p20

An\_Introduction\_Statistics\_Python\_c01

Thomas Haslwanter

An Introduction to Statistics with Python

With Applications in the Life Sciences

Part I Python and Statistics

\* statistics based on *Python*

\* This part is a kick-start for *Python*

through -

**Chapter 1**

**Why Statistics?**

Statistics is the explanation of variance in the light of what remains unexplained.

Every day we are confronted with situations with uncertain outcomes, and must make decisions based on incomplete data: “Should I run for the bus? Which stock should I buy? Which man should I marry? Should I take this medication? Should I have my children vaccinated?” Some of these questions are beyond the realm of statistics (“Which person should I marry?”), because they involve too many unknown variables. But in many situations, statistics can help extract maximum knowledge from information given, and clearly spell out what we know and what we don’t know. For example, it can turn a vague statement like “This medication may cause nausea,” or “You could die if you don’t take this medication” into a specific statement like “Three patients in one thousand experience nausea when taking this medication,” or “If you don’t take this medication, there is a 95 % chance that you will die.”

Without statistics, the interpretation of data can quickly become massively flawed. Take, for example, the estimated number of German tanks produced during World War II, also known as the “German Tank Problem.” The estimate of the number of German tanks produced per month from standard intelligence data was *1,550*; however, the statistical estimate based on the number of tanks observed was *327*, which was very close to the actual production number of *342* (http://en. wikipedia.org/wiki/German\_tank\_problem).

Similarly, using the wrong tests can also lead to erroneous results. In general, statistics will help to

* Clarify the question.
* Identify the variable and the measure of that variable that will answer that  
     
  question.
* Determine the required sample size.
* Describe variation.
* Make quantitative statements about estimated parameters.
* Make predictions based on your data.

**Reading the Book** Statistics was originally invented — like so many other things — by the famous mathematician C.F. Gauss, who said about his own work, “Ich habe fleissig sein müssen; wer es gleichfalls ist, wird eben so weit kommen.” (“I had to work hard; if you work hard as well, you, too, will be successful.”). Just as reading a book about playing the piano won’t turn you into a great pianist, simply reading this book will not teach you statistical data analysis. If you don’t have your own data to analyze, you need to do the exercises included. Should you become frustrated or stuck, you can always check the sample Solutions provided at the end of the book.  
   
**Exercises** Solutions to the exercises provided can be found at the end of the book. In my experience, very few people work through large numbers of examples on their own, so I have not included additional exercises in this book.  
   
If the information here is not sufficient, additional material can be found in other statistical textbooks and on the web:  
   
**Books** There are a number of good books on statistics. My favorite is Altman (1999): it does not dwell on computers and modeling, but gives an extremely useful introduction to the field, especially for life sciences and medical applications. Many formulations and examples in this manuscript have been taken from that book. A more modern book, which is more voluminous and, in my opinion, a bit harder to read, is Riffenburgh (2012). Kaplan (2009) provides a simple introduction to modern regression modeling. If you know your basic statistics, a very good introduction to Generalized Linear Models can be found in Dobson and Barnett (2008), which provides a sound, advanced treatment of statistical modeling.  
   
**WWW** In the web, you will find very extensive information on statistics in English at

* + http://www.statsref.com/
  + http://www.vassarstats.net/
  + http://www.biostathandbook.com/
  + http://onlinestatbook.com/2/index.html
  + http://www.itl.nist.gov/div898/handbook/index.htm

A good German web page on statistics and regulatory issues is http://www. reiter1.com/.  
   
I hope to convince you that *Python* provides clear and flexible tools for most of the statistical problems that you will encounter, and that you will enjoy using it.