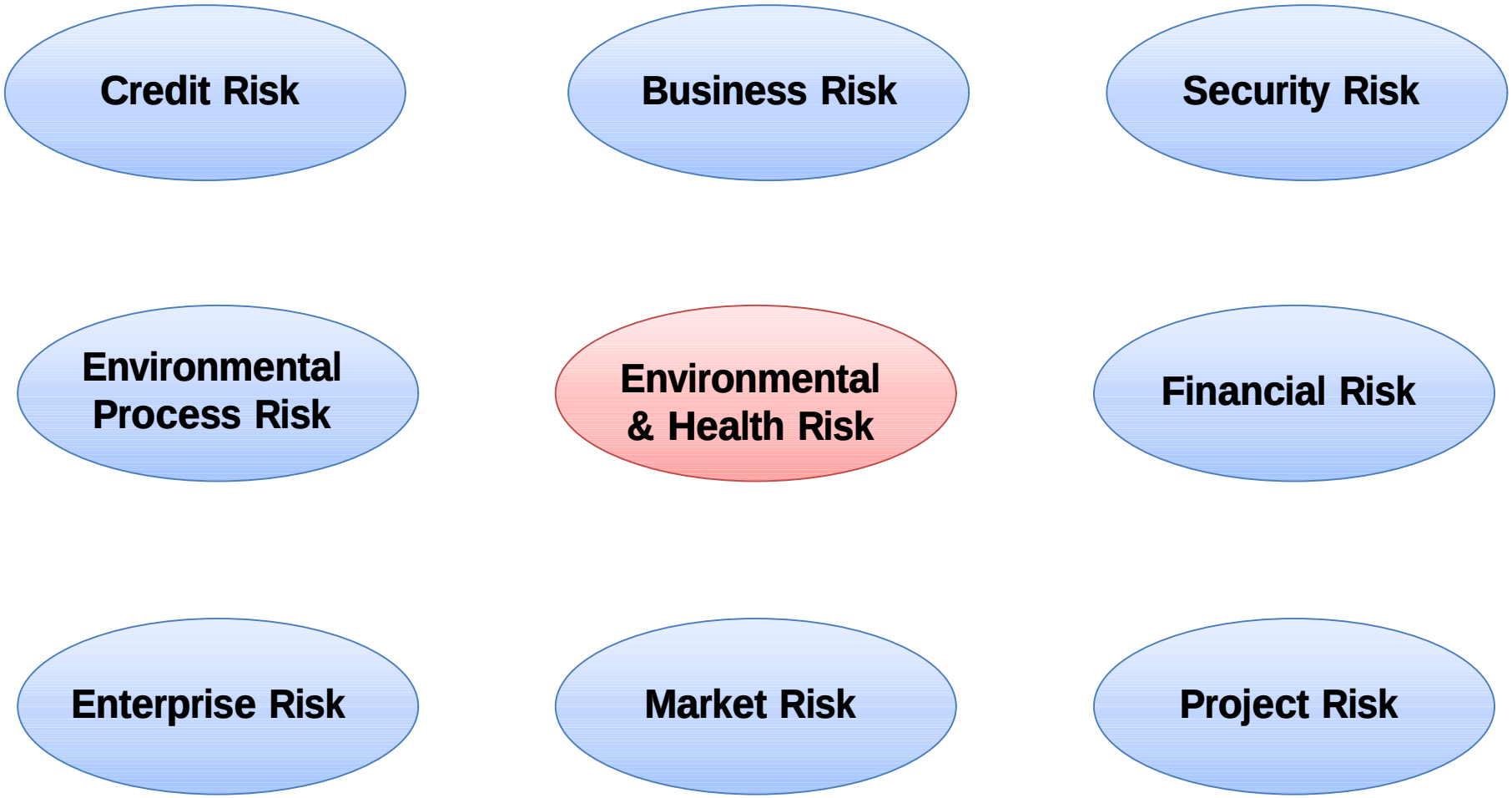
위해성 평가를 위한 불확실성의 이해

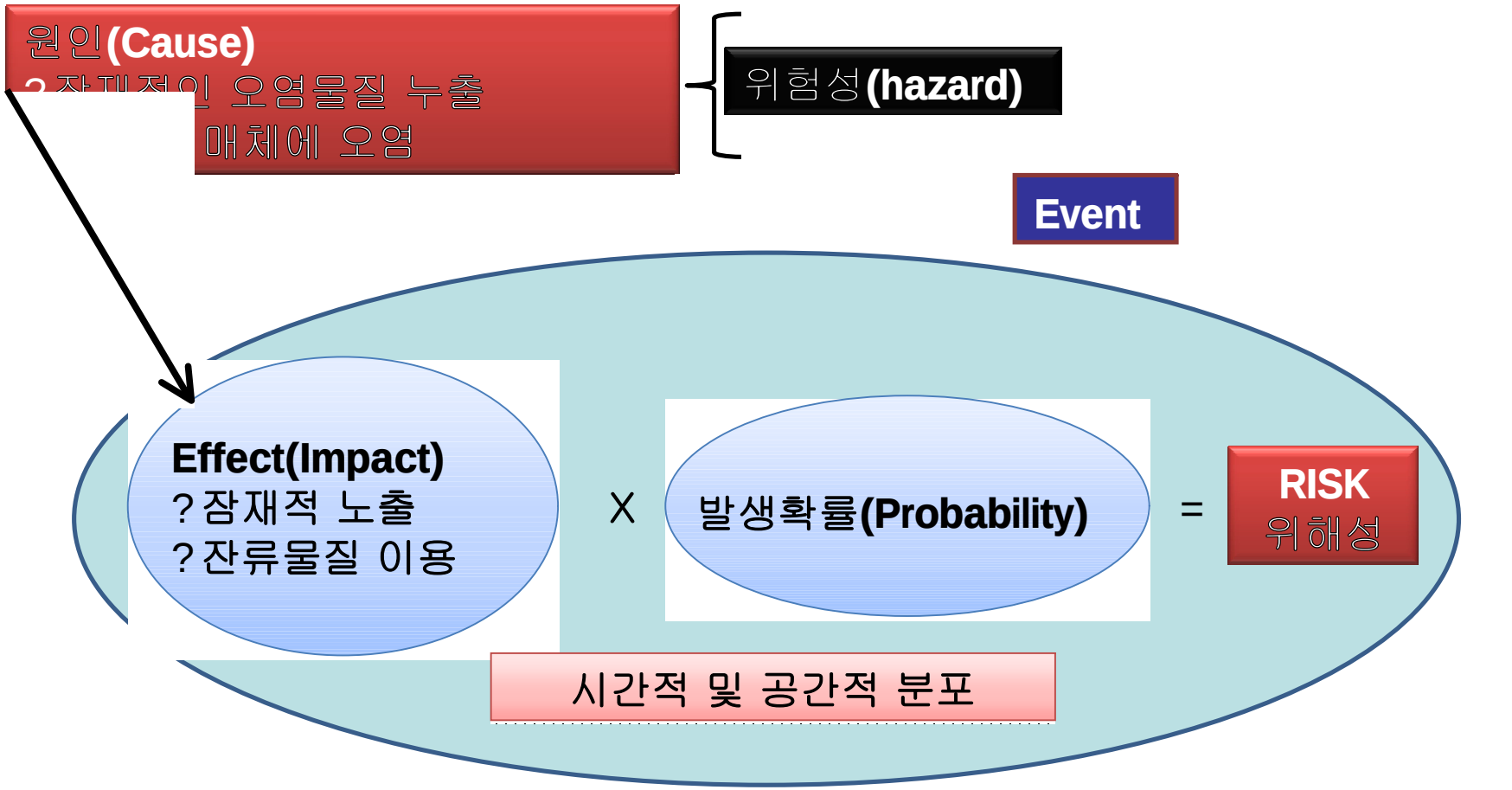
김호
서울대학교 보건대학원

리스크(Risk) ¡All Around¡



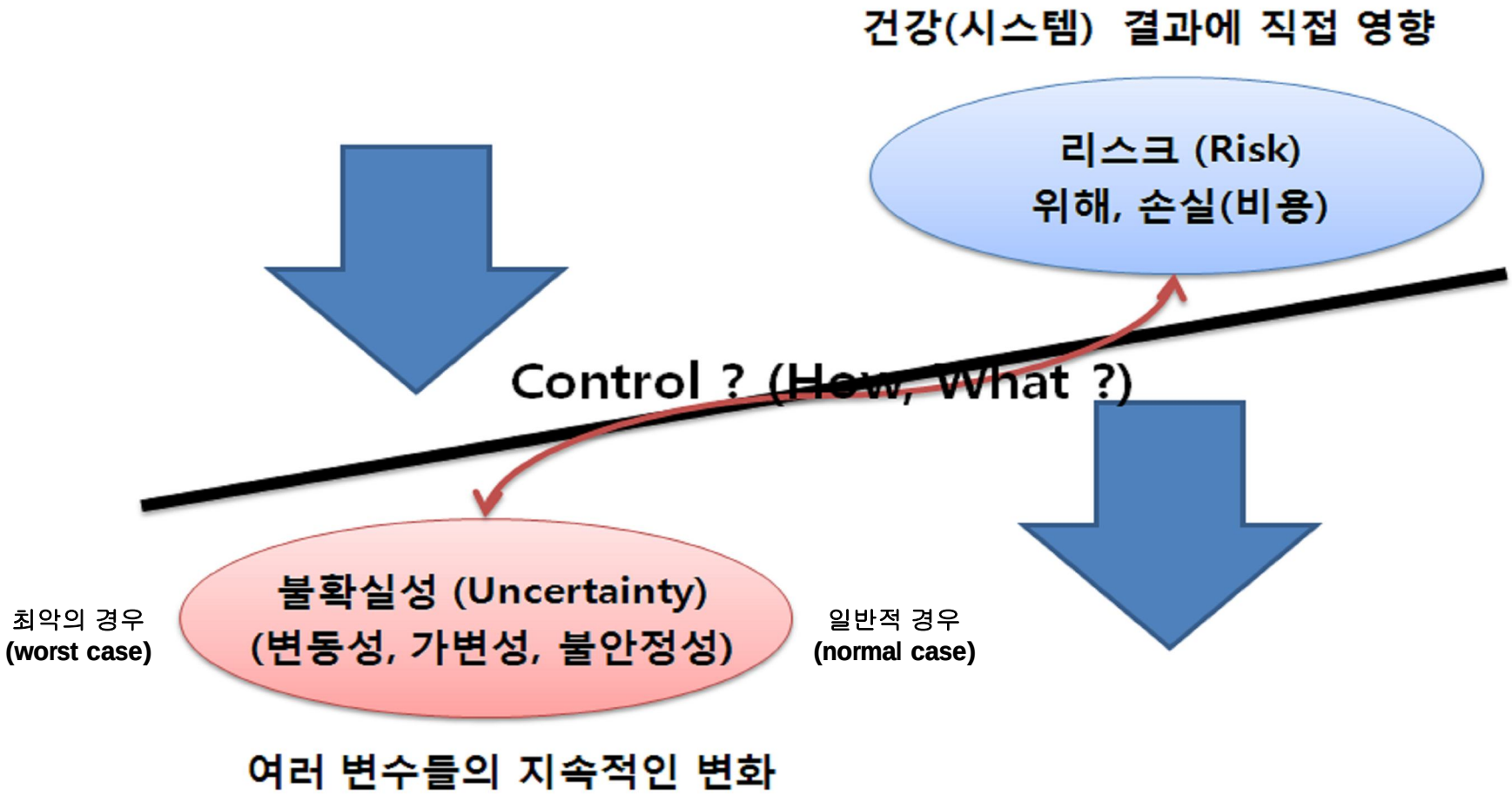
불확실성 환경에서 인간이 활동하는 모든 공간과 시간에
우리가 원하지 않은 리스크가 존재함

리스크(Risk) ?



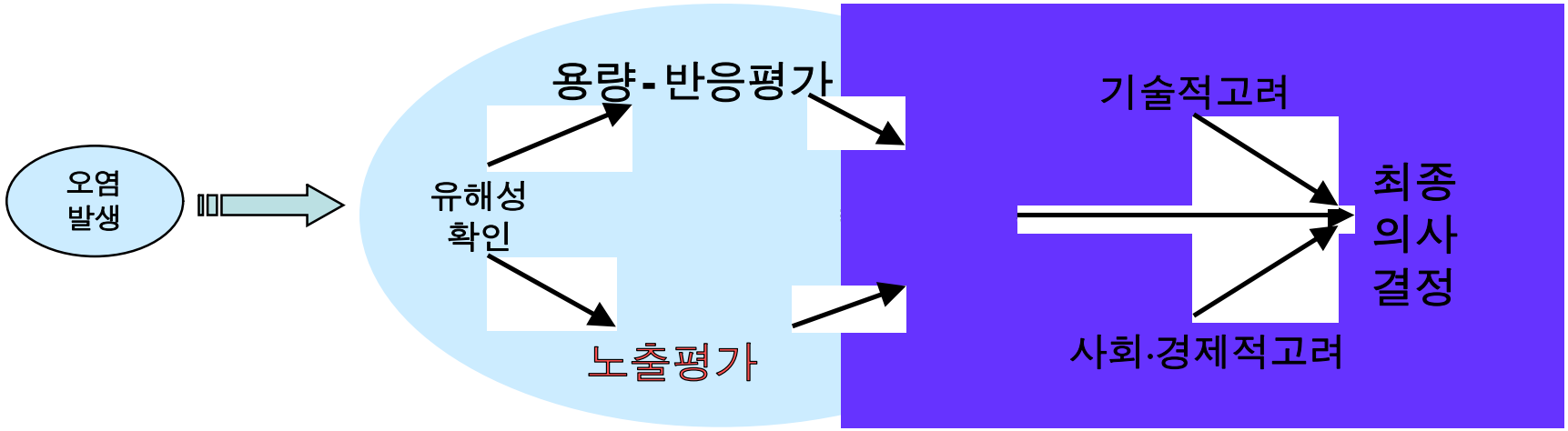
James Kimmance, 2005
U.S. EPA, 1997

불확실성(Uncertainty) vs 리스크(Risk)



불확실성 요인들을 효율적으로 평가 및 분석한다면
미래 잠재적인 리스크(risk)를 줄일 수 있음

위해성 평가 절차



1) Data Collection and Evaluation(유해성 확인)
① 관련 지역 data 수집 및 분석(Gather and analyze relevant site data)
② 잠재적으로 우려되는 오염물질 확인(Identify potential chemicals of concern)

2) Exposure Assessment(노출평가)
① 오염방출농도 분석(Analyze contaminant releases)
② 노출인구 확인(Identify exposed populations)
③ 잠재적 노출경로 확인(Identify potential exposure pathways)
④ 다매체 경로에 대한 노출농도들의 추정(Estimate exposure concentrations for pathways)
⑤ 다매체 경로에 대한 오염물질 섭취 추정(Estimate contaminant intakes for pathways)

3)Toxicity Assessment(용량-반응평가, 독성평가)
① 정량 및 정석적 독성정보 수집(Collect qualitative and quantitative toxicity information)
② 적절한 독성값 결정(Determine appropriate toxicity values)

4) Risk Characterization(위해도 결정)
① 일어날수 있는 건강상의 부작용에 대한 잠재적 위해 결정(Characterize potential for adverse health effects to occur)
② 발암 위해도 추정(Estimate cancer risks)
③ 비발암 위해도 추정(Estimate non-cancer hazard quotients)
④ 불확실성 평가(Evaluate uncertainty)
⑤ 위해 정보 요약(Summarize risk information)

목표위해성을 고려한 노출량 산정식 및 방법

노출시나리오에 따른 지하수 섭취, 경피 및 흡입노출 3가지 산정식(USEPA, 1996)

경구 노출에 의한 섭취 (Ingestion of Groundwater)

$$ADD = \frac{C \times IR \times AAF \times EF \times ED}{BW \times AT \times 365(day/year)}$$

C: chemical specific concentration in drinking water(mg/L)
IR: water ingestion rate(L/day)
AAF: chemical-specific oral-water absorption adjustment factor (mg/mg)
EF: exposure frequency(event/year)
ED : exposure duration (years)
BW: body weight (kg)
AT: average time (yr×day/yr), 70×365(carcinogenic), ED×365(non-carcinogenic)

경피 노출에 의한 섭취 (Dermal intake in the Shower)

$$ADD = \frac{C \times SA \times AAF \times ET \times PC \times EF \times ED}{BW \times AT \times 365(day/year)}$$

C: chemical specific concentration in drinking water(mg/L)
SA: total skin surface area(cm^2)
AAF: dermal-water chemical specific adsorption adjustment factor (mg/mg)
ET: bath or shower duration (hr/day)
PC: chemical-specific skin permeability constant (cm/hr)

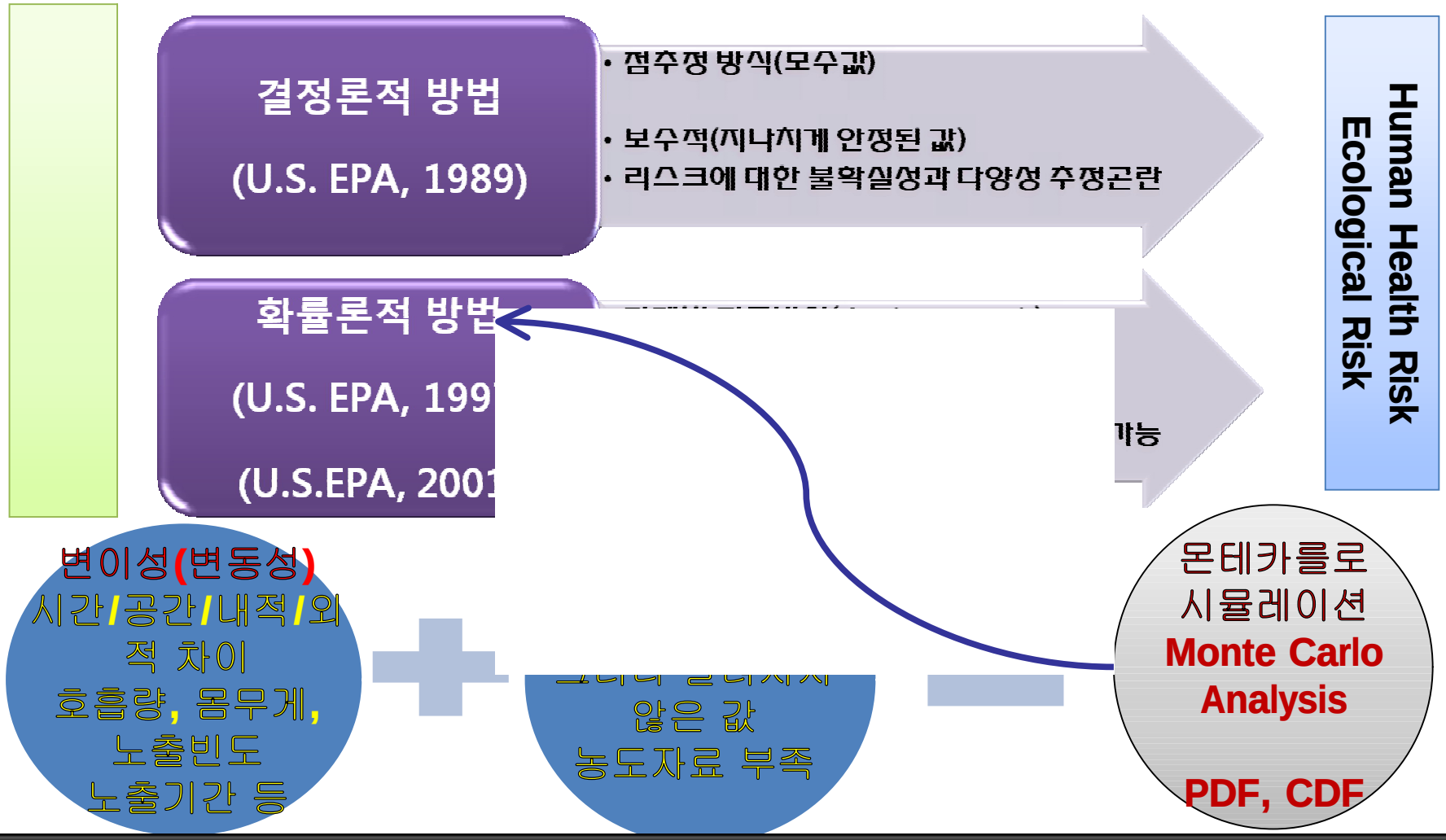
흡입 노출에 의한 섭취 (Inhalation in the Shower) (McKone, 1987; Foster and Chrostowski, 1986)

$$ADD = \frac{C \times InhR \times AAF \times ET \times LRF \times EF \times ED}{BW \times AT \times 365(day/year)}$$

C: chemical specific concentration in bathroom air(mg/m^3)
InhR: inhalation rate while showering (m^3/hr)
ET: shower duration (hr/day)
AAF: chemical-specific inhalation absorption adjustment factor (mg/mg)
LRF: lung retention factor (dimensionless)

- 목표위해성 수준
  발암물질 **(Target risk) : TR <10^{-6}(10^{-5}, 10^{-4})** (단 최대허용치**(RME) TR<10^{-4}**)
  비발암물질 **(Total Hazard Index) = HQ_1 + HQ_2 + ¡ . <1**

확률론적 방법론에 의한 리스크 평가



Risk = $R_{ingestion}$ + $R_{inhalation}$ + $R_{dermal\ contact}$ + ⅰ ⅰ R_{etc}

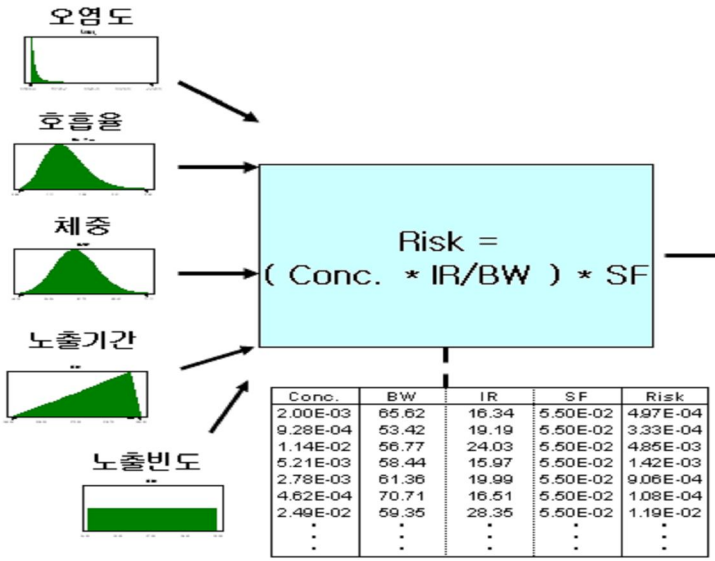
Risk = Prob(U, V) , U:불확실성 변수, **V:**변이성(변동성) 변수

= Prob(U=『농도』**, V=**『물섭취량, 공기호흡량, 피부노출면적, **ET, ED, WB,**

』

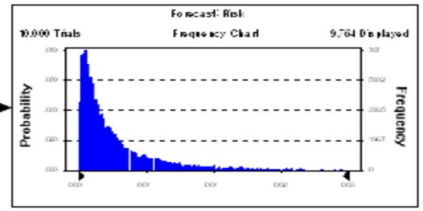
Bogen and Spear, 1987; Cohen et al., 1996)

리스크 계량화 방법

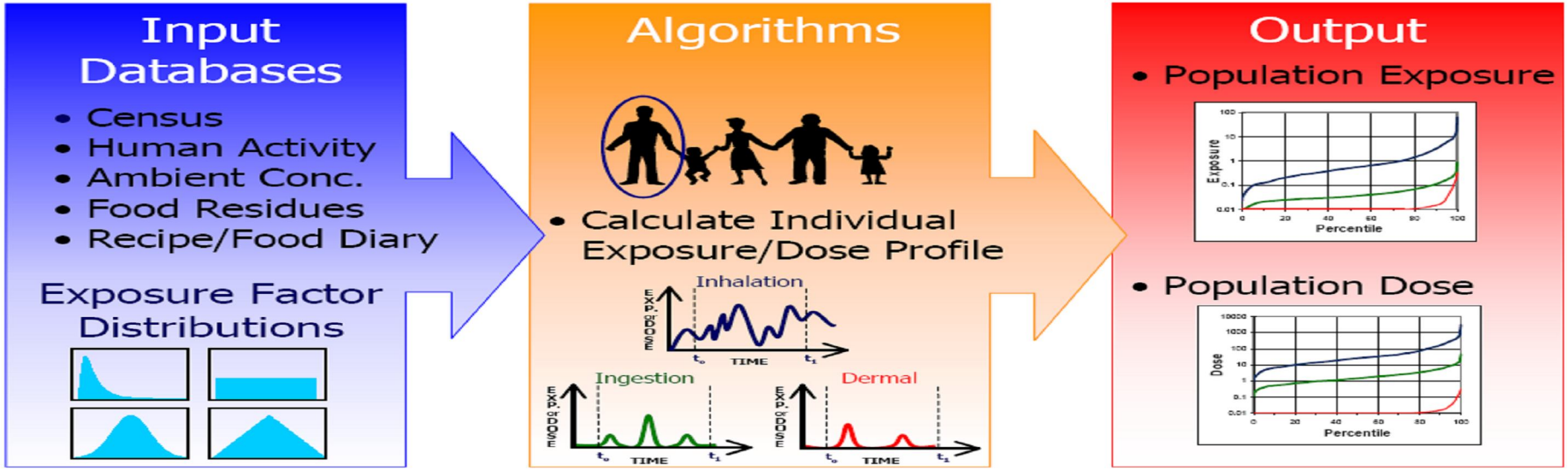


Conc.	BW	IR	SF	Risk
2.00E-03	65.62	16.34	5.50E-02	4.97E-04
9.28E-04	53.42	19.19	5.50E-02	3.33E-04
1.14E-02	56.77	24.03	5.50E-02	4.85E-03
5.21E-03	58.44	15.97	5.50E-02	1.42E-03
2.78E-03	61.36	19.99	5.50E-02	9.06E-04
4.62E-04	70.71	16.51	5.50E-02	1.08E-04
2.49E-02	59.35	28.35	5.50E-02	1.19E-02

인체 확률 위해도 분석
(PRA)

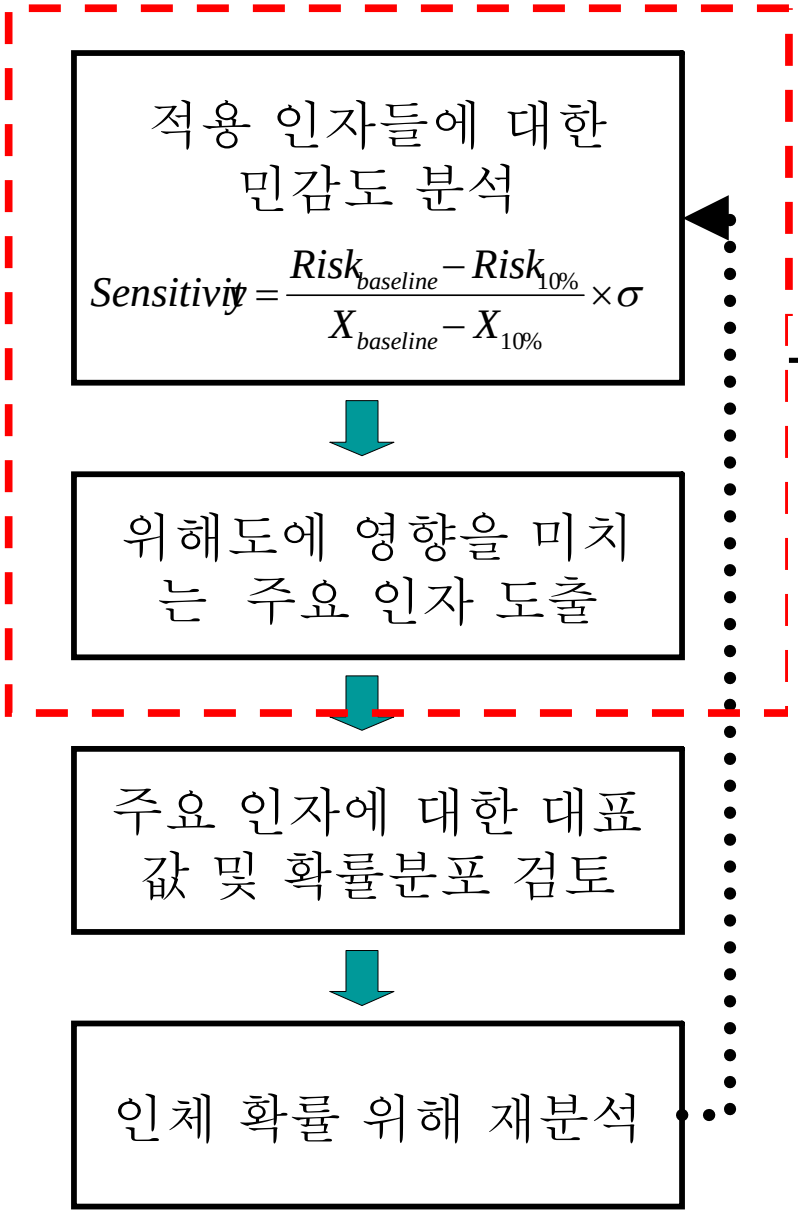


몬테카를로 방법**(Monte Carlo method)**은 난수 **(random number)**를 이용 문제를 푸는 수치적 방법 이용



Guiding Principles for Monte Carlo Analysis (EPA, 1997), EPA, 2001)
Policy for Use of Probabilistic Analysis in Risk Assessments (EPA,1997)
Suppl. Guide to RAGS: The Use of Probabilistic Analysis in Risk Assessment ? Part E

민감도 분석(Sensitivity Analysis)



정적 민감도 분석

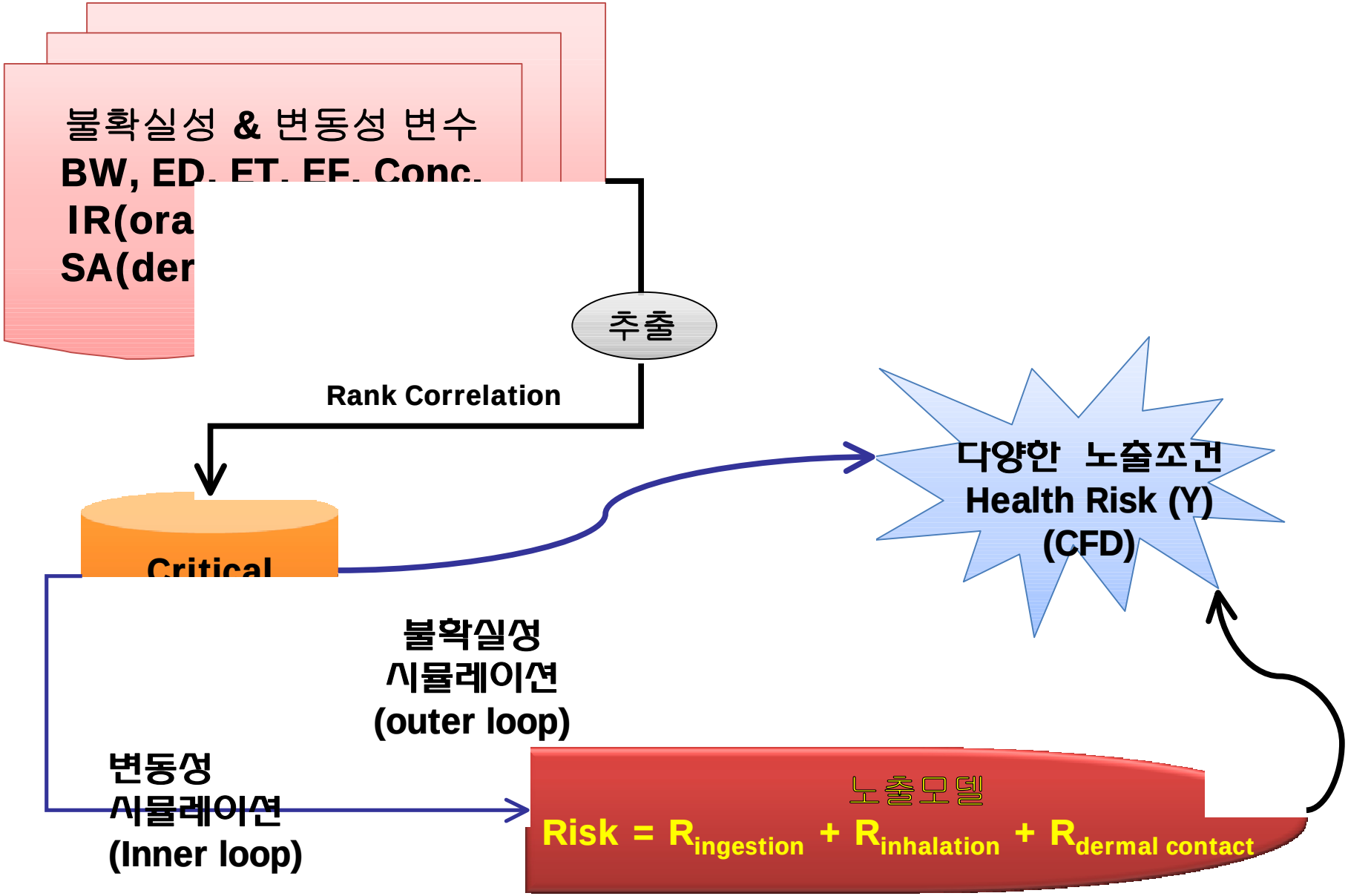
각 변수를 개별적으로 변화
최종 결과값의 변동분석을 통해 우선순위
상관관계 고려하지 않음

동적 민감도 분석

복수값을 동시에 변화시키면서 변수들의 상호효과 분석 및 파악
다중상관관계를 얽혀 있는 모델의 교호 작용 파악

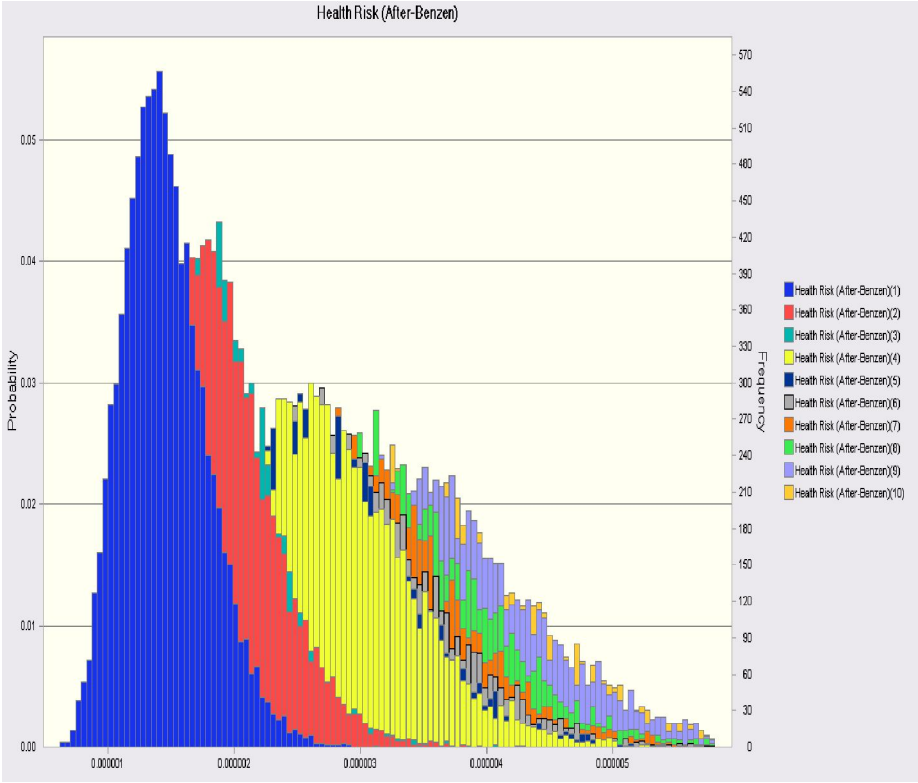
Spearman rank correlation
Contribution to variance

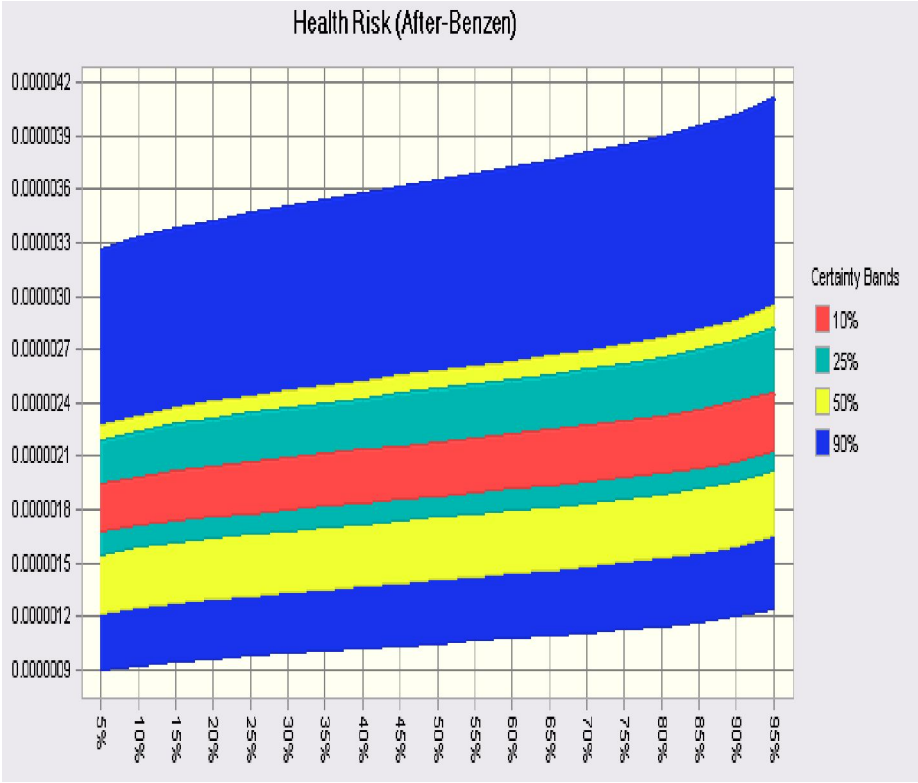
주요노출지수를 이용한 2차원 시뮬레이션
(2D Monte Carlo Simulation)



[6]-2. 2차원 시뮬레이션에 따른 다양한 노출조건 및 결과값 도출

Benzen (After)	Health Risk (Y)	Risk (1)	Risk (2)	Risk (3)	Risk (4)	Risk (5)	Risk (6)	Risk (7)	Risk (8)	Risk (9)	Risk (10)
		5.01E-07	1.47E-06	1.82E-06	2.06E-06	0.0000023	2.38E-06	2.68E-06	2.98E-06	3.00E-06	6.23E-06
Variables	BW (X1)	80.6	67.3	63.6	72.5	59.2	92.9	43.5	57.6	64.1	81.5
	ED (X2)	10.78	5.02	9.50	14.71	7.07	18.44	10.14	15.06	9.21	13.50
	EF (X3)	191	233	301	242	305	332	334	294	251	283
	IR(oral) (X4)	0.345	2.007	0.876	0.912	1.462	0.764	0.723	0.804	1.952	3.234

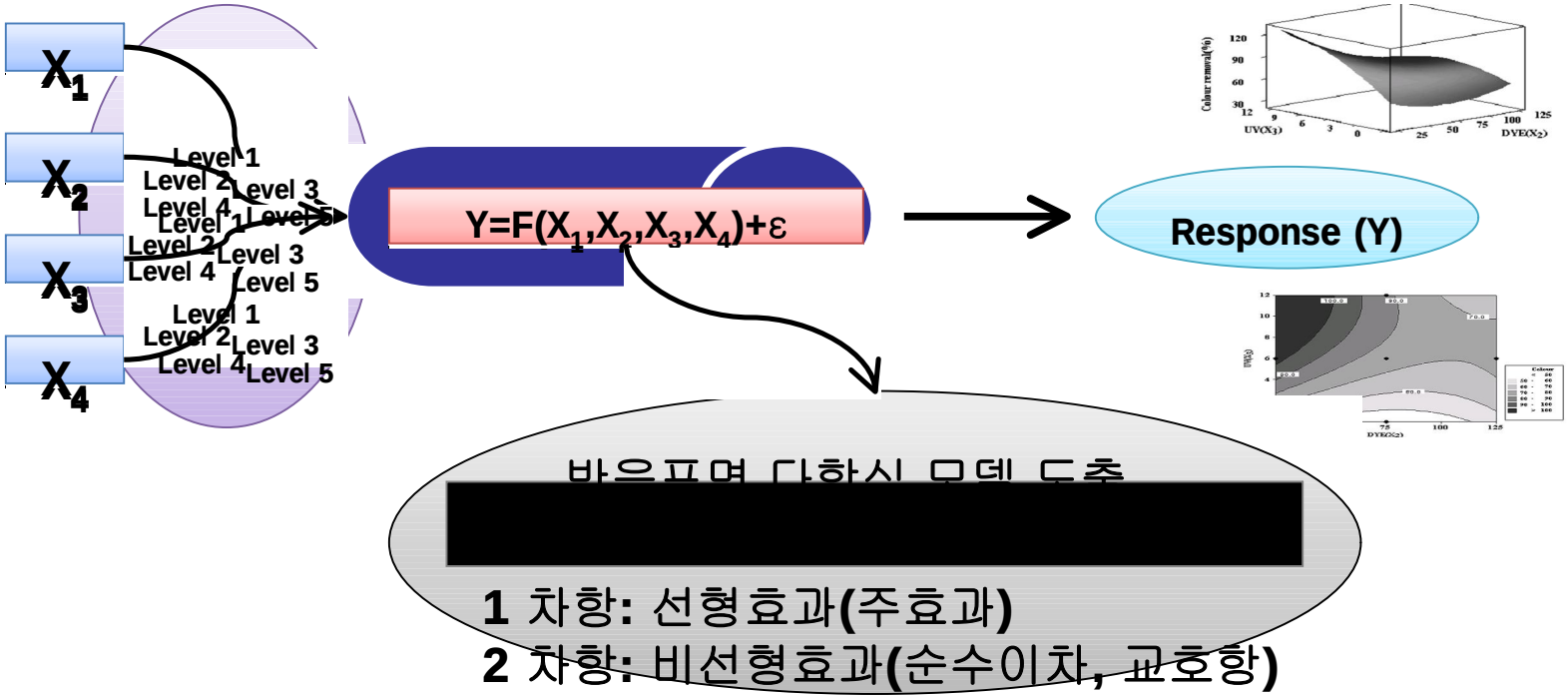


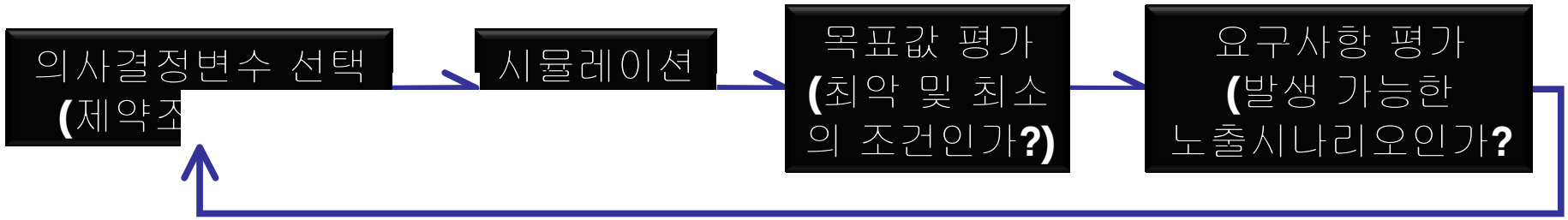


Action 5 복원공법에 다양한 시나리오를 적용시켜 그 결과들간의 분포 모양을 비교

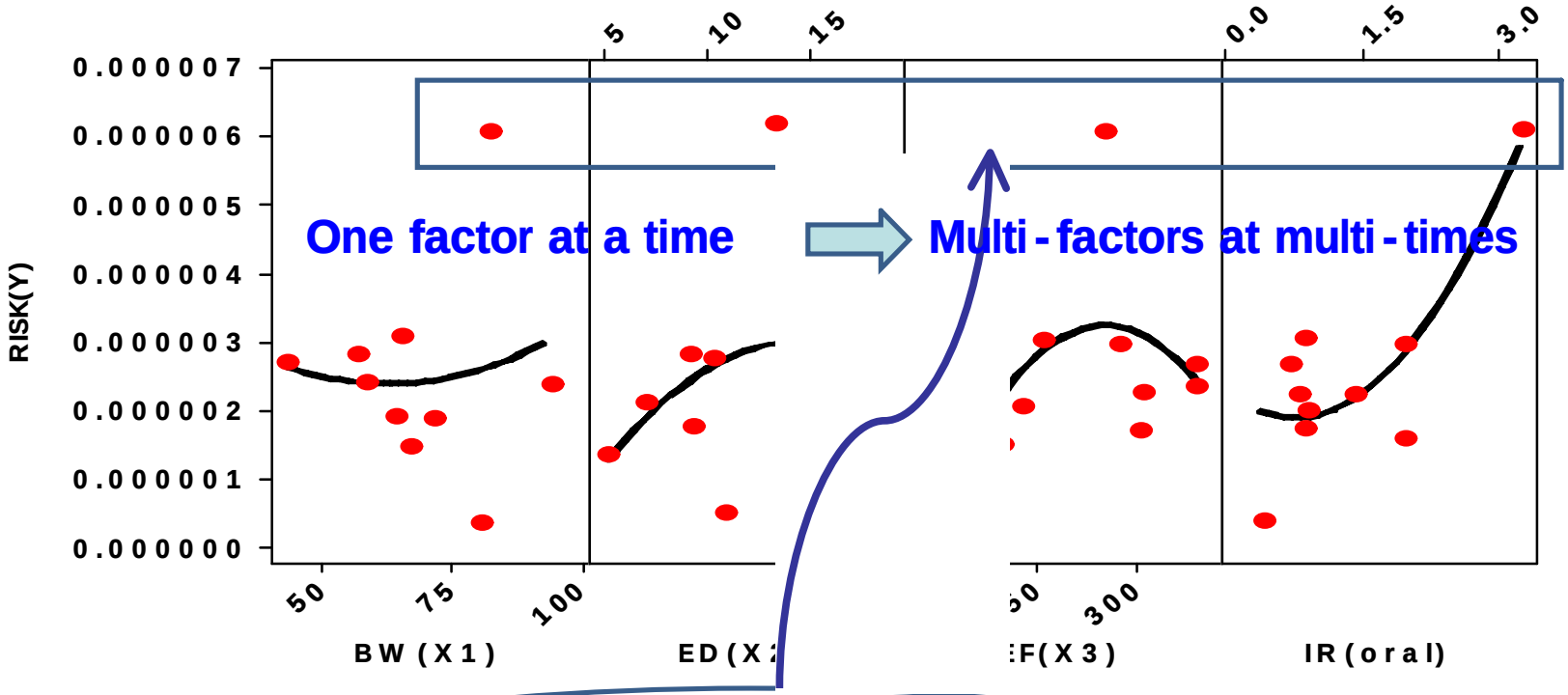
[7]-1. 인체위해성/인체위험도 노출지수 평가 및 모형수립

결과(**Health Risk/Hazard Index**)를 최소화 및 최대화시키는 노출지수(의사결정 변수) 조건 및 예측 모형을 수립





[7]-2. Benzen(After) Health Risk 대한 변수들 간의 관계 파악(산점도 그림)

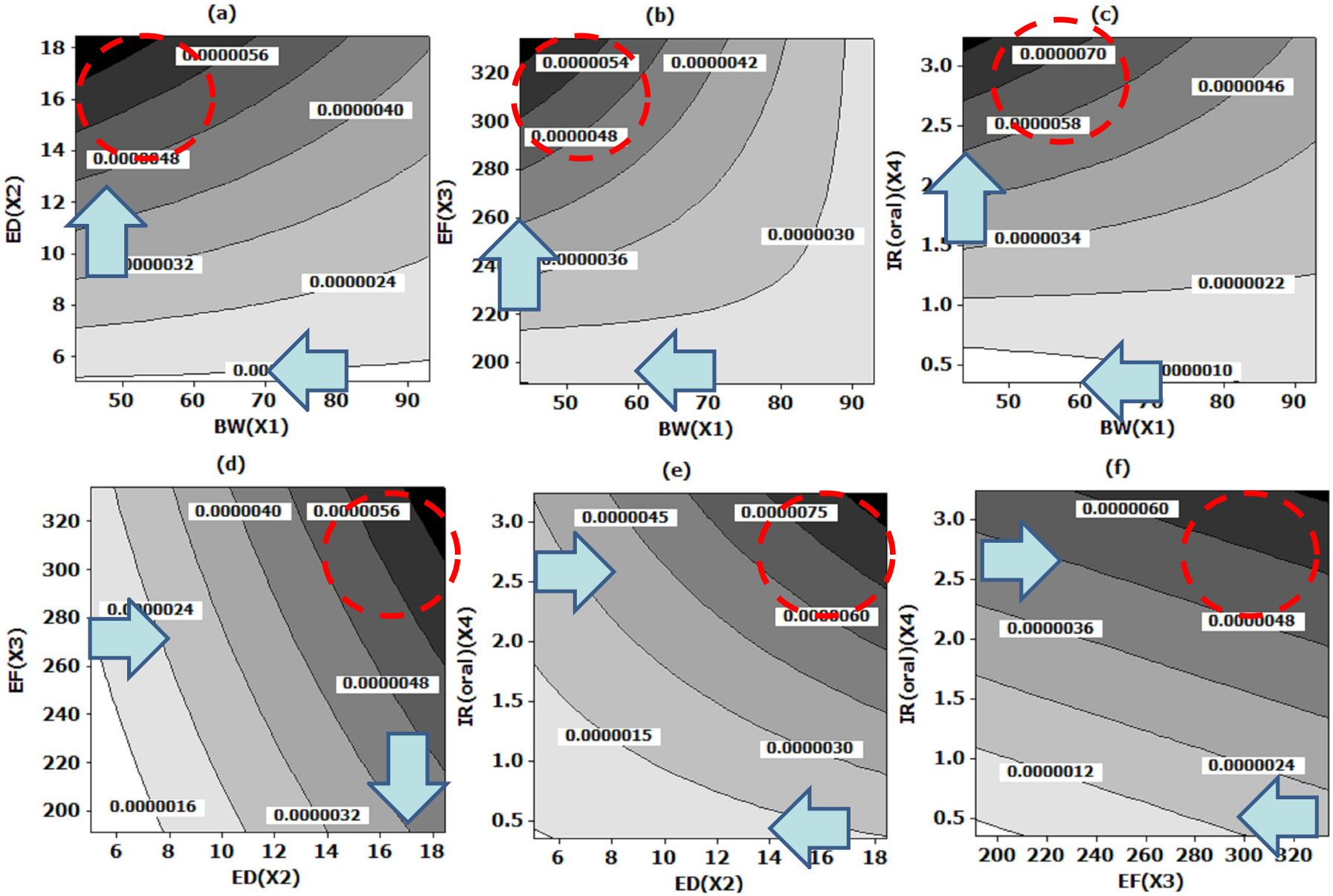


BW(Bodyweight; X_1) = 81.5 kg
ED(Exposure duration; X_2) = 13.5 years
EF(Exposure Frequency; X_3) = 283 event/year
IR(oral)(Ingestion rate; X_4) = 3.234 L/year

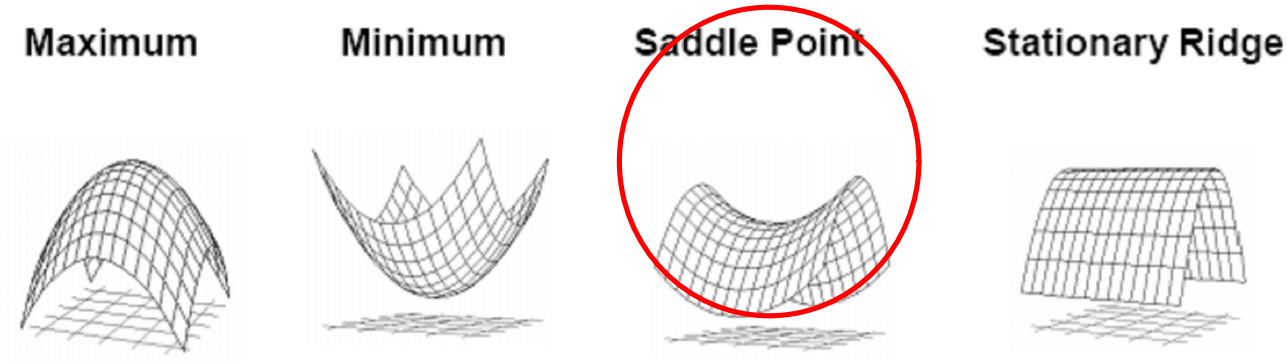
변수(X)와 결과(Y) 관계를 탐색, 단지 연관성 파악
한 변수가 다른 변수에 영향을 미치는 교호작용에 의한 인과관계 파악 아님

대부분 환경공학 오염물질 제거 특성관련 논문에서
이 그림을 보고 최적화라고 해석하는 등 오류를 범하고 있음

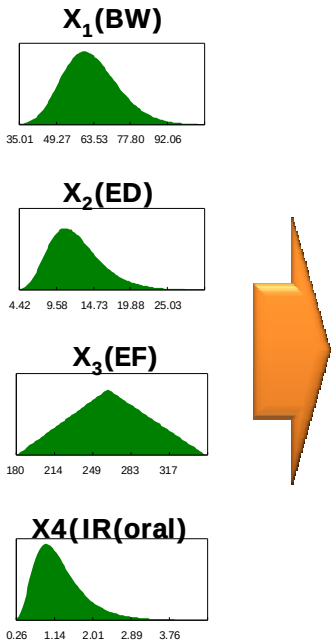
[7]-3. 주요 노출지수 상호작용에 따른 2차원 등고선도



[7]-4. 정준 및 능선분석을 통한 **Health Risk** 최악의 노출조건 수립



Stationary point



Coded Radius	Estimated Response	X_1(BW) [kg]	X_2(ED) [year]	X_3(EF) [event/yr]	X4(IR(oral) [L/hr]
0.0	2.895E-06	68.20	11.73	262.5	1.790
0.1	3.166E-06	66.29	11.87	258.9	1.840
0.2	3.737E-06	64.11	12.14	256.3	1.842
0.3	4.640E-06	61.97	12.41	253.8	1.844
0.4	5.877E-06	59.84	12.70	251.4	1.845
0.5	7.448E-06	57.73	12.67	249.0	1.846
0.6	9.354E-06	55.61	12.95	246.6	1.847
0.7	1.200E-05	53.50	13.22	244.2	1.847
0.8	1.42E-05	51.39	13.76	241.8	1.850
0.9	1.71E-05	49.28	14.03	239.4	1.850
1.0	2.03E-05	47.17	14.30	237.0	1.850

Health Risk에 대한 예측노출 다항식 수립

Health Risk(Y) =

-1.82E-05 + 2.37E-07X_1(BW) + 1.77E-07X_2(ED)

+ 4.36E-08X_3(EF) + 2.95E-06X_4(IR(oral))

선형(1차 식)

- 4.45E-09X_1(BW)* X_2(ED)-5.71E-10X_1(BW)* X_3(EF)

-3.72E-8X_1(BW) * X_4(IR(oral)) +7.44E-10 X_2(ED)* X_3(EF)

+1.344E-07X_2(ED)* X_4(IR(oral)) (R^2= 0.99)

비선형
교호항
(2차 식)

Health Risk 예측식을 이용하여 향후 **Action 5** 공법으로 유류오염물질 **(BTEX, TPH)** 제거 후 잔류오염물질 이용성에 대한 발암노출 예측이 가능

Uncertainty

? Risk : probability of an undesired or harmful event

? Decision 하는데 있어서 없애고 싶은 무엇, 하지만 결코 없앨 수 없는 무엇

? Uncertainty는 다음의 것들로 요약됨

  ? Bias

  ? Variance

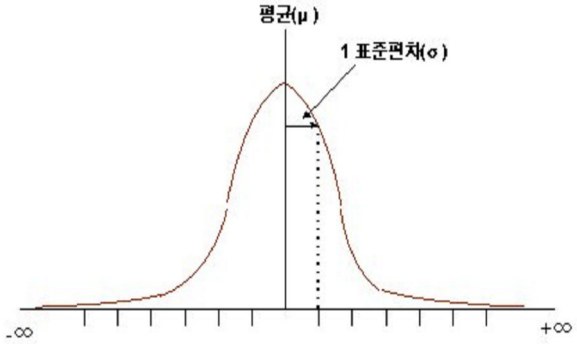
  ? 분포에 기초한 통계량들

모집단과 표본

? 모집단

? 모수

$N(\mu, \sigma^2)$

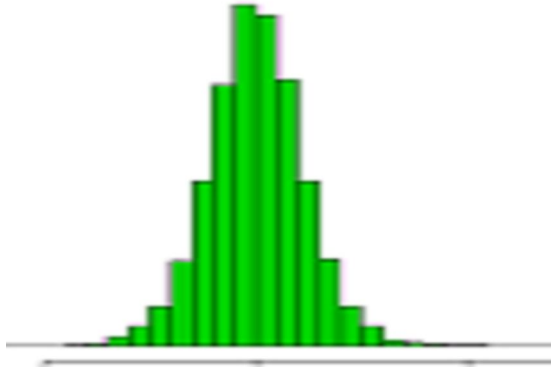


? 표본 $Y_1, \cdots, Y_n$

? 추정치

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

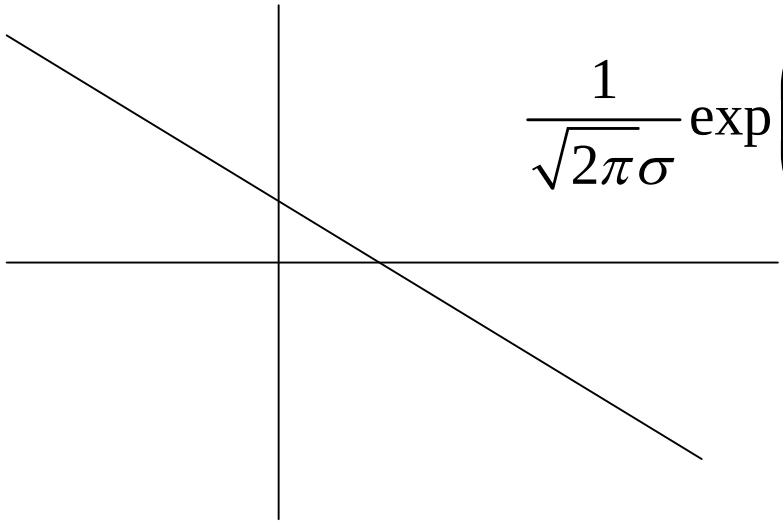
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \bar{Y})^2$$



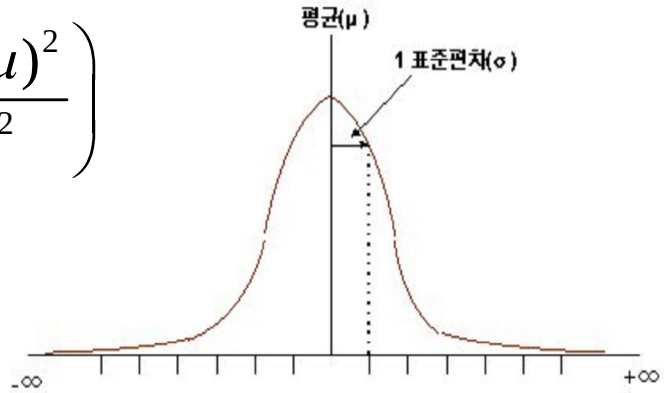
? **모수:** 가정한 모형의 통계적 성질을 완전히 결정하는 상수(들)

Y=a+b x

$$N(\mu,\sigma^2)$$



$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



편이(bias)와 효율(efficiency)

? 모수(θ)를 추정하는 추정치($\hat{\theta}$)가 있을 때

? Bias($\hat{\theta}$) = $E(\hat{\theta})-\theta$

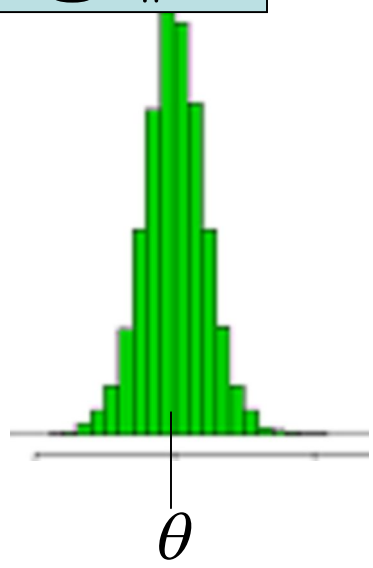
? Bias가 0인 경우 $\hat{\theta}$ 을 불편추정치 (unbiased estimator) 라고 부른다

? 한 모수(θ)를 추정하는 추정치가 두 개 ($\hat{\theta}_1,\hat{\theta}_2$) 가 있을 경우 두 추정치의 분산의 비율을 상대효율(relative efficiency)라고 한다

? RE $(\hat{\theta}_1,\hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$

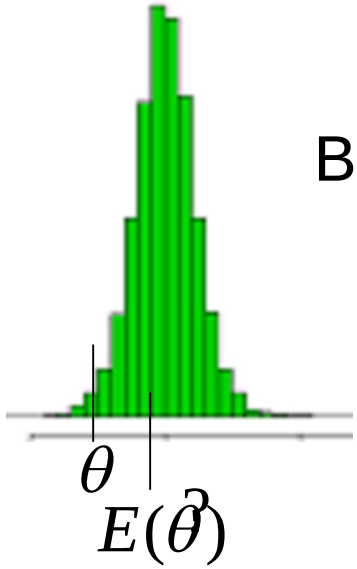
기초통계

추정치($\hat{\theta}$) 의 분포



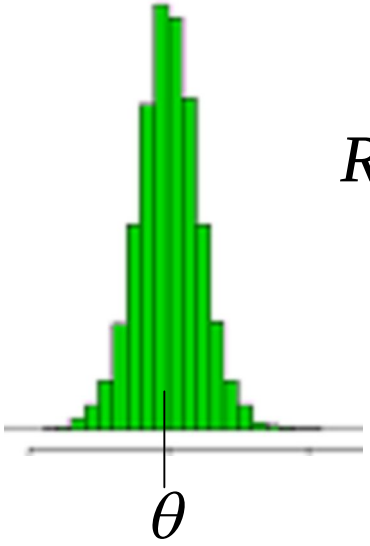
Bias=0

즉 불편추정량



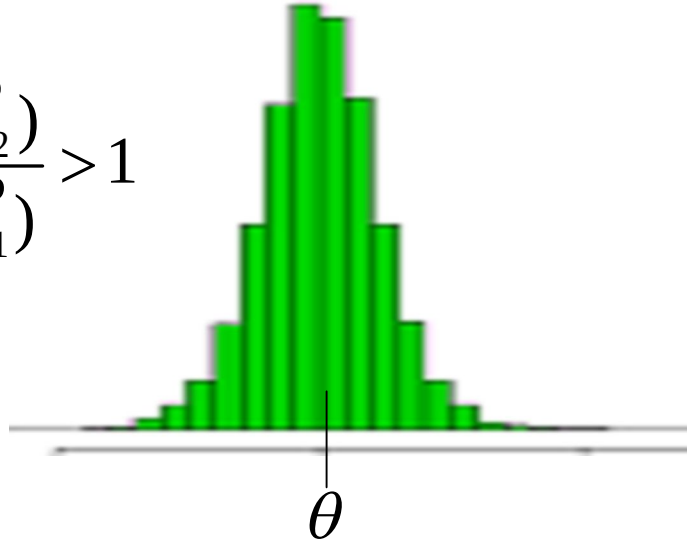
Bias가 있음

추정치($\hat{\theta}_1$) 의 분포



$$RE(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} > 1$$

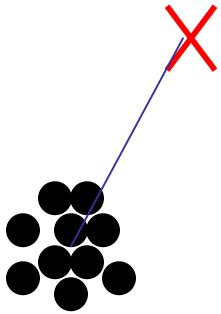
추정치($\hat{\theta}_2$) 의 분포



Case 1

Big bias

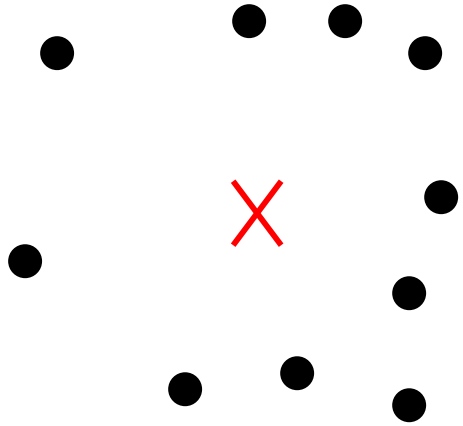
Small Variance

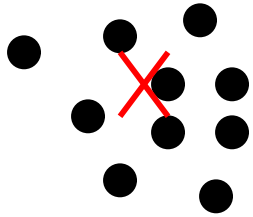


Case 2

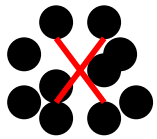
Small bias

Big Variance





Our Hope



Best Scenario!

- $E(\bar{Y}) = \mu, E(S^2) = \sigma^2$: 표본평균과 표본분산은 모평균과 모분산의 불편추정치 (unbiased estimator) 이다.
- 여러 번 반복하여서 평균을 구하면 원하는 값(모수)을 아주 근사하게 예측할 수 있다.
- 한 번의 분산으로는 신뢰구간을 이용하여서 그 정확성을 평가할 수 있다.

세 가지 모형

? response = deterministic model

? response = deterministic model + error : classical linear model

? response = deterministic model
+ stochastic components + error

: random effects model

? Bayesian model

Common Problems

? Uncertainty factor
? Extrapolation models and prediction
? Stochasticity
? Model : relationships between variables
? Decision making
? Combining information (to reduce uncertainty)
? Model uncertainty (Monte Carlo methods)

Uncertainty (safety) factor

Protection is ensured with high probability

? Acute to chronic endpoints

? Response on individuals to response on populations

? A single species response to a response in a group of species

? Laboratory to field conditions

? One to many exposure routes

? Direct to indirect effects

? A single location and time to multiple locations and/or rimes

? A test organism to a human

Extrapolation factor 1

Uncertainty factors by developing a model to extrapolate from one level to another.

Lower level의 uncertainty factor를 다른 factor의 uncertainty

예) pesticide to protect ecosystem

data from Ceriodaphnia dubia (a water flea), minnows, brook trout

Each lab produces no-observable-effect-level(Lowest-observed-adverse-effect level; LOAEL)

예제

? $y = \beta_0 + \beta_1 x$ deterministic model

? $\hat{y} = b_0 + b_1 x$ prediction

? Model 1) Random slope model

$$b_1 \sim N(\beta_1, \sigma_1^2)$$

? Model2) Random slope and random intercept model $b_0 \sim N(\beta_0, \sigma_0^2), b_1 \sim N(\beta_1, \sigma_1^2)$

? Model 3) —Model 2 + two parameters are correlated

? In a real situation, collect data and fit statistical model $y = \beta_0 + \beta_1 x + \varepsilon, \varepsilon \sim N(0, \sigma^2)$

Table 1 Twenty-two observations used to model the relationship between chlorophyll *a* and total phosphorus. Values are log transformed

Observation	Log total phosphorus	Log chlorophyll *a*
1.00	1.97	1.92
2.00	2.07	2.36
3.00	2.45	2.64
4.00	2.55	1.17
5.00	2.77	2.07
6.00	2.93	2.22
7.00	3.30	3.78
8.00	3.60	6.30
9.00	3.65	4.59
10.00	3.96	3.02
11.00	3.79	6.30
12.00	4.23	5.64
13.00	4.43	5.78
14.00	4.65	7.00
15.00	4.91	4.67
16.00	4.94	7.40
17.00	5.18	6.80
18.00	5.52	5.75
19.00	5.59	8.37
20.00	6.01	7.90
21.00	5.90	7.93

? estimated model $\hat{y} = -1.47 + 1.59x$

$\hat{\sigma}^2 = 1.21$

? Intercept $N(-1.47, 0.67)$

? Slope $N(1.59, 0.038)$

? Cov = 0.95

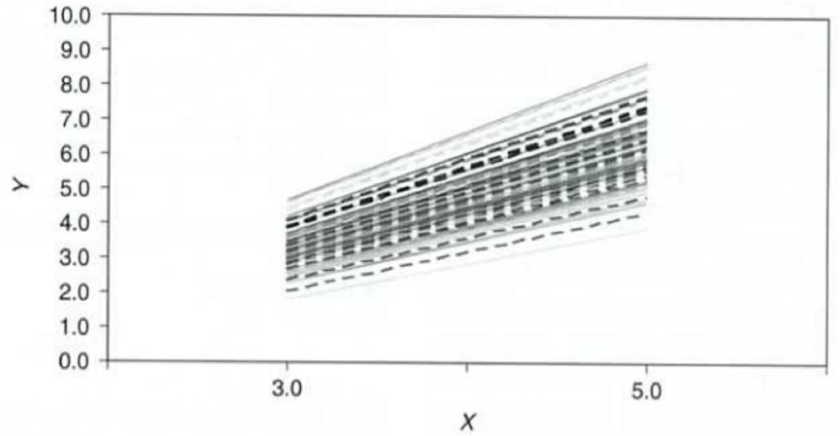


Figure 1 Sample lines assuming uncertainty only in the slope

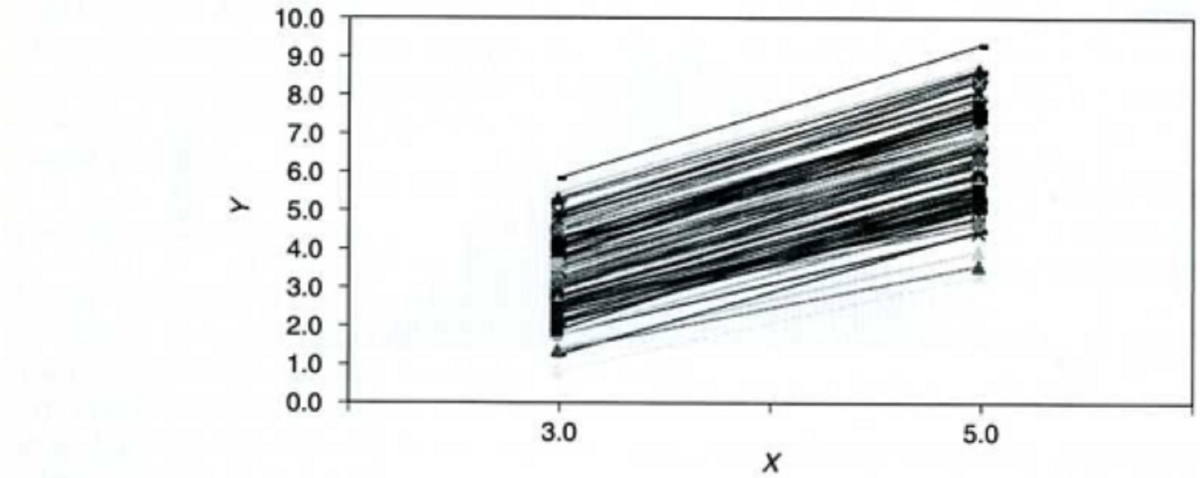


Figure 2 Sample lines assuming uncertainty in slope and intercept

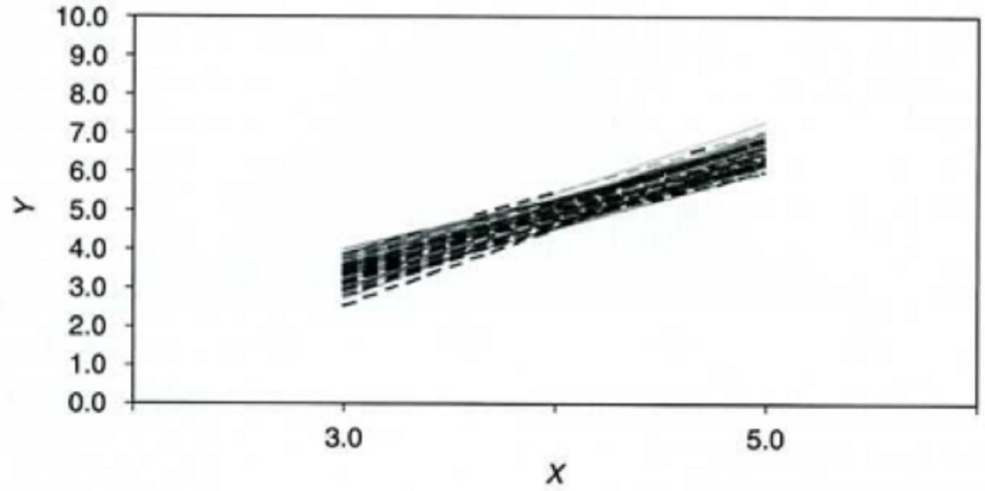


Figure 3 Sample lines assuming uncertainty in slope and intercept with correlation

? 모수들간의 correlation을 고려하지 않았을 때 (Figure 1, 2) -> 선들이 넓게 퍼진다.

? Correlation을 고려했을 때 (그림 3) : reduced variation

? 전형적인 회귀모형에서 **X**는 고정**(fixed non-random)**

Y는 확률적 **(random)**

? **X, Y**가 동시에 확률적 상관모형

이변량 정규분포 $(X,Y) \sim BN(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$

$$X \mid Y = y \sim N(\mu_{x|y}, \sigma_{x|y}^2)$$

$$Y \mid X = x \sim N(\mu_{y|x}, \sigma_{y|x}^2)$$

? ρ 는 모집단 상관계수로 다음과 같이 정의 된다.

$$\rho = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \cdot \sigma_Y}$$

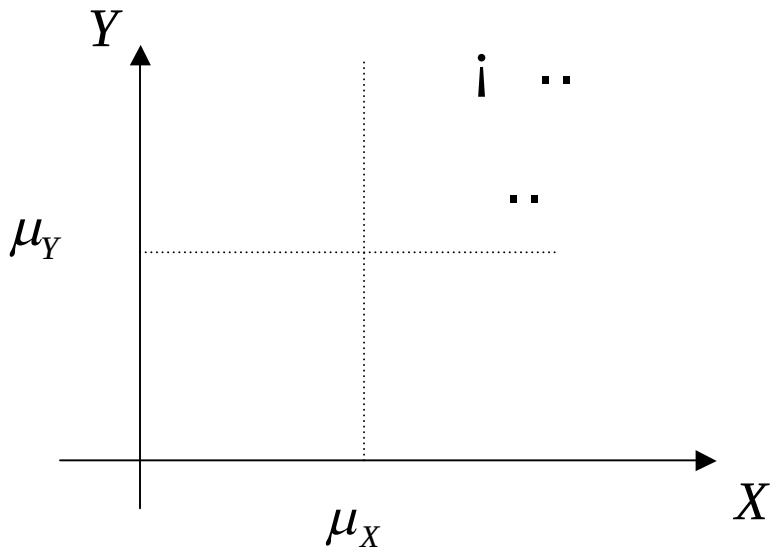
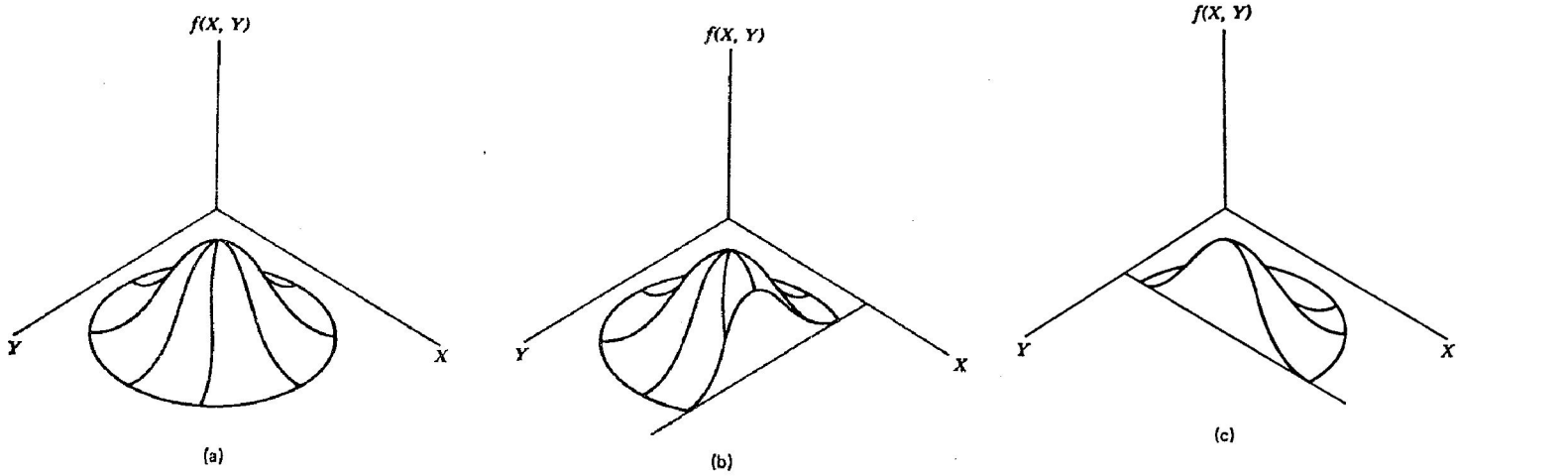


그림 이변량 정규분포



이변량정규분포 (a)이변량정규분포 (b)X에 대한 Y의 정규분포 하부모집단 (c)Y에 대한 X의 정규분포 하부모집단

- 표본상관계수 $r = \sqrt{r^2}$: 모집단 상관계수 ρ의 추정치

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

? 가설상의 $\rho = 0$ 일 때

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$$

회귀분석

- 영국의 과학자 Sir Francis Galton이 행한 유전에 관한 연구에서 비롯되었다.
- 부모의 신장과는 상관없이 2세의 신장이 일반 평균치에 복귀(revert)하는 특성을 발견하였다.
- 복귀(revert)는 회귀(regression)로 표현하기로 하였다.

? 회귀분석 기본모형

$$Y = \alpha + \beta x + \varepsilon$$

종속변수 독립변수 : 고전적인 모델에선 비확률

$$\varepsilon \sim N(0, \sigma^2) \quad iid$$

정규 동일분산 독립 (independently) 같은 분포 (identically distributed)

회귀모형 (Regression Model)

? 단순회귀분석법에서의 가정

Y : 종속변수, 반응변수

X : 독립변수, 설명변수

1. Y는 분포가 있는 확률변수
2. X는 고정된 값으로 오차 없는 통제 가능한 값
3. Y는 X값에 따라 하부모집단이 존재하고 하부모집단은 각각 정규분포를 하여야 한다.

?단순선형 회귀모형의 도식

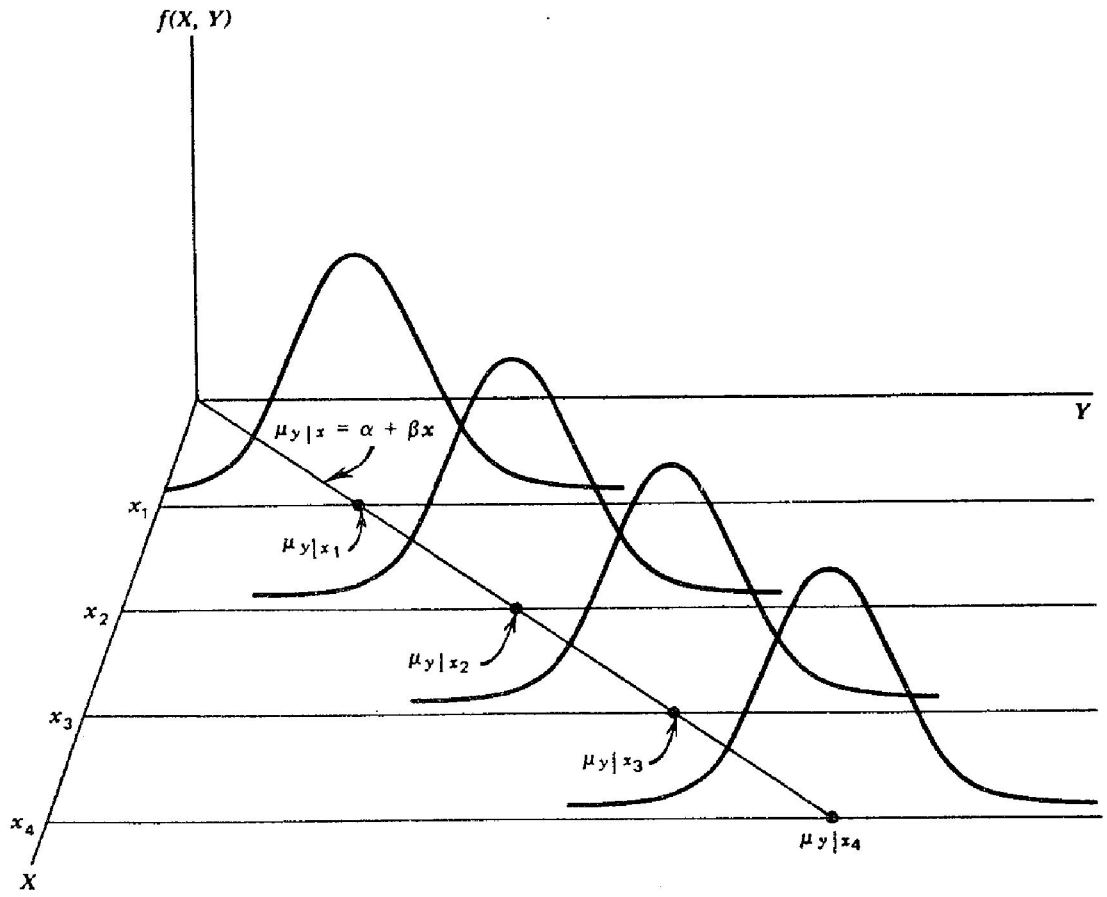


그림 8. 2. 1. 단순선형 회귀모형의 도식

4. 하부모집단의 분산은 동일
5. 선형 가정 $E_{y|x} = \alpha + \beta x$
6. Y값들은 통계적으로 독립이다.

$y_i = \alpha + \beta x + \varepsilon_i$ (선형관계)

$\varepsilon_i \sim N(0, \sigma^2)$, 독립, 정규성, x와 무관한 동일분산

-> 모든 가정은

Check 하는 것이 원칙

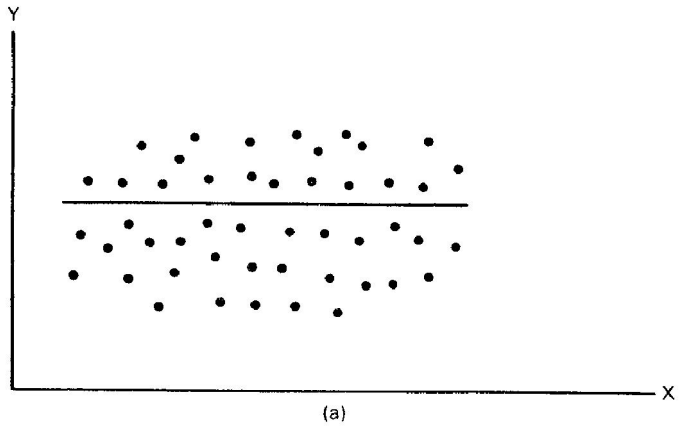


그림 8. 4. 1

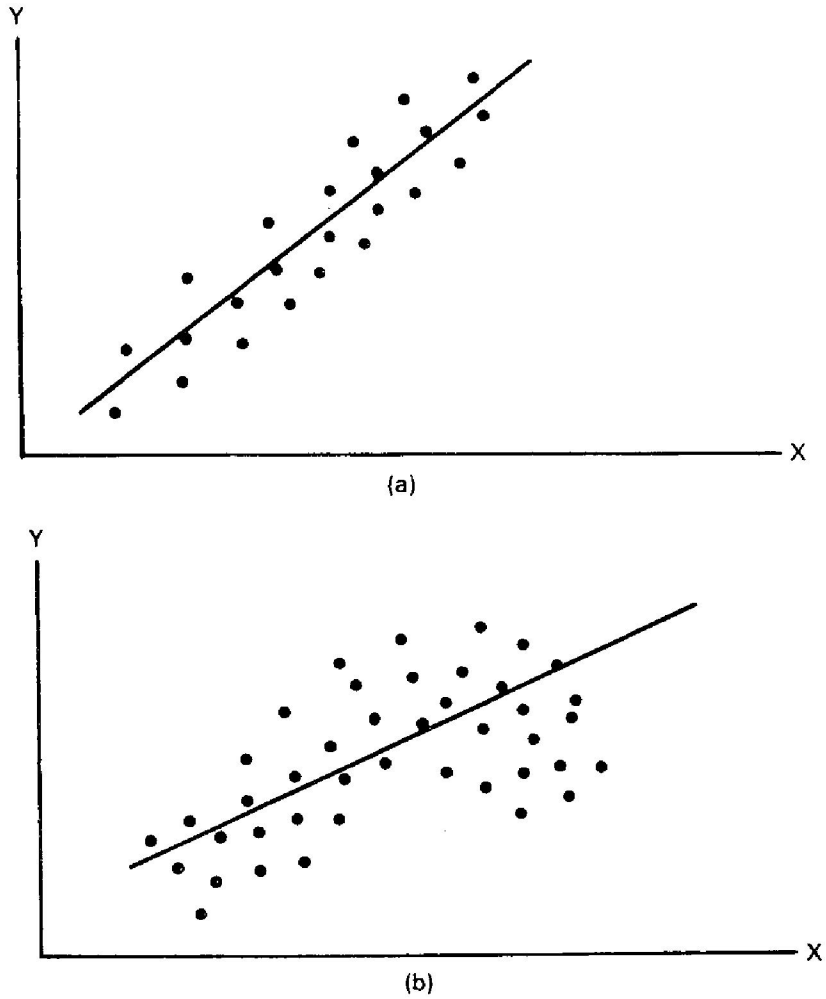


그림 8. 4. 2

회귀방정식의 사용

- 주어진 x에 대한 y의 예측

$$\hat{y} = a + bx$$

$$\text{신뢰구간} : \hat{y} \pm t_{(1-\alpha/2)} s^2_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

회귀방정식의 사용

- 주어진 x에 대한 y의 평균추정

(하부 모집단에 대한 평균 추정)

$$\hat{y} = a + bx$$

$$\text{신뢰구간} : \hat{y} \pm t_{(1-\alpha/2)} s^2_{y|x} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Monte Carlo 방법을 응용

? Crystal Ball을 사용 10,000번 simulation

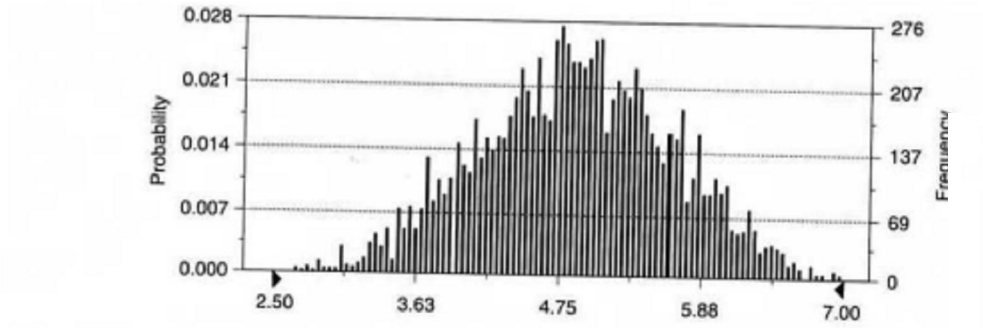


Figure 4 Predictions for the simulation in which the slope varies

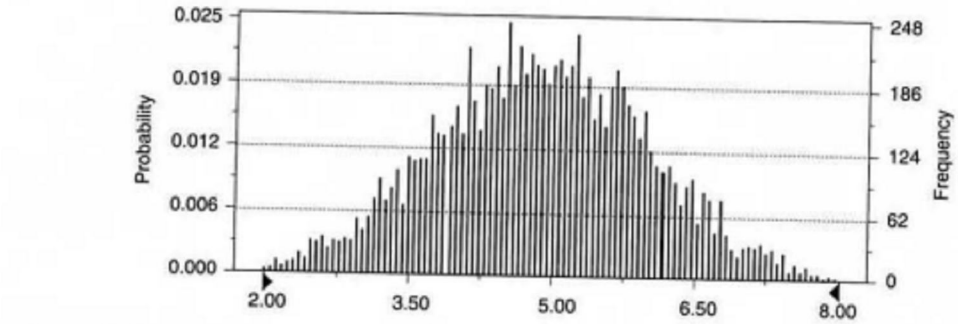


Figure 5 Predictions for the simulation in which the slope and intercept vary but are uncorrelated

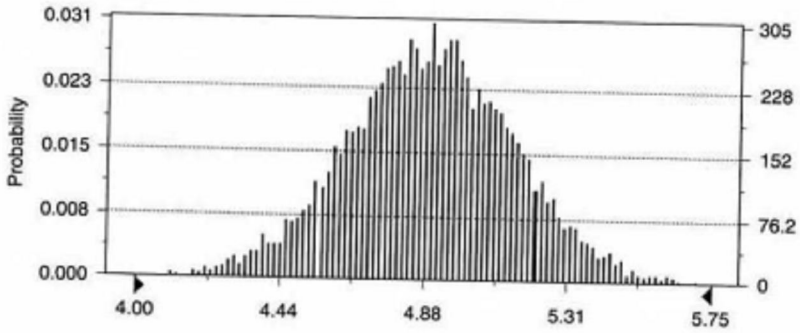


Figure 6 Predictions for the simulation in which the slope and intercept vary and are correlated

Monte Carlo 방법 응용 시 주의점

? 어떠한 모수(들)를 사용할 것인가 ?

? 어떠한 분포를 사용할 것인가 ?

? 모수들 간의 상관구조

? Sensitivity analysis

? 해석의 관점, 통계적 관점

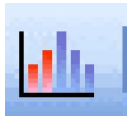
Monte Carlo simulation

- 불확실성의 효과를 측정하기 위해 난수 (random variable)을 이용하는 시스템
- 장점
  - 실행에 앞서 의사결정을 하는데 경제적임
  - 시스템에서 중요한 구성요소를 제시
- 단점
  - 결과가 입력값의 정확성에 민감함
  - 엑셀에 모델을 구축해야 함

모델구축의 6단계

1. 시스템의 순서도나 알고리즘 개발
2. Excel spread sheet에 모델을 구축
3. Crystal Ball을 이용하여 모델에서의 가정(assumption)과 예측값(forecast)정의
4. 시뮬레이션을 실행하고 결과를 분석
5. 모델평가
6. 모델개선 또는 의사결정

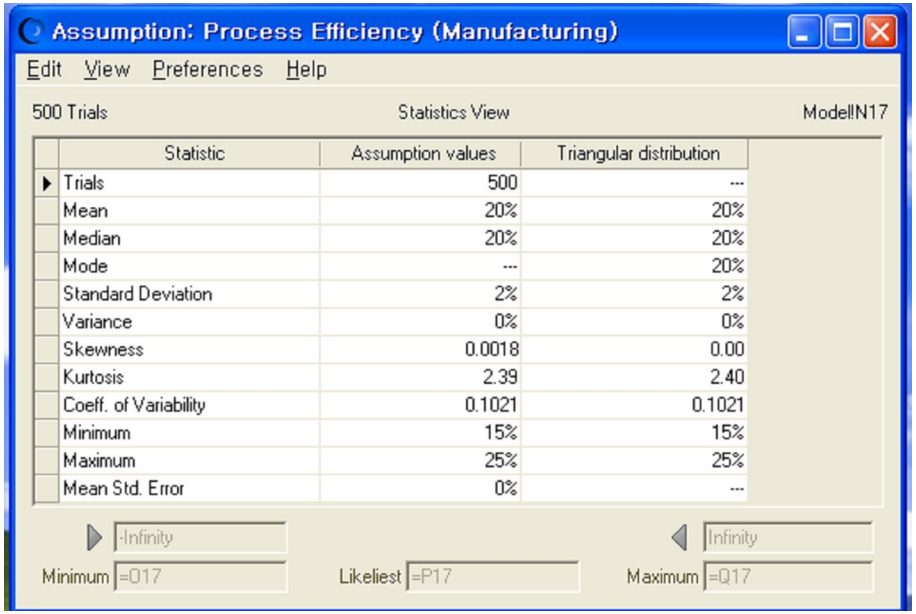
기본용어

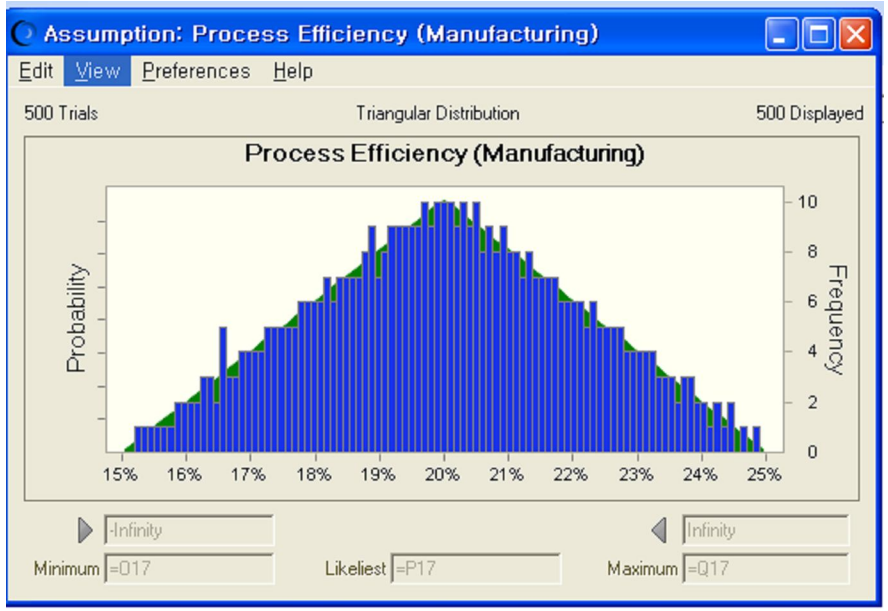
Crystal Ball 용어	일반적인 용어
Assumption 	입력값, X, 독립변수, 확률변수, 확률분포
Decision Variable 	통제변수
Forecast 	결과값, f(X), 종속변수

가정 차트(Assumption chart)

? 임의 누출과정의 노이즈 보기

? 이론적으로 입력한 분포와 표본 통계량비교



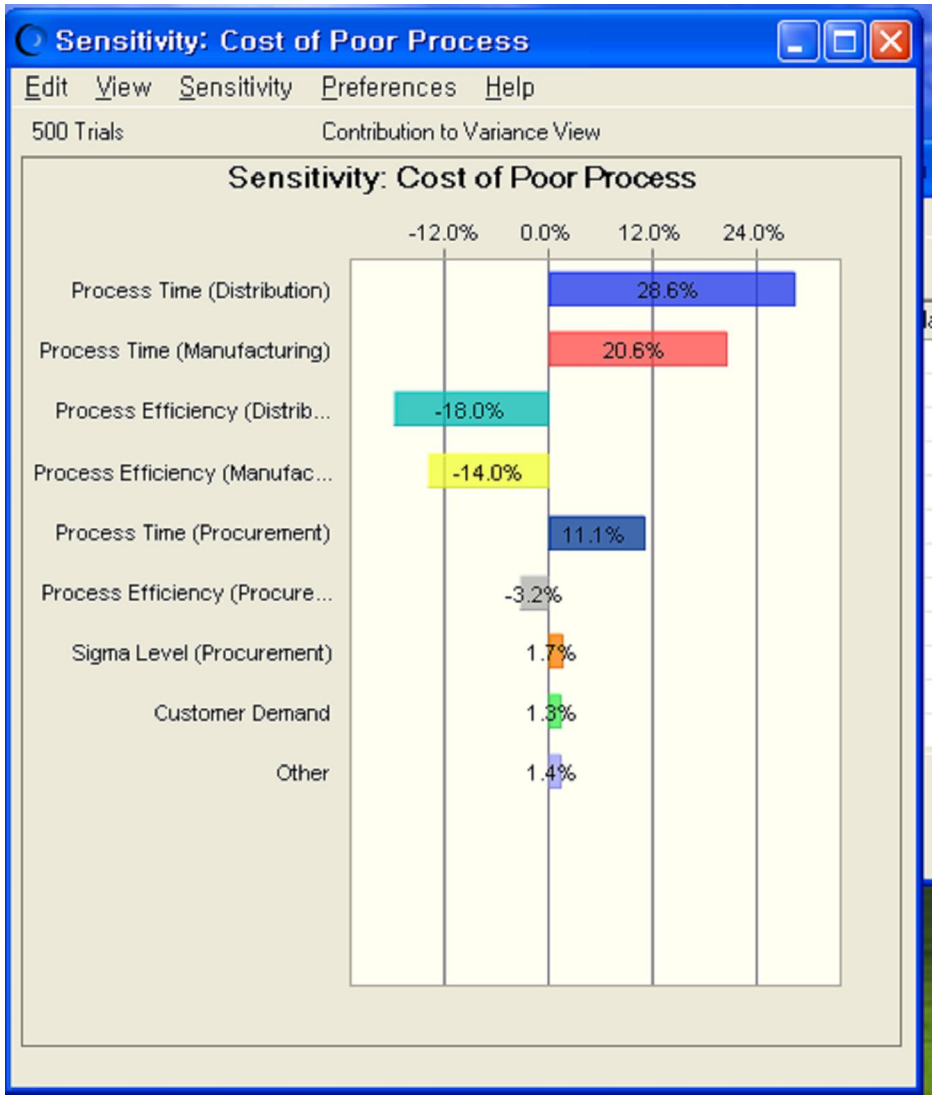


민감도 차트(sensitivity chart)

? 민감도 차트를 사용하는 이유

? 핵심적인 요소 확인

? 향상된 전략 수립

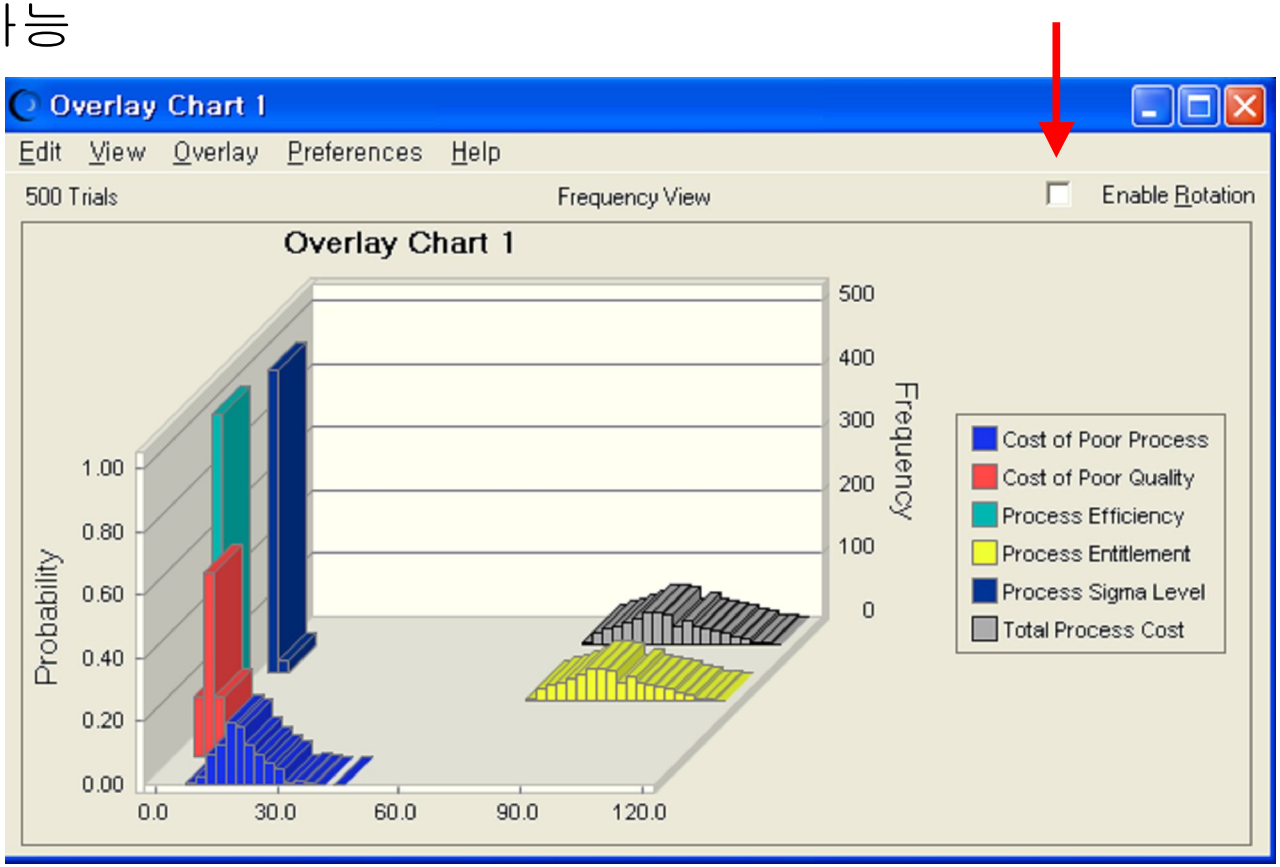


민감도 차트(sensitivity chart)

- 모델 검정
  - 현실적인가?
  - 모델에서 오류 찾아내기
  - 가정 재정의
- 향상된 전략 수립
  - 내가 통제할 수 있는 것인가?
  - 목표 수정/변동 축소
  - 자원의 할당
  - 프로세스나 시스템의 재구성

오버레이 차트(overlay chart)

- 여러 개의 예측 값을 동시에 분석
- 예측 값의 비교
- 동시에 여러 개의 예측 값을 분포비교
- 3D 그래프 , 회전 가능



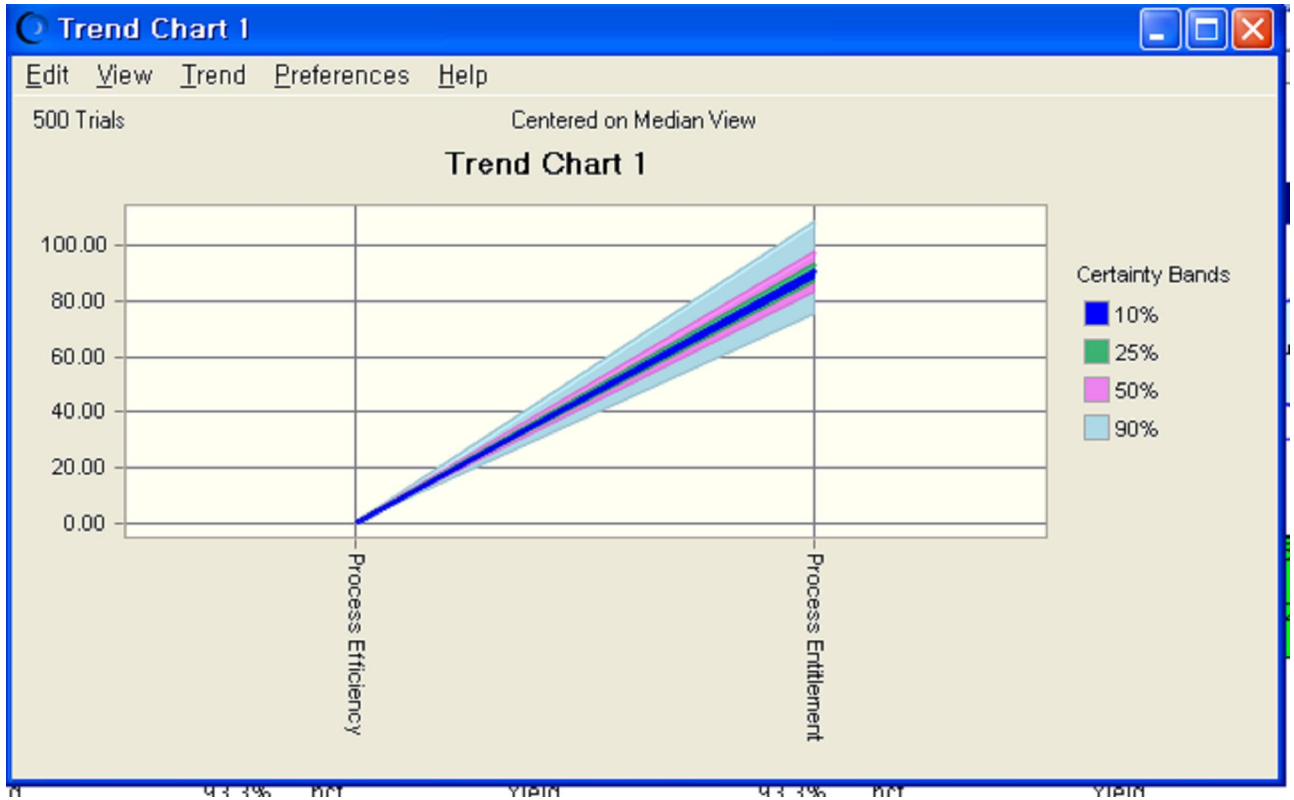
크리스탈 볼 데이터 편집

? 데이터(assumptions, decisions, forecasts)의 복사, 붙이기 또는 삭제 기능
? 엑셀의 값이나 수식에는 변화가 없음
? 셀 참조 이용
? 주의 : 크리스탈 복사볼 데이터가 아닌 엑셀의 색상복사와 혼동하지 말 것

추세차트(trend chart)

? 시계열 분석시에 주로 사용

? 신뢰구간 표시

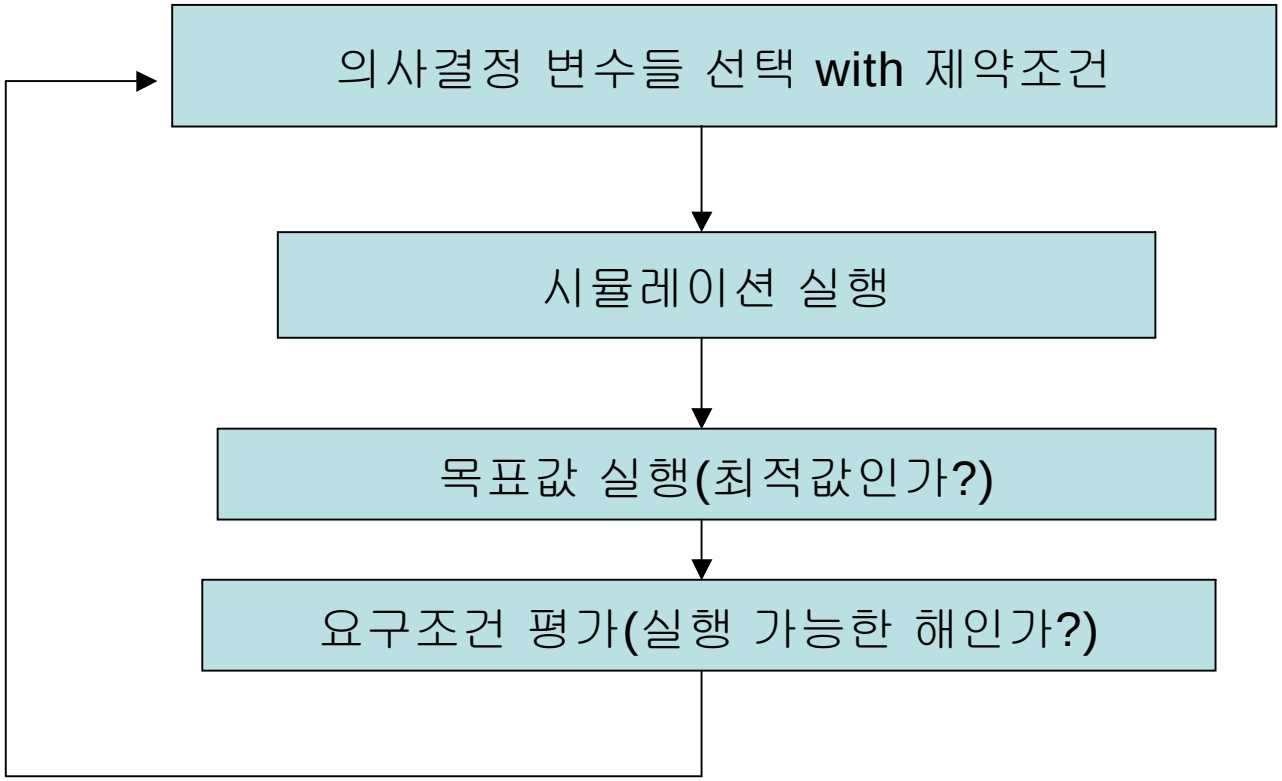


Optquest

최적화란 무엇인가?

- 결과(예측 통계량)을 극소화 또는 극대화 시키는 입력변수(의사결정변수)찾기
- 의사결정변수들->가정을 포함한 계산과정->예측값(들)
- X->함수->F(X)=Y

Optquest의 실행 프로세스

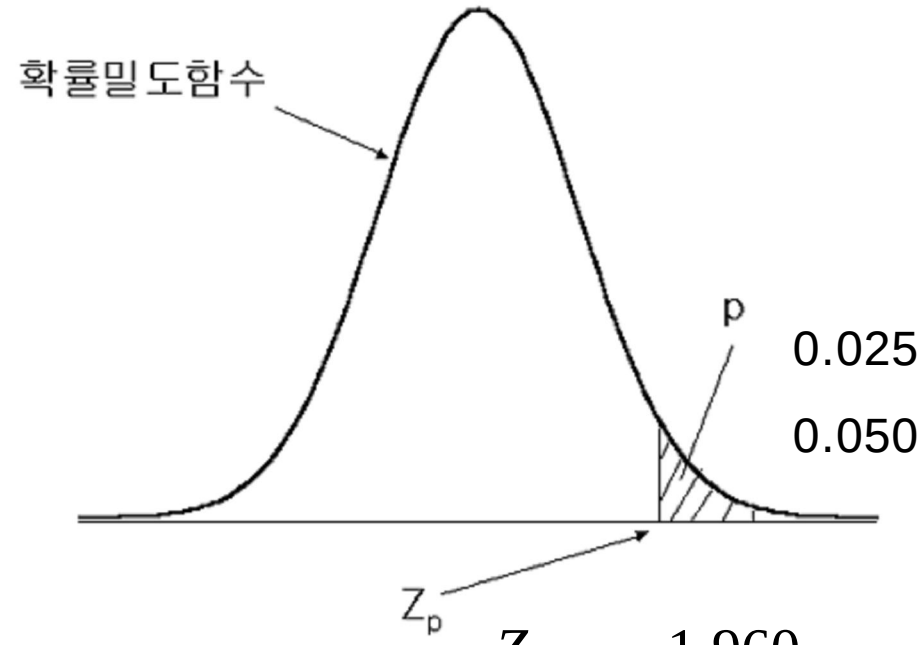


Optquest 결과

? 진행상태와 최적해(status and solution)
? 수행결과 그래프(performance graph)
? 막대그래프(bar chart)
? 최적화 로그파일(optimization log)
? 해 분석(solution analysis)
? 엑셀에 복사하기(copy to excel)

Terminology

분위수 : quantile



$$Z_{0.025} = 1.960$$

$$Z_{0.05} = 1.645$$

표 **4.** 표준정규분포 하에서의 **Z**값

제**1**종 오류 **(α)**	Z_α 단측	Z_α 양측	제**2**종 오류 **(β)**	검정력 **(power)**	Z_β
0.005	2.576	2.813			
0.010	2.326	2.576			
0.025	1.960	2.248			
0.050	1.645	1.960	0.05	0.95	1.645
0.100	1.282	1.645	0.10	0.90	1.282
0.200	0.842	1.282	0.20	0.80	0.842
0.300	0.524	1.036			
0.400	0.253	0.842			
0.500	0.000	0.674			

P-value (1)

? 연구목적 : 관심변수의 (모)평균이 두 집단에서 다르다.

? $\bar{Y}_1$ 첫 번째 집단에서의 표본 평균

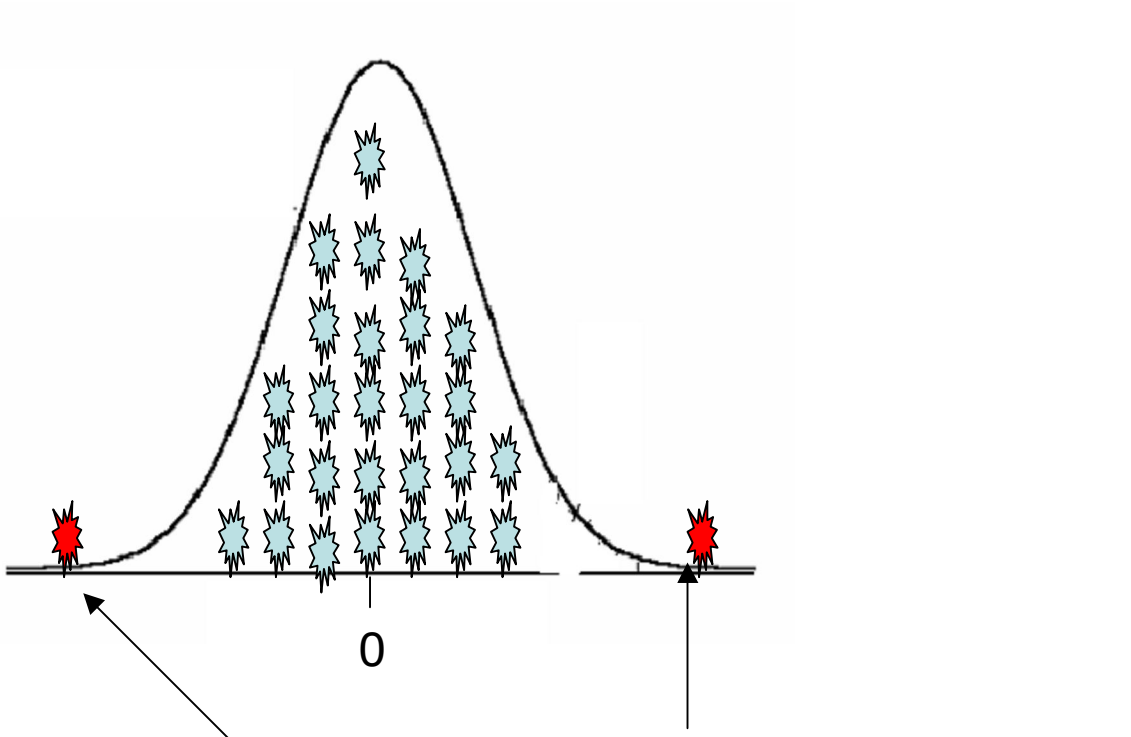
? $\bar{Y}_2$ 두 번째 집단에서의 표본 평균

? 만약 두 집단에서의 모평균이 같다고 하면

? 두 표본 평균은 비슷할 것이다.

? 표본평균의 차이를 반복적으로 구해보면

P-value (2)



P-value (3)

? P-value = 두 집단의 평균이 같다고 가정했을 때 우리의 자료, 혹은 더 차이가 나는 자료를 얻을 확률

? 작은 p-value : 위의 확률이 작다

➔ 통계적으로 가능하지 않은 일이 일어났다.

➔ 두 집단의 평균이 같다는 가정에 문제가 있다.

➔ 두 집단의 평균은 같지 않다고 결론 내린다.

P-value (3)

? 작지 않은 p-value : 두 집단의 평균이 같다고 가정하면 우리의 자료를 관측할 확률이 작지 않다.

➔두 집단의 평균이 같다는 가정에 문제가 없다.

양쪽검정, 한쪽검정

? A(얻은 자료) -> B (연구가설)

? -B -> -A

? 귀무가설 (-B) : 두 집단에 차이가 없다. (Ho)

? 대립가설 (B) : 두 집단에 차이가 있다. (Ha)

? 일종의 오류 : 옳은 귀무가설을 기각할 확률

= Pr (reject Ho | Ho is true) α

? 이종의 오류 : 틀린 귀무가설을 받아들일 확률

= Pr (Not reject Ho | Ha is true) β

? Power = 1- β, (있는 차이를 발견할 확률)

모수적 방법과 비모수적 방법 (1)

자료	평균	중앙값
1,2,3,4,5	3	3
1,2,3,4,5,100	19	3.5

?중앙값(median)은 평균에 비하여 이상치에 대해서 둔감(robust)하다.

?자료의 정규성 분포가정을 하면 평균과 분산을 통하여 모집단의 성질을 완전히 파악할 수 있다. (모수적 방법)

모수적 방법과 비모수적 방법 (2)

- 비모수적 방법은 자료의 (정규성) 분포가정을 하지 않는다
- 자료의 평균과 분산이 아닌 순위를 이용한 방법을 사용한다.
- 자료의 분포가정 (eg 정규성)이 만족되면 효율이 떨어진다.
- Robust 한 결과를 준다. (outlier에 둔감)

Likelihood and maximum likelihood estimator

? Likelihood

=Pr(getting our data|Given model)

e.g. 100명의 사람을 관찰했더니 30명의 환자를 관찰하였다.

Model : Y=100명의 사람 중 환자의 숫자

Binomial (n=100, p)

Likelihood and maximum likelihood estimator

? 확률밀도함수

=Pr(Y=y|Given model)= $\binom{100}{y} p^y (1-p)^{100-y}$

Maximum likelihood estimator

=확률밀도함수를 최대화 시켜주는 p

이 경우 MLE는 30/100=0.3 임을 보여줄 수 있다. How ?

Likelihood and maximum likelihood estimator

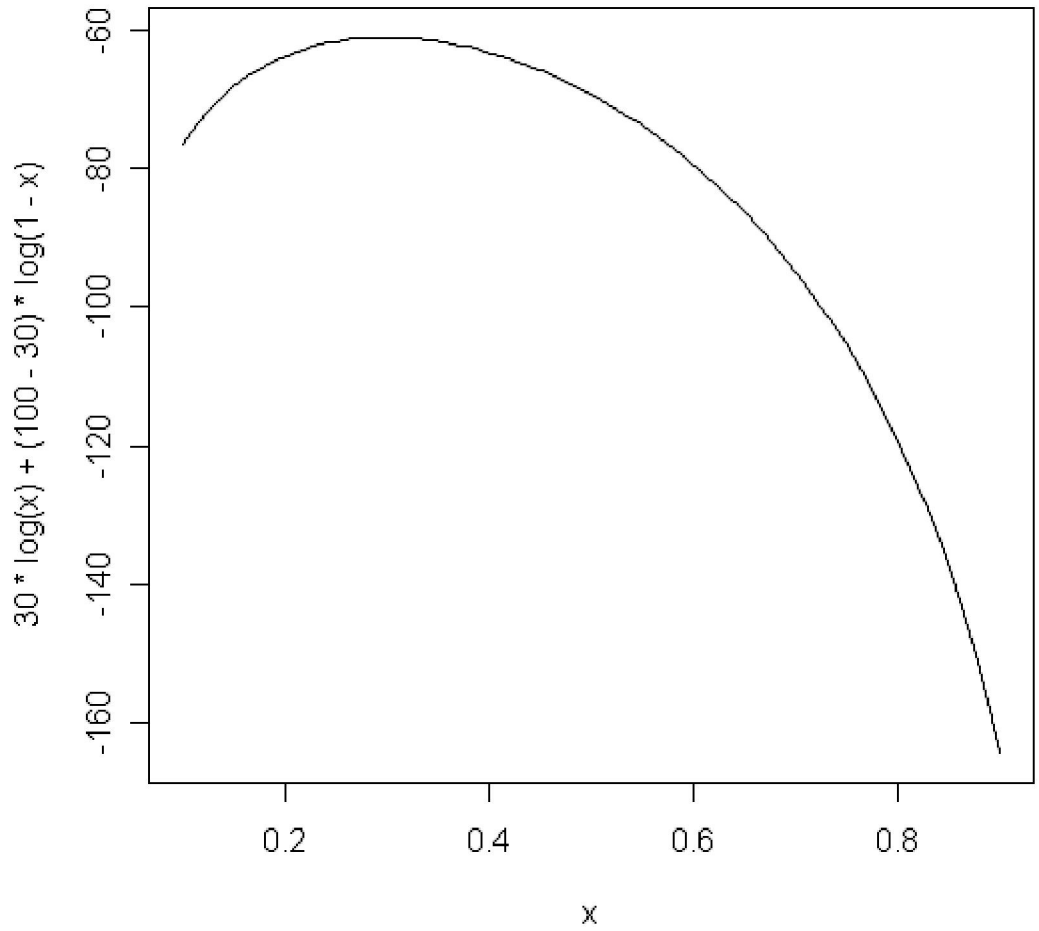
$$l = \log(likelihood) \propto y \log p + (100 - y)\log(1 - p)$$

$$\frac{\partial l}{\partial p} = 0$$ 을 만족시키는

$$\hat{p} = y/100$$

혹은

```
> curve(30*log(x)+(100-30)*log(1-x), 0.1, 0.9)
```


30 * log(x) + (100 - 30) * log(1 - x)
-60
-80
-100
-120
-140
-160
0.2
0.4
0.6
0.8
x

```
> f=function(x) 30*log(x)+(100-30)*log(1-x)
> optimize(f,c(0.1,0.9),maximum=TRUE)
$maximum
[1] 0.3000203

$objective
[1] -61.08643
```