심슨의 패러독스 개요

일반적으로 데이터 세트에서 내리는 결정은 데이터 세트에 적용하는 통계 측정 결과의 영향을 받습니다. 이러한 출력은 상관 관계 유형과 데이터 세트의 기본 시각화에 대해 알려줍니다. 그러나 때때로 데이터를 그룹으로 분리하여 통계적 측정값을 적용할 때 또는 함께 집계한 다음 통계적 측정값을 적용할 때 결정이 다릅니다. 동일한 데이터 세트의 결과에서 이러한 종류의 비정상적인 동작을 일반적으로 Simpson's paradox 라고 합니다. 간단히 말해서 Simpson의 역설은 데이터를 그룹으로 분리할 때와 데이터를 집계할 때의 두 가지 상황에서 데이터 세트를 분석할 때 분석 추세에 나타나는 차이입니다.

다음은 남성과 여성이 개별적으로 또는 결합한 두 개의 서로 다른 게임 콘솔에 대한 권장 비율을 나타내는 표입니다.

	추천 **PS4**	추천 엑스박스 원
남성	50/150=30%	180/360=50%
여자	200/250=80%	36/40=90%
결합	250/400=62.5%	216/400=54%

위의 표는 PS4와 Xbox One이라는 두 가지 게임 콘솔의 남성과 여성의 개별 및 결합 권장 비율을 보여줍니다.

최대 권장 사항이 있는 게임 콘솔을 구입한다고 가정합니다. 앞의 표에서 볼 수 있듯이 Xbox One은 PS4보다 남녀 모두 더 높은 비율로 권장됩니다. 그러나 동일한 데이터를 결합하여 사용할 경우 모든 사용자에 따르면 PS4의 권장 비율(62.5%)이 더 높습니다. 그래서, 당신은 어떤 것과 함께 갈 것인지 어떻게 결정할 것입니까? 계산은 괜찮아 보이지만 논리적으로 의사 결정이 좋지 않은 것 같습니다. 이것이 심슨의 역설입니다. 여기에서 동일한 데이터 세트는 두 가지 상반된 주장을 증명합니다.

음, 이 경우 주요 문제는 개별 데이터의 백분율만 볼 때 샘플 크기를 고려하지 않는다는 것입니다. 각 분수는 묻는 숫자만큼 게임기를 추천할 사용자의 수를 나타내므로 전체 샘플 크기를 고려하는 것이 적절합니다. 남성과 여성의 별도 데이터의 표본 크기는 많은 차이가 있습니다. 예를 들어, PS4의 남성 권장 사항은 50이고 Xbox One의 권장 사항은 180입니다. 이 숫자에는 큰 차이가 있습니다. Xbox One은 여성보다 남성의 반응이 훨씬 더 많지만 PS4의 경우는 그 반대입니다. 플레이스테이션을 추천하는 남성이 적기 때문에 데이터를 합치면 PS4에 대한 평균 평점이 낮아져 역설로 이어진다.

어떤 콘솔을 사용해야 하는지 단일 결정을 내리기 위해서는 데이터를 결합할 수 있는지 또는 별도로 봐야 하는지를 결정해야 합니다. 이 경우 남성과 여성 모두를 만족시킬 가능성이 가장 높은 콘솔을 찾아야 합니다 . 이러한 리뷰에 영향을 미치는 다른 요소가 있을 수 있지만 이 데이터가 없으므로 성별 편견에 관계없이 좋은 리뷰의 최대 개수를 찾습니다. 여기서 데이터를 집계하는 것이 가장 합리적입니다. 우리는 리뷰를 결합하고 전체 평균과 함께 갈 것입니다. 우리의 목표는 리뷰를 결합하고 총 평균을 보는 것이기 때문에 데이터 집계가 더 합리적입니다.

Simpson의 역설은 이론적으로 가능하지만 전체 사용 가능한 데이터에 대한 통계 분석이 정확 하기 때문에 실제로는 발생하지 않는 억지스러운 문제인 것 같습니다 . 그러나 현실 세계에는 심슨의 역설에 대한 잘 알려진 연구가 많이 있습니다.

심슨의 패러독스에서 얻을수 있는 교훈은 데이터만으로는 충분하지 않다는 것입니다. 데이터는 순전히 객관적이지 않으며 최종 플롯도 아닙니다.

따라서 우리는 데이터 세트를 다룰 때 전체 스토리를 얻고 있는지 여부를 고려해야 합니다.

논의를 통해 부분적인 것에 대한 결론을 종합하여 전체에 대한 결론을 얻는 것이 합리적이고

타당한 방법이며 이것이 논리와 직관이 충돌을 일으킬 때 문제를 해결하는 유일하고도 효과적인 방법입니다.