STATISTICS FOR BIOLOGY (AND BIOLOGISTS)

Xavier Prudent and Cihan Erkut

(prudent@mpi-cb.de / erkut@mpi-cbg.de)

Presentation and analysis of data are as important as collection of data. Statistics allows us to use our data for testing hypotheses and estimating quantities in a rigorous and quantified way. In recent years, with the help of improving computer technologies, statistical methods became more advanced and easier to apply. This eliminates the need to learn all the mathematics behind the procedures, but one should know very well which method to use and how to interpret its results.

In this 2-day course, our aim is first to introduce you to basic concepts of statistics (e.g. variables, data types, sample vs. population, central tendency and dispersion, distribution, transformation and error). Then, we will cover basic statistical tests to make comparisons between measurements (i.e. hypothesis testing, e.g. t-test, chi-square test, ANOVA and non-parametric tests). We will also introduce you to the key concepts of hypothesis testing (e.g. p-value and false discovery rate).

In the next part of the course, you will learn the basic ideas behind statistical modeling, which is used to make predictions based on your data. We will focus mainly on regression analysis and general linear models. In addition to these concepts, we will discuss a resampling method (i.e. bootstrapping), which is especially useful in cases with limited sample size.

On the second day of the course, you will get accustomed with more advanced techniques. These will start with multivariable tools (e.g. support vector machine, neural networks, boosted decision tree and Fisher discriminant). Then we will introduce a popular technique called principal component analysis, where one simplifies multidimensional data so that it can be analyzed more easily. After that, we will explain the concepts of Monte Carlo and optimal sampling, which lay behind most simulations. And finally, we will introduce Bayesian statistics, its assumptions and drawbacks, together with alternative statistics (e.g. concepts of prior and post probability, exploring the parameter space by Markov chains, etc.).

Because we are not statisticians or mathematicians, we will not bother you with formulas and abstractions. Instead, the course is designed in an intuitive way, so that you will learn the meaning of concepts rather than how they are represented in formulas. Because the course is designed for biologists, we will approach the topics from a biological perspective and give relevant examples.

There will be no limit on class size and everybody is welcome to join. If you have any questions about the content of the course, feel free to contact us.