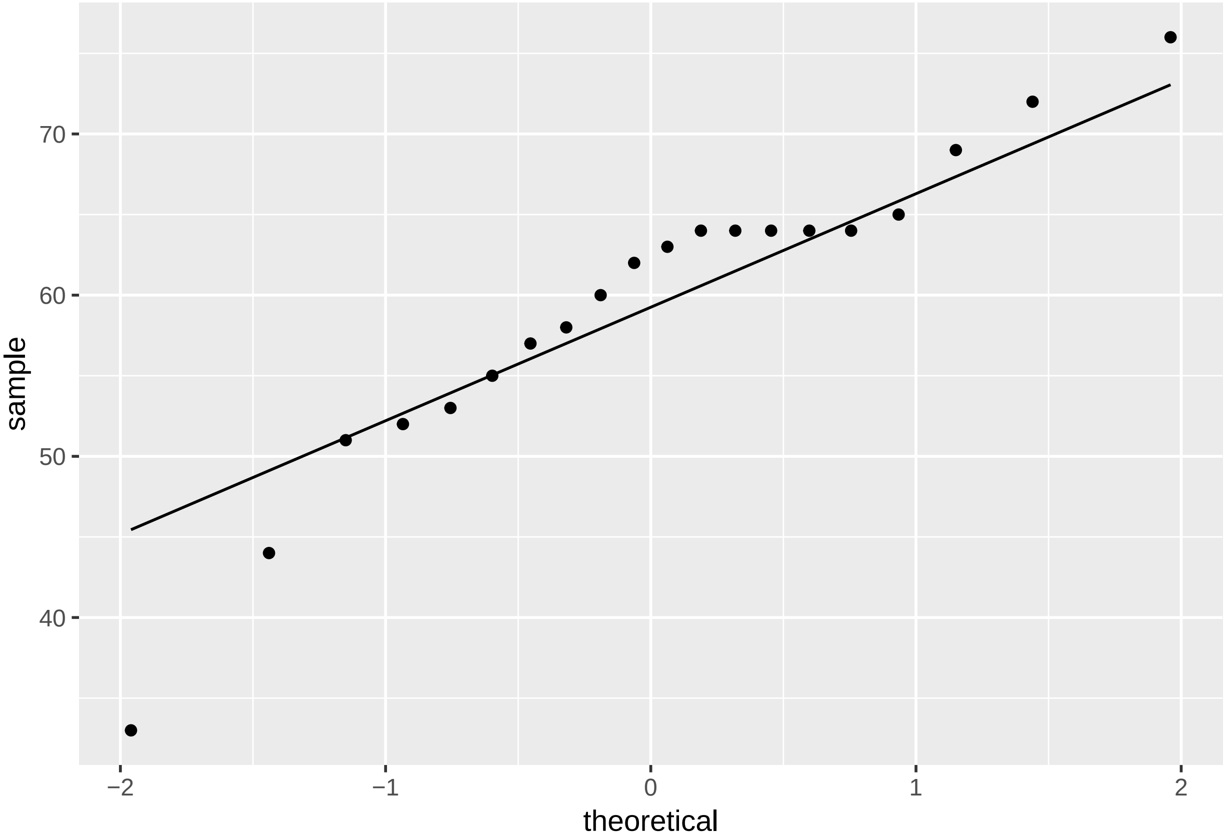
```
    Shapiro-Wilk normality test

data:  data$Age
W = 0.93058, p-value = 0.1584
```

4.4 Quantile-quantile plot

Quantile-quantile (Q-Q) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

```
# Q-Q plot
library(ggplot2)
ggplot(data, aes(sample = Age)) +
  stat_qq() +
  stat_qq_line()
```



4.5 Confidence interval

The interval of plausible estimates of the sample mean is known as the confidence interval. Generally, the interval encompasses 95% of all such sample means. It is calculated from Student's t distribution which is a family of distributions indexed by a parameter referred to as the degrees of freedom $(n-1)$. The upper and lower bounds of the interval are calculated as follows when the standard deviation of the population is unknown:

$$(\bar{x} - t_{n-1,1-\alpha/2}s/\sqrt{n}, \bar{x} + t_{n-1,1-\alpha/2}s/\sqrt{n})$$

where $t_{n-1,1-\alpha/2}$ is the the percentage points of the t distribution for a given degree of freedom $(n-1)$ and percentile $(1-\alpha/2)$. The length of the confidence interval is therefore proportional to the sample size (n), standard deviation (s), and confidence (α):

- As the sample size increases, the length of the confidence interval decreases
- As the standard deviation increases, the length of the confidence interval increases
- As the confidence desired increases, the length of the confidence interval increases

```r
# Determine the confidence interval of the mean from a t distriubtion
alpha = 0.05
mean = mean(data$Age)
sd = sd(data$Age)
n = length(data$Age)
CI = qt(p = 1-alpha/2, df = n-1)*sd/sqrt(n)
c(mean - CI, mean + CI)
```

```
[1] 54.93078 64.06922
```

Bibliography

Rosner, B. (2016). *Fundamentals of Biostatistics.* Cengage Learning, Boston, Massachusetts, 8th edition. ISBN 978-1-305-26892.