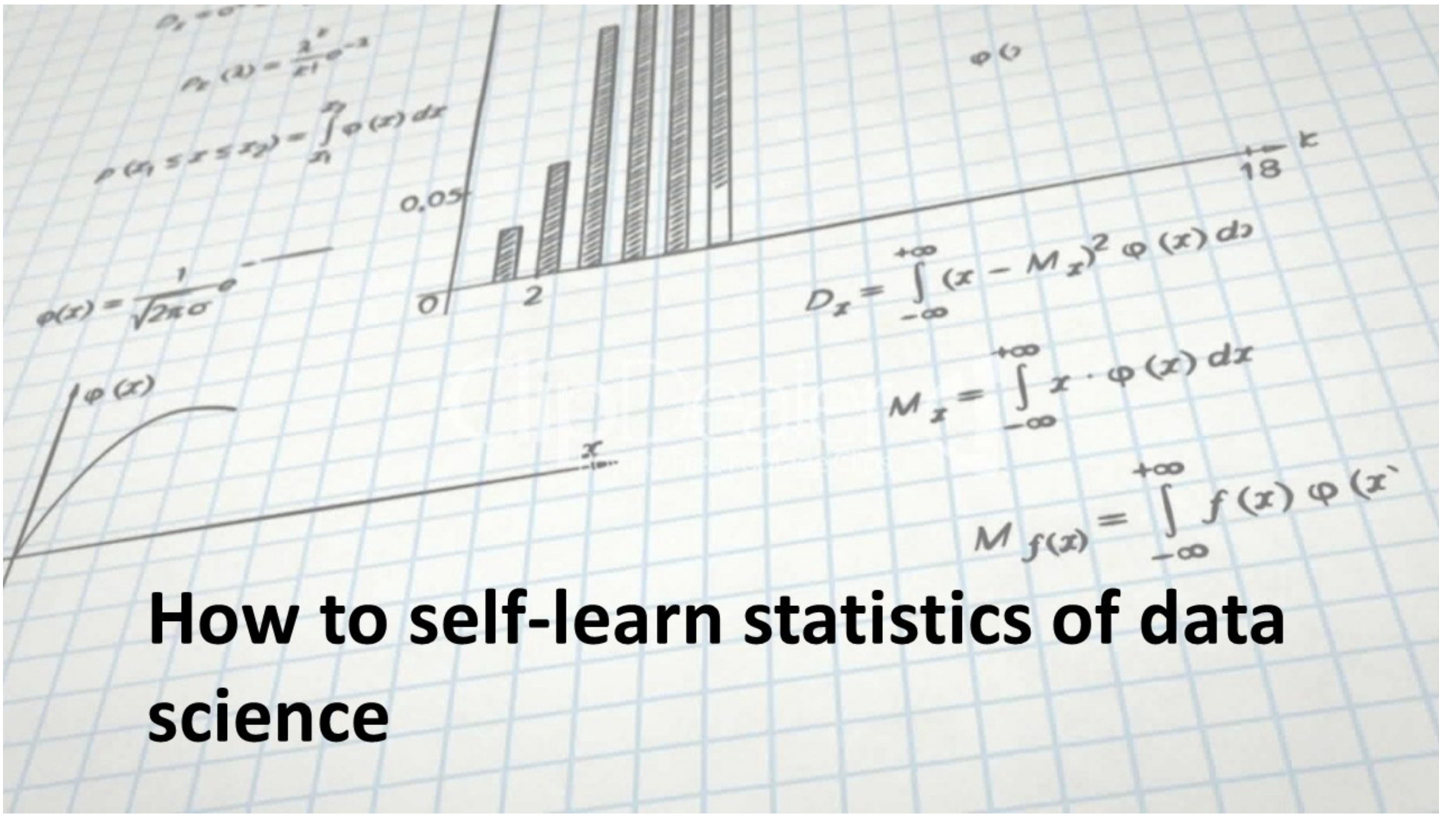
How to self-learn statistics of data science

Ashish Patel
Aug 12, 2018 · 7 min read

Learning Series!!!



Statistics: Understanding statistics, especially Bayesian probabilities, is critical to many machine learning algorithms.

Do you want to learn statistics quickly and inexpensively? Good news…, you can master core concepts, probabilities, Bayesian ideas, and even statistical machine learning through free online resources. Here are the best resources for self-study.

By the way, you don't need a math degree, but if you have a mathematical background, you will definitely like this fun, hands-on approach.

This tutorial will give you the statistical thinking you need in the data sciences, and it will make you more profitable than some aspiring data scientists without it.

You know, since you learned how to program, it always tempts you to use machine learning packages directly, even if you know what to do? If you want to start at the beginning, it is okay to learn how to go snowballing in a real project.

But if you do, you will probably never learn statistics and probability theory completely, and as a data scientist, these are a very necessary part of your career, which is why you have to learn.

1. First and foremost: basic Python skills

In order to complete this tutorial, you need the most basic Python programming skills, we will learn statistics by application, hands-on. If you don't have the relevant skills, you can learn python through self-learning through our tutorial. This is the fastest way to learn Python quickly. We recommend at least the second step of the tutorial. *Note: It* can be other languages, but the examples are all Python.

2. The necessity of statistics in data science

Statistics is a broad field that is used in many industries.

Its definition in Wikipedia is: *it is the collection, analysis, interpretation, presentation and organization of data. So it's not surprising that data scientists need to understand statistics.*

For example, data analysis requires at least descriptive statistics and probability theory. These theories will help you make better business decisions based on the data.

Key concepts include probability distribution, statistical significance, hypothesis testing, and regression.

Moreover, machine learning needs to understand Bayesian probabilities, and Bayesian probabilities are the engines of many machine learning modules.

Key concepts include conditional probabilities, prior probabilities, posterior probabilities, maximum likelihood estimates, and if these concepts make you fear, don't worry, once you roll up your sleeves and start learning, you'll understand.

3. The best way to learn data statistics in data science

So far, you may have discovered that the common way of “self-learning a certain knowledge X” is to jump out of the classroom and directly through hands-on methods, and the statistics in data science are no exception.

In fact, it is very interesting that we master the core concepts of statistics through programming.

If you don’t have a formal math-related educational background, you’ll find that this way you can more easily understand complex formulas. It will let you think about the logic of each calculation.

If you have some formal and relevant mathematical background, this approach will combine your theory with practice and give you a lot of interesting programming challenges.

Here are three steps to learning statistics and probability theory in the field of data science:

1. Statistical core concept

Descriptive statistics, distribution, hypothesis testing and regression.

2. Bayesian probability theory

Conditional probability, prior probability, posterior probability, maximum likelihood estimation

3. Introducing statistics in machine learning

Learn basic machine learning concepts and how to use statistics in machine learning

After completing these three steps, you will truly face and face more difficult machine learning problems and common data science applications.

The first step: the core concept of statistics

In order to know how to go to school statistics, first of all, to understand how it is used is very helpful for learning. Let’s look at some examples of real analysis or as an application that data scientists might use:

1. **Pilot design**: Your company starts a new product line, but sells it through offline retail. You need to design an A/B test to control the differences between the different areas. You also need to estimate some meaningful results from the store from a statistical perspective.

2. **Regression model:** Your company needs to be able to better predict what the demand for a personal product line will be in all its stores. Insufficient inventory and excess inventory can be costly. You consider building a series of regular regression models.

3. **Data conversion:** In the process you are testing, there are multiple machine learning models for you to use. Some models can generate corresponding data distributions by inputting data. You need to be able to identify them and convert the input data appropriately or know under what assumptions Correlation.

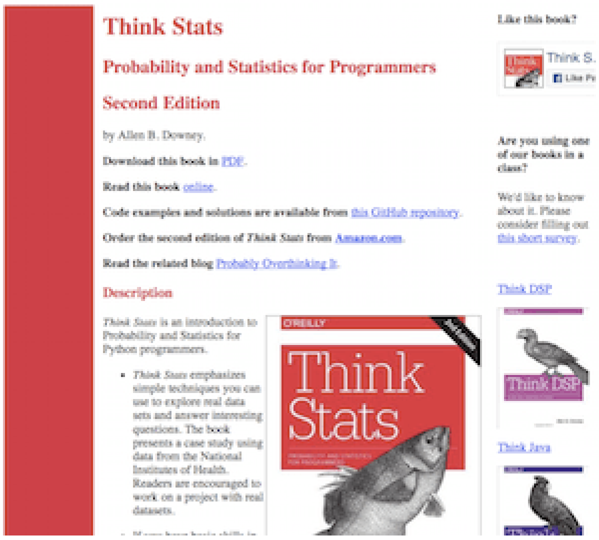
A data scientist has to make hundreds of decisions every day, ranging from a module challenge to a team R&D strategy.

Most decisions require a solid theoretical foundation of statistics and probability theory.

For example, data scientists need to constantly decide which data is deterministic and which data is random. In addition, they need to know if there are points of interest for further exploration.

These are the core things that are at the time of making an analysis decision (if you only know how to calculate the value, then just touch the surface).

Here are the resources of the best self-learning statistics we have found:



Think Stats

Probability and Statistics for Programmers

Second Edition

by Allen B. Downey.

Download this book in PDF.

Read this book online.

Code examples and solutions are available from this GitHub repository.

Order the second edition of *Think Stats* from Amazon.com.

Read the related blog Probably Overthinking It.

Description

Think Stats is an introduction to Probability and Statistics for Python programmers.

- *Think Stats* emphasizes simple techniques you can use to explore real data sets and answer interesting questions. The book presents a case study using data from the National Institutes of Health. Readers are encouraged to work on a project with real datasets.

Like this book?

Are you using one of our books in a class?

We'd like to know about it. Please consider filling out this short survey.

Think DSP

Think Java

Python, you can use them

Think like a Bayesian…

Think Stats is an excellent book (with a free PDF version) that introduces all the core concepts. What is the premise of reading this book? If you know how to program, then you can learn the statistics yourself in the process, and we found that this method is also suitable for those with a mathematical background.

Step 2: Bayesian Probability Theory

A philosophical debate on statistics is Frequentists and Bayesians, and Bayesian theory is more relevant when learning statistics in data science.

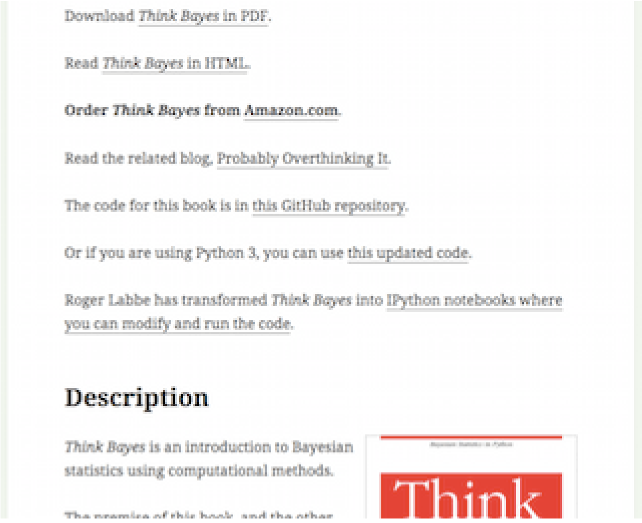
In short, frequency theory is used near sampling modules. This means that they will only be used to describe the data that has been collected.

Bayesian theory, on the other hand, is used not only for sampling modules, but also for data that is not known before collection. If you want to know more distinction between them, you can look at this post: the For A non-Expert, the What's at The -difference the BETWEEN frequentist and Bayesian Approaches? .

In Bayesian theory, the level of uncertainty before collecting data is called "a priori probability", and after the data, it is updated to "posterior probability". For some machine learning models, this is a very core concept. It is very important to master them.

Moreover, these concepts make sense after using them.

Here are the resources of the best self-learning Bayesian theory we have found:



Download *Think Bayes* in PDF.

Read *Think Bayes* in HTML.

Order *Think Bayes* from Amazon.com.

Read the related blog, Probably Overthinking It.

The code for this book is in this GitHub repository.

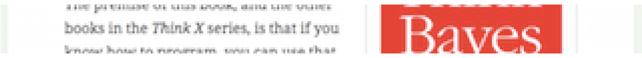
Or if you are using Python 3, you can use this updated code.

Roger Labbe has transformed *Think Bayes* into IPython notebooks where you can modify and run the code.

Description

Think Bayes is an introduction to Bayesian statistics using computational methods.

The premise of this book, and the other



Think like a Bayesian…

Think Bayes is an excellent book (with a free PDF version) that introduces all Bayesian theories. It is also a way to learn by programming. This is fun and simple. We found that this method is also suitable for those with a mathematical background.

Step 3: Introduce statistics in machine learning

If you want to learn statistics in data science, after you have completed the core concepts of statistics and Bayesian theory, there is no better way to use statistical analysis in machine learning modules.

The field of machine learning is closely related to statistics. Statistical machine learning is the most important way of machine learning now.

In this step, you will implement some machine learning modules from scratch, which will help you unravel a true understanding of its underlying technology.

At this stage, even if you copy the code directly line by line, it is ok.

It will help you to open up the black box of machine learning while consolidating your knowledge of statistical learning.

The following models were chosen because they illustrate the first few key concepts.

Linear regression

First we have an example of a predictive model…

- Linear Regression from Scratch in Python

Naive Bayes classifier

Again, we have a simple model that works very well…

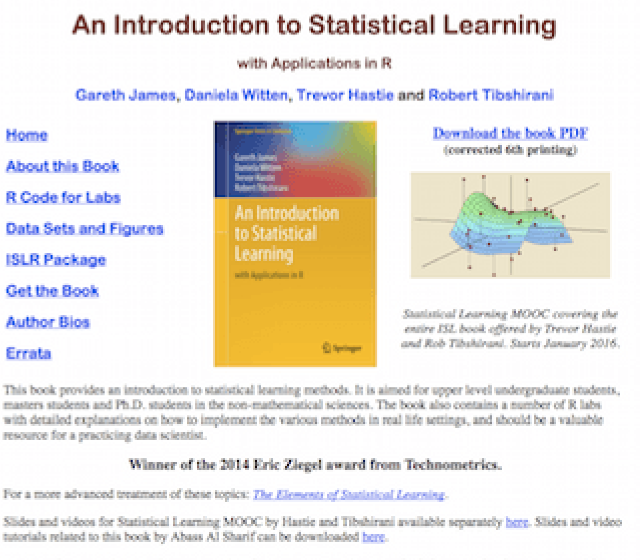
- Intuitive Introduction , Naive Bayes from Scratch in Python

Multi-arm gaming machine

Finally, we have the famous “20 lines of code to beat any A / B test!”

- Intuitive Introduction , Multi-Armed Bandits from Scratch in Python

If you are eager to learn more, we recommend the following resources.



An Introduction to Statistical Learning

with Applications in R

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

Download the book PDF
(corrected 6th printing)

Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani. Starts January 2016.

This book provides an introduction to statistical learning methods. It is aimed for upper level undergraduate students, masters students and Ph.D. students in the non-mathematical sciences. The book also contains a number of R labs with detailed explanations on how to implement the various methods in real life settings, and should be a valuable resource for a practicing data scientist.

Winner of the 2014 Eric Ziegel award from Technometrics.

For a more advanced treatment of these topics: *The Elements of Statistical Learning*.

Slides and videos for Statistical Learning MOOC by Hastie and Tibshirani available separately here. Slides and video tutorials related to this book by Abass Al Sharif can be downloaded here.

For your reference…

Introduction to Statistical Machine Learning is an excellent e-book (with free PDF version), the example is the use of R language, this book covers a wider range of topics, when you make more progress in machine learning This is a valuable tool. .

Thanks for reading this is my initiative to start this for all who are facing difficulties in this concept.Please share with all who need this article.

Data Science Statistics Ml Research Lab Bayesian Statistics Regression

About Help Legal