BOOTSTRAP

10.15.2018

RECAP

* Binomial test
  * Simulation
    * Analytic solution
* Mean & median
* Variance, standard deviation

DATA

* suppose we have a dataset: an electrophysiological recording from a spiking neuron

* we want to estimate the mean background spike rate of the neuron

* but we have a limited sample of data!

DATA

* we can estimate the mean directly (**data.mean()**)

* but our data is just a sample. how close is our data mean to the true mean?

* how would we know? we only have the data we have

  * …or do we?

STANDARD ERROR

* in statistics, *everything you compute is an estimate*

* we often want to know how well an estimate based on a finite sample of data approximates the “true” value

* the **standard error** is a number that says how variable an estimate is

STANDARD ERROR

* repeat the same experiment 100 times, and estimate your desired statistic (e.g. the mean) using each of the 100 datasets

* the standard deviation of those 100 estimates is the standard error

* can we get this number without actually running 100 experiments?

THE BOOTSTRAP

Brad Efron

* we don’t know what distribution the data comes from

* but we know one thing: the data

* so: we *sample* from the data to create “new” datasets, then use these to examine the variability of the mean

THE BOOTSTRAP

* creating a bootstrap sample:

  * suppose we have N numbers

* choose from among them N times *with replacement* (i.e. the same number can be chosen more than once)

THE BOOTSTRAP

* e.g.: data = [1,3,5,9]

* bootstrap_1 = [5,1,1,3]

* bootstrap_2 = [3,3,9,5]

* bootstrap_3 = [1,3,5,5]

* etc.

THE BOOTSTRAP

* each bootstrap sample *is a reasonable dataset that could have happened!*

THE BOOTSTRAP

* after taking many samples (with replacement), and computing the average of each, let’s look at their distribution

* on average, the *true population mean* should fall inside the middle 95 percentiles of the bootstrap statistics 95% of the time

BOOTSTRAP STANDARD ERROR

* the standard deviation of the bootstrap statistics is called the **bootstrap standard error**

* it’s a pretty good estimate of the standard error! we will see how good when we get to situations where we can compute an analytic standard error

CENTRAL LIMIT THEOREM

* let’s take a bunch of random values (from the same distribution)

* and take their average

* let’s do this many times

CENTRAL LIMIT THEOREM

* as the number of samples we take increases, the distribution of their averages converges to..

* a gaussian distribution!

END