Two Building Blocks of Statistics

통계가 모델과 데이타를 만나게 한다고 말씀드렸습니다. 도대체 어떻게 그게 가능한 것일까요? 그것을 가능하게 하는 두가지 Building Blocks은 Law of Large Numbers (LLN)과 Central Limit Theorem (CLT) 입니다.

앞으로 모든 개념은 Dualistic View of Model & Data를 기준으로 설명하겠습니다. 통계에서는 (Model, Data)를 (Population, Sample)이라고 부릅니다. 고등학교때에 배웠던 (모집단, 표본집단)이 (Population, Sample)에 해당합니다. 여러분들이 통계학을 배울때 퍽퍽하게 느껴지셨던 기억이 있나요? 저도 그런 기억이 있습니다. 그런 느낌이 드는 이유중에 하나는 통계학만 가지고서는 촉촉한 Model을 만들수가 없기 때문입니다. 통계는 Model과 Data를 연결시켜주는 고리로서의 역할을 하는 것이지 Model 자체에 대한 insight를 주는 것은 아닙니다. 전문분야에서 활동하며 얻은 지혜가 쌓이고 쌓여 Model이 만들어지는 것입니다.

아주 간단한 모델을 생각해보죠. 어떤 확률변수 X가 1 또는 -1의 값은 50%의 확률로 가지는 모델입니다. 간단하지만 '눈'에 보이지는 않고 우리가 '머리'로 상상만 하는 모델입니다.

이러한 모델에서 나오는 Data를 여러분들이 여러개 보았다고 합니다. '이러한 모델에서 나오는 Data'라고 표현을 했는데요. 좀 더 정확하게는 여러분이 이러한 모델에서 나왔다고 믿는 Data인 것이죠. 다시 말씀드립니다만 Model은 그 자체로는 뇌피셜입니다. 그것이 레알일 이유는 없습니다.

이제 여러분들이 눈으로 100개의 데이타를 봅니다, as in 1, -1, -1, 1, -1 ..... 여러분들이 눈으로 본 데이타가 여러분의 모델, 어떤 확률변수 X가 1 또는 -1의 값은 50%의 확률로 가지는 모델이 단순한 뇌피셜이 아니라는 증거가 될 수 있을까요?

1. LLN

LLN은 데이타를 잘 요리하면 데이타가 가질 수 있는 첫번째 좋은 특성을 보여줍니다. 저는 보통 LLN을 가르칠때 다음과 같은 문장을 큰소리로 복창시킵니다.

(the) **Sample mean converges to** (the) **true expectation**

Sample mean은 여러분들이 눈으로 본 데이타를 요리한 것이죠. True expectation은 model의 세계에서 계산할 수 있는 것이구요. 위에서 말씀드린 예를 가지고 적용해보면 다음과 같이 표현할 수 있습니다.

$$(1/100)*(\ 1 + (-1) + (-1) + 1 + (-1) \dots\ ) \text{ converges to } (½)*(-1) + (½)*1 = 0$$

너무 당연한가요? 그런데 이 간단한 특성이 여러분들이 배웠던 수많은 통계 test의 첫번째 building block 입니다.

(Digression)

LLN은 너무 단순해서 극적으로 드라이해보일수 있지만 실제로 여러가지 사회현상과 촉촉하게 연결되어 있습니다.

나중에도 말씀드릴 기회가 있겠지만 Finance라는 학문에서 가장 중요한 이론중의 하나인 diversification은 LLN과 거의 같은 컨셉입니다. 같이 움직이지 않는 risk component(idiosyncratic components)는 diversified된 portfolio에서는 사라지는 것을 diversification이라고 합니다. 이렇게 주장할 수 있는 이유가 LLN입니다.
또한 경제학에서 말하는 시장경제의 장점도 LLN과 밀접한 관련이 있습니다. 참여자들이 시장의 가격을 결정하는데에 점점더 많이 참여하면 가격은 경제시스템안의 정보를 반영하여 정확한 값을 나타내게 된다는 것이지요. 또한 경제변수를 예측할때에 아주 뛰어난 개별전문가의 예측보다 약간 뒤지는 분들의 예측을 average하는 것이 더 훌륭하다고 알려져 있습니다. 주식시장에서 실적보고를 할때마다 시장의 컨센서스를 beat했니 안했니하고 뉴스가 나오는데요. 시장의 consenus는 일반적으로 analyst forecast의 average또는 median으로 측정합니다. 조직이론에서도 점점 더 중요한 이슈가 되고 있는 Diversity and Inclusion의 Statistical background도 LLN에서 찾을 수 있다고 봅니다. 어떤개인도 정확하게 문제를 풀 수 없을 만큼 복잡한 문제를 다룰때에는 보다 많은 정보가 표현되고 반영될수 있는 조직이 성공할 확률이 높다는 것이죠. 싸이즈를 좀 더 크게 확장시켜보면 LLN은 정치학적으로 Democracy vs Dictatorship의 담론과도 연결지을 수 있습니다.

2. CLT

Normal distribution 이라는 표현을 들어보셨나요? Normal distribution은 아주 정확하게 define이 되어있는 분포입니다. 여러분도 Normal distribution의 bell-shaped curve의 그림은 자주 보셨을 겁니다. 왜 이렇게 생긴 distribution에 Normal이라는 이름이 붙었을까요? 마땅히 distribution이라면 따라야할 norm이라고 주장하고 있는 듯한 그 이름이 과연 적절한 것일까요?

CLT는 앞에 말씀드린 질문에 대한 해답을 제시합니다. CLT도 LLN과 마찬가지로 데이타를 잘 요리하면 만들어 낼 수 있는 특성을 보여줍니다만, 제가 느끼기에는 CLT가 LLN보다 더 신기합니다. 저는 CLT를 가르칠때 LLN과 라임을 맞춰 다음과 같은 문장을 큰소리로 복창시킵니다.

(the) **Boosted sample mean converges to** (the) **normal distribution**

Boosted sample mean은 좀 더 정확하게 말씀드리면 \sqrt{sample size} * sample mean 입니다. 위의 문장이 성립하려면 population mean 이 zero여야합니다만, 표현을 간단하게 할 수 있도록 그 부분은 생략했습니다. 그리고 normal distribution의 variance는 population의 variance와 같지만 그것도생략합니다.

CLT의 강력함은 위에서 말한 구호가 Model의 distribution과는 상관이 없다는 것입니다. Regardless of the population distribution, if you boost the sample mean properly, it converges to the same distribution of Normal! 정말 신기하죠? 그래서 Normal

distribution의 이름이 Normal인가 봅니다. 실제로 우리가 observe하는 많은 데이타들은 많은 결과들 합쳐져서 결과적으로 나타나는 경우가 많고 그것의 distribution을 보면 당연하게도 Normal distribution을 따르는 경우가 많이 있습니다.

LLN는 sample mean이 어디로 간다고 하고 CLT는 boosted sample mean이 저기로 간다고하는데, sample mean이나 boosted sample mean이나 똑같은 애들인 것같은데 왜 다른데로 간다는 거지.. 라고 하시는 분이 계신가요? 주어진 샘플이 하나 있을때에는 sample mean이나 boosted sample mean이 주는 정보량은 똑같겠죠.

LLN은 하나의 주어진 샘플이 사이즈가 커질때의 특성이고 CLT는 여러개의 가능한 샘플에 관한 특성을 말하고 있다고 보시면 됩니다. 이것에 대해서는 simulation exercise와 도움이 될 것 입니다.

(Digression)

혹시 CLT를 너무 심각하게 받아들여서 세상에서 보는 모든 distribution이 normal이여야 한다고 생각하실 분이 있을지도 모르겠습니다. 그런데 세상은 CLT를 가능하게 한 sum of something처럼 단순하지 않은 경우가 많습니다. 간단한 케이스를 생각해보면, sum이 아니라 multiple인 경우도 많이 있죠. Finance쪽에서 예를 찾는다면 복리로 받는 은행이자, 주식수익같은 것을 생각할 수 있습니다. 이런경우에는 단순하고 로그를 위해서 sum으로 바꾸는 trick을 부릴 수 있습니다.

세상에는 수많은 energy들이 모여서 협력하기도 하고 경쟁하기도 하면서 진화해갑니다. 그러면서 부자들이 더 부자가 되는 pareto distribution등이 생길수도 있고, 어느 정도 주어진 threhold밑으로는 다 소멸해버리는 trucated distribution등이 보일수도 있습니다.

조금 더 철학적인 측면에서 normal이라는 distribution이 존재하는 것과 normal한 개체가 존재하는 것은 좀 다른 부분이 있습니다. Normal distribution이라는 것은 a single random variable의 기준으로 보았을때 Normal하다는 것이죠. (여러분이 개체라고 관심을 가질만한) 어떠한 하나의 개체도 하나의 기준만으로 만들어지지 않습니다. 아주 간단하게 사람의 몸 사이즈만 본다고 해도 키, 몸무게, 팔길이, 허벅지 길이, 머리크기 ... 등등 여러개의 기준이 존재하죠. 미국공군이 비행기 운전석을 만들때 각 기준에서 normal하다고 볼수 있는 사람을 기준으로 운전석을 만들었는데 실제로 모든 기준으로 보았을때 normal한 조종사는 한명도 없었다는 이야기가 생각나네요.
https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html

3. An archetype of a statistical test

제가 학부때 계량경제학 수업을 들을 때에 많이 헷갈렸던 기억이 있습니다. 수많은 test들이 있는데 상황마다 다르게 이럴땐 이걸쓰고 저럴땐 저걸써야 했던 기억입니다. 지금 생각해보면 한가지 test에서 이런 저런 모습으로 변신했던 것을 모두 다른 test 였던 것으로 알고 따로따로 외우면서 괴로워했던 시절이었습니다.

모든 test라고 말하기는 그렇습니만, 여러분들이 보았던 test의 99%는 다음과 같은 형태를 갖습니다.

Boosted sample mean/sqrt(sample variance) VS N(0,1)

왜 이러한 test가 말이 될까요? 다음과 같은 세개의 step으로 나눠보죠.

Step1: Boosted sample mean converges to N(0,sigma_x^2), from CLT
Step2: Sample variance converges to sigma_x^2, from LLN
Step3: Boosted sample mean/sqrt(sample variance) VS N(0,1), from Steps 1 and 2

이제 왜 Boosted sample mean/sqrt(sample variance) 와 N(0,1) 을 비교하는 것이 reasonable한 것인지는 감이 오시죠?

(in depth)

그렇다면 이것은 도대체 무엇을 test하는 것을까요? 실제로 이러한 test는 **Model**의 모든 부분을 **test**하는 것이 아니라 **Model**의 '한' 측면만을 **test**하는 것입니다.

실제로 테스트가 말이 되는 비교가 될 수 있도록 implicit하게 들어간 부분은 true expection이 0이라는 것입니다. 그래야 Step1이 말이 됩니다. 이 부분이 틀리다면 Step1이 틀리겠죠. 그래서 Sample에서 계산한 값, Boosted sample mean/sqrt(sample variance)이 N(0,1)에서 왔다고 보기에 너무 익스트림한 값이 나오면, for example more than 1.96 or less than -1.95, true expection이 0이 아니었나보다 라고 결론짓는 것입니다.

모델의 모든 부분을 보는 것이 아니라 한 측면만 본다는 것은 test의 장점 일 수도 있고 단점 일수도 있습니다. 모델 specific하지 않기때문에 디테일 한 부분을 미스할 수 도 있지만 그렇게 때문에 an archetype of a statistical test를 찾을 수 있는 것이기도 합니다. 실제로 the archetype을 통해서 테스트 할 수 있는 것은 모델에서 뽑아낼 수 있는 것중에서 <있다 또는 없다>로 떨어지는 정도만 test가 가능합니다. 이렇게만 말씀을 드리면 너무 애매하기도 하고 그 정도로 간단한 것만 알아서 무슨 의미가 있나 싶은 의문이 생길수도 있지만 실제로 <있다 또는 없다>는 것만으로도 재미있는 해석을 할 수 있는 경우가 많이 있습니다. 이 부분에 대해서는 앞으로 여러번 반복할 기회가 있을 것입니다.

마지막으로 Dualistic view of Model and Data를 강조하면서 글을 마치겠습니다. 우리가 살고 있는 사회에서 어떤 분야의 전문가라고 인정을 받으려면 (i) 어떤 모델을 머릿속에 상상하고 있는지, (ii) 데이타를 통해서 모델의 어떤 부분을 보일 수 있고 보일 수 없는 지 정확하게 알고 있어야 합니다. 아무런 모델없이 데이타만 고문하다보면 인생이 허망해집니다. 데이타에서 거리를 두고 모델링을 상상할 수 있어야 하고, 모델에서 떨어져 어떠한 데이타가 모델을 뒷받침 할 수 있는지에 대해서도 고민해야 합니다.