Biostatistics

David Coffey

2018-06-29

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 (Table 1).

ID

Sex

Race

Age

Stage

SurvivalMonths

Status

DiagnosisYear

Treatment

TreatmentDurationMonths

BonyLesions

PlasmaCells

1q+

del13q

del17p

del1p

t(11;14)

t(14;16)

t(4:14)

t(6;14)

Albumin

B2M

Calcium

Creatinine

LightChainRatio

Hematocrit

LDH

MProtein

1

Male

Black

65

II

20.60

Unknown

2016

VRD

78

1

80.000

Normal

Normal

Normal

Normal

Normal

Normal

Abnormal

Normal

3.1

4.7

9.6

1.02

135.42

28

208

4.90

2

Female

White

44

I

16.23

Dead

2015

VRD

79

>3

20.000

Normal

Normal

Abnormal

Abnormal

Normal

Normal

Normal

Normal

4.8

1.5

9.6

0.72

3700.00

34

183

0.00

3

Male

Black

55

II

22.63

Alive

2016

VRD

173

>3

30.000

Normal

Normal

Normal

Normal

Normal

Normal

Normal

Normal

3.4

4.3

9.1

1.28

5.82

33

127

4.20

4

Female

White

64

I

22.63

Alive

2016

VRD

184

0

54.000

Normal

Normal

Normal

Normal

Normal

Normal

Normal

Normal

4.2

2.6

10.1

0.71

4.77

31

190

1.70

5

Female

White

62

III

21.30

Alive

2016

VRD

93

1

0.028

Normal

Normal

Normal

Abnormal

Normal

Normal

Normal

Normal

4.6

6.7

10.0

0.90

113.38

32

243

0.30

6

Male

White

64

III

17.57

Alive

2016

CyBorD

21

0

17.800

Normal

Abnormal

Normal

Normal

Normal

Normal

Normal

Normal

2.1

17.0

13.0

3.84

2105.97

20

100

7.30

7

Female

White

60

II

35.43

Unknown

2012

Not specified

528

0

0.000

Normal

Normal

Normal

Normal

Abnormal

Normal

Normal

Normal

4.3

4.4

11.4

1.01

15575.00

35

188

0.00

8

Male

White

58

II

27.83

Alive

2016

VRD

76

>3

5.000

Normal

Abnormal

Normal

Normal

Abnormal

Normal

Normal

Normal

4.1

2.3

9.3

0.93

44.80

44

97

1.10

9

Male

White

69

I

37.70

Alive

2015

Not specified

27

>3

9.600

Normal

Abnormal

Normal

Normal

Normal

Normal

Normal

Normal

4.5

2.3

9.2

0.91

82.20

40

205

0.00

10

Male

White

51

III

31.83

Alive

2015

CyBorD

192

0

43.000

Normal

Normal

Normal

Normal

Abnormal

Normal

Normal

Normal

4.3

8.6

16.0

3.90

148.15

37

253

0.70

11

Female

White

33

II

58.10

Unknown

2011

VRD

73

>3

0.000

Normal

Normal

Normal

Normal

Normal

Normal

Normal

Normal

2.6

2.4

8.6

0.79

1.94

31

180

6.50

12

Male

White

57

I

22.63

Unknown

2015

Not specified

24

>3

65.000

Normal

Abnormal

Normal

Normal

Normal

Normal

Normal

Normal

4.2

2.9

9.7

0.97

464.86

38

122

1.60

13

Female

Black

72

I

10.23

Unknown

2012

Not specified

55

1

21.000

Normal

Normal

Normal

Normal

Normal

Abnormal

Normal

Normal

4.0

5.5

9.9

1.10

1030.00

31

1551

0.00

14

Female

White

64

II

43.57

Alive

2014

Not specified

1979

>3

68.000

Normal

Normal

Normal

Normal

Normal

Normal

Normal

Normal

2.9

3.6

11.4

0.57

1860.00

32

89

2.90

15

Male

Not reported

63

III

52.77

Alive

2014

VRD

1016

0

38.000

Normal

Normal

Normal

Normal

Normal

Normal

Abnormal

Normal

2.6

5.7

9.2

1.07

75.82

30

104

4.30

16

Male

Asian

52

III

27.23

Alive

2016

VRD

162

0

55.000

Normal

Abnormal

Abnormal

Normal

Normal

Normal

Abnormal

Normal

3.7

3.9

9.2

1.08

741.18

22

172

0.25

17

Male

White

76

II

22.37

Dead

2014

Not specified

88

>3

25.000

Normal

Normal

Normal

Normal

Normal

Normal

Normal

None

3.0

5.3

9.9

1.27

21.05

34

130

3.90

18

Female

White

64

III

33.87

Unknown

2014

VRD

212

>3

0.000

Abnormal

Normal

Normal

Abnormal

Normal

Normal

Normal

Normal

4.4

1.9

10.0

0.83

1.93

43

241

0.00

19

Male

White

53

I

27.73

Unknown

2014

VRD

78

1

0.000

Normal

Abnormal

Normal

Normal

Normal

Normal

Normal

Normal

3.7

1.7

9.6

0.86

3.44

40

139

3.10

20

Male

White

64

II

84.70

Dead

2009

CyBorD

60

0

50.000

Normal

Normal

Normal

Normal

Normal

Normal

Normal

Normal

3.8

2.1

10.1

1.10

97.46

35

192

1.10

Chapter 2

Descriptive statistics

2.1 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.

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

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

```
[1] 59.5
```

2.2 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.

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

```
[1] 62.5
```

2.3 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
```

2.4 Geometric mean

The geometric mean ($l\bar{o}gx$) 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:

$$l\bar{o}gx = \frac{\sum_{i=1}^{n} logx_i}{n}$$

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 $l\bar{o}gx$ to form the geometric mean.

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

```
[1] 58.61499
```

Bibliography

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