



통계량 + 분산분석

2025 Spring

Industrial Data Science Lab & Unique AI

데이터의 값

- 수치형: 1,2,3,4,5,... 1.1,2.4,3.1,...
- 논리형: True or False
- 범주형: "합격" 또는 "불합격" 등
- 텍스트: "오늘의 뉴스는..."

정형 / 반정형 / 비정형 데이터

Structured / Semi Structured / Unstructured

1. 데이터의 이해

- 정형 데이터의 유형

- 1) 질적 자료(qualitative data=비계측자료 : nonmetric data)

- 범주형 자료(categorical data) : 예) 거주지, 성별, 등
 - 측정척도에 의한 구분
 - ① 명목자료(nominal data)
 - » 범주간 순서가 없음 : 예) 성별, 거주지, 등
 - ② 순서자료(ordinal data)
 - » 범주들 간에는 순서가 있는 자료, 사칙연산을 할 수 없음
예) 서비스 선호도(5점 척도) 등

- 2) 양적 자료(quantitative data=계측자료 : metric data)

- 수치자료, 사칙연산 가능
 - 합계, 평균, 최대값, 최소값, 분산 등으로 자료를 요약/정리 가능
 - 예 : 요금, 데이터 사용량 등

1. 데이터의 이해

- 분석 대상 데이터는 여러 개의 관측된 개체가 있고, 개체 마다 속성이 있음
 - 관측된 개체는 데이터의 행으로, 속성은 데이터의 열로 고려하며, 속성을 변수로 이해
- ① 관측된 개체
 - Observation=Case=individual=object
- ② 개체의 속성
 - Variable=attribute=feature=item(Qualitative variable: nominal, ordinal / Quantitative variable: interval)
 - 종속변수=Dependent variable=response variable=target variable=Y 변수
 - » 다른 변수에 의해 영향을 받는 변수
 - 독립변수=설명변수=Independent variable=explanatory variable=input variable=X 변수
 - » 종속 변수에 영향을 주는 변수

2. 통계학이란

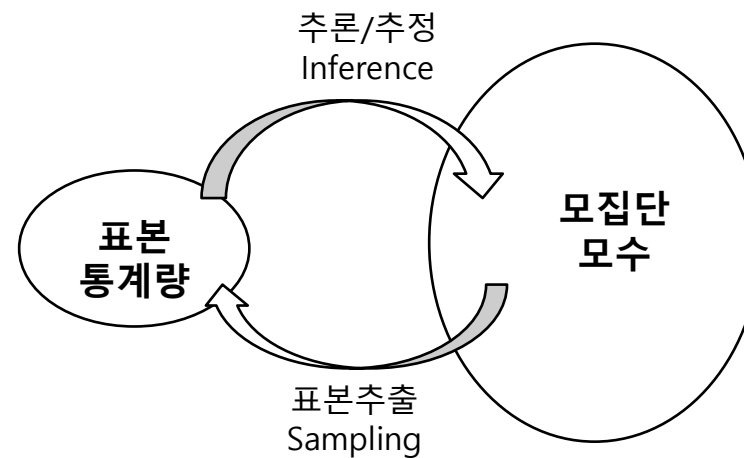
- 통계학?

- 1) 모수(parameter) : 불변

- 모집단의 특성을 수치로 나타낸 척도
 - 모수들은 전수조사(census)를 통해 얻은 자료로부터 구해짐.
 - 모집단의 평균인 모평균(population mean) : μ
 - 모집단의 분산인 모분산(population variance) : σ^2

- 2) 통계량(statistic) : 가변

- 표본자료로부터 얻어진 표본의 특성을 수치로 나타낸 척도
 - 표본의 평균인 표본평균(sample mean) : \bar{X}
 - 표본의 분산인 표본분산(sample variance) : S^2



3. 모집단과 표본

- 모집단과 표본

1) 모집단 (Population)

- 얻고자 하는 정보와 관련 있는 모든 개체로 부터 얻을 수 있는 **모든 관측값들의 집합**, 관심을 갖는 대상 전체
- 조사 및 관심의 대상이 되는 원소 하나 하나의 전체 집합(모임)
- 모집단은 일반적으로 매우 크고, 실제로 무한히 클 수도 있음
- 모집단 전체를 조사해서 얻은 통계자료를 모집단자료(Population data)

2) 표본 (Sample)

- 모집단의 일부분으로 원하는 정보를 얻기 위해 수행한 관측을 통해 얻어진 관측값
- 표본 공간 : 통계적 실험에서 모든 가능한 실험결과들의 집합
- 모집단의 특성을 파악하기 위하여 추출된 모집단의 일부. 즉, **모집단의 부분집합(subset)**
- 모집단의 일부분인 표본으로부터 조사된 자료를 표본자료(sample data)라고 함.

3. 모집단과 표본

- 통계량: 자료의 정리 및 요약

통계적 분석은 자료의 분포가 가지고 있는 특성을 찾아내서 그 특성을 숫자로 표시하기 위한 작업

- 분포의 특성

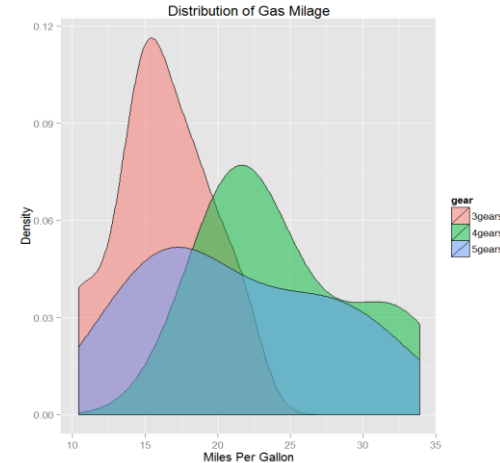
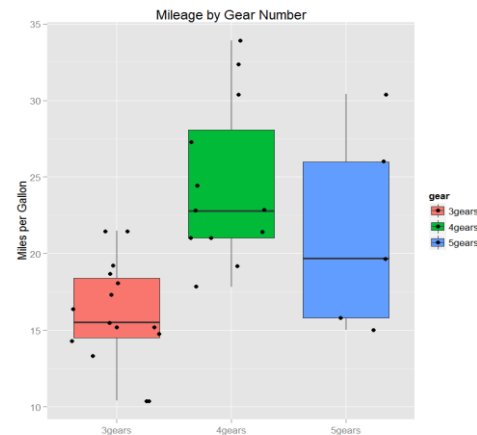
- 집중화 경향 (중심값 - 산술평균, 중앙값, 최빈값): 자료가 어느 위치에 집중되어 있는가를 나타냄
- 산포도 (범위, 분산, 표준편차, 백분위수): 자료가 산술평균을 중심으로 흩어져 있는 정도

| 이름 | 성별 | 나이 | 거주지 | 직업 | 요금 | 데이터 사용량 | 휴대폰 선호도 | 서비스 선호도 |
|-----|----|----|-----|-----|-------|------------|------------|------------|
| AAA | F | 20 | 서울 | 회사원 | 55000 | 3GB | LG | 5 |
| BBB | F | 19 | 인천 | 자영업 | 45000 | 9GB | 삼성 | 4 |
| CCC | M | 25 | 김포 | 회사원 | 35000 | 1GB | 샤오미 | 3 |
| DDD | F | 42 | 대전 | 회사원 | 75000 | 4GB | LG | 5 |
| EEE | F | 27 | 서울 | 자영업 | 65000 | 2GB | 소니 | 4 |
| FFF | M | 20 | 서울 | 회사원 | 55000 | 3GB | LG | 5 |
| GGG | M | 43 | 서울 | 자영업 | 45000 | 9GB | 삼성 | 4 |
| HHH | M | 25 | 대전 | 회사원 | 95000 | 11GB | 샤오미 | 3 |
| III | F | 42 | 김포 | 회사원 | 45000 | 3GB | LG | 5 |
| JJJ | F | 27 | 인천 | 자영업 | 40000 | 4GB | 소니 | 4 |

1. 기술통계 vs 추론통계

- 기술통계

- Descriptive Statistics
- 자료의 특성을 표, 그림, 통계량 등을 사용하여 쉽게 파악할 수 있도록 정리/요약, 자료를 요약하는 기초적 통계
- 예:
 - 중심위치의 측도: 표본평균, 중앙값
 - 산포의 측정: 분산, 표준편차, 사분위범위수, 백분위수, 변동계수, 평균의 표준오차
 - 분포의 형태에 관한 측도: 왜도(양수->왼쪽으로 치우친, 음수->우측으로 치우친), 첨도



1. 기술통계 vs 추론통계

- 추론통계

- Statistical Inference
- 모수 추정(Parameter estimation)
 - 점추정(Point Estimation): 모수가 특정한 값일 것이라고 추정, 표본의 평균/중위수/최빈값 등을 사용, 불편성/효율성/일치성/충족성
 - 표본 평균과 표본 분산
 - 구간추정(Interval Estimation): 점추정의 정확성을 보완하기 위해 확률로 표현된 믿음의 정도 하에서 모수가 특정한 구간에 있을 것이라고 선언
 - 추정량의 분포에 대한 전제 필요, 구해진 구간 내 모수가 있을 가능성의 크기(신뢰수준) 필요
 - 예: 95% 신뢰수준 하에서 모평균의 신뢰구간
- 가설검정(Hypothesis Test)
 - 모집단에 대한 어떤 가설을 설정한 뒤, 표본 관찰을 통해 가설의 채택여부 결정
 - 표본관찰이나 실험을 통해 귀무가설과 대립가설 중 택1
 - 귀무가설이 옳다는 전제 하 검정통계량을 구하여, 이 값이 나타날 가능성의 크기로 판단
 - 귀무가설/대립가설/검정통계량/유의수준/기각역/채택역
 - 제1종 오류/제2종 오류

2. 통계량의 이해

- 집중화 경향: 자료가 어느 위치에 집중되어 있는가를 나타냄, 평균, 중앙값 등이 있음
- 예를 들어, 평균은 집중화 경향에 대한 자료의 특성을 대표할 수 있는 값으로, 모든 관측값을 더해서 관측값 개수로 나누어 구하고, 중앙값은 어떤 주어진 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값

2. 통계량의 이해

- **Measure of central tendency**

- **평균(mean)**

- (산술)평균(mean; arithmetic mean; average), 균형점(자료의 중심), 모든 관측값의 크기(정보)를 반영, 이상값(outlier)의 영향을 받음.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

- **중앙값(Median)**

- 자료를 관측값의 크기순으로 배열하였을 경우, 중앙에 위치하는 관측값 이상값에 덜 민감
 - 관측값의 수가 짝수이면 중앙의 두 관측값을 평균함.
 - 50% percentile

- **최빈값(mode)**

- 자료 중에서 가장 많이 출현하는 관측값
 - 명목자료인 경우 평균과 중앙값은 무의미(예 : 허리사이즈)
 - 존재하지 않을 수도 있으며, 1개 이상 존재할 수도 있음.

2. 통계량의 이해

- **Measure of dispersion**

- 필요성

- 자료를 대표값으로 요약/정리하는 것 만으로 충분하지 않음
 - 자료에서 관측값들이 얼마만큼 퍼져있는가를 측정하는 척도인 산포도를 고려
 - 산포도는 자료에서 관측값들이 변화하는 크기인 변동량을 나타내는 값

- 범위(range)

- 자료의 관측값 중 가장 큰 값인 최대값(max)과 가장 작은 값인 최소값(min)과의 차이
 - 범위(range)=최대값(max)-최소값(min)
 - 자료들 중 두 관측값만 이용하며 관측값 하나 하나의 크기가 반영되지 못함.
 - 이상값에 의해 크게 영향 받음.

2. 통계량의 이해

- **Measure of dispersion**

- **백분위수와 사분위수**

- 평균과 표준편차는 자료의 분포에 대해 중요한 정보를 제공하지만, outlier 등의 영향을 받을 수 있으며, 자료 분포의 치우침 등에 대한 정보는 아님.
 - Outlier나 Skewness 등에 영향을 받지 않고 자료를 파악하기 위해 median이나 interquartile range 등을 활용

- **백분위수(Percentile)**

- 자료를 크기 순서에 따라 나열한 자료를 100등분하는 수로, X 분위값이란 자료가 X%보다 작거나 같게 되는 값

- **사분위수(Quartile)**

- 백분위수가 25%, 50%, 75%인 경우

2. 통계량의 이해

- **Measure of dispersion**

- **분산과 표준편차**

- 두 통계량 모두 각 자료가 평균에서 얼마나 퍼져있는지를 보는 정도, 관측값들이 자료의 중심인 평균으로부터 얼마나 떨어져 있는가의 척도, 즉 평균과 차이들의 평균

- **분산(Variance)**

- 각 자료가 평균에서 얼마나 퍼져있는지를 보는 정도. 각 자료의 평균과의 차이에 대한 평균
 - 모분산(population variance): 모집단으로부터 전수조사를 하여 얻은 관측값인 경우 모집단의 분산

- » $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$

- **표본 분산(sample variance)**

- $S^2 = \frac{1}{n-1} \sum_{i=1}^N (X_i - \bar{X})^2$

- **표준편차**

- 모분산이나 표본분산은 관측값들의 편차를 제공하여 계산하므로 모분산이나 표본분산의 측정단위는 관측값들의 측정단위와 일치하지 않으며, 그렇기 때문에 관측값의 측정단위와 일치시키기 위해서는 분산의 양의 제곱근을 사용

3. 공분산과 상관관계

- Measure of dispersion
 - 공분산 (Covariance)
 - 분산은 한 변수에 대한 자료의 퍼짐이며, 두 변수간의 관계를 알기 위해 공분산을 사용
 - 공분산은 X, Y가 각각의 평균들로부터 떨어진 거리, 변수의 편차간의 곱의 평균

$$\sigma(x, y) = E [(x - E[x])(y - E[y])],$$

3. 공분산과 상관관계

- 공분산
 - 두 변수의 상관(하나가 증가할 때 다른 하나는 감소하거나, 하나가 증가할 때 다른 하나도 증가하는 등의 관계) 정도를 나타내는 값
- 공분산의 해석
 - 공분산 > 0
 - 공분산 $= 0$
 - 공분산 < 0

3. 공분산과 상관관계

- Measure of dispersion

- 상관관계:

- 두 변수의 공분산을 각 변수의 편차로 나눠서 -1~1 사이로 조정한 값
 - Scaled version of the covariance between X and Y
 - Pearson correlation

$$\sigma(x, y) = E [(x - E[x])(y - E[y])],$$

- 상관관계는 -1~1사이의 값으로 Scale되며, 1은 두 변수 간 강한 상관관계가 있음을, 0은 두 변수가 관계가 없음을, -1은 두 변수간 음의 상관관계가 강하게 있음을 의미

$$R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

3. 공분산과 상관관계

- 기대값, 평균, 분산

- $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$

- $\sigma^2 = V(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (X - \mu)^2 xf(x)dx = E(X^2) - \mu^2$

- $E(aX) = aE(X)$

- $V(aX) = a^2V(X)$

4. 확률

- **확률(Probability)**

- 불확실성을 나타내는 측도(measure)로 0~1의 값으로 표현
- 확률(probability)은 어떤 특정한 사상이 발생할 가능성을 숫자로 나타낸 측정치임.
즉, 사상의 발생할 가능성을 말함.
- Random한 실험에서 실험결과가 항상 동일하지 않으므로 불확실성이 있으며, 이것을 확률로 측정

- **독립성**

- 두 사건 A, B에 대해 $P(A \text{ and } B) = P(A)P(B)$ 이면 A와 B는 서로 독립

(두 사건이 독립이 아닌 경우)

- **조건부 확률**

- 하나의 사건이 일어났을 때, 이 사건이 다른 사건과 **관련이 있는 경우** 사용되는 방법.
- A가 주어졌을 때, B가 일어날 조건부 확률은 $P(B|A)$ 로 표시하며 $P(A) > 0$ 면,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- A, B 가 서로 배반이면
 - $P(A \cup B) = P(A) + P(B)$ 이고, $P(A \cap B) = 0$

4. 조건부 확률과 베이즈 정리

- A, B에 대해서
 - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 - A, B가 exclusive $\Leftrightarrow A \cap B = \emptyset \Leftrightarrow P(A \cap B) = 0 \Leftrightarrow P(A \cup B) = P(A) + P(B)$
 - 조건부 확률
 - $P(A|B) = P(A \cap B) / P(B)$
 - 승법정리: $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$
 - A, B가 statistically independent $\Leftrightarrow P(A \cap B) = P(A)P(B) \Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(B|A) = P(B)$

4. 조건부 확률과 베이즈 정리

- 조건부 확률 (두 사건이 독립이 아닌 경우)
 - 하나의 사건이 일어났을 때, 이 사건이 다른 사건과 **관련이 있는 경우** 사용되는 방법.
 - A가 주어졌을 때, B가 일어날 조건부 확률은 $P(B|A)$ 로 표시하며 $P(A)>0$ 면,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

multiplication rule

- 표본공간이 S에서 A로 바뀐 것을 의미
- A, B 가 서로 배반이면

$$P(A \cup B) = P(A) + P(B) \text{이고, } P(A \cap B) = 0$$

$$P(A_1 \cup A_2 | B) = P(A_1 | B) + P(A_2 | B)$$

4. 조건부 확률과 베이지 정리

- 두 사건의 독립
 - 두 사건 A, B 에 대해 $P(A \cap B) = P(A)P(B)$ 이면 A 와 B 는 서로 독립
- 조건부 독립
 - $P(H) > 0$ 인 사건 H 와 두 사건 A, B 에 대해 $P(A \cap B | H) = P(A | H)P(B | H)$ 이면 A 와 B 는 H 가 주어진 조건하에서 서로 조건부 독립
 - H 가 주어졌을 때, A 가 추가되는 것은 B 에 대한 정보를 아는데 영향을 미치지 않음.

5. 확률변수와 확률 분포

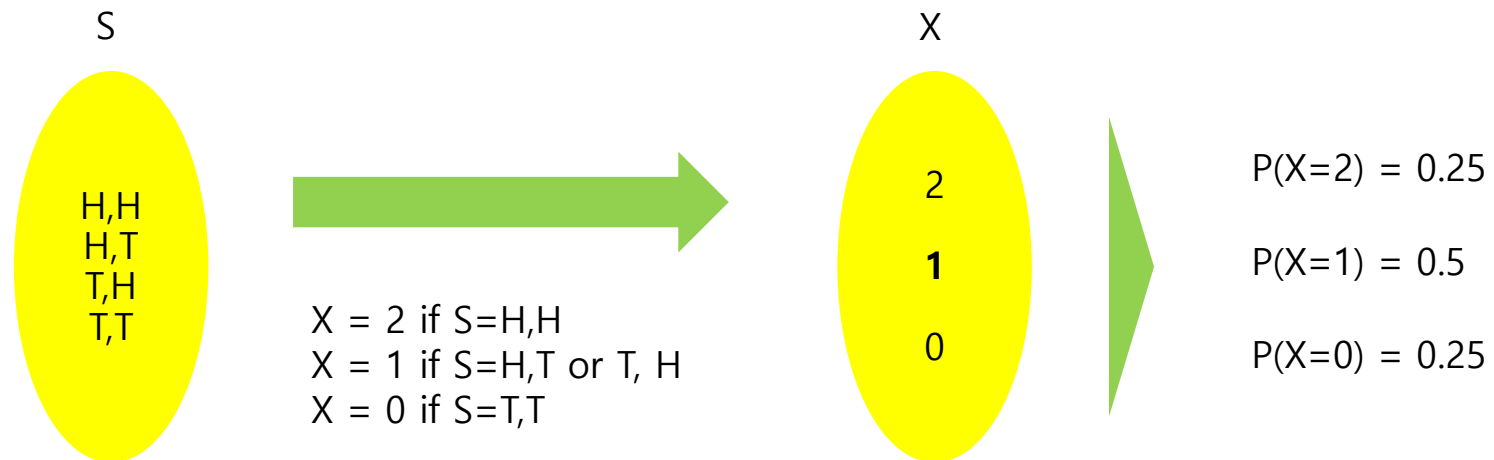
- **확률변수** : 측정치로부터 얻을 수 있는 값의 총 집합을 표본공간이라 하는데, 표본 공간상의 각각의 값에 실수를 부여하는 함수
- **확률분포** : 확률변수가 취할 수 있는 모든 값들에 대해 이들 값들이 취할 수 있는 확률을 그림이나 표(함수식)로 나타낸 것
 - 1) 이산확률분포 : 불량수나 결점수와 같이 셀 수 있는 확률변수에 대응되는 확률분포
 - 2) 연속확률분포 : 제품이 중량이나 치수와 같이 셀 수 없는 연속값을 갖는 확률분포

5. 확률변수와 확률 분포

- **Distribution**

- 예2: 동전을 두 번 던진다면?
 - 확률변수는 앞면의 횟수로 고려

| Sample Space(S) | R.V. (X) | P(X=x) |
|-----------------|----------|--------|
| H,H | 2 | 0.25 |
| H,T | 1 | 0.25 |
| T,H | 1 | 0.25 |
| T,T | 0 | 0.25 |



5. 확률변수와 확률 분포

- 조건부 확률의 확률분포

- 결합확률분포

| | X | | |
|---|-----|-----|-----|
| Y | 1 | 2 | 3 |
| 1 | 0.2 | 0.3 | 0.2 |
| 2 | 0.1 | 0.1 | 0.1 |

- 주변확률분포

| x | 1 | 2 | 3 |
|----------|-----|-----|-----|
| $f_X(x)$ | 0.3 | 0.4 | 0.3 |

| y | 1 | 2 |
|----------|-----|-----|
| $f_Y(y)$ | 0.7 | 0.3 |

- Y가 주어진 X의 조건부 분포 $f(x|y) = f(x,y) / f(y)$

| x | 1 | 2 | 3 |
|--------------|---------|---------|---------|
| $f(x y = 1)$ | 0.2/0.7 | 0.3/0.7 | 0.2/0.7 |
| $f(x y = 2)$ | 0.1/0.3 | 0.1/0.3 | 0.1/0.3 |

- X, Y는 독립?
- $f_{XY}(1, 1) = 0.2$
- $f_X(1) = 0.3$
- $f_Y(1) = 0.7$

5. 확률변수와 확률 분포



- 이항분포(binomial distribution)

베르누이 시행의 조건 -예) 동전 던지기

- 1) 시행의 결과는 한 사건은 성공(S), 다른 사건은 실패(F)로서 상호 배타적인 두 사건
- 2) 각 시행에서 성공이 나타날 확률은 $p=P(S)$, 실패가 나타날 확률은 $q=P(F)=1-p$
성공과 실패가 나타날 확률의 합은 $p+q=1$
- 3) 각 시행은 서로 독립적
- 4) 확률밀도함수: $p^x(1-p)^{n-x}$
 $X \sim \text{Ber}(p)$, $E(X) = p$, $\text{Var}(X) = p(1-p)$

이항분포는 여러 번의 베르누이 시행을 할 때 나타나는 분포

- 이항분포의 확률 밀도 함수 $X \sim B(n, p)$
- $E(X) = np$, $\text{Var}(X) = np(1-p)$

$$P(X=x|p) = \binom{n}{x} p^x (1-p)^{n-x} = {}_n C_x p^x (1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

5. 확률변수와 확률 분포

- 포아송분포(Poisson distribution)

단위시간이나 단위공간에서 특정 사건이 드물게 발생할 때

- 강판, 직물 등의 연속체에 평균 m 개의 흠이 있을 때,
- 랜덤하게 일정 단위를 채 취하여 흠을 조사할 때, 흠이 x 개 나타날 확률
- 단위시간 내에 은행에 찾아오는 고객의 수,
- 어느 지역의 하루 교통사고 수

포아송분포의 밀도함수

$$P(X=x) = \frac{e^{-m} m^x}{x!}$$

m : 평균발생횟수

x : 사건발생횟수

- 포아송 분포의 특성

- 이항분포에서 $p < 0.1$ 일 때, 포아송분포로 변화
- 포아송분포에서 $m > 5$ 일 때, 정규분포로 변화

5. 확률변수와 확률 분포

- 정규분포(Normal distribution)

- 1) 정규분포의 모양과 위치는 분포의 평균과 표준편차로 결정
- 2) 정규분포의 확률밀도함수는 평균을 중심으로 대칭인 종 모양
- 3) 정규곡선은 X축에 맞닿지 않으므로 확률변수 X가 취할 수 있는 값의 범위는 $-\infty < X < +\infty$ 이다
(관측값의 99.7%가 $\pm 3\sigma$ 안에 속해 있다)
- 4) 분포의 평균(μ)과 표준편차(σ)가 어떤 값을 갖더라도 정규곡선과 X축사이의 전체면적은 1이다

- 정규분포의 밀도함수

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$-\infty < x < \infty, \pi: 3.142(\text{원주율}), e: 2.7183$$

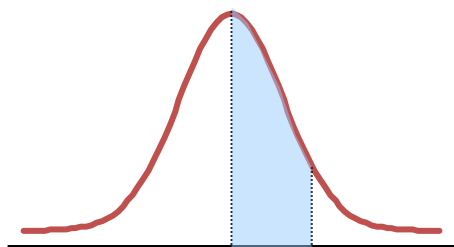
5. 확률변수와 확률 분포

- 표준정규분포

- 표준정규분포는 정규분포를 평균 $\mu=0$, 표준편차 $\sigma=1$ 이 되도록 **표준화** 한 것.
- 어떤 관측값 X 의 값이 그 분포의 평균으로부터 표준편차의 몇 배 정도나 떨어져 있는가를
- 다음과 같이 표준화된 확률변수 Z 로 나타내며, $N(0,1^2)$ 으로 표시한다.

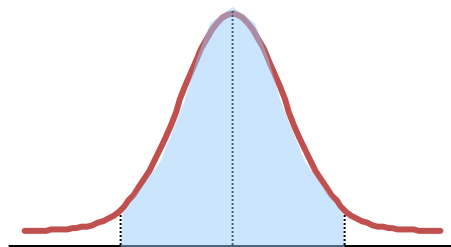
$$Z = \frac{X - \mu}{\sigma}$$

$Z=0$ 부터 $Z=1.5$ 사이에
확률변수가 있을 확률



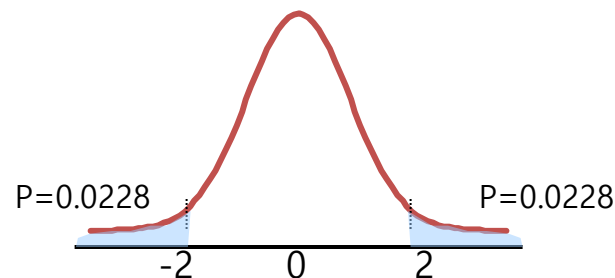
0
 $P(0 \leq Z \leq 1.5) = 0.4332$

$Z=0$ 에서 $\pm 45\%$ 에 해당하
는 Z 값



0
 $Z = \pm 1.6449$

$Z=-2$ 보다 작거나 $Z=2$ 보다 큰 사이에
확률변수가 있을 확률



$P(-2 \geq Z, Z \geq 2) = 0.0456$

5. 확률변수와 확률 분포

- Γ - 분포

- 양의 실수 영역에서 정의된 연속확률분포, 다양한 실세계 현상을 모델링할 때 유용
- PDF

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

- α : 형상 모수(shape parameter), β : 척도 모수(scale parameter), λ : 비율 모수(rate parameter, $1/\beta$)
- $\Gamma(\alpha)$ 는 감마 함수로, 정수 n 에 대해서는 $\Gamma(n) = (n-1)!$
- $E(X) = \alpha\beta$, $V(X) = \alpha\beta^2$
 - $\alpha = 1$ 이면 지수분포
- 활용:
 - 지수분포의 일반화: 지수분포는 감마분포의 특별한 경우
 - 대기시간 모델링: 사건이 여러 번 일어나는 데 걸리는 총 대기시간을 설명
 - 비대칭 데이터 모델링: 평균보다 큰 극단값이 자주 나타나는 데이터(오른쪽 꼬리)가 있는 경우 적합

5. 확률변수와 확률 분포

- 지수 분포

- 지수분포(Exponential Distribution)

- 어떤 사건이 발생할 때까지 걸리는 시간을 모델링
- 포아송 과정과 밀접한 관계가 있어, 시간 간격이나 대기 시간과 관련된 현상에서 널리 활용
- PDF
- $f(x; \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0$
- $\lambda > 0$, rate parameter이며 단위시간당 사건이 발생하는 평균 횟수
- $E(X) = \frac{1}{\lambda} \quad V(X) = \frac{1}{\lambda^2}$
- memoryless
 - $P(X > s+t \mid X > s) = P(X > t)$
- 활용
 - 다음 버스 올 때까지의 시간, 기계가 고장 나기까지의 시간, 패킷이 수신될 때까지의 시간, 사건 간의 시간 간격

5. 확률변수와 확률 분포

- 포아송, 감마, 지수 분포

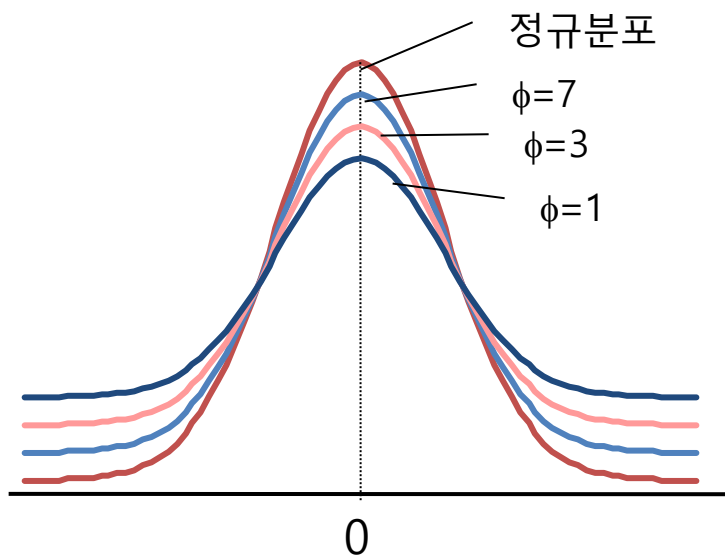
- 포아송 분포: 일정한 시간 내 발생한 사건의 횟수, 사건의 개수
- 지수 분포: 다음 사건이 발생할 때까지 걸리는 시간, 사건 간의 시간 간격
- 감마분포: n번째 사건이 발생할 때까지 걸리는 총 시간, 누적 대기시간(지수분포의 합)

- 예: 버스가 평균적으로 10분에 1대씩 오는 경우, $\lambda = \frac{1}{10}$
 - 다음 버스를 기다리는 시간: 지수 분포
 - 3번째 버스가 오기까지 걸리는 시간: 감마 분포
 - 30분동안 버스가 몇 대 왔는지: 포아송 분포

5. 확률변수와 확률 분포

- t - 분포(서로 다른 두 집단의 평균의 통계 검정)
- 정규분포로부터 확률표본이 크지 않고 표준편차(σ)를 모를 때

$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ 는 자유도 $n-1$ 인 t-분포를 따름.

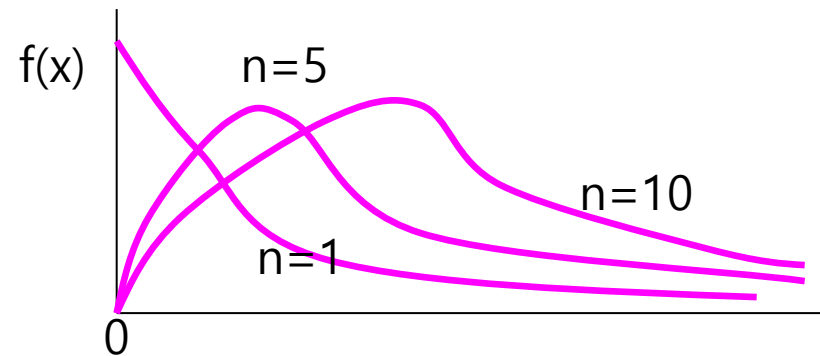


- T-분포의 특징

- t분포는 정규분포보다 퍼져 있으며 자유도(ϕ)가 커질수록 정규 분포에 근접
- 표본의 크기가 작아 표본의 표준편차(s)가 모집단의 표준편차(σ)보다 불확실성이 크기 때문
- 표본의 크기 n 이 커질수록 표본의 표준편차가 모집단의 표준편차에 접근하기 때문
- t분포는 자유도에 따라 달라지며 자유도는 표본의 크기에서 1을 뺀 것으로 $n-1$ 로 표시

5. 확률변수와 확률 분포

- χ^2 - 분포(서로 다른 2개 이상 집단의 비율의 통계 검정)
 - 단일 모집단의 분산의 표본분포는 χ^2 분포를 이용하여 나타낼 수 있음
 - 두 모집단의 분산의 표본분포는 F분포를 이용하여 나타낼 수 있음
- 정규모집단 $N(\mu, \sigma^2)$ 으로부터의 확률표본 X_1, X_2, \dots, X_n 에 대해
- $\chi^2 = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$ 의 분포를 자유도 $n-1$ 인 χ^2 분포라 한다.

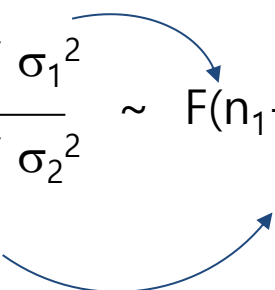


5. 확률변수와 확률 분포

- F - 분포(서로 다른 2개 이상 집단의 분산의 균질성 검증)
 - 모집단이 정규분포를 이루며 각각 σ_1^2, σ_2^2 라는 분산을 갖는 두 개의 모집단에서 각각 크기가 n_1, n_2 인 두 표본을 추출하여 표본분산을 계산
 - 두 표본분산이 S_1^2, S_2^2 이라고 할 때, 표본분산과 모분산의 비율로 이루어진 두 개의 χ^2 의 비율은 F분포를 이루며, F분포는 두 개의 자유도를 갖음

$$\chi_1^2 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi_1^2 (n_1-1)$$

$$\chi_2^2 = \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi_2^2 (n_2-1)$$

$$\frac{\chi_1^2 / (n_1-1)}{\chi_2^2 / (n_2-1)} = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1-1, n_2-1)$$


- 위의 식에서 F분포는 두 모집단의 분산을 비교하는 것임을 알 수 있음

6. 통계 검정

- 통계적 가설 검정(Hypothesis Testing)

- 가설 검정 = 가설(Hypothesis) + 검정(Testing)
- 가설검정이란?
 - 표본(Sample)으로부터 주어지는 정보를 이용하여, 모수에 대한 예상 혹은, 주장 또는 단순한 추측 등의 옳고 그름을 확률적인 개념을 이용하여 판정하는 과정
- 가설(Hypothesis)
 - 귀무가설(Null Hypothesis ; H_0)
 - "Equal(=) ; 같다"
 - 과거부터 알려져 왔던 모수에 대한 일반적인 내용
 - 대립가설(Alternative Hypothesis ; H_1)
 - "Not Equal(\neq) ; 다르다"
 - 자료로부터 강력한 증거에 의하여 입증하고자 하는 내용

6. 통계 검정

- 가설 검정 절차

- Step #1

가설 설정

귀무가설(H_0)과 대립가설(H_1)을 세운다.

- Step #2

검정 통계량(Test Statistics) 산출 or 유의 수준(α) 결정

- Step #3

기각치(Critical Value) 산출 or P-Value 산출


- Step #4

귀무가설(H_0)의 기각 여부 결정

If P-Value < 유의수준(α)이면, 귀무가설 (H_0) 기각

- Step #5

기술적 용어로 해석



Significance level, 귀무가설을 기각하게 되는 확률의 크기,
귀무가설이 옳음에도 이를 기각하는 확률의 크기

6. 통계 검정

- 통계적 가설 검정
 - 오류가 발생할 수 있음
 - 오류는 다음과 같이 제1종 오류와 제2종 오류로 나누어 고려
 - 제1종 오류: Type I error
 - 귀무가설(H_0)이 참이지만, 기각된 경우 발생
 - False positive로 이해할 수 있음
 - 제 1종 오류의 수준은 α (alpha)로 표시하며, alpha level 또는 유의수준(significance level)으로도 사용
 - 유의수준은 일반적으로 0.05 (5%)로 사용하며, (True임에도) 귀무가설을 잘못 기각하는 확률을 5%까지는 용인한다는 의미
 - 제2종 오류: Type II error
 - 귀무가설이 거짓임에도, 기각을 못한 경우 발생
 - False negative로 이해
 - 제2종 오류의 수준은 β (beta)로 표시하며, 검정력($1-\beta$)과 관련있는 것으로 고려

6. 통계 검정

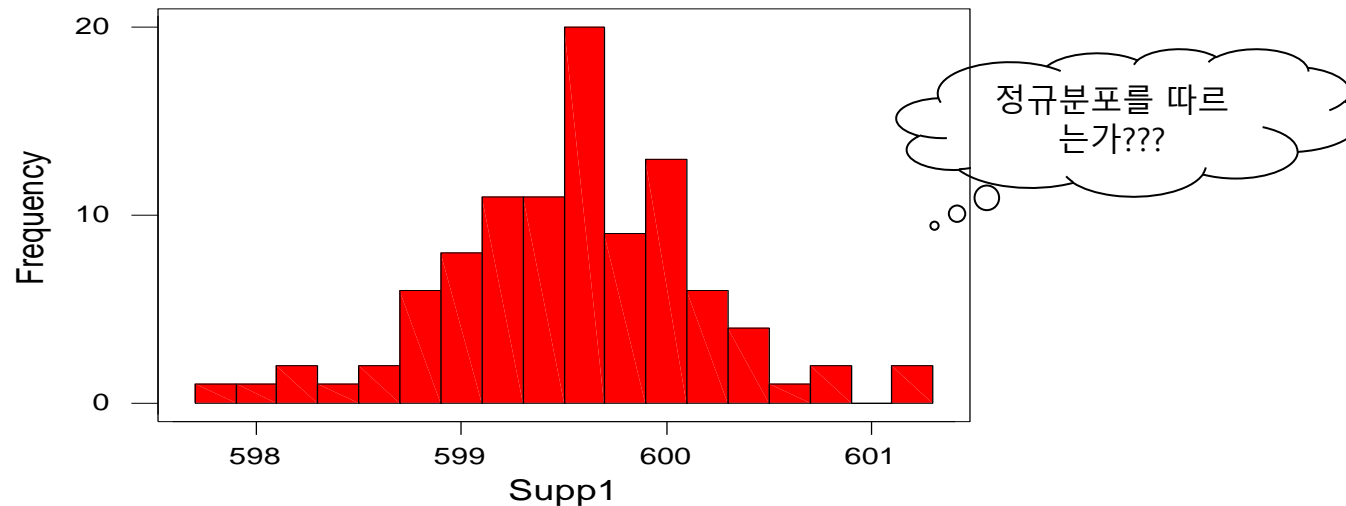
- 정규성 검정 - Normality Test

- 목적

- 대부분의 통계 분석에 있어, 데이터의 정규성을 가정한다.
 - 주어진 데이터의 정규성을 통계적인 방법으로 검정

- 가설

- 귀무가설(H_0) : 정규분포를 따른다.
 - 대립가설(H_1) : 정규분포를 따르지 않는다.



6. 통계 검정

- **t-검정 (t-Test)**
 - t-검정: 표본 데이터가 정규 분포를 따른다는 가정 하에 수행
 - 단일 표본 t-검정 (One-Sample t-Test): 한 집단의 평균이 특정 값과 차이가 있는지를 검정
 - 독립 표본 t-검정 (Independent Two-Sample t-Test): 두 독립적인 집단의 평균이 서로 차이가 있는지를 비교.
 - 쌍체 비교 (Paired Sample t-Test): 동일한 집단에 대해 두 번의 측정을 했을 때, 측정 전후 결과의 평균이 서로 다른지를 검정.

6. 통계 검정

- 쌍체 비교(Paired t-Test)

- 목적

- 동일한 모집단을 대상으로 처리 전, 후에 차이가 있는지 여부를 통계적인 방법을 이용해서 검정

- 가설

- 귀무가설(H_0) : $\delta = 0$ (두 모집단의 평균은 같다.)
 - 대립가설(H_1) : $\delta \neq 0$ (두 모집단의 평균은 다르다.)

| X | Y | d(=X-Y) |
|----------|----------|----------|
| X_1 | Y_1 | d_1 |
| X_2 | Y_2 | d_2 |
| \vdots | \vdots | \vdots |
| X_k | Y_k | d_k |

6. 통계 검정

- **쌍체 비교(Paired t-Test)**

- 쌍체 비교의 예

- 환자들에 대해 새로운 치료법을 적용한 실험 결과
 - 각 환자별로 기록된 수치는 환자들로부터 측정한 약의 효과(높을수록 좋은 수치)를 처치 전/후로 나누어 정리

- “약의 효과가 있을까?”

- 새로운 치료법 적용 전후의 효과 차이를 검정하기 위해 Paired T-test를 이용할 수 있음.
 - 귀무가설은 치료법 적용 전후에 효과의 평균 차이가 없는 것을 가정
 - 유의수준이 5%, P-value가 5% 유의수준보다 크면, 새로운 치료법으로 인한 개선 효과는 없다고 볼 수 있음

| | 환자1 | 환자2 | 환자3 | 환자4 | 환자5 | 환자6 | 환자7 | 환자8 | 환자9 | 환자10 |
|------|------|------|------|------|------|------|------|------|------|------|
| 처치 전 | 51.4 | 52 | 45.5 | 54.5 | 52.3 | 50.9 | 52.7 | 50.3 | 53.8 | 53.1 |
| 처치 후 | 50.1 | 51.5 | 45.9 | 53.1 | 51.8 | 50.3 | 52 | 49.9 | 52.5 | 53 |

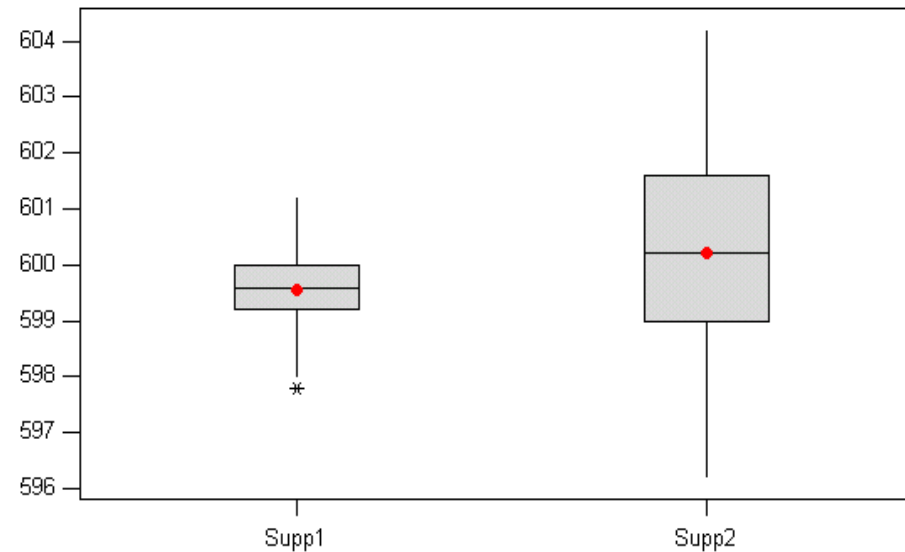
6. 통계 검정

- 범주형 데이터 분석에 활용
 - 표준 정규 분포를 따르는 독립적인 여러 확률 변수의 제곱합으로 정의
 - 비대칭적이며, 오른쪽으로 긴 꼬리
 - 자유도가 증가함에 따라 분포의 형태는 점점 더 정규 분포와 유사
- 활용
 - 적합도 검정 (Goodness-of-Fit Test):
 - 관찰된 빈도가 예상 빈도와 얼마나 잘 일치하는지를 평가, 실제 데이터 분포가 기대하는 이론적 분포를 따르는지 검정
 - 독립성 검정(Test of Independence):
 - 두 변수가 서로 독립적인지를 평가
 - 두 범주형 변수 간에 관련이 있는지를 확인 (예: 카이제곱 독립성 검정)
 - 분산 분석:
 - 두 집단의 분산이 동일한지 비교

6. 통계 검정

- F-Test

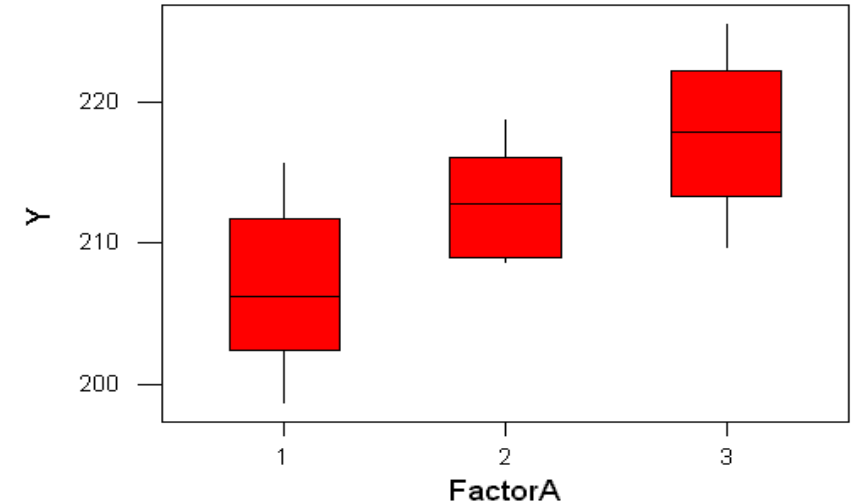
- 두 모집단 간에 산포 차이가 있는지 여부를 통계적인 방법을 이용해서 검정
- 귀무가설(H_0) : $\sigma_1^2 = \sigma_2^2$ (두 모집단의 산포는 같음)
- 대립가설(H_1) : $\sigma_1^2 \neq \sigma_2^2$ (두 모집단의 산포는 다름)



산포 차이가 있나?

6. 통계 검정

- 등분산 검정(Test of Equal Variances)
 - 등분산 검정: 두 모집단 간 혹은 세 집단 이상 간에 산포 차이가 있는지 여부를 통계적인 방법을 이용해서 검정
 - F-검정
 - 가장 기본적인 등분산 검정 방법
 - 두 집단의 분산 비율이 F-분포를 따르는지 검정하여, 두 집단 간 분산을 비교
 - 정규분포를 따라야 함
 - 그 외 Levene's Test, Bartlett's Test 등이 있음.
 - 가설
 - 귀무가설 $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$
 - 대립가설 $\sigma_i^2 \neq \sigma_j^2$



7. 분산분석

- 분산분석(ANOVA : Analysis Of Variance)
 - 여러 집단의 평균을 비교
 - 2개의 집단에 대한 평균을 비교하는 통계적 기법: t-Test
 - 3개 이상 집단의 평균 비교: 분산분석 (ANOVA : Analysis Of Variance)
 - 분산의 개념 기반하여 분석: 분산 계산 방식처럼 편차들의 제곱합을 해당 자유도로 나누어서 얻게 되는 값을 이용, 범주 각 수준의 평균들간의 차이 존재 여부 판단
 - 영국의 통계학자 피셔(R.A. Fisher)가 고안한 분석 기법, 농업 연구에서 사용이 시작되었으며, 사회과학, 공학, 의학 등 다양한 분야에 폭넓게 적용
 - 모형 구성: 설명변수는 범주형 자료(categorical data), 종속변수는 연속형 자료(continuous data)

7. 분산분석

- 일원분산분석 (One-way ANOVA)

- 독립변수에 하나의 요인(범주형 변수)만 고려하는 경우 사용
- 아래와 같은 자료에 대해 적용 (전체 Y 평균: \bar{Y})

| 요인(Factor) | Observation | Mean |
|-----------------|---------------------------------|------------------------------|
| Factor의 Level 1 | $Y_{11}, Y_{12}, \dots, Y_{1n}$ | Level 1 관측치의 평균, \bar{Y}_1 |
| ... | ... | ... |
| Factor의 Level r | $Y_{r1}, Y_{r2}, \dots, Y_{rn}$ | Level r 관측치의 평균, \bar{Y}_r |

- 이때, 측정된 각 값인 $Y_{ij} = \mu_i + \varepsilon_{ij}$ 로 나타나는데
 - $i=1,2,\dots,r$ 로 Factor의 여러 level 의미
 - $j=1,2,\dots,n$ 으로 각 level에 해당하는 관측치의 개수(level별로 동일하게 설정)
 - μ_i : i 번째 수준에서의 평균
 - Y_{ij}, ε_{ij} : i 번째 수준에서 측정된 j 번째 값이며, 이때의 오차. 특히, ε_{ij} 는 서로 독립이며, 정규분포 $N(\mu_i, \sigma^2)$ 을 따른다고 가정

7. 분산분석

- 일원분산분석 (One-way ANOVA)

- 분산분석은 관측치의 전체 변동(Sum of Squares)을, 비교하려는 요인 수준(Factor Level) 간 차이에 의해서 발생하는 변동(SSTR)과, 그 외 요인에 의한 변동(SSE)으로 나누어 분석
- 즉, 측정값과 전체 측정값 평균의 차이는 다음과 같이 나눌 수 있음

$$Y_{ij} - \bar{Y} = (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i)$$

- $Y_{ij} - \bar{Y}$: 관측치의 전체 변동
 - $\bar{Y}_i - \bar{Y}$: 수준 i에서 관측치와 전체 평균의 차이, 요인(범주형 변수)에 의한 변동
 - $Y_{ij} - \bar{Y}_i$: 관측치와 각 수준 평균 간 차이, 요인 수준i에 의해 설명될 수 없는 변동
- 위의 식을 양변을 제곱하여 더하면(Sum of Squares), 아래와 같이 $SST = SSTR + SSE$ 로 표현할 수 있음

$$\sum_{i=1}^r \sum_{j=1}^n (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^r n(\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^r \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2$$

$$SST = SSTR + SSE$$

- SST : 총제곱합(Total Sum of Squares)
- $SSTR$: 처리제곱합(Treatment Sum of Squares)
- SSE : 오차제곱합(Error Sum of Squares)

7. 분산분석

- 일원분산분석표 (One-way ANOVA Table)

| | Sum of Squares | degrees of freedom | Mean Square Error | F statistics |
|----|----------------|--------------------|-------------------|-----------------|
| 처리 | SSTR | r-1 | MSTR | <u>MSTR/MSE</u> |
| 오차 | SSE | nT - r | MSE | |
| 전체 | SST | nT - 1 | | |

F-통계량:
그룹간 변동성이 그룹내 변동성보다 커야하는 것을 의미!

- F 통계량으로 요인들의 수준 간 평균의 차이여부를 검정
- F 통계량: 처리평균제곱 (MSTR)이 커지면 오차평균제곱(MSE)은 작아지며, F 통계량 값이 커지게 됨. 즉, 수준 평균들간에 차이가 큼
- 가설 검정
 - 귀무가설 $H_0 : \mu_1 = \mu_2 = \dots = \mu_r$
 - 대립가설 $H_1 : \text{모든 } \mu_i \text{ 는 같지 않다. } i = 1, 2, \dots, r$
 - P-value < α 이면, H_0 기각 (H_0 reject)

7. 분산분석

- **다중비교 (Multiple Comparison)**

- 분산분석에서 한 요인(Factor) 내 세 개 이상의 수준(Levels)의 집단 간에 평균의 차이를 검정
- 그 결과 귀무가설 H_0 기각하는 경우, 각 인자 수준들의 평균이 같지 않다는 결론을 내리게 됨
- 이 때, 수준(Levels)에 따른 평균의 차이가 어떤 수준에서 발생했는지를 알고 싶을 때 사용
- Pairwise T-test, Tukey's HSD test를 사용할 수 있으나, Pairwise T-test는 1종오류(True->False 로의 오류)의 가능성이 커져서 Tukey's HSD test 사용 권장
- Tukey's HSD(honestly significant difference) test: Studentized range distribution 바탕으로 모든 가능한 두 수준들의 평균 차이 검정(pairwise post-hoc testing using Tukey HSD test)

7. 분산분석

- 다중비교 (Multiple Comparison)

- HSD 통계량

$$HSD_{ij} = q_{\alpha}(\gamma, -\gamma) \sqrt{\frac{MSE}{n}}$$

- α : 유의 수준, γ : 수준의 수, n_T : 관측치의 수
- MSE : error mean square
- $|\bar{Y}_i - \bar{Y}_j|$ 와 HSD 통계량 비교해서, $|\bar{Y}_i - \bar{Y}_j|$ 이 값이 더 크면 귀무가설(i와 j의 수준의 평균이 같음) 기각

| | group 1 | group 2 | group 3 |
|---------|---------|---------|---------|
| group 1 | - | 12 *** | -5 *** |
| group 2 | | - | -0.59 |
| group 3 | | | - |

7. 분산분석

- 이원분산분석(two-way ANOVA): 관측값 1개

- 2개의 요인(2 factors) 내의 요인 수준(factor levels) 간의 조합(combination)을 각각 개별 개별 집단으로(groups, treatments)로 고려하여, 요인수준 간 조합의 평균 차이 비교
- 예를 들어, 요인(factor) A '크기'가 3개의 요인 수준(factor levels, 대,중,소)이 있고, 요인(factor) B '맛'이 2개의 요인 수준(factor levels, 맛있음, 맛없음)이 있으면, 총 그룹의 수는 $3 \times 2 = 6$ 개
- (1) 관측값이 하나일 경우와 (2) 관측값이 2개 이상일 경우 (반복 실험을 할 경우)로 구분

| | 수준 1 | ... | 수준 b | 평균 |
|------|---------------------|-----|---------------------|---------------------|
| 수준 1 | Y_{11} | ... | Y_{1b} | $\overline{Y}_{1.}$ |
| ... | ... | ... | ... | |
| 수준 a | Y_{a1} | ... | Y_{ab} | $\overline{Y}_{a.}$ |
| 평균 | $\overline{Y}_{.1}$ | | $\overline{Y}_{.b}$ | $\overline{Y}_{..}$ |

- 이원분산분석(two-way ANOVA): 관측값 1개, Interaction 고려 안 하는 경우
 - 이원분산분석모형의 편차 ($Y_{ij} - \bar{Y}_{..}$)는 다음과 같이 구분할 수 있음
 - $Y_{ij} - \bar{Y}_{..} = (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{.j} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$
 - $Y_{ij} - \bar{Y}_{..}$: 편차
 - $\bar{Y}_{i.} - \bar{Y}_{..}$: 요인 A 편차
 - $\bar{Y}_{.j} - \bar{Y}_{..}$: 요인 B 편차
 - $Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$: 오차
 - 위의 식의 값을 제곱합한 후 아래식과 표처럼 나타낼 수 있음
 - SST: 총제곱합 (자유도 $ab-1$)
 - SSA: 요인A 수준 평균 간 제곱합 (자유도 $a-1$)
 - SSB: 요인B 수준 평균 간 제곱합 (자유도 $b-1$)
 - SSE: 오차 제곱합 (자유도 $(a-1)(b-1)$)

7. 분산분석

- 이원분산분석표(two-way ANOVA Table)

| 요인 | squared sum | degrees of freedom | mean squared | F statistics |
|------|-------------|--------------------|--------------|--------------|
| 요인 A | SSA | a-1 | MSA | MSA/MSE |
| 요인 B | SSB | b-1 | MSB | MSB/MSE |
| 오차 | SSE | (a-1)(b-1) | MSE | |
| 계 | SST | ab-1 | | |

F-통계량:
그룹간 변동성이 그룹내 변동
성보다 커야하는 것을 의미!

- F 통계량 검정
 - 요인A 효과 검정 = MSA/MSE (F통계량)
 - H_0 : A의 수준에 따라 평균 차이가 없다
 - H_1 : A의 수준에 따라 평균 차이가 있다
 - 요인B 효과 검정 = MSB/MSE (F통계량)
 - H_0 : B의 수준에 따라 평균 차이가 없다
 - H_1 : B의 수준에 따라 평균 차이가 있다

7. 분산분석

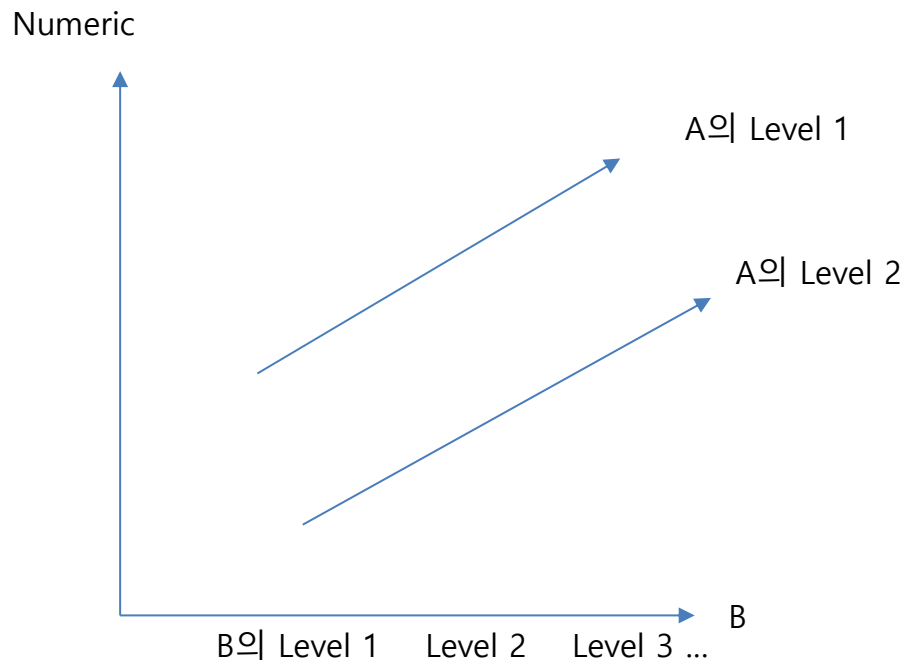
- 이원분산분석표(two-way ANOVA Table): Interaction 고려 하는 경우

| 요인 | squared sum | degrees of freedom | mean squared | F statistics |
|------|-------------|--------------------|--------------|--------------|
| 요인 A | SSA | a-1 | MSA | MSA/MSE |
| 요인 B | SSB | b-1 | MSB | MSB/MSE |
| A*B | SSAB | (a-1)*(b-1) | MSAB | MSAB/MSE |
| 오차 | SSE | (a-1)(b-1) | MSE | |
| 계 | SST | ab-1 | | |

- 수식: $SST=SSA+SSB+SSAB+SSE$
- F 통계량 검정
 - 요인A 효과 검정 = MSA/MSE (F통계량)
 - H_0 : A의 수준에 따라 평균 차이가 없다
 - H_1 : A의 수준에 따라 평균 차이가 있다
 - 요인B 효과 검정 = MSB/MSE (F통계량)
 - H_0 : B의 수준에 따라 평균 차이가 없다
 - H_1 : B의 수준에 따라 평균 차이가 있다
 - 요인A와 B의 조합에 의한 Interaction 효과 검정 = $MSAB/MSE$ (F통계량)
 - H_0 : AB의 수준에 따라 평균 차이가 없다
 - H_1 : AB의 수준에 따라 평균 차이가 있다

7. 분산분석

- **Interaction Effect?**

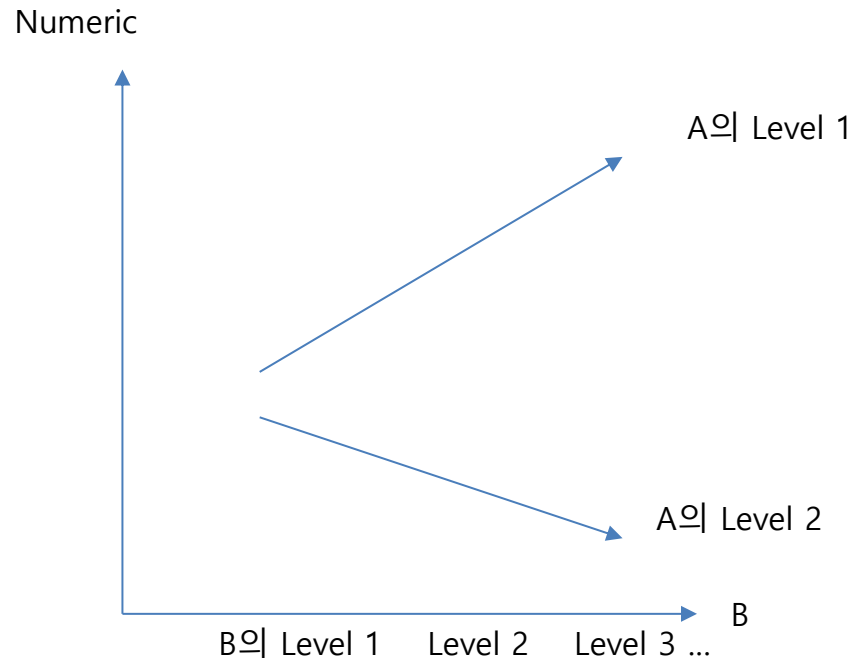


- A의 Effect (Main Effect)
- B의 Effect (Main Effect)
- AB의 Interaction Effect는 없음

- Two way Anova의 Interaction effect: A, B 두 변수 간의 조합이 만들어내는 결과
 - A, B 모두 Numeric 값에 영향을 주고 있음
 - 변수 간 조합에 의한 패턴으로 해석할 부분이 없고 그래서 평행이동하는 패턴

7. 분산분석

- Interaction Effect?



- A의 Effect (Main Effect)
- B의 Effect (Main Effect)는 없음
- AB의 Interaction Effect

- Two way Anova의 Interaction effect: A, B 두 변수 간의 조합이 만들어내는 결과
 - B가 단일 변수로는 효과가 없어 보이지만, A 변수와 함께 고려 시, 다른 패턴을 보여줌

7. 분산분석 – Taguchi Desin

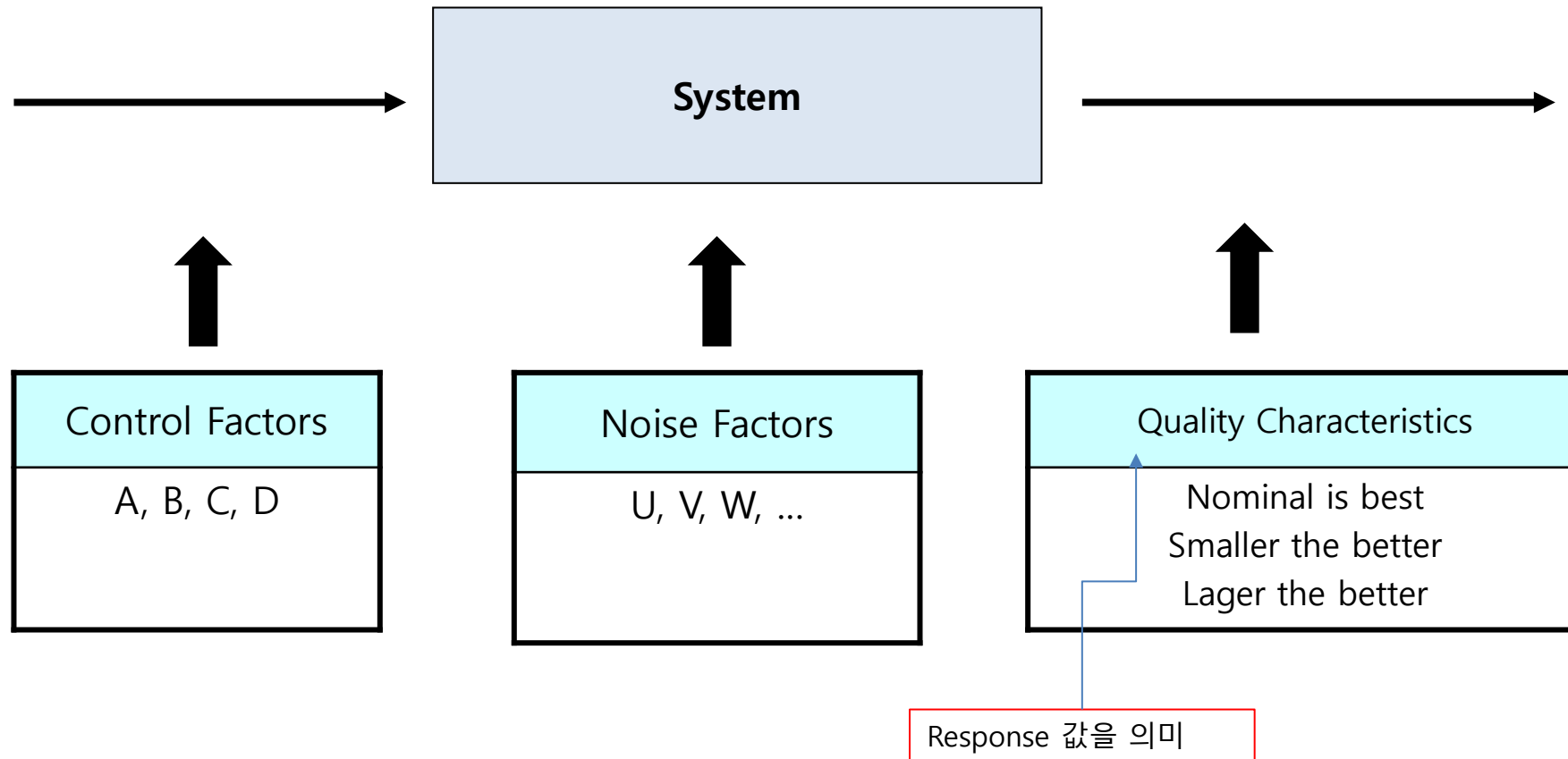
Dr. Genichi Taguchi

- Born in Japan in 1924.
- Japan's Telephone Communications System
- Industrial Quality Movement in US
- Father of the "Taguchi Method", "Robust Engineering".
- Cost Driven Quality Engineering
- Applications of Engineering Strategies VS Advanced Statistical Techniques



Source: Sung Hyun Park, Robust Design and Analysis for Quality Engineering, Chapman Hall

Factor Characteristics Relation Diagram



7. 분산분석 – Taguchi Design

- **Taguchi Design의 특징**
 - 직교배열표(orthogonal arrays)를 이용한 부분요인 실시법(Fractional Factorial Design)
 - 실험 시 잡음요인(noise factor)의 인식: Control 인자와 함께 Control 안되는 인자를 고려
 - 분산분석과 함께 손실개념을 도입한 성능통계량으로서 신호 대 잡음비(signal-to-noise ratio)의 사용
- **Orthogonal?**
 - 직교, 독립
 - Control 인자들의 독립

7. 분산분석 – Taguchi Desin

- 직교배열표

- 제품 품질/생산성을 향상시키거나 불량률 감소하려는 실험에서 일반적으로 고려해야 할 인자의 수는 많은 경우 사용

- ① 인자의 수가 많은 경우(보통 4개 이상)에 큰 "그물"을 쳐서 주효과(main effect)와
 - ② 기술적으로 보아서 있을 것 같은 2인자 교호작용(interaction)을 검출하고,
 - ③ 기술적으로 없으리라고 생각되는 2인자 교호작용 및 고차의 교호작용을 희생시켜서,
 - ④ 실험회수를 적게 할 수 있는 실험계획을 짤 수 있도록 만들어 놓은 표

7. 분산분석 – Taguchi Design

Loss Function

- **Nominal is best**

$$L(Y) = k (y - m)^2$$

$$SN_i = 10 \log \left[\frac{(\bar{y}_i)^2}{v_i} \right]$$

- **Smaller is better**

$$L(Y) = k y^2$$

$$\hat{E}(y^2) = \frac{\sum y_i^2}{n}$$

$$SN_i = -10 \log \left(\frac{1}{n} \sum_{j=1}^n y_{ij}^2 \right)$$

Response 증가 → SN 증가

- **Larger is better**

$$L(Y) = k \frac{1}{y^2}$$

$$\hat{E}\left(\frac{1}{y^2}\right) = \frac{\sum \frac{1}{y_i^2}}{n}$$

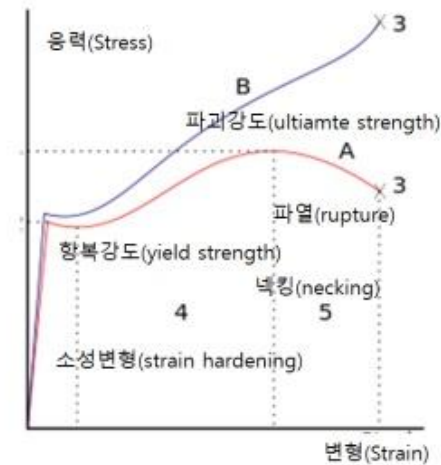
$$SN_i = -10 \log \left(\frac{1}{n} \sum_{j=1}^n \frac{1}{y_{ij}^2} \right)$$

Response 증가 → SN 감소

7. 분산분석 – Taguchi Design

- 예

- 자동차의 점화 케이블의 코어 인장력을 규격 40에 맞도록 하는 다구치 디자인



- 인장강도?
 - 재료의 세기를 나타내는 힘으로, 재료가 절단되도록 끌어당겼을 때 견뎌내는 최대 하중을 재료의 단면적으로 나눈 값

7. 분산분석 – Taguchi Desin

- 예
 - 주요 실험인자와 잡음인자를 다음과 같이 정의

| 인자구분 | 인자명 | 수준1 | 수준2 | 수준3 |
|------|------------|------|------|-----|
| 제어인자 | A: 압출장치 | A1 형 | A2 형 | |
| | B: 생산라인 속도 | 저속 | 중속 | 고속 |
| | C: 가열온도 | 저온 | 보통 | 고온 |
| | D: 절연재료 | D1 | D2 | D3 |
| | F: CV 압력 | 저압 | 중압 | 고압 |
| | G: CV 속도 | 저속 | 중속 | 고속 |
| | H: 편조기 장력 | 낮다 | 보통 | 높다 |
| | I: 릴리스 도포 | I1 | I2 | I3 |
| 잡음인자 | S: 작업 샘플 | S1 | S2 | |
| | P: 샘플내의 위치 | P1 | P2 | |

7. 분산분석 – Taguchi Design

- 예

- 직교 배열

- A의 수준 : 1, 2
 - B~I의 수준 : 1, 2, 3
 - $L_{18}(2^1 \times 3^7) \Rightarrow L_{18}(2^1 \times 3^2)$

- nominal is best

$$SN_i = 10 \log \left[\frac{\frac{1}{n} (S_{m(i)})}{V_i} \right] = 10 \log \left[\frac{(\bar{y}_i)^2}{V_i} \right]$$

Where

$$V_i = \sum_{j=1}^n \frac{(y_{ij} - \bar{y}_i)^2}{n-1} = \text{sample variance for the } i\text{th row}$$

$$\bar{y}_i = \sum_j \frac{y_{ij}}{n} = \text{sample mean for the } i\text{th row}$$

$$S_{m(i)} = \frac{1}{n} \left(\sum_{j=1}^n y_{ij} \right)^2 = n (\bar{y}_i)^2 = \text{correction term for the } i\text{th row}$$

- SN ratio는 변동의 크기에 영향을 받음(y의 평균이 아닌) / 영향을 주는 인자를 찾기 위해서 sensitivity 계산: 민감도? 1번째 행의 n개의 값에 대해 다음과 같이 정의

$$S_{n(i)} = 10 \log [S_{m(i)}] = 10 \log [n(\bar{y}_i)^2] = 20 \log [\sqrt{n} \bar{y}_i]$$

7. 분산분석 – Taguchi Design

- 직교배열표

| 관측치 | A | B | C | D | F | G | H | I | S(=0) | | S(=1) | |
|-----|---|---|---|---|---|---|---|---|-------|-------|-------|-------|
| | | | | | | | | | P(=0) | P(=1) | P(=0) | P(=1) |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 30 | 40 | 38 | 49 |
| 2 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 10 | 15 | 25 | 25 |
| 3 | 0 | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 49 | 53 | 53 | 55 |
| 4 | 0 | 1 | 0 | 0 | 1 | 1 | 2 | 2 | 62 | 58 | 52 | 68 |
| 5 | 0 | 1 | 1 | 1 | 2 | 2 | 0 | 0 | 30 | 50 | 49 | 62 |
| 6 | 0 | 1 | 2 | 2 | 0 | 0 | 1 | 1 | 10 | 25 | 29 | 36 |
| 7 | 0 | 2 | 0 | 1 | 0 | 2 | 1 | 2 | 58 | 42 | 41 | 50 |
| 8 | 0 | 2 | 1 | 2 | 1 | 0 | 2 | 0 | 28 | 29 | 32 | 31 |
| 9 | 0 | 2 | 2 | 0 | 2 | 1 | 0 | 1 | 110 | 74 | 94 | 115 |
| 10 | 1 | 0 | 0 | 2 | 2 | 1 | 1 | 0 | 76 | 86 | 66 | 103 |
| 11 | 1 | 0 | 1 | 0 | 0 | 2 | 2 | 1 | 52 | 37 | 54 | 59 |
| 12 | 1 | 0 | 2 | 1 | 1 | 0 | 0 | 2 | 55 | 79 | 62 | 98 |
| 13 | 1 | 1 | 0 | 1 | 2 | 0 | 2 | 1 | 5 | 35 | 16 | 42 |
| 14 | 1 | 1 | 1 | 2 | 0 | 1 | 0 | 2 | 52 | 96 | 79 | 91 |
| 15 | 1 | 1 | 2 | 0 | 1 | 2 | 1 | 0 | 50 | 70 | 56 | 65 |
| 16 | 1 | 2 | 0 | 2 | 1 | 2 | 0 | 1 | 15 | 20 | 18 | 21 |
| 17 | 1 | 2 | 1 | 0 | 2 | 0 | 1 | 2 | 51 | 62 | 59 | 70 |
| 18 | 1 | 2 | 2 | 1 | 0 | 1 | 2 | 0 | 77 | 83 | 66 | 74 |

7. 분산분석 – Taguchi Design

- 실습

40을 목표로

하는 망목 특성 /

각각의 SN비와

민감도를 구하기

| Experiment number | Total(T_i) | Mean(\bar{y}_i) | Variance(V_i) | SN = $10 \log \left[\frac{(\bar{y}_i)^2}{V_i} \right]$ | Sensitivity $S_n = 10 \log \left[\frac{T_i^2}{4} \right]$ |
|-------------------|----------------|---------------------|-------------------|--|---|
| 1 | 157 | 39.25 | 60.92 | 14.0 | 37.90 |
| 2 | 75 | 18.75 | 56.25 | 7.8 | 31.49 |
| 3 | 210 | 52.50 | 6.33 | 26.4 | 40.41 |
| 4 | 240 | 60.00 | 45.33 | 19.0 | 41.58 |
| 5 | 191 | 47.75 | 174.92 | 11.0 | 39.60 |
| 6 | 100 | 25.00 | 120.67 | 6.9 | 33.98 |
| 7 | 191 | 47.75 | 62.92 | 15.6 | 39.60 |
| 8 | 120 | 30.00 | 3.33 | 24.3 | 35.56 |
| 9 | 393 | 98.25 | 341.58 | 14.5 | 45.87 |
| 10 | 333 | 83.25 | 254.25 | 14.3 | 44.42 |
| 11 | 202 | 50.50 | 89.67 | 14.5 | 40.09 |
| 12 | 294 | 73.50 | 368.33 | 11.6 | 43.34 |
| 13 | 98 | 24.50 | 289.67 | 2.6 | 33.80 |
| 14 | 318 | 79.50 | 387.0 | 12.1 | 44.03 |
| 15 | 241 | 60.25 | 80.25 | 16.5 | 41.61 |
| 16 | 74 | 18.50 | 7.00 | 16.9 | 31.37 |
| 17 | 242 | 60.50 | 61.67 | 17.7 | 41.64 |
| 18 | 300 | 75.00 | 50.00 | 20.5 | 43.52 |
| Total | 3779 | 944.75 | | 266.2 | 709.81 |

7. 분산분석 – Taguchi Desin

해석

- Inner array의 각 행에 대한 SN비와 민감도 계산
- 분산분석으로 SN에 유의한 Control 인자 / 민감도에 유의한 영향 주는 Control 인자 발견
- Control인자를 다음 3가지 카테고리로 분류
 - 1) Dispersion control factor : significant factors for SN ratios
 - 2) Mean adjustment factor : significant factors for sensitivities
 - 3) Insignificant factor : other control factors

*인자가 1)이면서 2)인 경우, dispersion control factor.
- Optimum condition 발견
 - (1) dispersion control factor: SN 비 크게 해주는 수준
 - (2) mean adjustment factor: 추정된 응답이 목표 응답과 가장 가까운 수준
 - (3) insignificant factor : 경제성이나 간결함, 운용가능성 등으로 고려한 인자

7. 분산분석 – Taguchi Design

| Experiment number | Total(T_i) | Mean(\bar{y}_i) | Variance(V_i) | SN = $10 \log \left[\frac{(\bar{y}_i)^2}{V_i} \right]$ | Sensitivity $S_n = 10 \log \left[\frac{T_i^2}{4} \right]$ |
|-------------------|----------------|---------------------|-------------------|---|--|
| 1 | 157 | 39.25 | 60.92 | 14.0 | 37.90 |
| 2 | 75 | 18.75 | 56.25 | 7.8 | 31.49 |
| 3 | 210 | 52.50 | 6.33 | 26.4 | 40.41 |
| 4 | 240 | 60.00 | 45.33 | 19.0 | 41.58 |
| 5 | 191 | 47.75 | 174.92 | 11.0 | 39.60 |
| 6 | 100 | 25.00 | 120.67 | 6.9 | 33.98 |
| 7 | 191 | 47.75 | 62.92 | 15.6 | 39.60 |
| 8 | 120 | 30.00 | 3.33 | 24.3 | 35.56 |
| 9 | 393 | 98.25 | 341.58 | 14.5 | 45.87 |
| 10 | 333 | 83.25 | 254.25 | 14.3 | 44.42 |
| 11 | 202 | 50.50 | 89.67 | 14.5 | 40.09 |
| 12 | 294 | 73.50 | 368.33 | 11.6 | 43.34 |
| 13 | 98 | 24.50 | 289.67 | 2.6 | 33.80 |
| 14 | 318 | 79.50 | 387.0 | 12.1 | 44.03 |
| 15 | 241 | 60.25 | 80.25 | 16.5 | 41.61 |
| 16 | 74 | 18.50 | 7.00 | 16.9 | 31.37 |
| 17 | 242 | 60.50 | 61.67 | 17.7 | 41.64 |
| 18 | 300 | 75.00 | 50.00 | 20.5 | 43.52 |
| Total | 3779 | 944.75 | | 266.2 | 709.81 |

- 각 요인의 수준별 합계간 차이 제곱 평균
- S_A : (A 수준0의 합-수준1의 합) 제곱 평균
- S_B : B 각 수준 제곱 합 평균-전체 제곱합
- ...



$$\begin{aligned}
 S_T &= \frac{266.2^2}{18} = 3,936.8 \\
 S_T &= 14.0^2 + 7.8^2 + \dots + 20.5^2 - 3,936.8 = 582.6 \\
 S_A &= \frac{1}{18} (126.7 - 139.5)^2 = 9.1 \\
 S_B &= \frac{1}{6} (88.6^2 + 68.1^2 + 109.5^2) - 3,936.8 = 142.8 \\
 S_C &= \frac{1}{6} (82.4^2 + 87.4^2 + 96.4^2) - 3,936.8 = 16.8 \\
 S_D &= \frac{1}{6} (96.2^2 + 69.1^2 + 100.9^2) - 3,936.8 = 98.2 \\
 S_E &= \frac{1}{6} (83.6^2 + 96.1^2 + 86.5^2) - 3,936.8 = 14.3 \\
 S_F &= \frac{1}{6} (77.1^2 + 88.2^2 + 100.9^2) - 3,936.8 = 47.3 \\
 S_H &= \frac{1}{6} (80.1^2 + 78.8^2 + 107.3^2) - 3,936.8 = 86.3 \\
 S_I &= \frac{1}{6} (100.6^2 + 63.2^2 + 102.4^2) - 3,936.8 = 163.3 \\
 S_e &= S_T - (S_A + S_B + S_C + S_D + S_E + S_F + S_H + S_I) = 4.5
 \end{aligned}$$

7. 분산분석 – Taguchi Desin

- SN 비에 대한 ANOVA
 - Dispersion Control Factor(산포 제어 인자) 선택
 - SN비를 크게 해주는 Level을 선택
- 민감도에 대한 ANOVA
 - Mean Control Factor(평균조정인자) 선택
 - 산포 제어 인자로 이미 선택된 인자를 제외한 나머지가 평균조정인자
 - 목표 수치와 가까운 값이 되는 Level을 선택
- Other Control Factor(기타제어인자)
 - 경제적 요인 등을 고려한 Level 선택

7. 분산분석 – Taguchi Design

Analysis of SN ratios

| factor | | A | B | C | D | F | G | H | I | |
|----------------|---|-------|-------|-------|-------|-------|-------|-------|-------|---------------|
| level (sum) | 0 | 139.5 | 88.6 | 82.4 | 96.2 | 83.6 | 77.1 | 80.1 | 100.6 | SUM= 266.2 |
| | 1 | 126.7 | 68.1 | 87.4 | 69.1 | 96.1 | 88.2 | 78.8 | 63.2 | |
| | 2 | | 109.5 | 96.4 | 100.9 | 86.5 | 100.9 | 107.3 | 102.4 | |
| level (avg) | 0 | 15.5 | 14.77 | 13.73 | 16.03 | 13.93 | 12.85 | 13.35 | 16.77 | AVG= 14.79 |
| | 1 | 14.08 | 11.35 | 14.57 | 11.52 | 16.02 | 14.70 | 13.13 | 10.53 | |
| | 2 | | 18.25 | 16.07 | 16.82 | 14.42 | 16.82 | 17.88 | 17.07 | |

| Factor | S | ϕ | V | F ₀ |
|--------|--------|--------|-------|------------------|
| A | 9.1 | 1 | – | |
| B | 142.8 | 2 | 71.4 | 11.2** |
| C | 16.8 | 2 | – | |
| D | 98.2 | 2 | 49.1 | 7.7* |
| F | 14.3 | 2 | – | |
| G | 47.3 | 2 | 23.7 | 3.7 ^Δ |
| H | 86.3 | 2 | 43.2 | 6.8* |
| I | 163.3 | 2 | 81.7 | 12.8** |
| e | 4.5 | 2 | – | |
| (e) | (44.7) | (7) | (6.4) | |
| T | 582.6 | 17 | | |

인자들의 최적 수준 조합은
각 인자에 대해 SN비를 크
게 해주는 수준들의 조합

Optimal condition for the dispersion control factors: B₂D₂G₂H₂I₂

7. 분산분석 – Taguchi Design

Analysis of sensitivity

| factor | | A | B | C | D | F | G | H | I | |
|----------------|---|--------|--------|--------|--------|--------|--------|--------|--------|----------------|
| level (sum) | 0 | 345.99 | 237.65 | 228.67 | 248.69 | 239.12 | 226.22 | 242.11 | 242.61 | SUM= 709.81 |
| | 1 | 363.82 | 234.60 | 232.41 | 231.35 | 224.95 | 250.91 | 232.74 | 216.60 | |
| | 2 | | 237.56 | 248.73 | 229.77 | 245.74 | 232.68 | 234.96 | 250.60 | |
| level (avg) | 0 | 38.44 | 39.61 | 38.11 | 41.45 | 39.85 | 37.70 | 40.35 | 40.44 | AVG= 39.43 |
| | 1 | 40.42 | 39.10 | 38.74 | 38.56 | 37.49 | 41.82 | 38.79 | 36.10 | |
| | 2 | | 39.59 | 41.46 | 38.30 | 40.96 | 38.78 | 39.16 | 41.77 | |

| Factor | S | Φ | V | F ₀ |
|--------|---------|--------|-------|----------------|
| A | 17.66 | 1 | 17.66 | 2.03 |
| B | 1.00 | 2 | – | |
| C | 37.93 | 2 | 18.97 | 2.19 |
| D | 36.73 | 2 | 18.37 | 2.12 |
| F | 37.60 | 2 | 18.80 | 2.17 |
| G | 54.65 | 2 | 27.33 | 8.15 |
| H | 7.99 | 2 | – | |
| I | 105.35 | 2 | 52.68 | 6.07 |
| e | 43.08 | 2 | – | |
| (e) | (52.07) | (6) | 8.68 | |
| T | 341.99 | 17 | | |

Significant factors are A,C,D,F,G and I

7. 분산분석 – Taguchi Desin

- **Control 인자 분류:**
 - Dispersion control factors : B,D,G,H,I
 - Mean adjustment factors : A,C,F
 - Other control factors : 없음

| Level | A | B | C | D | F | G | H | I |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0 | 46.58 | 52.96 | 45.54 | 61.64 | 52.83 | 42.13 | 59.46 | 55.92 |
| 1 | 58.39 | 49.50 | 47.83 | 47.88 | 43.50 | 69.13 | 49.25 | 39.25 |
| 2 | 55.50 | 64.08 | 48.13 | 61.13 | 46.21 | 48.75 | 62.29 | |
| Average | 52.49 | 52.49 | 52.49 | 52.49 | 52.49 | 52.49 | 52.49 | 52.49 |

- ACF의 최적 수준으로 추정된 평균이 목표값인 40에 가까운 A_i, C_j, F_k 를 찾기: $A_0C_0F_1$

$$\begin{aligned}\hat{u}(A_0C_0F_1) &= \overline{A_0} + \overline{C_0} + \overline{F_1} - 2\overline{T} \\ &= 46.58 + 45.54 + 43.50 - 2 \times 52.49\end{aligned}$$

7. 분산분석 – Taguchi Design

Dispersion control과 mean adjustment factors을 고려한 전체 Optimum condition:

$$A_0B_2C_0D_2F_1G_2H_2I_2$$

- SN비

$$\begin{aligned} &= \overline{A_0} + \overline{B_2} + \overline{C_0} + \overline{D_2} + \overline{F_1} + \overline{G_2} + \overline{H_2} + \overline{I_2} - 7\overline{T} \\ &= \frac{139.5}{9} + \frac{109.5}{6} + \frac{82.4}{6} + \frac{100.9}{6} + \frac{83.6}{6} + \frac{100.9}{6} + \frac{109.3}{6} + \frac{102.4}{6} - 7 * \frac{266.2}{18} \\ &= 26.97dB \end{aligned}$$

- 추정된 core pulling force: 예) A인자 0수준의 측정값 평균

$$\begin{aligned} \hat{\mu} &= \overline{A_0} + \overline{B_2} + \overline{C_0} + \overline{D_2} + \overline{F_1} + \overline{G_2} + \overline{H_2} + \overline{I_2} - 7\overline{T} \\ &= 46.58 + 55.50 + 45.54 + 48.13 + 43.5 + 46.21 + 48.75 + 62.29 - 7 * 52.4 \\ &= 38.40kgcm^{-2} \end{aligned}$$

- ✓ 추정된 core pulling force가 40에 근접
- ✓ 추정된 SN비는 높은 수준으로 Noise 인자에 Robust하며 작은 변동을 갖음



Industrial Data Science Lab

Contact:

won.sang.l@gwnu.ac.kr

<https://sites.google.com/view/idslab>