제1강 통계학 개요

1. 통계학이란?
2. 통계자료 및 통계분석
3. 자료의 표현

1. 통계학이란?

(1) 통계학(statistics) 정의

- 불확실한 현상을 대상으로 자료를 수집하고 정리하며, 자료가 수집된 대상에 적절한 모형을 설정하고 추정, 검정 및 예측을 하는 학문
- 통계를 구하는 과정과 구하여진 통계를 모형화하여 어떠한 추론(추정,가설검정,예측)을 연구하는 학문
- 불확실성 하에서의 올바른 의사결정 방법을 연구하는 학문으로 자료의 수집, 분류, 분석과 해석의 체계를 갖고 있음
- 불확실성을 최소로 감소시키는 방법을 연구하는 학문

(참고)통계학의 응용: 인구통계, 사회통계, 경영통계, 경제통계, 산업통계, 교육통계. 기상통계, 물리통계

(2) 통계학 분류

① 기술통계학(descriptive statistics) : 모집단에 대한 추론이나 어떤 결론을 도출함이 없이 수집된 정보를 간단 명료하고 유용하게 정리하는 문제를 다룸.

- 통계조사: 전수조사, 표본조사(통계조사의 3요건:신속성,정확성,저렴성)
- 자료정리:숫자, 그림, 표
- 자료특성: 평균, 분산, 왜도, 첨도

(자료 표현방법)

- 숫자:평균,분산,표준편차,중위수,최빈값....
- 그림:막대그래프,원그래프,...
- 표:도수분포표...

② 추리통계학(inferential statistics)

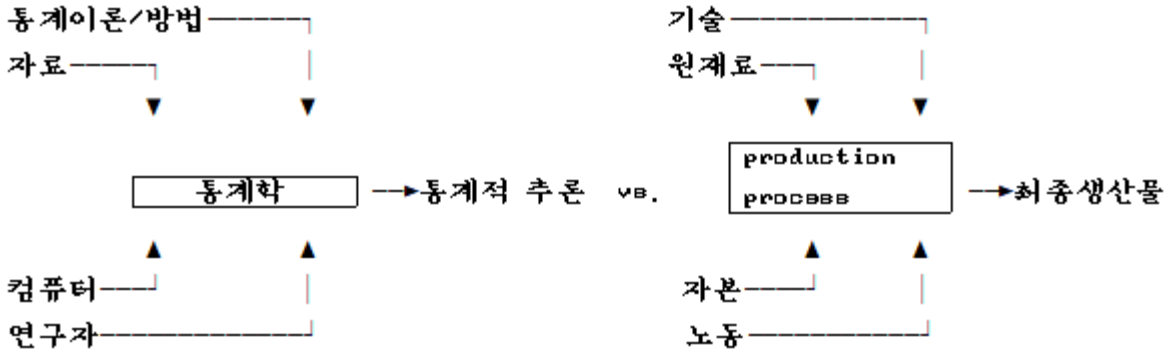
- 표본을 기초로 하여 모집단의 특성을 추정하고 일반화하며 또한 예측하는 문제를 다룸.
- 어떤 현상에 대해 얻은 정보를 이용하여 그 현상에 대한 보다 일반적인 의사결정을 내리는 문제를 다룸.(예: 여론조사)

(목적)추리통계학은 표본에 의해 얻은 정보를 기초로 하여 모집단의 특성(평균,분산 등)에 대해 일반화하는 것을 목적으로 함

(분류)추정(estimation), 가설검정(hypothesis testing), 예측(prediction)

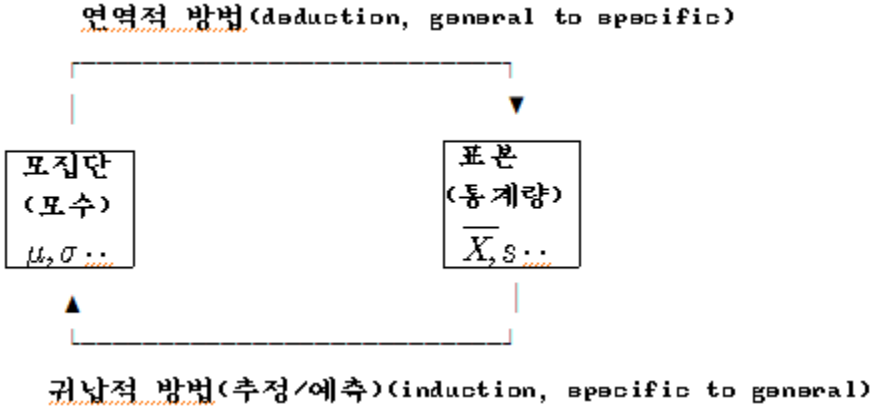
- 추정: 점추정, 구간추정
- 가설검정: t, F, chi-square 검정 등
- 예측: 점예측, 구간예측

(3) 통계적 추론(statistical inference)의 과정



(5) 통계학 기본용어

- 모집단(population):연구자의 관심대상이 되는 모든 개체의 집합
- 표본(sample):모집단에서 조사대상으로 채택된 일부 집단(부분집합)
- 추론(inference):표본정보를 바탕으로 모집단에 관한 의사결정,추정,예측을 하는 것.
- 신뢰도(reliability):추론에 대한 신뢰성의 측도.
- 변수(variable):관측 때마다 서로 다른 값을 취하는 것으로 변수를 표현하는 방법으로 일반적으로 X,Y,Z와 같은 영문자로 표현.
- 모수(parameter):모집단의 수량적인 특성(평균,분산,표준편차..)
- (표본)통계량(statistic):표본의 수량적인 특성. 표본에 담긴 정보를 요약하는 공식
- 추정량(estimator):모수를 추정하는 공식을 나타내는 통계량
- 추정치(estimate):추정량(공식)에 실제의 관찰값을 넣어 계산한 통계량의 값
- 표본오차(sampling error):모집단을 조사하지 않고 표본조사의 결과만을 가지고 모집단의 특성을 추정할 때 발생하는 오차
- 비표본오차(non-sampling error):관찰오류, 누락, 오기 등 표본추출과정에서 오류로 인하여 발생하는 오차



2. 통계자료 및 통계분석

(1) 통계자료의 종류

① 횡단면자료와 시계열자료

- 횡단면자료(cross section data):한 시점에서 측정된 자료
- 시계열자료(time series data):일정한 기간을 두고 측정된 자료
- 종적자료(longitudinal data):횡단면자료와 시계열자료가 결합된 자료로 합동횡단면자료(pooled cross section data)라고도 함
- 패널자료(panel data):종적자료의 특수한 경우

② 질적자료와 양적자료

- 질적자료(qualitative data):개개의 단위가 어떤 성질을 갖는 지를 개별적으로 식별하여 계수(counting)에 의해(또는 범주에 의해) 관찰할 수 있는 자료로 수치로 나타낼 수 없으며 범주자료(categorical data)라고도 함(예:인구의 남여별,불량품과 우량품의 구별,학생들의 성적 평가)
- 양적자료(quantitative data):각 단위가 갖는 특정한 양적 성질을 측정하여 계량(measuring)에 의해 관찰할 수 있는 자료로 수치로 나타낼 수 있음(예:무게,길이,연령..)

③ 양적자료의 종류

- 연속자료(continuous data):일정한 범위내의 모든 값을 택하면서 연속적으로 변하는 양으로 측정 가능한(measurable) 자료(예:연령,키,무게..)(analog)
- 이산자료(discrete data):항상 정수의 값을 택하면서 변하는 양으로 세는 것이 가능한(countable) 자료 (예:총 인원수,방수,가족수..)(digital)

(2) 통계분석

① 통계분석

- 통계분석이란 수집된 자료의 특성을 일목요연하게 파악하기 위하여 통계량을 산출하는 과정을 말함.
- 통계분석에는 그래프(산포도, 히스토그램)나 통계량(위치, 분포의 정도)이 이용됨
- 통계분석기법에는 그래프분석, 빈도분석, T-검증, 분산분석, 상관분석, 회귀분석, 시계열분석 등이 있음
- 통계패키지에는 SAS, SPSS, EViews, STATA, Rats, Gauss, R, Python 등이 있음

② 통계분석의 단계

- 문제의 파악
- 표본설계(모집단조사 또는 표본조사)
- 자료수집(published data or survey data)

(참고)유용한 Website : 한국은행 경제통계시스템, 통계청 국가통계포털

- 자료의 분석(모수의 추정 및 가설검정)
- 의사결정

3. 자료의 표현

(1) 표/그래프

① 도수분포표(frequency distribution table)

-양적 정보를 그 양적 성질의 크기에 따라 분류한 결과를 표로 나타낸 분류통계표

-수집된 자료를 상호배타적인 계급들로 집단화하여 요약한 표

(용어)

변수(X):분류의 표준이 되는 양적 성질(변수의 값(X1,X2,X3...)

계급(class):분류한 각 구간

계급구간(class width):각 계급의 폭

계급한계(class limit):계급의 양 끝 점

계급값(class mark):각 계급의 중앙값(Yi:i번째 계급의 계급값)

도수(frequency):각 계급에 속하는 관찰 단위의 개수

(예)경제학과 1994년 신입생의 수능시험 성적분포표

수능시험 성적	도수(인원수)	누적도수	Y_i
140-144	3	3	142
135-139	7	10	137
130-134	10	20	132
125-129	8	28	127
120-124	1	30	122

(참고)도수분포표의 작성 방법

- 자료들을 집단화할 수 있도록 계급을 설정한다.
- 계급의 개수는 관찰총수의 제곱근 or 5-20개 정도가 적당하다.
- 계급구간은 통일, 계급의 구간=(자료의 최대값-자료의 최소값)/계급의 수
- 관찰치가 중복됨이 없이 한 계급에만 속하도록 계급한계를 정할 것

② Histogram:각 계급값에 대하여 계급의 도수를 수직적인 막대기로 나타낸 도수분포의 그래프

(2) 숫자(통계량)

- 변수의 분포의 중심이 어디에 있는지→대표값(mean)
- 변수가 어떤 범위에 어느 정도로 분산되어 있는지-->분산도(dispersion)
- 변수의 분포가 대칭에서 어느 정도 벗어났는지-->비대칭도(skewness) 또는 편도

① 대표값

- 분포의 중심위치(central location)를 나타내는 측정치(넓은 의미의 분포의 중심으로서 도수 전체를 대표하는 값)으로 산술평균이 대표적인 값

(모평균) $$\mu = \frac{1}{N}(X_1 + X_2 + \ldots + X_N) = \frac{\sum_{i=1}^{N} X_i}{N}$$

(표본평균) $$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{\sum_{i=1}^{n} X_i}{n}$$

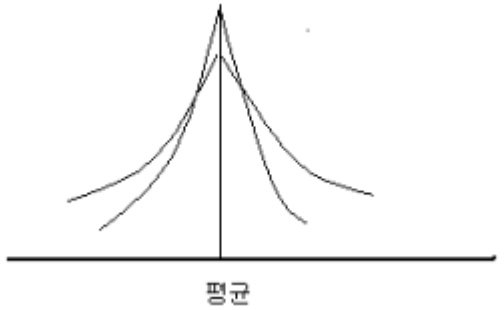
- 기하평균의 log값(상용로그 또는 자연로그)은 비율로 나타낸 변수의 log값들의 산술평균이다. 따라서 이 값의 anti-log가 기하평균 G임

$$G = \sqrt[n]{X_1 X_2 .. X_n} = (X_1 X_2 .. X_n)^{\frac{1}{n}}$$

또는 $$\log G = \frac{1}{n}\sum_{i=1}^{n} \log X_i$$

② 분산도

- 자료가 어떤 범위에 어느 정도로 분포되어 있으며 또 대표값 주위에 얼마나 가까이 분포되어 있는지를 나타내는 측도를 분산도(measure of dispersion or variation)라 함
- 대표값이 같더라도 분포의 정도는 다를 수 있음
- 분산도를 측정하는 방법에는
  변수의 크기에 의한 방법과 편차에 의한 방법이 있음



● 변수의 크기에 의한 방법 :

- 범위(range)=변수의 최대값-최소값
- 사분위수(Quartile):관측치를 작은 값으로부터 올림차순으로 배열하였을 경우 아래사분위수(lower quartile:Q1)는 관측치의 25% 순서에 있는 값이며 위사분위수(upper quartile:Q3)는 관측치의 75% 순서에 있는 값을 말함

-중위수의 위치=(n+1)/2

-Q1의 위치=([중위수 위치]+1)/2

-Q3의 위치=[중위수 위치]+([중위수 위치]+1)/2 (단, [ ]는 최대정수

(참고)상자그림(Box Plot)

- 상자그림이란 최소값, 아래사분위수, 중위수, 위사분위수, 최대값 등 5개의 순서통계량을 이용하여 자료를 그린 그림을 말함.

- 평균값에 대한 변수의 편차에 의한 방법:

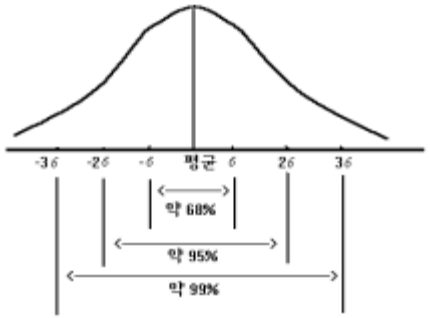
- 평균편차(mean absolute deviation):변수 X의 값 X1,X2,...Xn의 평균값을 $\bar{X}$라 할 때 각 Xi의 평균값으로부터의 차이를 편차(deviation)라 하며, 편차의 합은 항상 0이기 때문에 절대치의 합의 평균을 평균절대편차라 함

$$M.A.D = \frac{1}{n}\sum_{i=1}^{n} | X_i - \bar{X} |$$

- 분산(variance):변수 X의 값 X1,X2,...Xn의 평균값을 $\bar{X}$, 모평균을 μ라 할 때 평균값에 대한 편차의 제곱값을 더하여 얻은 값을 N으로 나누어 준 값을 모분산(σ^2)이라 하고 n-1로 나누어 준 값을 표본분산이라 하고 s^2으로 표시함.

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

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



- 표준편차:분산에 제곱근을 택하여 그 중에서 양(+)의 값을 표준편차(standard deviation)라 하고 s로 표시하는데 최초 변수의 단위와 일치시키기 표준편차를 구함.

- (변동)계수:분산의 상대적 측정치로 가장 많이 이용되는 것으로 표준편차의 평균에 대한 백분율을 분산계수(coefficient of variation)라 하고 이를 C.V로 나타낸다. 일반적으로 평균이 증가함에 따라 표준편차가 증가하므로 평균을 무시하고 표준편차만을 가지고 분산도를 비교하는 것보다 표준편차의 평균에 대한 백분율을 구해 비교하는 것이 바람직함

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2} \qquad C.V. = \frac{s}{\bar{X}} \times 100(\%)$$

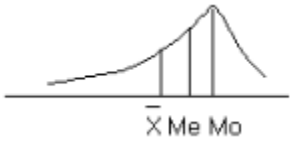
③ 편도(skewness)

- 분포의 비대칭의 정도를 편도 또는 비대칭도라고 하는데 어떤 분포가 한쪽 방향으로 꼬리 부분이 늘어지면 평균값(산술평균, 중위수)은 그쪽 꼬리 방향으로 이동함
- 비대칭의 정도를 측정하는 방법으로는 Pearson이 개발한 비대칭계수(coefficient of skewness; C_s)가 있음

$$C_s = \frac{3(\mu - Me)}{\sigma} \quad \text{또는} \quad C_s = \frac{3(\bar{X} - Me)}{s}$$

(해석)

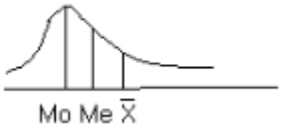
- $C_s < -1$: 왼쪽으로 아주 긴 꼬리를 갖는 분포(skewed to the left
- $-1 \le C_s < 0$: 왼쪽으로 약간 긴 꼬리를 갖는 분포(skewed to the left)



->산술평균은 outlier에 영향을 받음

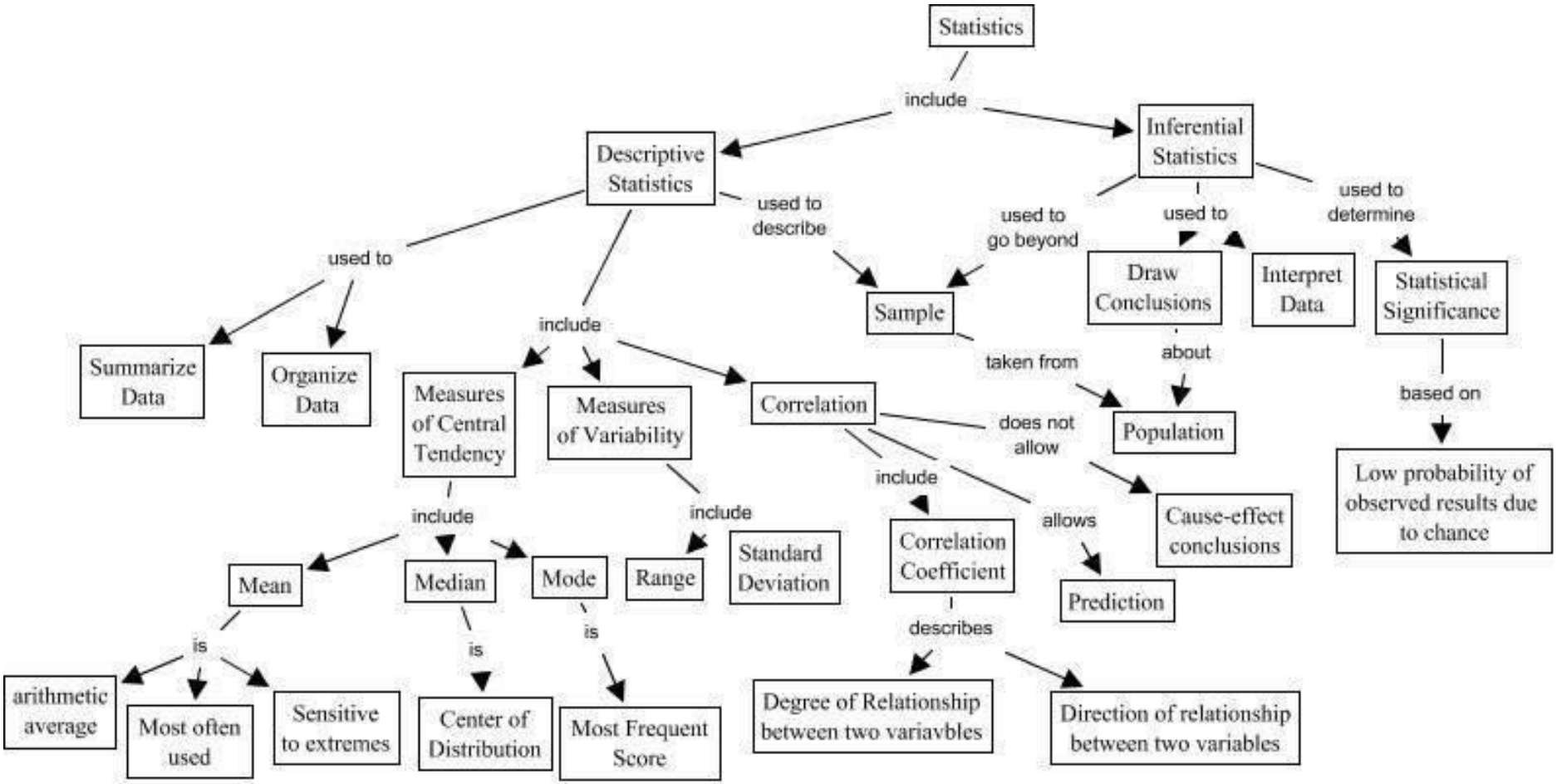
- $C_s = 0$: 좌우대칭분포

- $0 < C_s \le 1$: 오른쪽으로 약간 긴 꼬리를 갖는 분포(skewed to the right)
- $C_s > 1$: 오른쪽으로 아주 긴 꼬리를 갖는 분포(skewed to the right)



->산술평균은 outlier에 영향을 받음

(3) 통계분석의 분류



- 실증분석 시 유의사항
  - 시계열자료의 경우 안정성 여부에 대한 통계적 검정을 반드시 할 것
  - ‘t-통계량’ 또는 ‘유의수준’에 근거한 변수의 중요도에 대한 해석
  - R-square에 대한 해석(크기 및 크기 비교)

(참고) 계량경제의 분류

구분	회귀분석		시계열분석		패널분석	
방정식	단일방정식	연립방정식	일변량 (univariate)	다변량 (multivariate)	회귀분석	시계열분석
데이터	횡단면/시계열	횡단면/시계열	시계열	시계열	pool/패널	pool/패널
모형	①단순모형	-	④AR(p) ⑤MA(q) ⑥ARMA(p,q) ⑦ARIMA(p,d,q)	⑧VAR(p) /VECM	⑨패널모형	⑩패널VAR(p) / 패널 VECM
	②다중모형	③연립방정식모형				
	정태/동태모형	정태/ 동태모형	동태모형	동태모형	정태/ 동태모형	동태모형
추정방법	LS(OLS, WLS, ILS), MLE	LS(OLS, 2SLS, 3SLS)	OLS, NLS, MLE	OLS	OLS, LSDV, DID	OLS, GMM
software	Excel, Stata, R, Python					

①단순모형 : $y_i = \alpha + \beta x_i + \epsilon_i \quad (i = 1,2,...,N)$ $\quad y_t = \alpha + \beta x_t + \epsilon_t \quad (t = 1,2,...,T)$

②다중모형 : $y_i = \alpha + \beta X_i + \epsilon_i$
$X_{ki} = (x_{1i}, x_{2i}, ..., x_{k-1i})$
$(i = 1,2,...,N)$

③연립방정식모형(예시) : $y_{1i} = \beta_{10} + \beta_{11} y_{2i} + \gamma_{11} x_{1i} + u_{1i}$
$y_{2i} = \beta_{20} + \beta_{21} y_{1i} + \gamma_{21} x_{1i} + u_{2i}$

④AR(p) : $y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t$
$(p = 1,2,...,p)$

⑤MA(q) : $y_t = e_t - \theta_1 e_{t-1} - ... - \theta_q y_{t-q}$
$(q = 1,2,...,q)$

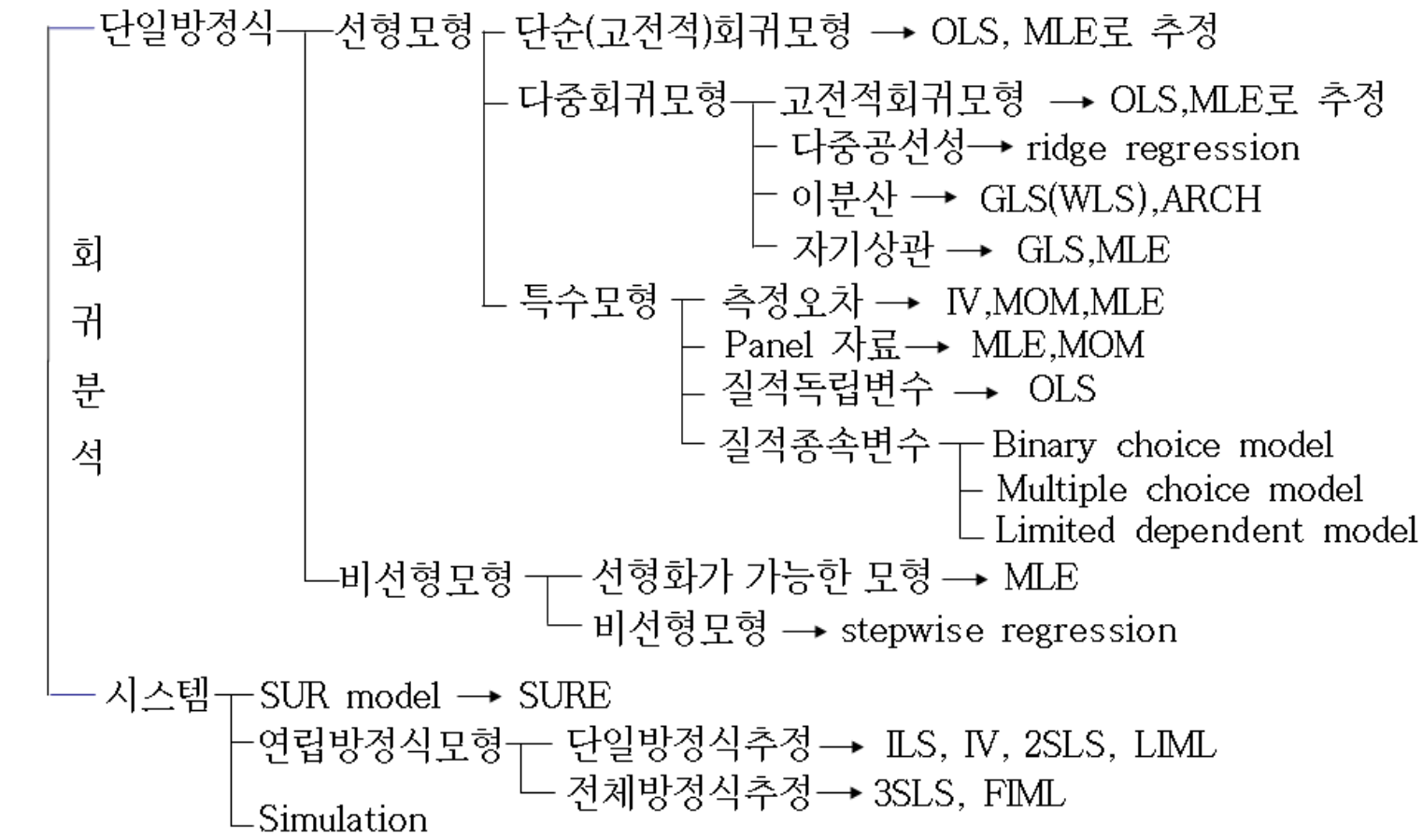
⑥ARMA(p,q) : $y_t - \phi_1 y_{t-1} - ... - \phi_p y_{t-p} = e_t + \theta_1 e_{t-1} + ... + \theta_q y_{t-q}$
$(p = 1,2,...,p, q = 1,2,...,q)$

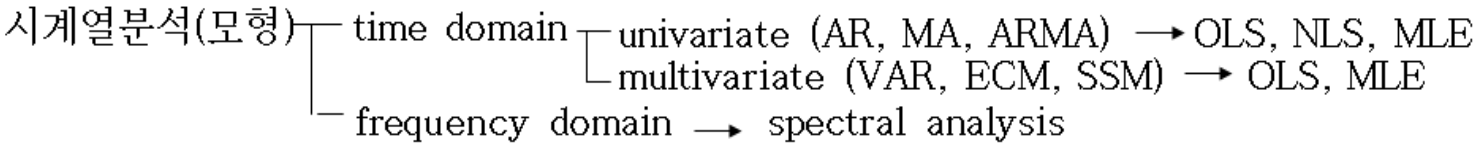
⑦ARIMA(p,d,q) : $\Delta y_t - \phi_1 \Delta y_{t-1} - ... - \phi_p \Delta y_{t-p} = e_t + \theta_1 e_{t-1} + ... + \theta_q y_{t-q}$
$(p = 1,2,...,p, q = 1,2,...,q)$

⑧VAR(p) : $X_t = \mu + A_1 X_{t-1} + ,...,+ A_k X_{t-k} + u_t$ 또는 $X_t = \mu + A(L) X_{t-1} + u_t$
$(t = 1,2,...,T)$ $\quad (t = 1,2,...,T)$

⑨패널모형 : $y_{it} = \alpha + \sum_{k=1}^{K} \beta_k X_{kit} + u_{it}$ 또는 $y_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k X_{kit} + \epsilon_{it}$, (단 $\alpha_i = \alpha + \mu_i$)

⑩패널 VAR(p) : $x = \mu_i + A_i(L) X_{t-1} + u$
$(i = 1,2,...,N, t = 1,2,...,T)$


회귀분석
단일방정식
선형모형
단순(고전적)회귀모형 → OLS, MLE로 추정
다중회귀모형
고전적회귀모형 → OLS,MLE로 추정
다중공선성 → ridge regression
이분산 → GLS(WLS),ARCH
자기상관 → GLS,MLE
특수모형
측정오차 → IV,MOM,MLE
Panel 자료 → MLE,MOM
질적독립변수 → OLS
질적종속변수
Binary choice model
Multiple choice model
Limited dependent model
비선형모형
선형화가 가능한 모형 → MLE
비선형모형 → stepwise regression
시스템
SUR model → SURE
연립방정식모형
단일방정식추정 → ILS, IV, 2SLS, LIML
전체방정식추정 → 3SLS, FIML
Simulation


시계열분석(모형)
time domain
univariate (AR, MA, ARMA) → OLS, NLS, MLE
multivariate (VAR, ECM, SSM) → OLS, MLE
frequency domain → spectral analysis

(참고)

1.연구모형 및 활용 software

연구모형	해당 논문 번호	해당 연구보고서 번호
Multi-sectoral model	2, 4, 8, 12, 14, 17, 18, 20	
ECM	1, 10	
VAR / 구조 VAR / 패널 VAR	7, 9, 15, 40	
I-O	11, 19	26,30,38,47,50,81,84,86,87,92,93,94,98
Nested Logit	16, 22	
SSM	5, 6, 13, 29, 37	
SEM	30, 31	31, 36
PSM	34	
UCM	35, 38, 39	24, 80
지역경제분석모형	26, 27	
지수개발	25, 42	9, 17, 18, 28, 57, 82
시계열예측	32, 43	15, 48,
Survey	3	
경제타당성분석		11, 12
DEA		22, 38, 50
정성분석		32, 43
판별분석		37
패널모형		59
Probit모형		68

구분	내용
OA	DOS / Windows / Internet / Word processing / Spreadsheet / Presentation / Database
통계 패키지	MSTAT / SAS / RATS / EViews / Stata (Limdep, Shazam, Minitab, Matlab, SPSS, DEAP)
Language	Html / Gauss / R / Python

2.연구경쟁력 강화

① 자가 진단

- 당신이 설정한 목표는 무엇입니까?
- 당신이 특히 자신 있는 분야는 무엇입니까?
- 당신이 특히 약한 분야는 무엇입니까?
- 당신은 남과 무엇을 다르게 하고 있나?(Uniqueness)
- 당신은 남보다 무엇을 앞서고 있나?(Excellence)
- 당신은 무엇을 준비하고 있나?(Next, Future)

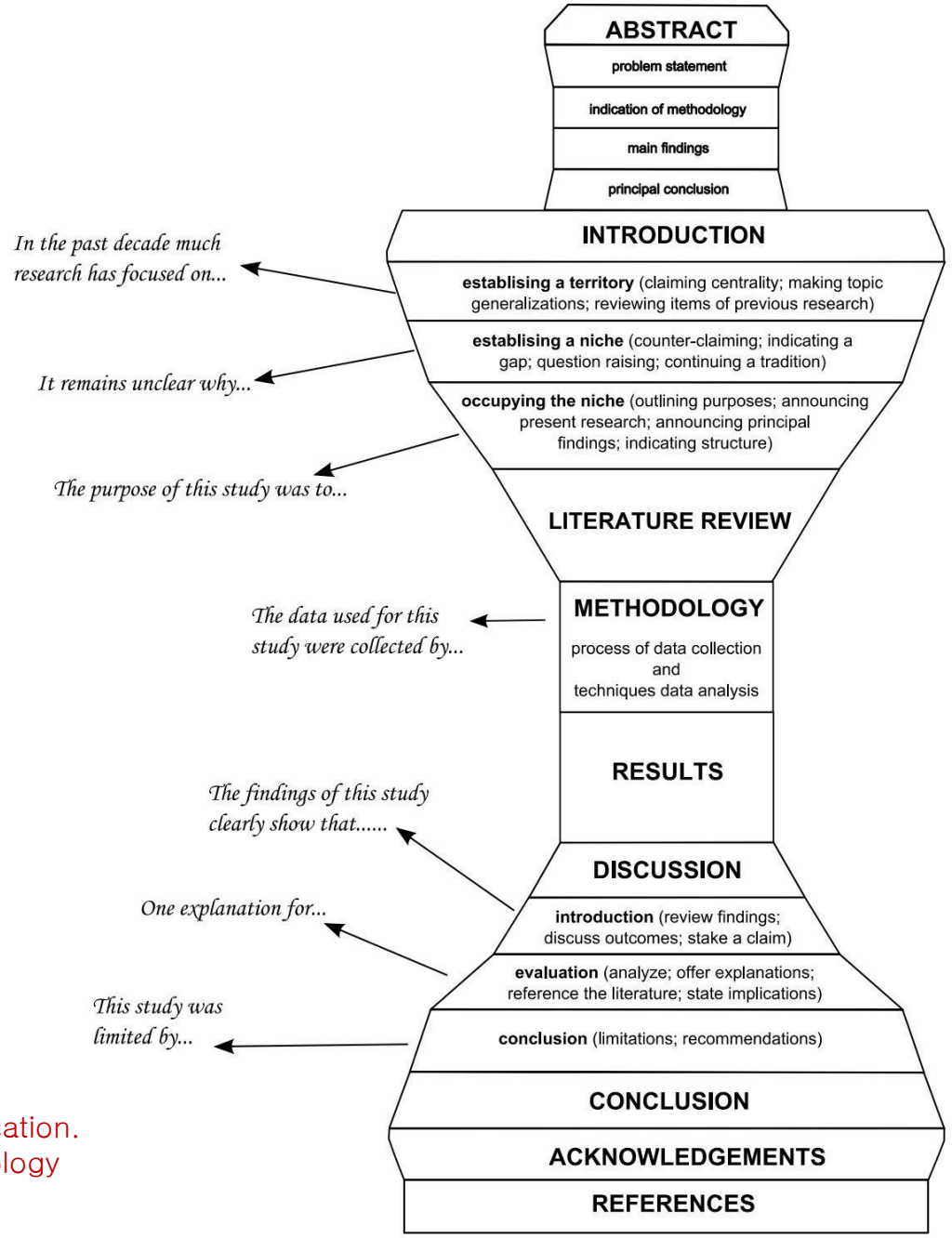
② 생존전략

- 극한을 경험하라
- Career Jump / Research Jump를 경험하라
- 차별적인 Expertise를 가져라(창의성, 우수성, 선제성)(자기관리 : 건강관리, 시간관리, 기록관리)
- 무식하고 무식하게 연구하라
- 지도교수 / 연구분야 / 연구방법론 선정에 고민하라

③ 연구영역 확장

- 연구모형의 적용 대상 확장 : 산업별/ 지역별/ 국가별
- 연구모형의 확장 1 : VAR / SVAR / PVAR
- 연구모형의 확장 2 : 정태모형 / 동태모형
- 연구모형의 다양화 : VAR/SSM/DEA/S(Structural)EM/UCM/S(Simultaneous)EM/PSM/ 지역경제분석모형
- 패키지 및 language의 선택 및 활용

3.논문 구조 및 작성 예시



Burrows, T. (2011). Writing research articles for publication. Unpublished manuscript, the Asian Institute of Technology Language Center, Khlong Luang, Thailand