SIMPSON PARADOX

Sampson Paradox is an example of one of the counter intuitive properties of probability distributions. It states that, “*The trend or result that appears in several different groups of data but it reverses or disappears when the data is combined.*” The main cause of Simpson Paradox is due to unhandled heterogeneity of data.

Simpson Paradox can be expressed using probability function. Let's consider a small example to understand the inequality generated due to Simpson Paradox.
Suppose, we are trying out two new flavours of Red Bull i.e “Tropical” and “Green Apple” and for evaluating which one is better we perform a small experiment of asking opinions of 1000 people(Men and Women Combined).

Red Bull Flavour	**Sample Size**	**#Liked Flavour**
Tropical	1000	800
Green Apple	1000	700

Here, the probability of liking Tropical flavour is more than people liking Green Apple. Such that, **P(G|P) < P(T|P) ;** where G=Green Apple, T=Tropical and P= Total People.
But, if we split the people into Man and Woman, there exists few changes in the probability of liking the flavour.

Red Bull Flavour	**Men**	**#Liked Flavour(Men)**	**Women**	**#Liked Flavour(Women)**
Tropical	900	760 (84.5%)	100	40 (40%)
Green Apple	700	600 (85.7%)	300	150 (50%)

As we can see, by splitting the people into Men and Women, the probabilities of Men and Women liking Green Apple flavour increases than Men and Women liking Tropical flavour due to mismatch sample sizes. Such that, **P(G/M) > P(T/M) and P(G|W) > P(T|M) ;** where M= Men ,W=Women.

Hence, it can be proved that the probability or result which occurred in the combined dataset reverses when data is split or exists separately.