The Neyman-Pearson Lemma and exponential families of random variables

April 24, 2020

This is just one more note on how the Neyman-Pearson Lemma works in a somewhat general case. It also serves to connect what we're doing in Chapter 10 with Chapter 9.

Recall in Chapter 9 we introduced the exponential family of random variables. Namely random variables of the form:

$$f_Y(y|\theta) = \begin{cases} a(\theta)b(y)e^{-[c(\theta)d(y)]} & \text{if } y \in [a,b] \\ 0 & \text{otherwise} \end{cases}.$$

where θ is an unknown parameter and the interval $[a,b]$ doesn't depend on θ.

We say that given a random sample, $Y_1, Y_2, \ldots, Y_n$ from this distribution then $\sum_{i=1}^n d(Y_i)$ is a sufficient statistic for the estimation of the θ. So this is family of distribution has nice properties.

In this section, we'll make the additional assumption that the function $c(\theta)$ is decreasing. The case $c(\theta)$ is increasing is very similar. The case $c(\theta)$ is sometimes increasing and sometimes decreasing can a be a little weird, this could the likelihood function is also increasing and decreasing, leading to multiple local extrema.

Let's apply the Neyman-Pearson Lemma to derive the most powerful test of $H_0 : \theta = \theta_0$ against the simple hypothesis $H_a : \theta = \theta_a$ for some $\theta_a > \theta_0$ at level α. (The opposite inequality will be very similar, so we focus on this one for concreteness.)

We consider the ratio of likelihood functions:

$$\frac{L(\theta_0)}{L(\theta_a)} = \frac{a(\theta_0)\prod_{i=1}^n b(y_i)e^{-[c(\theta_0)\sum_{i=1}^n d(y_i)]}}{a(\theta_a)\prod_{i=1}^n b(y_i)e^{-[c(\theta_a)\sum_{i=1}^n d(y_i)]}} = \frac{a(\theta_0)}{a(\theta_a)} e^{-[(c(\theta_0)-c(\theta_a))\sum_{i=1}^n d(y_i)]}$$

for $a \le y_i \le b$.

Then the Neyman-Pearson Lemma tells us the most powerful RR is of the form

$$\frac{L(\theta_0)}{L(\theta_a)} = \frac{a(\theta_0)}{a(\theta_a)} e^{-[(c(\theta_0)-c(\theta_a))\sum_{i=1}^n d(y_i)]} < k$$

We now solve for the sufficient statistic $\sum_{i=1}^n d(y_i)$. We begin by multiplying both sides by $\frac{a(\theta_a)}{a(\theta_0)}$ and then taking the ln of both sides to get:

$$-[(c(\theta_0)-c(\theta_a))\sum_{i=1}^n d(y_i)] < \ln\left(\frac{a(\theta_a)}{a(\theta_0)}k\right)$$

Then using our assumption that $c(\theta)$ is decreasing and that $\theta_a > \theta_0$ we have that $-(c(\theta_0) - c(\theta_a))$ is negative, so when we divide by it, the inequality changes and the rejection region is then of the form:

$$\sum_{i=1}^{n} d(y_i) > k'$$

for some constant k'. To determine k' we need to know the distribution of $\sum_{i=1}^{n} d(Y_i)$, sometimes this can be done, sometimes it can be quite hard. If n is large, since it is the sum of independent random variables, we can use that it will be approximately normal. In any of the cases where we know or can approximate the distribution of $\sum_{i=1}^{n} d(Y_i)$, we can then pick k' such that the probability under H_0 that

$$\sum_{i=1}^{n} d(Y_i) > k'$$

is α.