DS 102 Discussion 6

Wednesday, October 6, 2020

Introduction

In this discussion, we'll explore more of the wonders of the bootstrap. In particular, we'll see how it can be used for hypothesis testing, by providing an estimate of the null distribution in situations where the null distribution is far too complicated to derive in closed-form (which is common in real-world problems).

First, we'll review the set-up of a generic hypothesis testing problem. Suppose we have a sample $X \sim P$ from some unknown distribution P. Our goal is to test whether P satisfies a certain condition:

$$H_0 \colon P \text{ does not satisfy the condition} \tag{1}$$
$$H_A \colon P \text{ does satisfy the condition.} \tag{2}$$

For example, the condition could be that the mean of the distribution is greater than zero ($H_0 \colon \mathbb{E}_P[X] > 0, H_A \colon \mathbb{E}_P[X] = 0$), or that the distribution is different from some reference distribution Q ($H_0 \colon P = Q, H_A \colon P \neq Q$).

Suppose we have some **test statistic** T (that is, some function of X) that we think is appropriate for distinguishing H_A from H_0, and we observe the value $T = t$ for our sample X. By definition, the test will have **significance level** α if

$$\mathbb{P}_0(T \geq t) \leq \alpha$$

where $\mathbb{P}_0$ denotes probability under the **null hypothesis** H_0, and $\mathbb{P}_0(T \geq t)$ is called the ***p*-value**. Therefore, to test H_0 vs. H_A with significance level α, we need to compute the p-value. That requires us to know *the distribution of the test statistic under the null distribution*—a common bottleneck in hypothesis testing, because this distribution can be difficult or impossible to derive in closed-form.

However, we can use the bootstrap to estimate the distribution of T under the null distribution, by simulating a situation where the null is true.

Application of Bootstrap for Hypothesis Testing

The application we'll consider is the analysis of the velocities (in km/sec) $X_1, \ldots, X_{82}$ of $n = 82$ galaxies measured during a survey of the Corona Borealis region of the sky. The distribution of galaxy velocities provides information about the structure of the far universe—in particular, astrophysicists interpret a *multimodal* distribution of velocities as evidence for the existence of voids and superclusters.

Our goal is therefore to perform a hypothesis test of whether or not the distribution of velocities is multimodal:

$$H_0 \colon m(p) = 1 \tag{3}$$
$$H_A \colon m(p) > 1 \tag{4}$$

where p is the distribution of galaxy velocities, and m gives the number of modes of a distribution.

To develop an appropriate test statistic for this problem, we have to get creative. We want a test statistic that somehow quantifies how suitable a unimodal distribution is for modeling this data. If such a test statistic takes on a low value, we could then reject the null. One way we can devise such a test statistic is a following: we'll model our data using *kernel density estimation*:

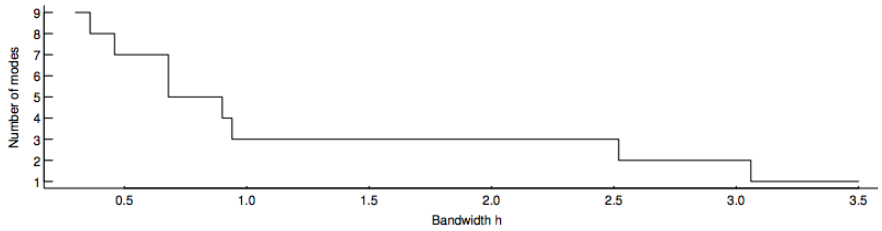
$$\hat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{5}$$

where K is some non-negative *kernel function* that captures the influence of each data point X_i on the density of an arbitrary point x. A common choice of kernel is the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

The parameter $h > 0$ is a *bandwidth parameter* that captures how close data points X_i must be to x to influence its density. For larger values of h, more data points have an influence on the density at x, whereas for smaller values of h, only data points very close to x influence it. Both K and h are user-selected.

It can be shown that the number of modes of $\hat{p}_h(x)$ decreases monotonically as h increases. For our galaxy data, the relationship between $m(\hat{p}_h)$ and the bandwidth h is shown in the following figure:



Let H_1 be the minimal bandwidth value for which $\hat{p}_h$ is unimodal:

$$H_1 = \min\{h \colon m(\hat{p}_h) = 1,\ m(\hat{p}_{h'}) > 1 \text{ for all } h' < h\}. \tag{6}$$

We will use H_1 as the test statistic. For our galaxy data, the observed value of the test statistic is $h_1 = 3.05$ (as you can see from the figure).

(a) In order to perform a hypothesis test with significance level α, what do we need to compute? What distribution does this require knowledge of?

(b) Propose a distribution we can use for the null hypothesis. Hint: restrict yourself to distributions of the form of a kernel density estimate.

(c) Let $Z^* = (Z_1^*, \ldots, Z_{82}^*)$ denote a bootstrap sample from the dataset. It can be shown that $Z_i^* + h_1\epsilon_i$ for $\epsilon_i \sim \mathcal{N}(0, 1)$ gives independently and identically distributed samples from $\hat{p}_{h_1}$ (we won't worry about proving this here). Using this fact, write down a bootstrap procedure for computing the p-value.