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/

Chapter 1

Preliminary Mathematics

1.1 Mathematical notations, sets, functions, exponents and logarithms

This section contains some basic important notations and conventions used in mathematics and statistics.

1.1.1 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
- **Rreciprocal** 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. $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$

1.1.2 Variables, constants and symbols

Mathematics gives us a precise language to communiate 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 beggining of the alphabet is used to reprsent constants
- π, ω and Greek letteres below are used frequently used to represent common constat, e.g. $\pi = 3.14159$

Table 1.1: 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	Υ	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

I am referring to Table 1.1

ss ## Differentiation ## Integration ## Vectors and Matrices

Chapter 2

Introduction to R, R Studio and R markdown

2.1 R

2.2 R Studio

2.3 R markdown

Chapter 3

Probability: reasoning under uncertainty

Learning outcomes

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

3.1 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.

3.2 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.

3.3 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,** A', 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) $\{1, 2, 3\}$

3.4 Exercises

3.5 Answers to exercises

Chapter 4

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

4.1 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.

4.2 Discrete random variables

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

–>

–>

Chapter 5

Summarising and visualising data

Chapter 6

Linear regression

6.1 Simple regression

6.2 Multiple regression