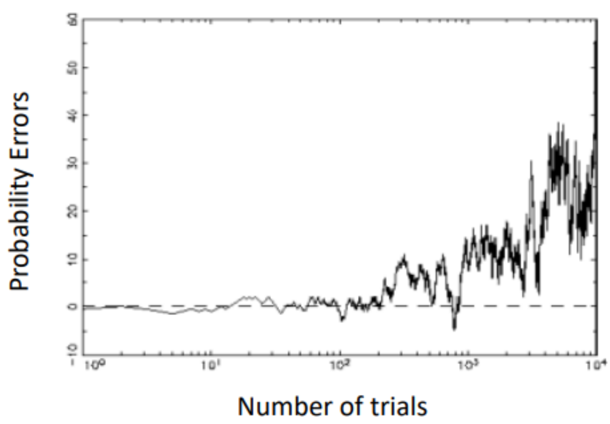
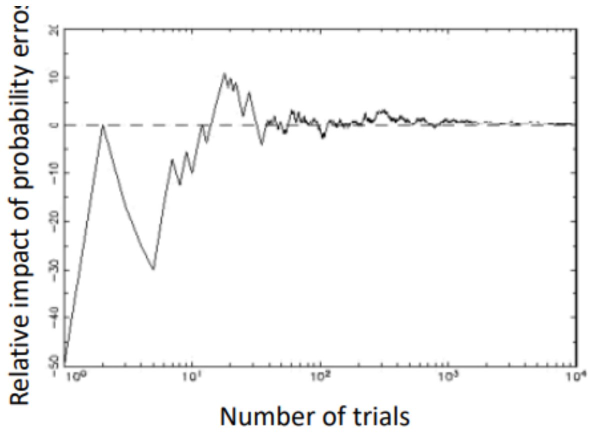
Probability error(확률 오차)

관측값과 예측값 사이의 차이 ex) 동전을 10번 던져서 앞면이 나올 횟수를 5라고 예측하겠지만, 실제 관측값은 6일 수 있다. 이 경우 probility error는 +1이다.

- The law of large numbers (거대수의 법칙) -시행 횟수가 아주 많아질수록 확률 오차의 절댓값은 커진다. (계속해서 누적되므로) -시행 횟수가 아주 많아질수록 이론적인 확률(예측값)에 가까워지므로, 확률 오차의 상대적인 영향력은 줄어든다.





Expected Value(기댓값)

어떤 확률 과정을 무한히 반복했을 때 얻을 수 있는 값들의 평균으로 기대하는 값
$E[X] = \sum_i x_i p_i$, where x_i represents the ith possible value of X, and p_i is the probability of X taking the value x_i.

ex1) 동전을 무한히 던져서 나오는 면(앞:1, 뒤:0)의 기댓값 구하기 X(확률변수): 동전을 던져서 나오는 면 xi: 1, 0 pi: 앞면, 뒷면 모두 0.5 E[X]=(1×0.5)+(0×0.5)=0.5

ex2) 주사위를 무한히 던져서 나오는 눈의 기댓값 구하기 X: 주사위를 던져서 나오는 눈 xi: 1, 2, 3, 4, 5, 6 pi: 모두 1/6 E[X]=(1×(1/6) + 2×(1/6) + 3×(1/6) + 4×(1/6) + 5×(1/6) + 6×(1/6)) = 3.5

- Mean(평균)과의 차이??

기댓값과 평균은 결과적인 계산값은 같지만 자료에 대한 관점이 다르다. 평균은 **이미 나와있는 정확한 자료**에 대해 그 값을 모두 더하여 도수로 나눈 값이지만, 기댓값은 주어진 사건에 대한 확률을 반영하는 시행에 대하여 평균으로 **기대하는 값**을 말한다.

Standard Error(표준 오차, SE)

수많은 표본들 통계의 편차를 구함으로써 모수의 통계를 추정하는데, 대부분 평균을 다루기 때문에 SEM이라고도 한다. 아래는 SEM에 대한 공식이다.

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

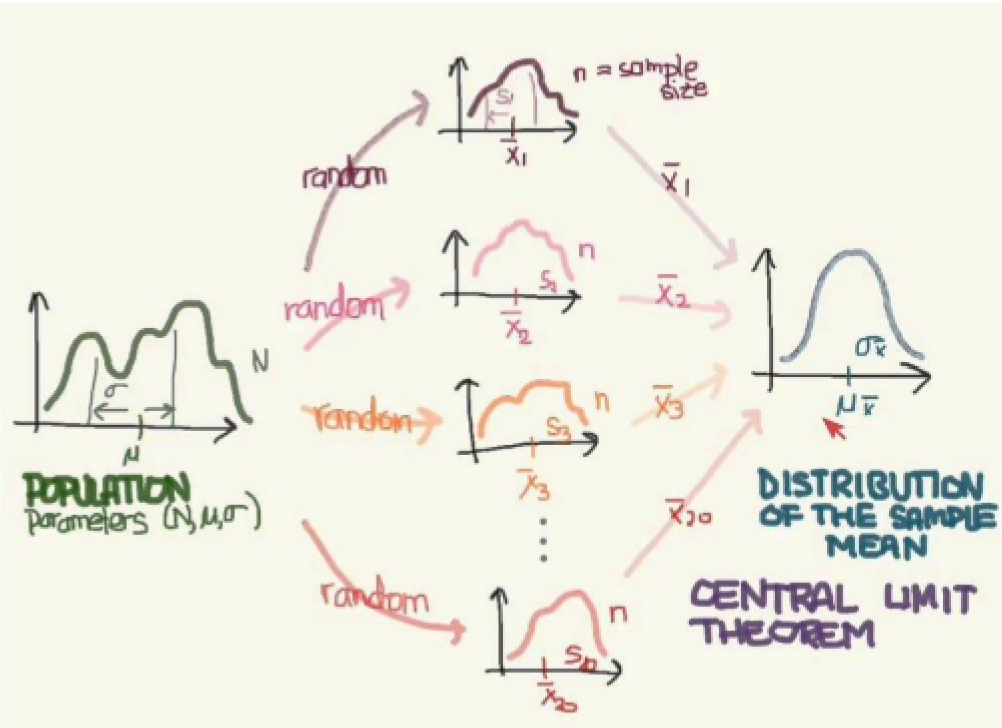
$$\text{variance} = \sigma^2$$

$$\text{standard error } (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

where:

$\bar{x}$ = the sample's mean

n = the sample size



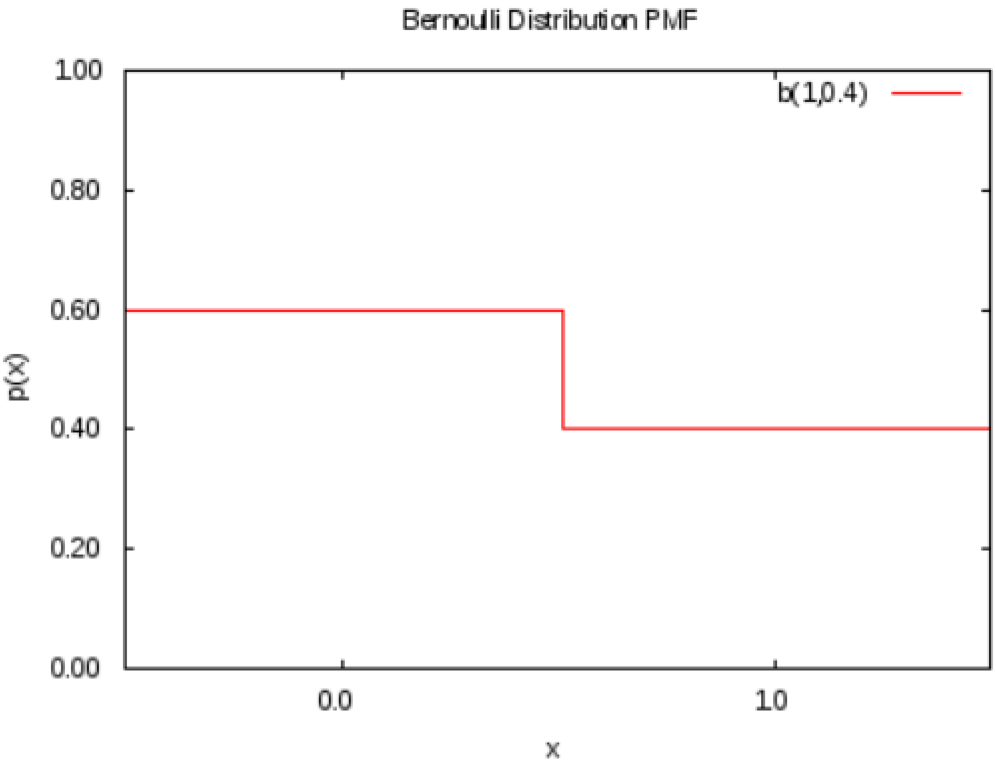
ex) 주사위를 36번 던지는 경우의 SEM 주사위를 한 번 던질 때의 표준편차 σ =√{(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2}/6 ≈ 1.7078 **SEM** = 1.7078/√36 = 0.2846

=> 주사위를 던질 때 3.5가 나올 가능성이 가장 크고, 그 위아래로 0.28 떨어진 값이 나올 가능성이 크다.

Binomial Distribution(이항 분포)

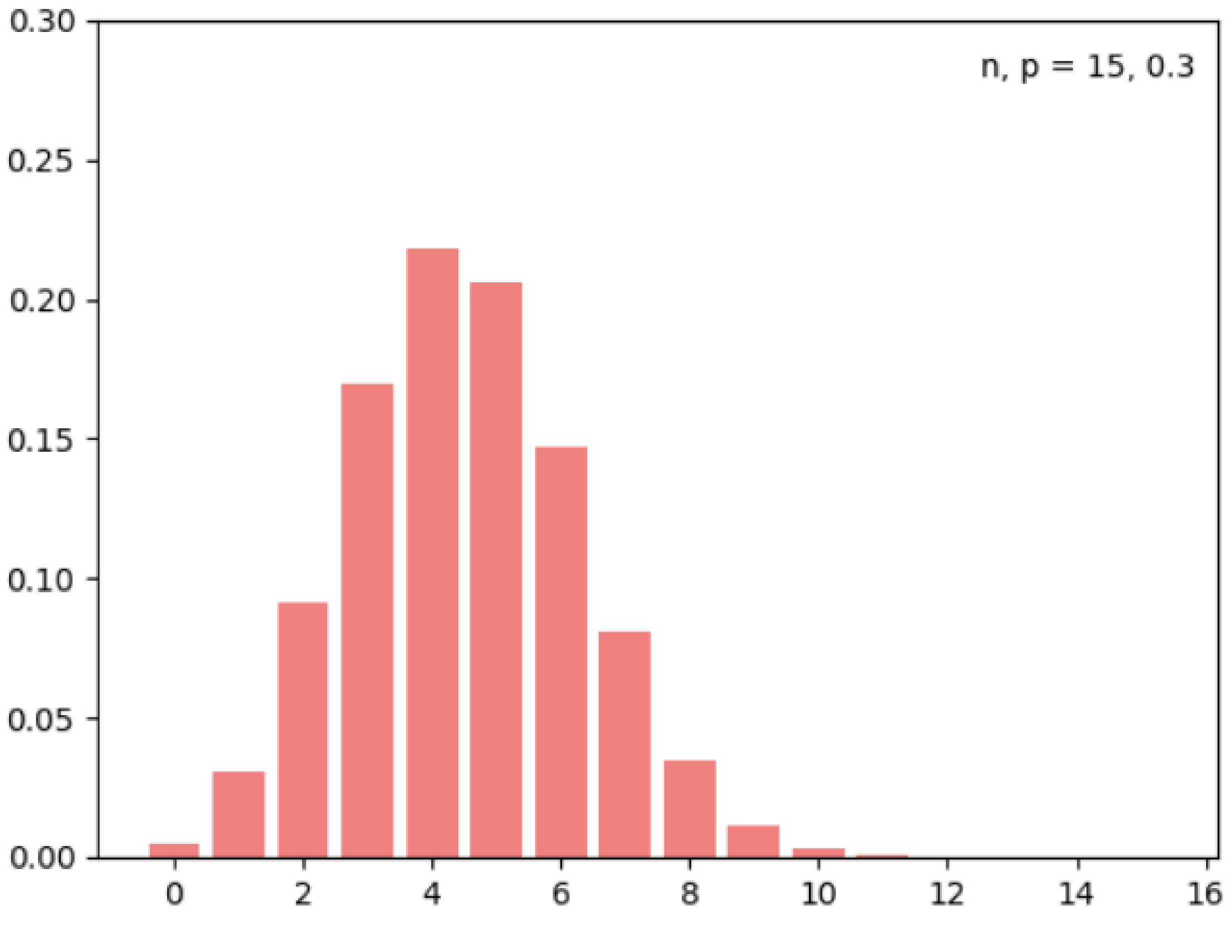
Bernoulli trials: 결과가 2가지 중 하나로만 나오는 실험 ex)증가/감소, 성공/실패, 생존/사망 **Bernoulli distribution**: 베르누이 시행의 결과를 나타내는 이산 확률 분포로, 결과가 0이나 1로 표현됨 -1개의 매개변수

p(성공확률, 1-p는 실패확률)로 정의



Binomial distribution(이항 분포): 베르누이 시행을 독립적으로 *반복*하여 결과를 관찰하는 경우에 사용되는 확률 분포 -베르누이 분포에서 확률변수 X 가 성공 1, 실패 0 두 개의 값만을 갖는다면, 이항 분포에서 확률변수 X는 성공의 횟수를 가짐 -2개의 매개변수 p(성공확률), n(시행횟수)으로 정의 -**X~B(n,p)**로 표현

- 이항분포의 확률질량함수



$$f(x) = P(X=x) = B(n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$= \frac{n!}{x!(n-x)!} \cdot p^x (1-p)^{n-x}$$

- $P(X=x)$는 x번 성공할 확률입니다.
- $\binom{n}{x}$는 이항 계수(binomial coefficient)로, n개의 시행 중에서 x번의 성공이 발생할 수 있는 조합의 수입니다.
- p, $(1-p)$, n는 각각 성공확률, 실패확률, 시행 총 횟수입니다.

ex) 동전을 10번 던지는 시행의 이항 분포

$$k=0: \frac{10!}{0! \cdot 10!} \left(\frac{1}{2}\right)^0 \cdot \left(1-\frac{1}{2}\right)^{10} = 0.0010$$

$$k=1: \frac{10!}{1! \cdot 9!} \left(\frac{1}{2}\right)^1 \cdot \left(1-\frac{1}{2}\right)^{9} = 0.0098$$

$$k=2: \frac{10!}{2! \cdot 8!} \left(\frac{1}{2}\right)^2 \cdot \left(1-\frac{1}{2}\right)^{8} = 0.0439$$

$$\vdots$$

$$k=10: \frac{10!}{10! \cdot 0!} \left(\frac{1}{2}\right)^{10} \cdot \left(1-\frac{1}{2}\right)^{0} = 0.0010$$

X가 이항분포를 따르는 확률변수일 때 E(X)와 Var(X) 구해보기

- 우선 첫 번째 시행(X1)에 대한 평균과 분산을 구한다.
  $P(X_1 = x) = p^x(1-p)^{(1-x)}, \qquad x = 0, 1$

  $E(X_1) = \sum_x xP(X_1 = x) \quad = 0 \times (1-p) + 1 \times p = p$

  $Var(X_1) = E(X_1 - E(X_1))^2 = \sum (x-p)^2 P(X_1 = x) \quad = (0-p)^2 (1-p) + (1-p)^2 p = p(1-p)$
- X1, X2, ..., Xn에 대한 평균과 분산을 구한다.

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = nE(X_1) = np$$

$$Var\,(X) = Var(X_1 + X_2 + \cdots + X_n) = nVar(X_1) = np(1-p)$$

Ex) 동전을 100번 던지는 시행의 E(X)와 V(X) 구하기

1. 동전을 1번 던질 때 앞면이 나올 확률은? P(X1=1) = (1/2)^1(1-1/2)^(1-1) = 1/2 = 0.5

2. 동전을 1번 던질 때의 기댓값은? E[X1] = p = 1/2 = 0.5

3. 동전을 1번 던질 때 앞면이 나올 횟수의 분산은? Var(X1) = p(1-p) = 1/2×1/2 = 1/4 = 0.25

4. 동전이 앞면이 나올 횟수의 기댓값은? E(X) = np = 100×0.5 = 50

5. 동전이 앞면이 나올 횟수의 분산은? Var(X) = np(1-p) = 100×0.5×0.5 = 25

=> 동전 던지기를 100회 시행했을 때 앞면이 나올 횟수의 기댓값은 50 정도이고, 그 분포는 25이다.

Sample Size에 따른 MSE 감소

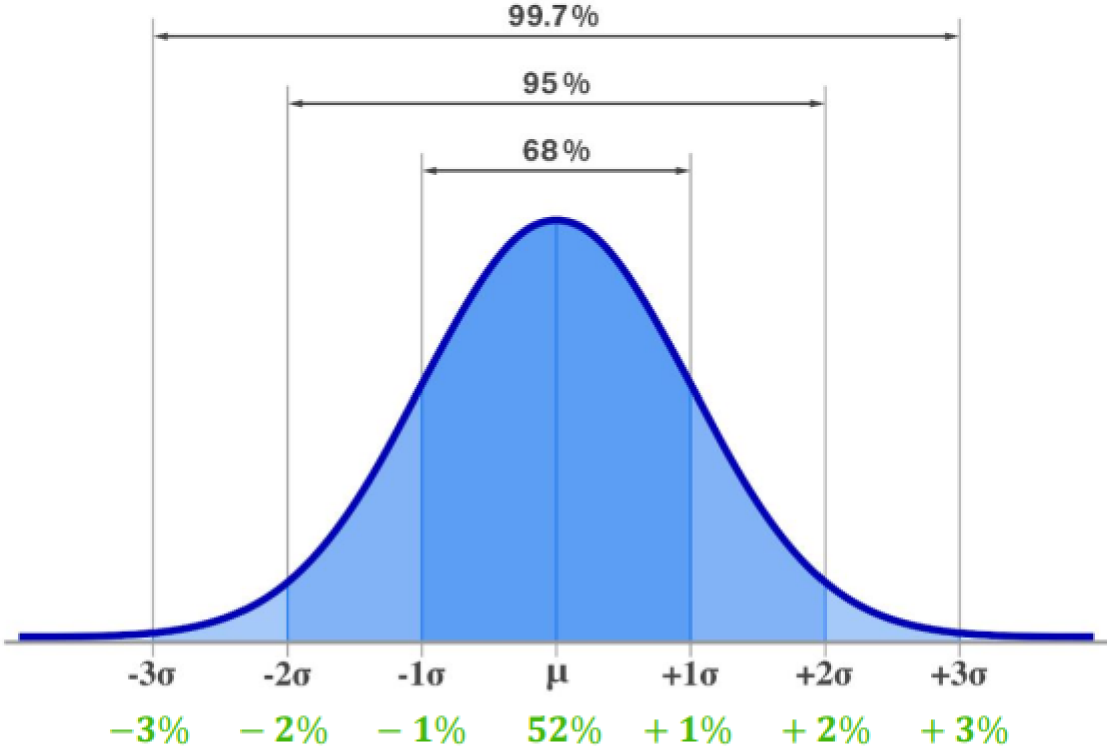
sample size가 클 수록 MSE는 감소한다. Ex) 100,000명의 투표자 중 100명의 투표자를 샘플링해서 조사했더니 특정 후보자를 뽑을 확률이 53%라는 것을 알게 되었다. 이 데이터로 우리는 이 후보자가 선거에서 이길 것이라고 확신할 수 있을까?

- X1은 베르누이 시행을 따른다. Var(X1) = p(1-p) = 0.53 × 0.47 = 0.2491 Std(σ) = √0.2491 = 0.4991 SEM = σ/√n = 0.0499 (약 5%) -> 후보자의 우승 확률 = 48~58% -> 우승은 보장되지 않음

Ex) 같은 상황에서, 100명이 아니라 2,500명의 투표자를 샘플링해서 조사했더니 특정 후보자를 뽑을 확률이 52%라는 것을 알게 되었다. 이 경우에는 후보자의 승리를 확신할 수 있을까?

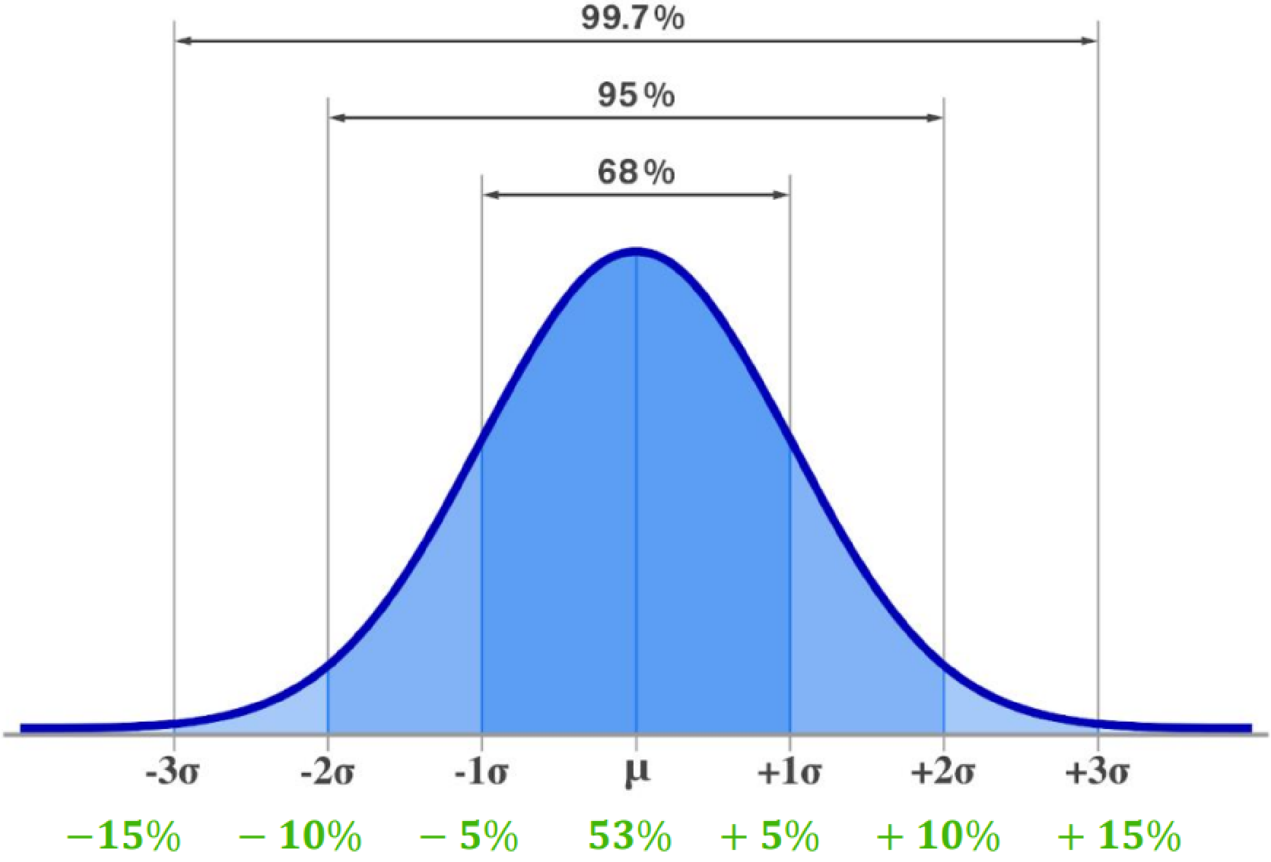
- Var(X1) = p(1-p) = 0.52 × 0.48 = 0.2496 Std(σ) = √0.2496 = 0.4996 SEM = σ/√n = 0.01 (약 1%) -> 후보자의 우승 확률 = 51~53% -> 우승 보장!

Confidence Interval(신뢰 구간)



- 후보자가 당선될 확률이 51~53%라는 것에 68%의 자신감이 있다.
- 후보자가 당선될 확률이 50~54%라는 것에 95%의 자신감이 있다.
- 후보자가 당선될 확률이 49~55%라는 것에 99.7%의 자신감이 있다.

- ex1의 경우, 95%의 자신감이 있는 신뢰 구간은? -> 43~63%



Bootstrap

모수를 *복원*하면서 *반복적*으로 샘플링하는 것-> 모수를 다 알지 못할 때 유용

Simpson's paradox

데이터를 그룹으로 나누어서 봤을 때의 trend가 데이터 전체로 봤을 때는 반전되거나 나타나지 않는 현상