Biostatistics

David Coffey

2018-07-02

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 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.

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

```
[1] 64
```

2.4 Geometric mean

The geometric mean ($\bar{logx}$) 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:

$$\bar{logx} = \frac{\sum_{i=1}^{n} logx_i}{n}$$

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 $\bar{logx}$ to form the geometric mean.

```r
# 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.