제9장 조사데이터의 분석

전광희 교수
jkh96@cnu.ac.kr

학습내용

- 본 강의에서는 조사 데이터를 통계 분석할 때 널리 사용되는 분석기법에 대해서 살펴보고, 조사를 통해서 얻어진 데이터를 분석하여 그 결과로서 현상을 설명한다, 조사 데이터를 분석할 때 가장 기초가 되고 널리 사용되고 있는 기법에는 기술, 관계표현, 요약, 비교, 예측 등이 있다.

학습내용

1. 데이터분석을 위한 준비
2. 측정의 수준
   1) 명목척도 2) 순서척도(서열척도) 3) 구간척도(등간척도) 4) 비율척도(비척도)
   5) 측정수준과 관련된 참고사항
3. 타당도의 개념과 측정방법
   1) 타당도의 개념 2) 표면타당도 3) 내용타당도 4) 기준관련타당도: 동시타당도, 예측타당도
   5) 구성체타당도
4. 신뢰도의 개념과 측정방법
   1) 신뢰도의 개념 2) 재검사법(검사-재검사법) 3) 동형방법(복수양식법) 4) 반분법
   5) 내적일관성법(내적합치도법)

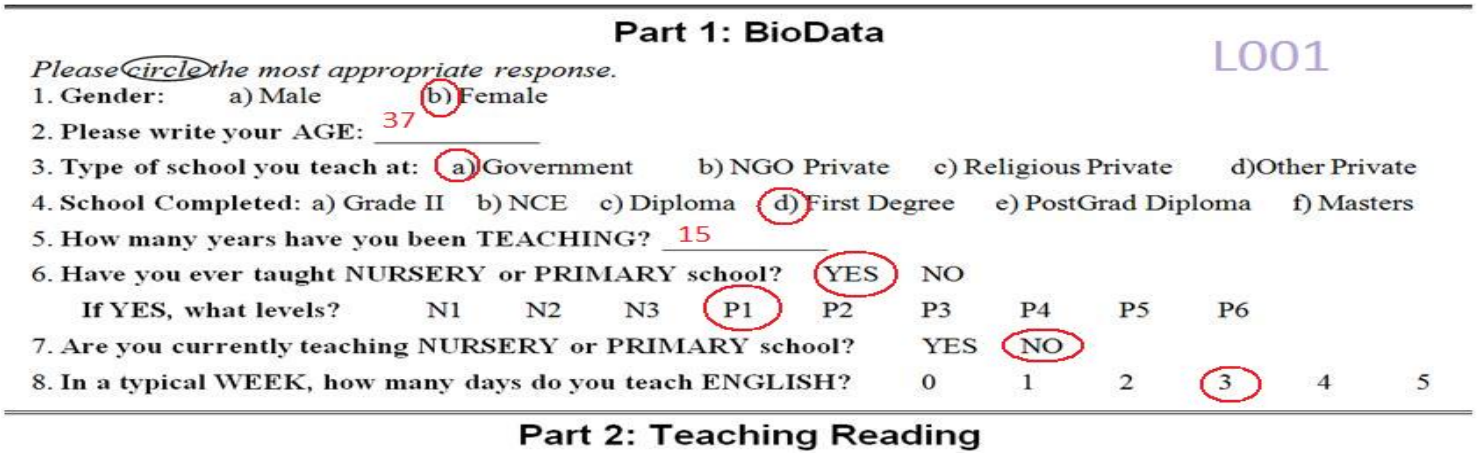
데이터 분석을 위한 준비

Ⅰ.데이터 입력

- 조사 데이터는 컴퓨터를 이용해서 분석하기에 알맞은 형태로 입력되어야 함
- 데이터 입력과정에서는 각 응답범주에 대해서 어떤 수치를 대응시킬 것인가에 대한 명확한 규칙을 마련해야 함
- 데이터 입력과정에서는 코드 작업이 명확해야 함

코드(code)란?

- 코드(code): 응답결과를 수치로 변환해 주는 규칙
- 코드작업, 곧 코딩(coding)을 위해서는 각 응답에 대해서 대응되는 숫자와 응답을 입력할 위치에 대한 정보를 사전에 결정해야 한다.



Part 1: BioData

L001

Please circle the most appropriate response.

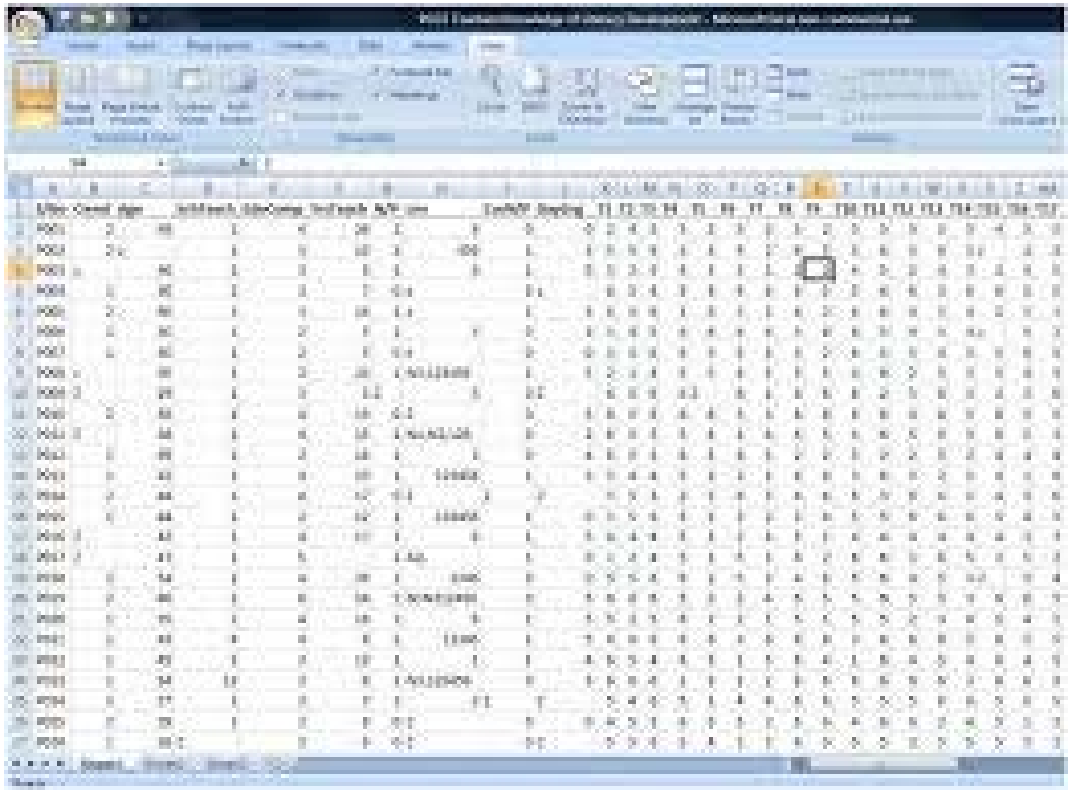
1. **Gender:** a) Male (b) Female
2. **Please write your AGE:** 37
3. **Type of school you teach at:** (a) Government b) NGO Private c) Religious Private d)Other Private
4. **School Completed:** a) Grade II b) NCE c) Diploma (d) First Degree e) PostGrad Diploma f) Masters
5. **How many years have you been TEACHING?** 15
6. **Have you ever taught NURSERY or PRIMARY school?** (YES) NO
   If YES, what levels? N1 N2 N3 (P1) P2 P3 P4 P5 P6
7. **Are you currently teaching NURSERY or PRIMARY school?** YES (NO)
8. **In a typical WEEK, how many days do you teach ENGLISH?** 0 1 2 (3) 4 5

Part 2: Teaching Reading

데이터를 입력하는 과정

① 데이터를 어떤 형식의 파일로 구조화할 것인가에 대하여 결정한다.
② 데이터 입력을 위한 코드를 결정한다.
③ 응답결과를 표준적인 범주와 연결시키는 데이터 코딩 작업을 수행한다.
④ 데이터를 입력하여 데이터 파일로 만든다.
⑤ 통계분석에 앞서 입력된 데이터의 정확성,완결성,일관성 등을 점검한다.

- 대부분의 경우에는 직사각형 형태로 데이터 파일을 만들게 됨
- 데이터는 엑셀과 같은 스프레드시트(spreadsheet)에 바로 입력할 수 있음
  - 스프레드시트에 입력된 데이터에서 행(row)은 케이스(cases), 열(column)은 변수(variables)를 나타냄
- 응답자가 응답하지 않아서 빈칸이라면 해당 칼럼에 특정 숫자(9 또는 99 등)또는 마침표(.)등을 입력할 수 있음



부호화 지침서

- 데이터를 입력하고 나면 부호화 지침서(code book)를 만듦
- 입력된 데이터를 다른 분석자가 이용하고자 할 때 데이터세트를 이해할 수 있게 한다
- 데이터에 대해서 충분히 점검하여 완벽한 형태로 데이터를 정리한 후에 최종적으로 마련하는 것이 바람직하다
- 각 변수가 가질 수 있는 값과 입력된 위치, 변수에 대한 이름, 각 코드의 의미에 대한 간단한 설명 등을 포함하게 된다

데이터 분석을 위한 준비

- 적합한 분석기법을 선택해야 한다
- 분석기법을 선택할 때에는
  - 분석목적을 명확하게 해야 하고,
  - 독립변수와 종속변수는 어느 변수이며,
  - 각 변수의 측정 척도가 무엇인가를 살펴야 한다.
- 통계분석기법은 변수의 수와 변수에 대한 측정수준(measurement level)에 따라 결정된다.

1. 측정의 척도

- 각 변수의 측정 척도에 따라 적용할 수 있는 분석기법에 차이가 있다
- 일반적으로 측정 척도에 따라서 변수가 갖고 있는 정보의 양(量)은 명목척도, 순서척도, 구간척도, 비율척도의 순으로 증가한다
- 비율척도로 측정된 자료는 구간척도나 순서척도 또는 명목척도로 변환하여 분석할 수 있음

  ☞ 비율척도로 측정된 데이터에 적용할 수 있는 분석기법이 가장 다양하고, 반대로 정보의 양이 가장 적은 명목척도로 측정된 자료에 적용할 수 있는 분석기법이 가장 적다.
- 측정의 척도는 변수의 내재적인 특성으로 고정된 것이 아니라, 변수를 측정하고 개념화하는 방법의 결과로 나타난 것이다

2. 독립변수와 종속변수

- 변수(variable)란? 조사단위로부터 측정된 특성

  예) 학생들의 체중, 키, 제품에 대한 만족도

  일반적으로 조사 데이터를 분석할 때 분석방법의 선택은 측정 척도와 변수의 수에 깊은 관련이 있음

- 조사데이터를 분석함에 있어서 어떤 변수는 독립변수로, 어떤 변수는 종속변수로 정하게 됨
- 독립변수는 설명변수(說明變數)라고도 함

  ☞이들 변수들이 결과나 종속변수를 설명하거나 예측하는 데 사용되기 때문

- 독립변수와 종속변수는 조사목적과 목표모집단을 설정함으로써 명확해짐

3. 분석기법

- 연구주제,변수의 수,각 변수에 대한 측정 척도에 따라 결정됨
- 분석기법을 선택할 때 점검해야 할 사항
  - ① 독립변수의 수는 몇 개인가?
  - ② 독립변수의 데이터 유형을 살핀다.
  - ③ 종속변수의 수는 몇 개인가?
  - ④ 종속변수의 데이터 유형을 살핀다.
  - ⑤ 사용 가능한 데이터 분석기법은 무엇인가?
  - ⑥ 측정 척도와 분석기법의 가정 등을 고려해서 조사목적에 알맞은 분석기법을 선택한다.

- 만약 변수가 하나인 경우라면 일변량 분석기법(univariate analytic technique)을, 두 개인 경우라면 이변량 분석기법(two-variate analytic technique)을 사용해야 함
- 통계분석은 크게 기술적 분석(descriptive analysis)과 추론적 분석(inferential analysis)으로 구분됨
  - -기술적 분석: 데이터에 대한 기초적인 분석

    예: 도수분포표(frequency table), 중심위치(central location) 및 산포도(dispersion) 측도 등의 수치요약, 히스토그램(histogram) 등의 그래프, 산점도(scattergram)와 상관계수(correlation coefficient) 등
  - -추론적 분석: 가설검증(hypothesis testing)이나 통계모형을 적합하여 분석하는 것

    예:회귀분석(regression analysis),모평균 검증(population mean test), 카이제곱 검증(chi-square test), 상관분석(correlation analysis), 로그선형모형(log-linear model) 분석

4.데이터 분석에 앞서

- 데이터 분석 전 입력된 전체 데이터를 점검해야 함
- 점검의 첫 번째 단계는 이상치를 검출하고 부정확한 값을 찾아내는 것
  - 이상치:나머지 데이터와 비교할 때 지나치게 크거나 작은 값
  - 두 번의 분석결과를 비교하여 이상치의 영향을 알아보고,이를 통해서 이상치를 어떻게 처리할 것인가에 대해서 알 수 있음
- 또한, 데이터에 있는 부정확한 관측값은 선별해야 함
- 무응답에 대한 처리문제
  - 단위 무응답(unit non-response): 표본으로 추출된 개인이나 추출단위로부터 응답을 얻지 못한 경우에 발생한다
  - 항목무응답(item non-response): 설문항목 중에서 어느 항목에 대해 응답하지 않는 부분적인 무응답의 경우

통계 소프트웨어

❖통계 소프트웨어란?

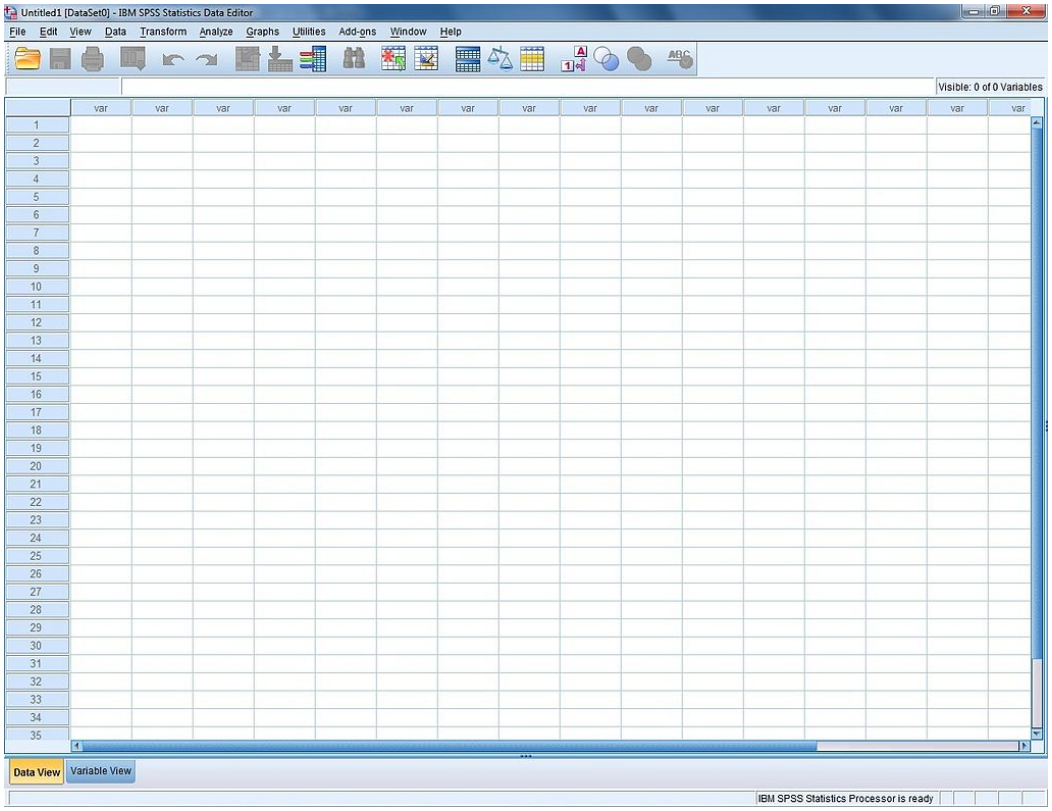
광범위한 분야에서 데이터의 처리와 통계분석을 쉽게 할 수 있도록 개발된 소프트웨어라고 할 수 있으며, 이를 통계패키지(statistical package)라고 한다.

1. SPSS (Statistical Package for the Social Science)

- 1970년대 초 시카고대학을 중심으로 처음에는 사회과학에서 발생하는 통계 문제를 분석하기 위해서 개발됨
- 오늘날에는 광범위한 분야에 대한 데이터 입력, 관리 및 통계분석을 목적으로 사용되고 있음
- 모든 분석절차가 메뉴 방식으로 구성됨
- 데이터의 입력과 관리가 간편
- 한글로 각종 설명이 지원되므로 일반 사용자도 쉽고 편하게 이용할 수 있는 특징

IBM-SPSS Software 안내 홈페이지

SPSS running on Window 7

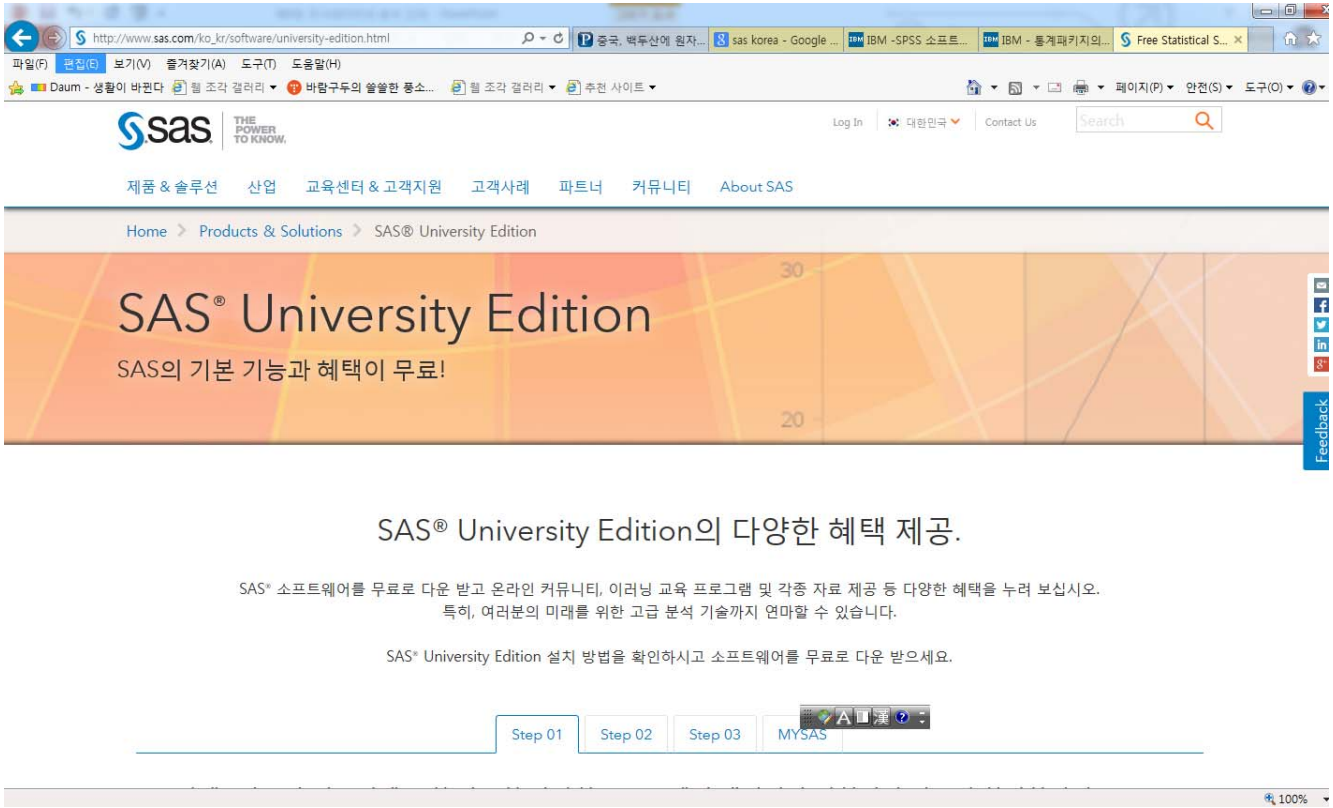


SPSS Inc. Logo

2. SAS (Strategic Applications Software)

- 1970년대 초에 미국의 노스캐롤라이나 주립대학(North Carolina State University)에서 일반적인 데이터 분석을 위하여 개발된 것
- 프로그램 자체의 융통성이 크고, 데이터베이스와 데이터웨어하우징, 그리고 각종 응용 프로그램과의 호환 등이 편리한 통합 패키지로서 데이터의 처리기능이 뛰어남
- 오늘날 데이터의 단순한 통계적 처리보다는 방대한 양의 데이터 처리와 여러 가지 데이터 분석을 통한 의사결정에 도움을 주는 시스템으로 인식되고 있음

SAS Software 안내 홈페이지



SAS Output

The FREQ Procedure

Table of CHD by xsrum

CHD	xsrum				
Frequency Expected Deviation Cell Chi-Square Percent Row Pct Col Pct	0-199	200-159	220-259	260+	Total
chd	12	8	31	41	92
	22.083	17.583	32.536	19.798	
	-10.08	-9.583	-1.536	21.202	
	4.6037	5.223	0.0725	22.764	
	0.90	0.60	2.33	3.09	6.92
	13.04	8.70	33.70	44.57	
	3.76	3.15	6.60	14.34	
nochd	307	246	439	245	1237
	296.92	236.42	437.46	266.2	
	10.083	9.5831	1.5357	-21.2	
	0.3424	0.3885	0.0054	1.6886	
	23.10	18.51	33.03	18.43	93.08
	24.82	19.89	35.49	19.81	
	96.24	96.85	93.40	85.66	
Total	319	254	470	286	1329
	24.00	19.11	35.36	21.52	100.00

Output - (Untitled)

The REG Procedure
Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3898.77279	1949.38639	69.51	<.0001
Error	27	757.16737	28.04324		
Corrected Total	29	4655.94016			

Root MSE	5.29559	R-Square	0.8374
Dependent Mean	307.97792	Adj R-Sq	0.8253
Coeff Var	1.71947		

Parameter Estimates

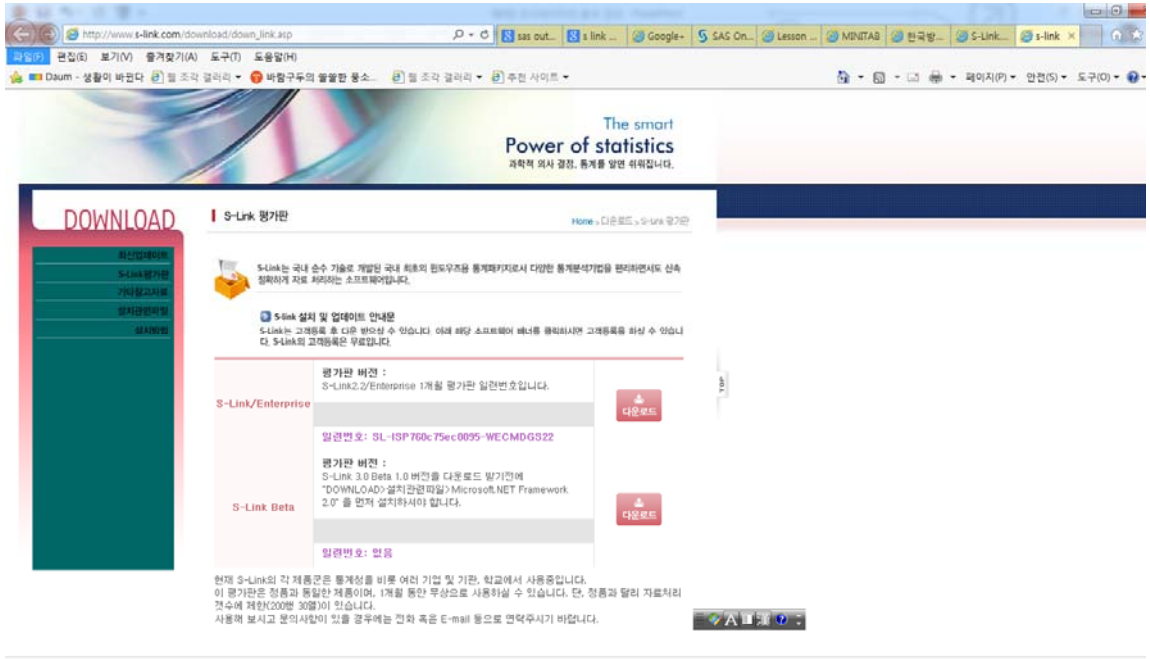
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	221.71079	51.80520	4.28	0.0002
x	1	-1.03425	0.09110	-11.35	<.0001
z	1	0.94752	0.25568	3.71	0.0010

3. 미니탭(Minitab)

- 각종 데이터의 분석을 메뉴 방식으로 처리할 수 있도록 개발한 통계 소프트웨어

4. S-Link

- 순수한 국내 기술로 개발된 통계패키지로서 스프레드시트 형태로 데이터를 입력하여 메뉴 방식으로 분석하는 통계 소프트웨어
- 모든 체계가 한글로 되어 있고, 대부분의 데이터 분석기법을 모두 포함하고 있어서 쉽고 편리하게 활용 가능



5. 엑셀

- 엑셀(EXCEL): 마이크로소프트사에 의해서 개발된 스프레드시트 프로그램
- 스프레드시트(spreadsheet)란 컴퓨터에서 계산,자료의 정리 및 관리,차트의 작성 등을 효율적으로 할 수 있게 만든 응용 프로그램
- 통계분석을 목적으로 개발된 통계패키지는 아니지만 조사된 데이터의 기초적인 분석이 가능하다.

4. 표, 그래프, 데이터의 수치 요약

1. 그래프와 표

(1) 원그래프(pie graphs): 각 범주에 속한 상대도수에 비례하도록 원의 조각을 나누어 그래프로 표현한 것이다.

- **원그래프 작성시 참고할 사항**
  - ① 각 범주에 속한 비율이나 백분율을 표현하고자 할 때 사용한다.
  - ② 그래프의 제목은 간단하게 정하여 그래프 아랫부분에 적는다.
  - ③ 그래프 작성에 외부의 데이터를 이용하였다면 데이터의 출처를 밝힌다.
  - ④ 8개 이상의 조각으로 구분하는 것은 피한다.
  - ⑤ 필요할 경우에 가장 작은 조각들은 묶어서 '기타'항목으로 처리한다.
  - ⑥ 어떤 조각을 강조하기 위해 나머지 부분에서 떼어 내어 나타내거나 가장 어두운 색, 밝은 색 또는 무늬를 이용해서 표현한다.

(2) 막대그래프(bar graphs)와 꺾은선그래프(line graphs)

- **막대그래프(bar graphs)** : 각 범주에 속한 비율을 하나의 막대로 표현하여 시각적 비교가 유용함. 해석이 쉬워서 조사보고서에 널리 사용. 수직 막대그래프와 수평 막대그래프로 구분.
- **꺾은선그래프(line graphs)** : 구간척도나 비율척도로 측정된 데이터에 대해서 시간의 흐름에 따른 변화추이의 집단간 비교를 목적으로 작성된다.

**** 막대그래프와 꺾은선그래프 작성시 참고할 사항**

① 그래프의 제목(title)을 붙인다.
② x축과 y축 값의 의미를 설명한다.
③ 외부의 데이터를 이용하는 경우에는 데이터의 출처(source)와 조사기관(survey agency)을 밝힌다.
④ 집단 간의 비교 또는 시간 흐름에 따른 변화를 보기 위해서는 수직 막대그래프를 이용한다. 막대의 수가 6개 보다 많은 경우에는 수평 막대그래프나 꺾은선그래프를 이용한다.
⑤ 시간의 흐름에 따라 여러 지점에서 조사결과를 비교할 목적이라면 꺾은선그래프를 이용한다.
⑥ 전체적으로 조사결과가 잘 나타날 수 있도록 y축의 값을 정한다. 만약 y축이 0에서 시작하지 않는다면 이를 그래프에 명확하게 밝힌다.

(3) 표(table)

표는 응답자의 분포 또는 응답결과를 정리하거나 시간 흐름에 따른 조사 결과를 비교할 때 사용된다. (보고서에서 유용하게 사용됨.) 조사결과를 수치적으로 표현하고자 한다면 표를 이용하는 것이 적당하다.

-표 작성시 참고할 사항

① 모든 표에서 백분율과 분석대상이 된 사례 수를 표시해야 한다.

② 두 변수에 대해서 표를 작성하는 경우에 열에는 비교하려고 하는 가장 중요한 변수의 항목을 위치하세 하는 것이 좋다.

③ 필요하다면 항목 값(item value)을 응답자 수를 기중으로 오름차순 또는 내림차순으로 정리한다.

④ 정형화된 기호를 사용해서 보고서를 읽는 사람이 쉽게 이해할 수 있도록 배려한다. 예를 들어, 통계적으로 유의한 정도(statistical significance)를 표현할 때 유의수준 5%에서 통계적으로 유의하면 *, 유의수준 1%에서 통계적으로 유의하면 ** 등으로 표시하는 것이 일반적이다.

2. 데이터의 수치 요약

① 중심위치의 측도, 산포의 측도, 상관계수, 왜도와 첨도 등
② 주로 연속형 변수의 분포 특징을 수치로 요약할 때 사용됨

1. 중심 위치의 측도

(1) 평균 (mean)

-데이터의 산술평균을 말하며, 중심위치의 측도 중에서 가장 널리 사용 됨
-연속형 변수에 대해서 좌우대칭인 분포를 갖는 경우에 사용됨

평균 : 변수 값의 총합을 표본 크기로 나누어준 값

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

(2) 중앙값 (median)

- 연속형이나 순서형 변수에 대해서 분포가 한쪽으로 치우친 경우에 사용됨
- 자료를 크기 순으로 나열할 때 가장 가운데 오는 값

데이터를 크기 순으로 나열하여 데이터의 수가 홀수이면 중앙에 위치하는 자료 값이 중앙값이 되고, 짝수이면 중앙에 위치하는 2개 데이터의 평균이 중앙값이 된다. 데이터가 한쪽으로 치우치거나 이상값(outliers)이 있는 경우에 중심위치의 측도로 이용된다.

$$\tilde{x} = x_{\frac{n+1}{2}}$$

(3) 최빈값 (mode)

어느 변수에 대해서 가장 빈도가 많은 관찰 값

2. 산포의 정도에 대한 측도

(1)범위(range)

- 연속형 데이터에 대해서 사용되고, 조사된 변수 값의 최대값(maximum)에서 최소값(minimum)을 뺀 것
- 간단하게 퍼진 정도를 구할 수 있지만, 이상값이 있는 경우에는 그 영향을 크게 받아 변동을 나타내는 데 적합하지 않음

(2)사분위수범위(interquartile range)

- 범위와 유사한 형태이지만 양극단위 값들에 의한 영향을 거의 받지 않는 산포의 측도
- 전체 데이터를 순서대로 정리하여 4등분한 후 자료의 가운데에 위치한 50% 부분에 대해서 범위를 구한 것

중앙값(Q2)를 구한 후 중앙값을 중심으로 두 덩어리의 데이터에서 각각 중앙값(Q1,Q3)을 구하여 자료 전체를 4등분하는 사분위수(Q1,Q2,Q3)를 구하게 된다. 여기서 1 사분위수와 3사분위수의 차이가 사분위 간 범위가 되는 것이다.

$$\text{사분위수범위}(IQR) = Q_3 - Q_1$$

(3) 표준편차(standard deviation)

- 평균이 자료의 중심위치를 나타내는 통계량으로 쓰일 때 산포의 측도로 가장 널리 사용되는 통계량
- 데이터가 평균을 중심으로 얼마나 넓게 분포하고 있는가를 나타내는 통계량

표준편차는 분산의 양의 제곱근으로 데이터의 흩어진 정도를 표현하고자 할 때, 원 자료의 측정단위와 같은 단위이기 때문에 널리 사용된다.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}x_i^2 - \frac{(\sum_{i=1}^{n}x_i)^2}{n}}{n-1} : \text{분산}$$

$$s = \sqrt{s^2} : \text{표준편차}$$

3. 왜도와 첨도

(1) 왜도 (skewness)

- 연속형 데이터에 대해서 좌우대칭인지 아니면 어느 한쪽으로 치우친 형태의 분포를 갖는지를 나타내는 측도로, 만약 분포가 완벽히 좌우대칭이라면 왜도 값은 0을 나타냄
- 분포의 모양이 오른쪽으로 꼬리가 긴 분포에서 왜도는 양수의 값을 나타낸다. 반대로 왼쪽으로 꼬리가 긴 분포에 대한 왜도 값은 음수 값을 나타낸다.

(2) 첨도 (kurtosis)

- 분포가 얼마나 뾰족한지 아니면 평평한지를 나타내주는 통계량
- 분포가 중심이 집중되어 뾰족한 형태를 갖는 경우의 첨도는 양수 값을 갖고, 반대로 중심에 덜 집중되어 평평한 형태의 분포를 갖는 경우의 첨도는 음수 값을 나타낸다. 정규 분포의 경우에 첨도 값은 0이다.

3. 상관관계

1. 연속형 데이터와 상관계수

(1)상관관계

변수들 사이의 일관적인 연관성을 의미함

(2)상관관계의 사례

- 공부시간과 시험점수는 어떤 관계인가?
- 소득이 높은 사람은 삶의 만족도도 높은가?

(3)상관계수(correlation coefficient)

① 두 변수가 모두 연속형 변수인 경우에 상관관계를 살펴보기 위하여 사용

② 피어슨의 적률상관계수(Pearson product-moment coefficient)라고도 함

③ 두 변수 X,Y에 대한 관측값을 각 $x_1, x_2, \cdots, x_n$ 과 $y_1, y_2, \cdots, y_n$ 라고 할 때, 상관계수는 다음과 같이 계산할 수 있다.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- 두 변수의 상관관계는 우선 두 변수 사이의 산점도(scattergram)를 그려서 살펴본다.
- 상관계수 값은 -1과 1사이의 값을 갖는다. 두 변수의 상관관계가 강할수록 (-1이나 1 근처의 값을 갖는 경우) 직선관계가 더욱 뚜렷해진다.

2. 순서형 데이터의 상관계수

(1) 사례

제품에 대한 만족도와 교육정도가 모두 순서형으로 측정된 경우에 두 변수의 관련성 정도를 평가하는 경우

(2) 스피어만 순위상관계수(Spearman rank correlation)

- 두 변수 모두가 순서형이거나 한 변수는 순서형이고 다른 변수는 연속형인 경우에 두 변수 사이의 상관관계를 나타내기 위해서 사용됨
- 두 변수 X,Y에 대한 관측 값을 각각 $x_1, x_2, \cdots, x_n$ $y_1, y_2, \cdots, y_n$

$$R_i = x_1, x_2, \cdots, x_n \text{에서 } x_1 \text{의 순위}$$
$$S_i = y_1, y_2, \cdots, y_n \text{에서 } y_1 \text{의 순위}$$
$$\Rightarrow r_s = \frac{\sum_{i=1}^{n}(R_i - \overline{R})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}}$$

3. 두 명목형 변수 사이의 관계

- 명목형 두 변수의 관계에 대한 통계적 유의성 여부에 관심이 있는 경우 → 카이제곱 검증 이용

5. 자주 사용되는 데이터 분석기법

- 조사를 통해서 얻어진 데이터는 크게 연속형인 경우와 범주형인 경우로 구분 할 수 있다.
- 일반적으로 독립변수가 순서형인 경우는 명목형 변수로 간주하여 분석하고, 종속변수가 순서형인 경우에는 연속형 변수처럼 간주하여 분석할 수 있을 것이다.
- 어떤 통계분석을 적용하고자 한다면 반드시 통계분석에 대한 교재를 참고하여 그러한 통계분석이 가능하게 되는 가정에 대해서 살펴보아야 한다.
- 만약 조사된 데이터가 적용하고자 하는 통계분석에 필요한 가정을 만족하지 못한다면 다른
- 통계분석기법을 찾아야 한다. 통계분석을 진행하면서 분석결과를 다시 검토하여 분석에 필요한 가정을 만족하는지 확인해야 한다.

1. t-검증

- 통계조사를 통해서 종종 두 그룹 또는 그 이상의 집단을 대상으로 비교를 행하게 된다. 이와 같은 두 집단 간의 비교문제에서 가장 널리 사용되는 방법은 두 집단의 t-검증이다.

< 통계적 가설검증의 절차>

(1) 귀무가설을 서술한다.

▪ 일반적으로 귀무가설(null-hypothesis)은 두 집단의 평균에서 차이가 없다 등으로 서술됨

▪ <사례> 연구문제(research hypothesis) : A학교와 B학교 사이의 시험성적에 차이가 있는가?

→ 귀무가설 : A학교와 B학교 사이의 시험성적에 차이가 없다.

*만약 평균에 차이가 없는 것으로 검증되었으면 "귀무가설을 기각시키지 못했다"로 표현한다. 만약에 두 집단의 모평균 차이에 대한 가설검증에서 귀무가설이 기각되었으면 차이가 있다고 한다.

(2) 통계적 검증의 유의수준(significance level)을 정한다.

▪유의수준은 가설검증이 시행되기 전에 결정함

▪유의수준은 귀무가설이 옳은데도 불구하고 기각될 확률을 의미

▪일반적으로 유의수준 0.05, 0.01등으로 작은 값이 이용됨

(3) 검증통계량(test statistics)의 값을 구한다.

▪ 검증통계량의 값과 기각역(rejection region)을 비교하여 귀무가설의 기각 여부 결정

▪ 오늘날 검증방법은 통계패키지로 계산한 p값을 이용하여 실행함

▪"계산된 p값이 유의수준보다 작으면 귀무가설 기각"

* 조사데이터가 한쪽으로 치우쳐진 형태이거나 이상값이 있어서 정규분포를 따른다고 가정할 수 없다면 비 모수적 방법(non-parametric methods)을 적용하는 것을 검토해야 한다. 비 모수적 방법은 관측된 데이터의 분포에 대해서 가정을 하지 않고 통계적 추론(statistical inference)을 한다.

- T-검증을 적용하기 위해서는 조사 데이터가 정규분포(normal distribution)를 만족해야 한다.
- 먼저 일표본 검증문제에서 t-검증을 적용하기 위해서는 조사 데이터가 정규분포를 따라야 한다. 조사 데이터가 정규분포를 따르는가를 살펴보기 위해서는 조사 데이터에 대한 히스토그램(histogram)이나 줄기-잎 그림(stem-leaf chart)을 그려서 살펴보거나, 통계패키지에서 제공하는 정규 확률도표를 이용할 수 있다.
- 만약 조사 데이터가 한쪽으로 치우쳐진 형태이거나 이상치가 있어서 정규분포를 따른다고 가정할 수 없다면 비 모수적 방법을 적용하는 것을 검토해야 한다. 비 모수적 방법은 관측된 데이터의 부호에 대해서 가정을 하지 않고 통계적 추론을 한다.

*T-검증 사례

① 서울시의 어느 구는 이 지역의 가구당 평균 소득액이 서울시 전체 평균 소득액과의 차이를 검증하는 경우 → 한 표본 t-검증

② 어떤 다이어트 프로그램에 참여하기 전과 후의 체중변화를 살펴보는 경우 → 대응비교 t-검증

③ 성인 대상으로 하는 독서 성향 조사 결과를 분석하여 남자와 여자의 독서량에 차이가 있는가를 검증하는 경우 → 독립표본 t-검증

** 대응비교(pairwise comparison)의 문제는 같은 조사대상자에게서 대개 두 차례에 걸쳐서 관측하여 모평균 차이의 유무를 확인하게 된다.

- 대응 비교의 문제는 같은 조사대상자에게서 대개 두 차례에 걸쳐서 관측하여 모평균 차이의 유무를 확인하게 된다. 이 경우에도 분석에 필요한 가정은 두 변수가 모두 정규분포를 따라야 한다는 점이다. 만약 정규분포를 따르지 않는 다면 비 모수적 방법인 **윌콕슨 부호 순위 검증(Wilcoxon signed-ranks test)**을 이용하는 것이 바람직하다.
- 세 번째 검증(독립표본 t-검증)에 대한 기본적인 가정도 마찬가지로 관측 값들이 정규 분포를 따른다는 것이다. 이 경우에는 두 집단의 분산이 동일하다고 가정할 수 있는 경우와 서로 다른 경우로 구분되어 검증통계량에 약간의 차이가 있다. 만약 조사 데이터가 정규분포를 만족하지 못하거나 이상치가 있는 경우에는 비 모수적 검증법인 **윌콕슨 순위합검(Wilcoxon rank-sum test)**을 이용하는 것이 바람직하다.

2. 분산분석(ANOVA: analysis of variance)

- 분산분석에서 다루는 분야는 세 개 이상의 그룹에 대한 평균 차이 유무에 대한 검증이다. 분산분석의 결과로 알 수 있는 것은 그룹간의 전반적인 차이에 대한 검증 결과 이다. 차이가 있는 것으로 검증되었어도 그 차이가 구체적으로 어느 쌍이나 그룹 간의 차이로 발생한 것인지는 알 수 없다.
- 차이의 구체적인 원인 규명을 위해서는 추가로 **다중비교(multiple comparison)** 분석을 해야 한다.
- 다중비교를 위한 대표적인 검증 법으로는 **피셔(Fisher)의 LSD(least significant difference) 검증, 터키(Tukey)의 HSD(honest significant difference) 검증. 셰페(Scheffe) 검증** 등이 있다.

3. 카이제곱 검증(chi-square test)

- 표와 같이 조사된 데이터를 정리하여 얻은 이차원 분할표 자료에서 두 변수 간의 연관성 여부를 검증할 때 사용된다.
- 카이제곱 통계량의 값이 크면 클수록 두 변수가 독립적이라는 귀무가설이 옳지 않다는 증거가 된다. 일반적으로 카이제곱 검증을 통계패키지를 이용해서 하는 경우에는 p값(유의확률)을 이용한다.
- 계산된 p값이 유의수준보다 작게 되면 귀무가설을 기각하고 대립가설을 채택하게 되며, 반대로 p값이 유의 수준보다 크면 귀무가설을 기각하지 못한다.

*카이제곱검증사례

어느 대학교에서 남녀 대학생 각각 150명씩을 표본으로 추출하여 학교생활에 대한 만족도(만족, 보통, 불만족)을 조사하였다. 이 경우에 제기 될수 있는 문제는 학생의 성별 구분과 학교생활에 대한 만족도가 서로 연관이 있는지에 대한 것이다.

	남 자	여 자	합 계
만족한다	65	40	105
보통이다	65	70	135
불만족이다	20	40	60
합 계	150	150	300

H_0 : 행과 열의 변수는 서로 독립이다.

$$x^2 = \sum_{i=1,j=1}^{n}\sum^{n} \frac{(f_{ij} - E_{ij})^2}{E_{ij}}$$

위 식에서 f_{ij}는 표의 분할표에서 (i,j)칸의 관측도수이고, E_{ij}는 두 변수가 서로 독립이라는 가정하에서 얻을 수 있는 (i,j)칸의 기대도수이다. 총 응답자수가 n이고, i번째 행의 응답자가 r_i, j번째 열의 응답자가 c_i인 경우에 (i,j)칸의 기대도수는 다음과 같이 정의된다.

$$E_{ij} = \frac{r_i \times c_i}{n}$$

행과 열의 수가 각각 r과 c인 경우에 앞서 정의한 카이제곱 검증통계량 x^2은 두 변수가 서로 독립이라는 가정하에서 자유도가 $(r-1)(c-1)$인 카이제곱 분포를 따르게 된다. 따라서 유의 수준 α에서 검증인 경우에

$$x^2 > x^2[(r-1)(c-1);\alpha]$$

이면 두 변수가 서로 독립이라는 귀무가설을 기각한다.

4. 분석결과 해석에 대한 유의사항

(1) 결과에 대한 지나친 해석

- 지나치게 해석하는 방식 중 가장 흔히 하는 것은 표본오차(sampling error)를 무시하는 경우, 의미 없는 관계에 지나치게 중요성을 부여하는 경우, 인과관계를 지나치게 언급하는 경우 등이다.
- 조사 데이터를 분석하여 얻은 결론은 모집단에 대한 것으로 일반화 되며 이렇게 표본조사의 결과를 모집단의 결과로 일반화하기 위해서는 조사된 표본이 확률표본이고, 조사과정 전체가 엄밀하게 관리되었을 때 가능하다.

(2) 조사과정에서 오차들이 개입될 수 있음을 고려할 것

- 무작위 오차(random error)가 보사결과에서 나타나는 비체계적인 변동이라면 편향(편의, bias)는 어떻게든 조사결과에 체계적으로 영향을 주는 것을 말한다.
- 일반적으로 오차는 조사방법, 질문지 내용, 면접과정, 표본추출, 응답자의 규모, 면접원의 특성, 결과분석 등 전체 진행과정에서 발생할 수 있다.
- 조사 데이터를 분석하는 데 가장 어려운 점은 표본오차(표집오차, sampling error)를 제외한 대부분의 오차를 정확히 측정할 수 없다는 점인데, 이것은 조사결과 얼마나 많은 오류가 개입되어 있는지 정확하게 알 수 있는 방법이 없다는 것을 의미한다.

(3) 통계적 유의성과 실제적인 유의성

- 통계적 검증결과는 유의하게 나타났지만 실제로 적용할 만큼 뚜렷하지 못한 경우를 볼 수 있다.
- 일반적으로 표본의 크기가 큰 경우에는 작은 차이라도 통계적으로 유의미한가의 여부만을 가지고 조사 데이터를 평가하는 것은 부적절하고 신뢰구간도 함께 사용되어야 한다. 현실의 문제에서는 모수에 대한 신뢰구간(confidence)과 현실적인 측면을 함께 고려해야 할 것이다.

예) 새로운 생산공정을 도입하여 제품의 질을 향상시키고자 하는 경우에는 새로운 공정의 도입으로 나타나는 질의 향상 정도와 공정 도입에 따른 현실적 비용 등을 고려해야 한다. 이는 경우에 따라서는 공정 도입 전과 후에 제품의 질에는 통계적으로 유의한 차이가 있지만, 현실적으로 유의한 차이라고 볼 수 없는 경우가 발생할 수 있기 때문이다.

감사합니다