정규분포, 그리고 "평균으로의 회귀"
<기초 통계학의 숨은 원리 이해하기>, 다 하지 못한 이야기들

동일한 분포를 따르는 독립적인 확률변수들의 평균은, 확률변수의 수가 늘어남에 따라 정규분포로 수렴한다는 중심극한정리Central Limit Theorem는 다음과 같이 나타낼 수 있다.

X_i : 서로 독립이고, 동일한 확률분포를 따르는 확률변수들[1)]

(공통 평균 : μ, 공통 분산 : σ^2)

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

중심극한정리는 자연계의 많은 변수들이 왜 정규분포를 따르는지에 대해 한 가지 설명을 제시한다. 사람을 비롯하여 동식물의 많은 변수들이 정규분포와 비슷한 분포를 따른다.[2)] 사람의 키, 몸무게가 대표적인 예이다. 만약 사람의 키를 결정하는 요인들(유전적 요인과 환경적 요인들)이 굉장히 많고, 서로 비슷한 분포를 따르고, 그들이 독립적으로 사람의 키에 영향을 미친다면, 사람의 키는 정규분포를 따르게 될 것이다.

유전 현상에서 정규분포는 단지 하나의 변수가 따르는 분포 이상의 의미가 있다. 19세기 영국의 학자 골튼Galton의 실험을 살펴보자. 스위트피sweet pea란 콩의 무게는 정규분포를 따른다. 그는 동일한 무게의 부모에서 나온 자식들의 무게는 어떤 분포를 따르는지 궁금했다. 그는 스위트피의 종자들을 무게에 따라 일곱 단계로 분류하여, 비슷한 무게의 종자를 70개씩를 모았다. 따라서 한 무게 단계에서 70개씩 총 490개의 종자를 모은 것이다. 그리고 그들을 모두 비슷한 환경에서 심어 길렀다. 그렇게 해서 새로 얻은 자식 종자의 무게의 분포를 살펴보았더니 놀랍게도 동일한 무게의 부모에게서 나온 자식 종자들은 (다른 무게의 부모에게서 나온 자식 종자들의 무게와) 분산이 동일한 정규분포를 따르고 있었다.

이 결과는 다음과 같이 해석될 수 있다. 종자의 무게는 무게를 결정하는 유전자에 의해 상당 부분 결정된다. 하지만, 그 밖의 다른 여러 유전자들과 여러 가지 환경요소들도 작지만, 독립적으로 종자의 무게에 영향을 미친다. 여기서 유전자는 부모에게서 자식에게 전달된다. 따라서 자식 종자의 무게는 많은 부분이 부모에 의해 결정된다. 반면, 부모에 의해 결정되지 않는 많은 요소들의 영향에 의해, 같은 부모에게서 나온 자식 종자들의 무게가 같지 않고, 정규분포를 띄게 되는 것이다.

한 가지 주목할 만한 현상은 만약 부모 세대의 무게의 분포와 자식 세대의 무게의 분포가 동일하게 유지된다면(세대 간에 무게 분포의 차이가 거의 없다면)[3)], 그리고, 부모의 무게를

---

1) 보통 통계학에서는 독립적이고 동일한 분포(independent and identically distributed)라는 의미에서 i.i.d라고 표시한다.
2) 제한된 수의 관찰된 값을 통해 정규분포라는 것을 증명하기는 불가능에 가깝다. 따라서, 이후에는 "정규분포와 유사한 분포"를 보이는 경우, 단순히 "정규분포를 따른다."라고 표현할 것이다.

통해 자식의 무게를 정확하게 예측할 수 없다면(동일한 무게의 부모에게서 나온 자식의 무게가 앞 선 스위트피의 경우와 같이 정규분포를 따른다면), 회귀 분석을 통해 예측되는 자식의 무게는 언제나 부모의 무게보다 평균으로 치우치게 된다. 이 현상을 현대의 통계학자들은 “평균으로의 회귀”라고 부른다.

골튼은 위의 조건에서 논리적으로 “평균으로의 회귀 현상”이 일어날 수 밖에 없음을 다음과 같이 설명하였다. 만약 동일한 무게의 부모에게서 나온 자식의 무게가 한 값으로 정해지지 않고, 일정한 값을 중심으로 정규분포를 이루고 있고, 무게가 동일한 부모에게서 나온 자식의 무게의 평균이 부모의 무게와 같다면, 자식의 분포는 부모의 분포보다 넓어지게 된다.(그림 1)

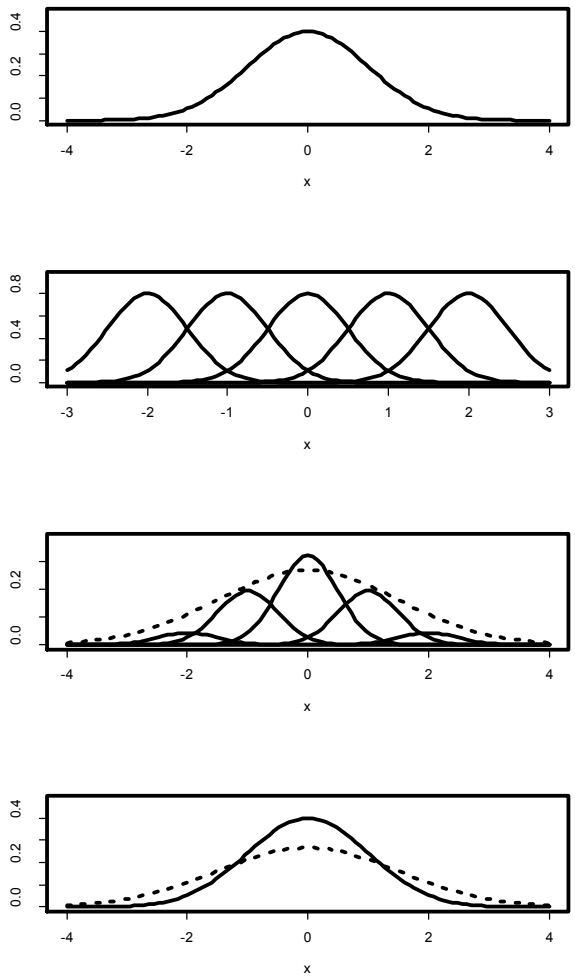

그림 1 부모의 키의 분포/부모의 키가 -2, -1, 0, 1, 2일 때 자식의 키의 분포, 자식의 키의 분포(점선), 부모의 키의 분포(실선)과 자식의 키의 분포(점선)

3) 부모 세대와 자식 세대의 유전자 구성이 비슷하고, 환경 또한 크게 다르지 않다면, 부모 세대의 분포와 자식 세대의 분포는 거의 비슷할 것이다.

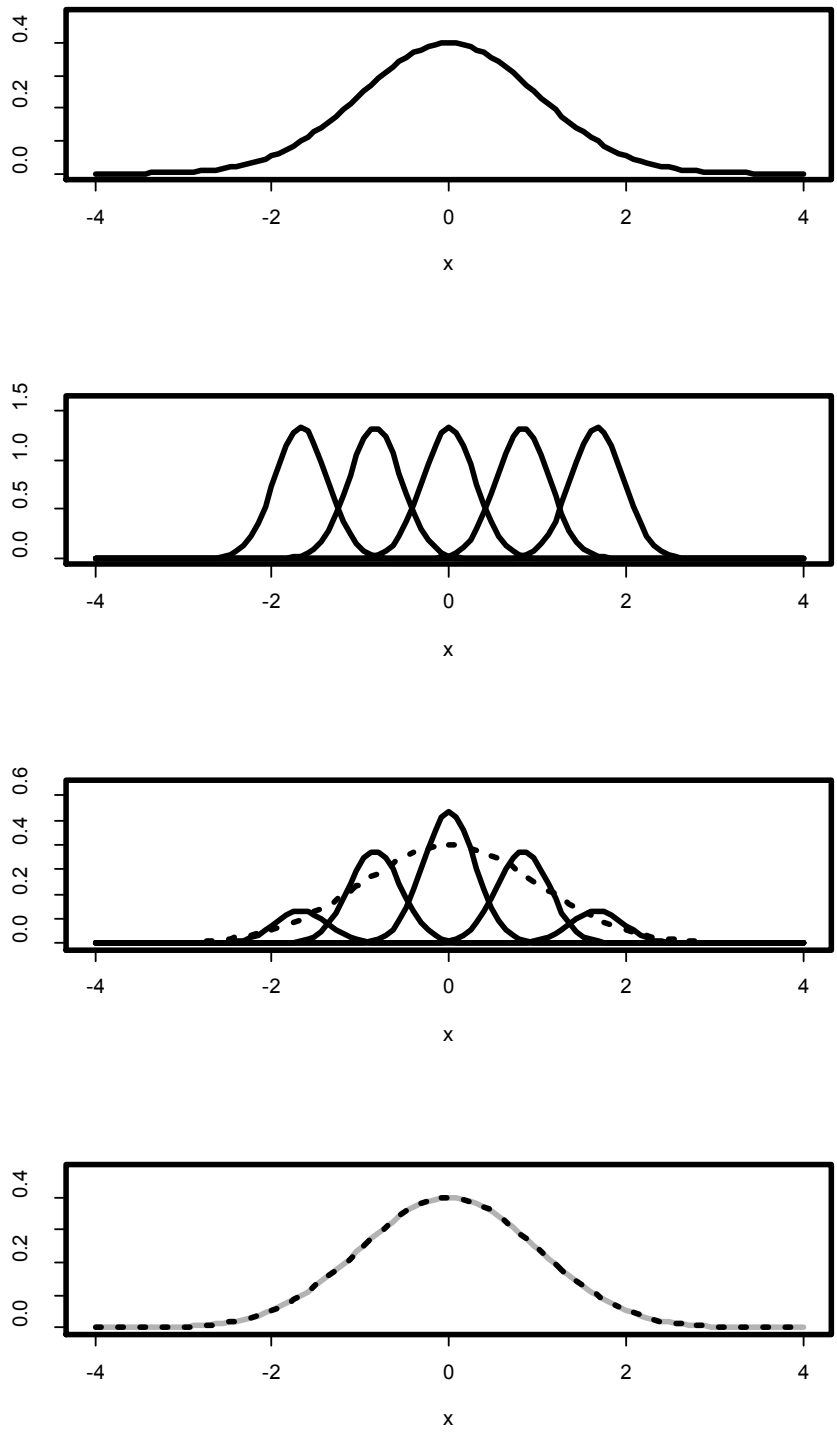

그림 2 평균으로의 회귀 현상이 일어날 때: 부모의 키의 분포, 부모의 키가 -2, -1, 0, 1, 2일 때 자식의 키의 분포, 자식의 키의 분포(점선), 부모의 키의 분포(실선)과 자식의 키의 분포(실선)

따라서, 자식의 무게가 하나의 값으로 정해지지 않고 일정한 분산의 정규분포를 띄는 상황에서 자식 세대의 분포가 부모 세대의 분포와 같기 위해서는 “평균으로의 회귀” 현상이 일어날 수 밖에 없다.[4)]
설명변수(X)과 종속변수(Y)가 모두 동일한 정규분포(평균과 분산이 같은)를 따를 때, 예측 오차에 의해 “평균으로의 회귀 현상”이 필연적으로 일어난다는 것을 염두해 둔다면, "평균으로의 회귀“를 해석하는 데 좀 더 신중할 필요가 있다.

대부분의 심리학적 테스트나 측정값은 신뢰도가 0.7~0.9 정도이다. 여기서 신뢰도가 1보다 낮다는 것은 측정값이 실제값을 중심으로 넓게 분포되어 있다는 의미이다.[5)] 따라서, 같은 테스트나 측정을 시간을 달리하여 실시하면, 그 값은 일정하지 않다.[6)] 그렇게 신뢰도가 1보다 낮은 테스트를 시간을 두고 2번 실시했을 때, 첫 번째 결과와 두 번째 결과의 분포가 동일한 평균과 분산을 갖는 정규분포라면, “평균으로의 회귀” 현상으로 어쩔 수 없이 일어난다.

구체적인 예를 들어 보자. 여기 A라는 시험이 있다. 이 시험의 신뢰도는 0.8이다. 그 말은 이 시험을 (시험 내용을 다 잊어 버릴 만큼) 충분히 긴 시간적 간격을 두고 두 번 실시했을 때, 첫 번째 시험 결과와 두 번째 시험 결과의 상관계수가 0.8이라는 얘기다. 이제 100명의 학생들에게 이 시험을 충분한 시간 간격을 두고 실시하였다. 결과를 보니, 첫 번째 시험 성적의 분포와 두 번째 시험 성적의 분포는 그리 다르지 않았다. 그리고 첫 번째 시험 성적을 통해 두 번째 시험 성적을 예측하는 회귀 분석을 해 보니, 기울기가 1보다 작았다. 그 말은 첫 번째 점수가 평균보다 높았던 학생들은 두 번째 점수가 낮아질 것으로 예측되고, 첫 번째 점수가 평균보다 낮았던 학생들은 두 번째 점수가 높아질 것으로 예측된다는 얘기이다. 그 결과를 본 연구자는 이것이 첫 번째 시험을 잘 못 본 학생들은 열심히 공부를 했고, 첫 번째 시험을 잘 본 학생들은 공부를 게을리 했기 때문에 나온 결과라고 해석하는 할 수도 있다. 하지만, 앞 서 본 바와 같이 시험의 신뢰도가 1보다 낮기 때문에(같은 능력을 가진 학생의 시험 성적이 언제나 동일하게 나오지 않기 때문에) 나타나는 “평균으로의 회귀” 현상일 수도 있다. 즉, 학생들의 능력은 전혀 변하지 않고도 시험의 낮은 신뢰도에 의해 “평균으로의 회귀” 현상이 일어날 수 도 있다.

회귀분석에서 사람들이 궁금해하는 또 다른 현상을 살펴보자. 회귀 분석에서, 대부분의 경우, X값으로 예측한 Y값과, Y값으로 예측한 X값이 일치하지 않는다. 예를 들어서 스위트피의 부모의 무게를 X, 자식의 무게를 Y로 놓고, 회귀 분석을 통해 다음과 회귀선을 얻었

---

4) 사실 평균으로의 회귀 현상은 비단 설명변수와 종속변수의 분포가 동일할 때만 적용되는 것은 아니다. 비록 두 분포가 평균이나 분산이 다른 정규분포를 따를 때에도, 우리는 평균을 뺀 후 표준편차로 나눠줌으로써 두 분포가 모두 $N(0,1)$를 따르도록 변형시킬 수 있다[1)]. 이 때 두 분포는 모두 $N(0,1)$을 따르게 되고, 평균으로의 회귀 현상이 일어난다. 따라서 설명변수와 종속변수의 분포의 평균과 분산이 다를 때, 평균으로의 회귀 현상은 표준 점수의 평균으로의 회귀라고 말할 수 있다.

5) 만약 특정한 테스트가 수리 능력을 측정한다고 치자. 어떤 사람의 수리 능력이 실제로 9인데 반해, 그 사람의 수리능력 테스트 결과는 항상 9가 나오는 것이 아니라, 그 사람의 컨디션, 주변 조건 등에 의해 8.2, 9.2, 8.8 등으로 테스트를 볼 때마다 변할 것이다.

6) 사실 신뢰도는 첫 번째 측정값과 두 번째 측정값의 평균의 위치에 대해서는 말하고 있지 않다. 만약 두 번째 측정값들이 첫 번째 측정값들보다 일정하게 높다면, 신뢰도는 변하지 않는다. 하지만, 회귀 계수 역시 두 값(두 번째 측정값과 첫 번째 측정값)의 차이가 일정한 경우에 영향을 받지 않는다.

다고 치자.(X, Y는 모두 평균 5인 정규분포를 따른다.)

$$Y = 0.8X + 1$$

위의 식을 X에 대해 풀면, 다음이 된다.

$$X = \frac{5}{4}Y - \frac{5}{4} (\text{혹은 } X = 1.25Y - 1.25)$$

하지만 실제로 X를 Y에 회귀분석을 시키면, 다음과 같은 결과가 나온다.

$$X = 0.8Y + 1$$

이 현상을 어떻게 설명할 수 있을까?

처음 얻은 회귀식 $Y = 0.8X + 1$는 특정한 X(부모의 무게)가 주어졌을 때, Y(자식의 무게)의 기댓값(혹은 특정한 X일 때, Y의 분포의 평균)을 나타낸다. 부모의 무게가 1인 경우에 자식의 무게의 분포를 살펴보면, 그 기댓값(평균)이 1.8이라는 의미이다. 그리고 회귀식 $X = 0.8Y + 1$ 역시 특정한 Y(자식의 무게)가 주어졌을 때, X(부모의 무게)의 평균(기댓값)을 나타낸다.

차이는 여기에 있다. 우리가 흔히 $Y = 0.8X + 1$이면, $X = \frac{5}{4}Y - \frac{5}{4}$라고 생각하는 이면에는 X과 Y값이 하나의 값으로 주어졌을 때이다. 만약 X, Y값이 $Y = 0.8X + 1$를 만족한다면, 반드시 $X = \frac{5}{4}Y - \frac{5}{4}$을 만족하게 되어 있다. 하지만 회귀직선 $Y = 0.8X + 1$를 생각할 때, 상관계수가 1이 되는 매우 극히 드문 경우를 제외하고는 (X, Y)의 값이 모두 $Y = 0.8X + 1$를 만족하지 않는다.

질문을 다시 상기해 보자. 왜 특정한 X값으로 예측한 Y값으로 X값을 다시 예측하면 원래의 X값으로 되돌아 오지 않는가? 구체적으로 $X = 3$일 때, Y의 기댓값은 3.4이다. 그리고, $Y = 3.4$일 때, X의 기댓값은 3.72이어서, 처음의 $X = 3$과 같지 않다.

그것은 앞 서 설명했던 것처럼 회귀직선이 의미하는 바가 일대일 대응이 아니라, 집단의 중심을 나타내기 때문이다. $X = 3$일 때, Y의 분포는 3.4를 정점으로 넓게 퍼져 있다. 하지만, $X = 3.1$일 때에도 Y의 분포는 3.48을 중심으로 퍼져있어서, $Y = 3.4$를 포함하고 있다. 그리고 X의 평균이 5이기 때문에, $X < 3$일 확률보다 $X > 3$일 확률밀도가 더 높아서, $P(Y = 3.4|X < 3)$보다 $P(Y = 3.4|X > 3)$이 더 높을 수도 있다! 따라서 $X = 3$일 때 Y의 기댓값이 3.4라고 해서, $Y = 3.4$를 만족하는 X값에서 가장 높은 확률을 가진 값이 3이라고 장담할 수 없는 것이다.(정규분포에서 평균, 혹은 기댓값은 확률밀도가 가장 높은 값이다.)

결론적으로 특정한 X값으로 예측한 Y값으로 X값을 다시 예측하면 원래의 X값으로 되돌

아 오지 않는 것은 특정한 X값일 때, 가능한 Y값이 하나로 정해지지 않고, 넓게 퍼져 있기 때문이다! 만약 그것이 넓게 퍼져 있지 않고, 하나의 값으로 고정된다면, 특정한 X값으로 예측한 Y값으로 X값을 다시 예측하면 원래의 X값으로 되돌아온다!(사실 이것은 일대일 대응인 상황이고 상관계수가 1인 특수한 상황이다!)

<요약>

평균으로의 회귀 현상은 설명변수, 종속변수, 그리고 (설명변수의 값이 주어졌을 때,) 조건부 종속 변수의 분포가 정규분포를 따를 때, 필연적으로 일어난다. 설명변수와 독립변수를 바꿔서 회귀분석을 했을 때, 원래의 회귀식과 일치하지 않는 이유는 회귀식이 일대일 대응을 의미하는 것이 아니기 때문이다.