Introduction to biostatistics and machine learning

Olga Dethlefsen, Eva Freyhult, Bengt Sennblad, Payam Emami

2020-10-02

Contents

Preface

This "bookdown" book contains teaching and learning materials prepared and used during "Introduction to biostatistics and machine learning" course organised by NBIS, National Bioinformatics Infrastructure Sweden. The course is open for PhD students, postdoctoral researcher and other employees in need of biostatistical skills within Swedish universities. The course is geared towards life scientists wanting to be able to understand and use basic statistical and machine learning methods. It also suits those already applying biostatistical methods but who have never gotten a chance to reflect on or truly grasp the basic statistical concepts, such as the commonly misinterpreted p-value.

More about the course https://nbisweden.github.io/workshop-mlbiostatistics/

Part I

Preliminary Mathematics

Chapter 1

Mathematical notations

Aims

- to recapitulate the basic notations and conventions used in mathematics and statistics

Learning outcomes

- to recognize natural numbers, integrals and real numbers
- to understand the differences between variables and constants
- to use symbols, especially Sigma and product notations, to represent mathematical operations

1.1 Numbers

- **Natural numbers, N**: numbers such as 0, 1, 3, ...
- **Integers, Z**: include negative numbers ..., -2, -1, 0, 1, 2
- **Rational numberes**: numbers that can be expressed as a ratio two integers, i.e. in a form $\frac{a}{b}$, where a and b are integers, and $b \neq 0$
- **Real numbers, R**: include both rational and irrational numbers
- **Reciprocal** of any number is found by diving 1 by the number, e.g. reciprocal of 5 is $\frac{1}{5}$
- **Absolute value** of a number can be viewed as its distance from zero, e.g. the absolute value of 6 is 6, written as $|6| = 6$ and absolute value of -5 is 5, written as $|-5| = 5$
- **Factorial** of a non-negative integer number n is denoted by $n!$ and it is a product of all positive integers less than or equal to n, e.g. $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$

1.2 Variables, constants and letters

Mathematics gives us a precise language to communicate different concepts and ideas. To be able to use it it is essential to learn symbols and understand how they are used to represent physical quantities as well as understand the rules and conventions that have been developed to manipulate them.

- **variables**: things that can vary, e.g. temperature and time
- **constants**: fixed and unchanging quantities used in certain calculations, e.g. 3.14159
- in principle one could freely choose letters and symbols to represent variables and constants, but it is helpful and choose letters and symbols that have meaning in a particular context. Hence, we
- x, y, z, the end of the alphabet is reserved for variables
- a, b, c, the beginning of the alphabet is used to represent constants
- π, ω and Greek letters below are used frequently used to represent common constant, e.g. $\pi = 3.14159$

Table 1.1: Uppercase and lowercase letters of the Greek alphabet

Letter	Upper case	Lower case	Letter	Upper case	Lower case
Alpha	A	α	Nu	N	ν
Beta	B	β	Xi	Ξ	ξ
Gamma	Γ	γ	Omicron	O	o
Delta	Δ	δ	Pi	Π	π
Epsilon	E	ϵ	Rho	P	ρ
Zeta	Z	ζ	Sigma	Σ	σ
Eta	H	η	Tau	T	τ
Theta	Θ	θ	Upsilon	Y	υ
Iota	i	ι	Phi	Φ	ϕ
Kappa	K	κ	Chi	Γ	γ
Lambda	Γ	γ	Psi	Ψ	ψ
Mu	M	μ	Omega	Ω	ω

1.3 A precise language

- Mathematics is a precise language meaning that a special attention has to be paid to the exact position of any symbol in relation to other.
- Given two symbols x and y, xy and x^y and x_y can mean different things
- xy stands for multiplication, x^y for superscript and x_y for subscript

1.4 Using symbols

If the letters x and y represent two numbers, then:

- their **sum** is written as $x + y$
- subtracting y from x is $x - y$, known also as **difference**
- to multiply x and y we written as $x \cdot y$ or also with the multiplication signed omitted as xy. The quantity is known as **product of x and y**
- multiplication is **associative**, e.g. when we multiply three numbers together, $x \cdot y \cdot z$, the order of multiplication does not matter, so $x \cdot y \cdot z$ is the same as $z \cdot x \cdot y$ or $ycdotz \cdot x$
- division is denoted by $\frac{x}{y}$ and mans that x is divided by y. In this expression x, on the top, is called **numerator** and y, on the bottom, is called **denominator**
- division by 1 leaves any number unchanged, e.g. $\frac{x}{1} = x$ and division by 0 is not allowed

Equal sign

- the equal sign = is used in **equations**, e.g. $x - 5 = 0$ or $5x = 1$
- the equal sign = can be also used in **formulae**. Physical quantities are related through a formula in many fields, e.g. the formula $A = \pi r^2$ relates circle area A to its radius r and the formula $s = \frac{d}{t}$ defines speed as distance d divided by time t
- the equal sign = is also used in identities, expressions true for all values of the variable, e.g. $(x - 1)(x - 1) = (x^2 - 1)$
- opposite to the equal sign is "is not equal to" sign $\neq$, e.g. we can write $1 + 2 \neq 4$

Sigma and product notation

- the Σ notation, read as **Sigma notation**, provides a convenient way of writing longs sums, e.g. the sum of $x_1 + x_2 + x_3 + ... + x_{20}$ is written as $\sum_{i=1}^{i=20} x_i$
- the Π notation, read as **product notation**, provides a convenient way of writing longs products, e.g. $x_1 \cdot x_2 \cdot x_3 \cdot ... \cdot x_{20}$ is written as $\prod_{i=1}^{i=20} x_i$

1.5 Inequalities

Given any two real numbers a and b there are three mutually exclusive possibilities:

- $a > b$, meaning that a is greater than b
- $a < b$, meaning that a is less than b
- $a = b$, meaning that a is equal to b

Strict and weak

- inequality in $a > b$ and $a < b$ is **strict**
- as oppose to **weak** inequality denoted as $a \geq b$ or $a \leq b$

Some useful relations are:

- if $a > b$ and $b > c$ then $a > c$
- if $a > b$ then $a + c > b$ for any c
- if $a > b$ then $ac > bc$ for any positive c
- if $a > b$ then $ac < bc$ for any negative c

1.6 Indices and powers

- **Indices**, also known as **powers** are convenient when we multiply a number by itself several times
- e.g. $5 \cdot 5 \cdot 5$ is written as 5^3 and $4 \cdot 4 \cdot 4 \cdot 4 \cdot 4$ is written as 4^5
- in the expression x^y, x is called the *base* and y is called *index* or *power*

The laws of indices state:

- $a^m \cdot a^n = a^{m+n}$
- $\frac{a^m}{a^n} = a^{m-n}$
- $(a^m)^n = a^{m \cdot n}$

Rules derived from the laws of indices:

- $a^0 = 1$
- $a^1 = a$

Negative and fractional indices:

- $a^{-m} = \frac{1}{a^m}$ e.g. $5^{-2} = \frac{1}{5^2} = \frac{1}{25}$ for negative indices
- e.g. $4^{\frac{1}{2}} = \sqrt{4}$ or $8^{\frac{1}{3}} = \sqrt[3]{8}$ for fractional indices

1.7 Excercises

Exercise 1.1. Classify numbers as natural, integers or real. If reall, specify if they are rational or irrational.

a) $\frac{1}{3}$
b) 2
c) $\sqrt{4}$
d) 2.3
e) π
f) $\sqrt{5}$
g) -7
h) 0
i) 0.25

Exercise 1.2. Classify below descriptors as variables or constants. Do you know the letters or symbols commonly used to represent these?

a) speed of light in vacuum
b) mass of an apple
c) volume of an apple
d) concentration of vitamin C in an apple
e) distance from Stockholm central station to Uppsala central station
f) time on the train to travel between the above stations
g) electron charge

Exercise 1.3. Write out explicitly what is meant by the following:

a) $\sum_{i=1}^{i=6} k_i$
b) $\prod_{i=1}^{i=6} k_i$
c) $\sum_{i=1}^{i=6} i^k$
d) $\prod_{i=1}^{i=3} i^k$
e) $\sum_{i=1}^{i=4} (i+1)^k$
f) $\prod_{i=1}^{i=4} (k+1)^i$

Answers to Excercises

Exr. 1.1

a) real, rational
b) natural and integers, integers include natural numbers
c) $\sqrt{4} = 2$ so it is a natural number and/subset of integers
d) real number, rational as it could be written as $\frac{23}{10}$
e) real number, irrational as it cannot be explained by a simple fraction
f) real number, irrational as it cannot be explained by a simple fraction
g) integer, non a natural number as these do not include negative numbers
h) nautral number, although there is some argument about it as some define nautral numbers as positive integers starting from 1, 2 etc. while others include 0.
i) real, rational number, could be written as $\frac{25}{100}$

Exr. 1.2

a) constant, speed of light in vacuum is a contant, denoted c with $c = 299792458 \frac{m}{s^2}$
b) variable, mass of an apple is a variable, different for different apple sizes, for instance 138 grams, denoted as $m = 100g$
c) variable, like mass volume can be different from apple to apple, denoted as V, e.g. $V = 200cm^3$

d) variable, like volume and mass can vary, denoted as ρ_i and defined as $\rho_i = \frac{m}{V}$. So given 6.3 milligrams of vitamin C in our example apple, we have $\rho_i = \frac{0.0063g}{200cm^3} = 0.0000315\frac{g}{cm^3}$ concentration of vitamin D

e) constant, the distance between Stockholm and Uppsala is fixed; it could be a variable though if we were to consider an experiment on a very long time scale; distance is often denoted in physics as d

f) variable, time on the train to travel between the stations varies, often denoted as t with speed being calculated as $s = \frac{d}{t}$

g) constant, electron charge is $e = 1.60217663 \cdot 10^{-19}C$

Exr. 1.3

a) $\sum_{i=1}^{i=6} k_i = k_1 + k_2 + k_3 + k_4 + k_5 + k_6$

b) $\prod_{i=1}^{i=6} k_i = k_1 \cdot k_2 \cdot k_3 \cdot k_4 \cdot k_5 \cdot k_6$

c) $\sum_{i=1}^{i=3} i^k = 1^k + 2^k + 3^k$

d) $\prod_{i=1}^{i=3} i^k = 1^k \cdot 2^k \cdot 3^k$

e) $\sum_{i=1}^{i=4} (i+1)^k = (1+1)^k + (2+1)^k + (3+1)^k + (4+1)^k$

Chapter 2

Sets

Part II

Introduction to R, R Studio and R markdown

Chapter 3

R

bla bla bla

Chapter 4

R Studio

bla bla bla

Chapter 5

R Markdown

bla bla bla

Part III

Probability

Chapter 6

Probability: reasoning under uncertainty

Learning outcomes

- understand the concept of probability
- manipulate probabilites by their rules
- assign probabilities in very simple cases

6.1 Introduction

Some things are more likely to occur than others. Compare:

- the chance of the sun rising tomorrow with the chance that no-one is infected with COVID-19 tomorrow
- the chance of a cold dark winter in Stockholm with the chance of no rainy days over the summer months in Stockholm

We intuitively believe that the chance of sun rising or dark winter occurring are enormously higher than COVID-19 disappearing over night or having no rain over the entire summer. **Probability** gives us a scale for measuring the likeliness of events to occur. **Probability rules** enable us to reason about uncertain events. The probability rules are expressed in terms of sets, a well-defined collection of distinct objects.

6.2 Basic set definitions

- **set**: a well-defined collection of distinct objects, e.g. $A = \{2, 4, 6\}$

- **subset,** $\subseteq$: if every element of set A is also in B, then A is said to be a subset of B, written as $A \subseteq B$ and pronounced A is contained in B, e.g. $A \subseteq B$, when $B = \{2, 4, 6, 8, 10\}$. Every set is a subset if itself.
- **empty set,** $\emptyset$: is a unique set with no members, denoted by $E = \emptyset$ or $E = \{\}$. The empty set is a subset of very set.

6.3 Basic set operations

- **union of two sets,** $\cup$: two sets can be "added" together, the union of A and B, written as $A \cup B$, e.g. $\{1, 2\} \cup \{2, 3\} = \{1, 2, 3\}$ or $\{1, 2, 3\} \cup \{1, 4, 5, 6\} = \{1, 2, 3, 4, 5, 6\}$
- **intersection of two sets,** $\cap$: a new set can be constructed by taking members of two sets that are "in common", written as $A \cap B$, e.g. $\{1, 2, 3, 4, 5, 6\} \cap \{2, 3, 7\} = \{2, 3\}$ or $\{1, 2, 3\} \cap \{7\} = \emptyset\}$
- **complement of a set,** At**,** A^C: are the elements not in A
- **difference of two sets,** $\setminus$: two sets can be "substracted", denoted by $A \setminus B$, by taking all elements that are members of A but are not members of B, e.g. $\{1, 2, 3, 4\} \setminus \{1, 3\} = \{2, 4\}$. This is also in other words a relative complement of A with respect to B.
- **partition of a set**: a partition of a set S is a set of nonempty subset of S, such that every element x in S is in exactly one of these subsets. That is, the subset are pairwise *disjoint*, meaning no two sets of the partition cotain elements in common, and the union of all the subset of the partition is S, e.g. Set $\{1, 2, 3\}$ has five partitions: i) $\{1\}, \{2\}, \{3\}$, ii) $\{1, 2\}, \{3\}$, iii) $\{1, 3\}, \{2\}$, iv) $\{1\}, \{2, 3\}$ and v) $\{1, 2, 3\}$

6.4 Exercises

Exercise 6.1. Here is my exercise.

6.5 Answers to exercises

Chapter 7

Probability: random variables

Learning outcomes

- understand the concept of random discrete and continous variables
- to be able to use probability density/mass functions and cumulative distribution functions and to understand the relationship between them
- describe properties of binomial, geometric, Poisson, uniform, exponential and normal distributions and identify which distributions to use in practical problems

7.1 Random variables

The outcome of a random experiment can be described by a random variable.

Example random variables:

- The weight of a random newborn baby
- The smoking status of a random mother
- The hemoglobin concentration in blood
- The number of mutations in a gene
- BMI of a random man
- Weight status of a random man (underweight, normal weight, overweight, obese)
- The result of throwing a die

Whenever chance is involved in the outcome of an experiment the outcome is a random variable.

A random variable is usually denoted by a capital letter, $X, Y, Z, \ldots$. Values collected in an experiment are observations of the random variable, usually denoted by lowercase letters $x, y, z, \ldots$.

A random variable can not be predicted exactly, but the probability of all possible outcomes can be described.

The population is the collection of all possible observations of the random variable. Note, the population is not always countable.

A sample is a subset of the population.

7.2 Discrete random variables

A discrete random variable can be described by its *probability mass function*.

–>

–>

Chapter 8

Summarising and visualising data

Chapter 9

Linear regression

9.1 Simple regression

9.2 Multiple regression