A short introduction to biostatistics (and R)

Jean-Baka Domelevo Entfellner

BecA-ILRI Hub, Nairobi, Kenya

ABCF Bioinformatics Community of Practice, 30 April 2018

Foreword

Definition (statistics)

Statistics is the branch of **mathematics** enabling scientists to deal with statements including some kind of **randomness** or **uncertainty**.

Underlying maths: **probability theory** (measure theory, integration, special distributions, moment-generating functions...)

Sources of randomness or uncertainty for a scientist:

1. finite *a priori* knowledge
2. limited accuracy of the measuring tools (scales, measuring tapes, human eyes): $\mathrm{dist}(\text{measurement}, \text{true value})$ is a random variable > 0
   $\hookrightarrow$ measurement error, either systematic or random
3. **randomness is inherent to complex interacting systems** (e.g. biological systems)

Uncertainty/randomness is everywhere

- What will be the temperature tomorrow at noon in Hermanus?
- How many kids will you have in your life?
- How many people are living in SA right now?
- How tall is an adult sequoia tree?
- How many mitochondria in an mature human liver cell?
- How many cells in my body?
- How large will a growing colony of bacteria be at $t_0 + 24$h?

Vocabulary: random variables, observations

To deal with randomness, statisticians talk about **random variables**.

Definition (random variable)

A random variable is a mathematical variable whose value originates from some random process. It can be discrete or continuous, and is usually subject to sampling (trials or observations).

The possible values of a r.v. X usually describe $\mathbb{R}$ (*continuous* r.v.) or $\mathbb{N}$ (*discrete* r.v.), or a subset thereof.

Some examples of random variables

Discrete random variables:

1. the number of kids you will have takes its value out of $\mathbb{N} = [\![0, \infty[\![$, but it is very unlikely to be in $[\![20, \infty[\![$;
2. the number of mitochondria in a mature human liver cell also lies in $\mathbb{N}$, but very unlikely outside of $[\![1000, 100000]\!]$.

Continuous random variables:

1. the temperature tomorrow at 12:00 in Hermanus is a continuous random variable taking its values out of $]-273.15, \infty[$, but the bulk of the probability density is certainly on $[10, 40]$ (all measures in $^\circ$C);
2. the height of an adult sequoia tree is somewhere on $\mathbb{R}_+$, but most probably in $[10, 120]$ (in metres).

More vocabulary: samples

- Our first tool to study random variables: **repeated observations**
  $\hookrightarrow$ e.g. recordings of temperatures in a weather station, cell counts in an organism, etc.
- A **sample of size *n*** is a set of n observations of the same random variable.
- A **datapoint** is a single observation taken from a sample.
- A sample is a subset of an **underlying population** (e.g. all cells of all human beings, all sequoia trees, past, present and future, etc). That population is partially unknown or simply too big to be dealt with.

Descriptive & inferential statistics

Descriptive statistics: direct calculations of useful quantities from a sample
$\hookrightarrow$ e.g. sample mean, sample variance, sample frequencies, quantiles, etc.

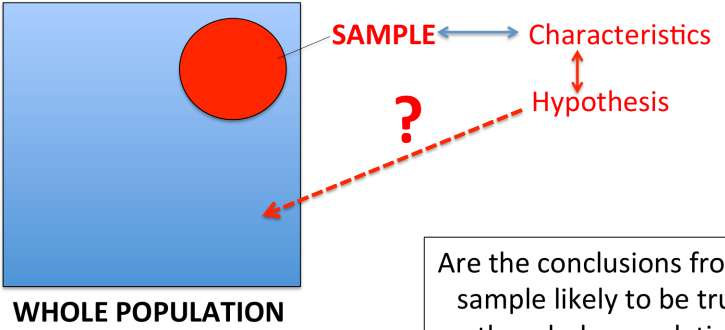
Descriptive & inferential statistics

Descriptive statistics: direct calculations of useful quantities from a sample
$\hookrightarrow$ e.g. sample mean, sample variance, sample frequencies, quantiles, etc.

Inferential statistics: from a sample, try to infer knowledge about its underlying population.

1. get a sample from repeated measurements/observations;
2. manipulate a model of the generative process that gave rise to what you have just measured;
3. test hypotheses, infer statements containing a certain level of uncertainty;
4. transfer those results into knowledge about the underlying population.

Observations on a sample, hypotheses on the whole population



Are the conclusions from the sample likely to be true in the whole population?

(Bio)stats: a set of tools

Biostats are made of a collection of techniques and tools:

- descriptive statistics: **describe a sample** and its properties (range, mean, variance, etc)
- estimation: **estimate parameters** of the underlying distribution, providing **confidence intervals**
- inferential statistics: perform **hypothesis testing** on one or more sample(s)
- correlation studies: measure the **association** between two variables
- analysis of variance (**ANOVA**): to model the sources of variance in a multidimensional dataset
- **regression** analysis: probabilistic modeling of a *response variable* through the use of a set of *predictors*
- techniques for **dimensionality reduction** to represent and analyze high-dimensional datasets

R, that huge toolbox in your computer

R: software to perform statistical analyses. It is:

- multi-platform (GNU/Linux, MacOSX, Windows, etc)
- highly modular, many contributors worldwide (libraries exist for virtually every type of studies/data)
- easy to use (interactive interpreter + integrated development environment, RStudio)
- able to output high-quality graphics
- free software (GNU GPL)

R, that huge toolbox in your computer

R: software to perform statistical analyses. It is:

- multi-platform (GNU/Linux, MacOSX, Windows, etc)
- highly modular, many contributors worldwide (libraries exist for virtually every type of studies/data)
- easy to use (interactive interpreter + integrated development environment, RStudio)
- able to output high-quality graphics
- free software (GNU GPL)

R is no magic: you have to know **what you want to do** and **why** you want to do it before asking R to compute it.

A few R screenshots