



## Week03

# 비즈니스 애널리틱스 개요 (II)

국민대학교 비즈니스IT전문대학원  
안 현 철 교수 (hcahn@kookmin.ac.kr)

## 목 차



국민대학교  
KOOKMIN UNIVERSITY

### 1

## 통계학 개요

- 통계 및 통계분석
- 통계의 기본 개념들
- 통계의 2가지 분야
- 통계분석의 절차
- 통계분석과 컴퓨터

### 2

## SPSS Statistics 활용

- SPSS Statistics 기본 구성
- 자료의 입력
- 기술통계 실습
- 카이제곱 검정: 교차분석
- 독립표본 t검정
- 대응표본 t검정
- 일원배치 분산분석: One-way ANOVA

# 통계 및 통계분석

## ❖ 통계학(Statistics)이란?

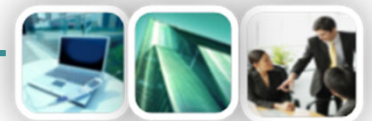
진실된 현황을 가능한 정확히 알아내고, 미래에 대한 적절한 예측을 하기 위해 자료를 효율적으로 수집 정리하고, 분석하는 학문분야

## ❖ 통계분석의 목적

- 1. 진실된 현황을 제대로 파악하기 위해
- 2. 미래를 예측하기 위해

## ❖ 통계학이 적용되는 대표적인 예들

- 공학자 : 현재 생산되는 제품의 품질이 어떤 상태에 있는가?
- 경영자 : 제품에 대한 소비자의 선호도가 어떻게 나타나고 있는가?
- 사회과학자 : 정치에 대한 국민들의 의견은 어떻게 나타나고 있는가?



# 통계의 기본 개념들 : 모집단과 표본 (1)

우리나라 국회의원들이 소유하고 있는 주택의 평균가격은 얼마인가?

## ❖ 모집단 (population)

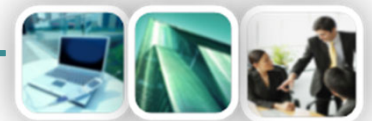
- 관심의 대상이 되는 집단 전체
- (예) 우리나라 국회의원들이 소유하고 있는 모든 주택

## ❖ 모수 (parameter)

- 모집단이 가지고 있는 특징을 나타내는 수치
- (예) 평균 가격

## ❖ 전수조사 (census)

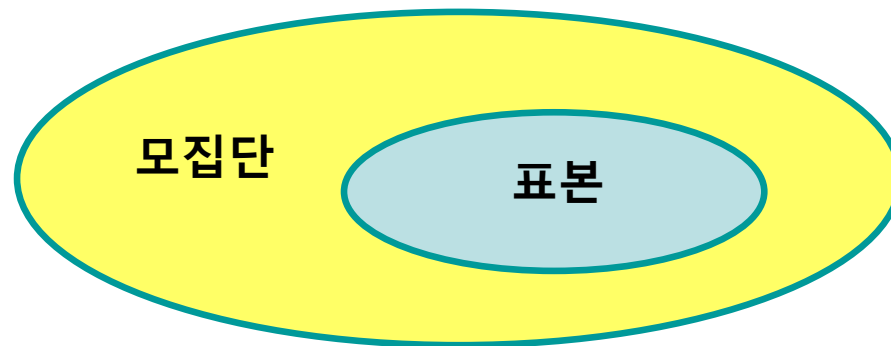
- 모집단의 모든 개체를 전부 조사하는 방법
- 모수의 값을 가장 정확하게 알아내는 방법 : 하지만 시간과 비용은?



# 통계의 기본 개념들 : 모집단과 표본 (2)

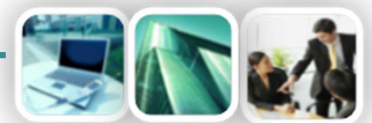
## ❖ 표본 (Sample)

- 모수의 값을 알아내기 위해 추출된 모집단의 일부분
- 일반적인 통계분석에서는 모집단의 일부만 조사하여, 모집단의 모수를 추측!
- 100% 정확하지는 않지만, 거의 근접한 값을 제공해 줄 수 있음
- 대신 시간과 비용을 크게 절약할 수 있는 방법



## ❖ 표본 추출

- 표본이 모집단을 대표할 수 있도록 추출하는 것이 중요
- 표본추출이 잘못되어 발생한 해프닝 사례  
: 미국의 1936년 대선 → 민주당 Franklin Roosevelt vs. 공화당 Alf Landon



# 통계의 2가지 분야

## ❖ 기술통계학

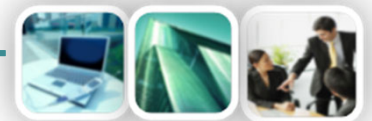
수집된 자료들이 가지고 있는 정보를 손쉽게 파악할 수 있도록 표현하는 방법에 대한 내용을 다루는 학문분야

- 통계조사를 마치고 나면, 대부분 많은 양의 정보가 발생  
→ 수집 자료들이 가지고 있는 정보를 쉽게 파악할 수 있도록 하는 방안 필요
- 그래프를 이용하거나, 평균치 등의 대표값을 구하여 제공하는 방법 등을 포함

## ❖ 추리통계학

수집된 자료들을 분석해 볼 때, 알아내고자 했던 모집단의 모수는 얼마로 추정되는지, 미래의 예측치는 얼마라고 예상되는지 등의 결론을 도출하는 학문 분야

- 수집된 자료를 통하여 원래 알아보고자 했던 문제에 대한 결론을 도출
- 통계학의 대부분 내용들이 여기에 해당



# 통계분석의 절차

## ❖ 1. 문제의 올바른 파악

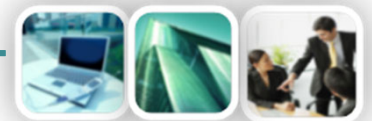
- 대상이 되는 모집단에 대해 정확하게 설정
- 알아내려는 사항이 무엇인지 제대로 파악

## ❖ 2. 표본의 추출과 자료의 수집

- 비표본오차(nonsampling error)가 발생하지 않도록 설계
  - 표본오차 : 모집단을 표본으로 추정하는데서 비롯되는 오차 (피할 수 없음)
  - 비표본오차 : 표본추출이 잘못할 경우, 발생하는 오차 (피할 수 있음)
- 자료 수집의 2가지 방법
  - 실험 (experiment) : 자연과학 분야
  - 설문 (survey) : 사회과학 분야

## ❖ 3. 수집된 자료의 적절한 정리 (전처리) : 기술통계학

## ❖ 4. 모집단에 대한 추론 : 추리통계학



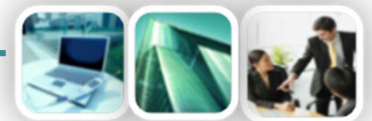
# 통계분석과 컴퓨터

## ❖ 통계분석과 컴퓨터는 불가분의 관계

- 통계는 많은 데이터를 수집, 정리, 분석해야 함 → **컴퓨터 활용이 필수**
- 많은 통계분석용 컴퓨터 패키지 존재

## ❖ 대표적인 통계 패키지

- **SAS (Statistical Analysis System)** : 공학자들에게 선호 / 방대한 자료 처리
  - “SAS라는 통계상자”, 김충련, 데이터리서치, 1993
- **SPSS (Statistical Package for the Social Science)** : 사회과학자들이 선호
  - 우수하고 편리한 Interface로 유명한 패키지
  - 본 실습은 SPSS를 중심으로 강의를 전개
  - “SPSS17.0 사회과학 통계분석”, 강병서, 김계수 공저, 한나래 아카데미, 2009
- Minitab
- IBM Scientific Subroutine Packages
- eViews, Statistica 등





# SPSS Statistics의 기본 구성 (1)

## ❖ SPSS Statistics의 기본 화면 구성 : 데이터 편집기

1. 메뉴바 및 도구모음

2. 데이터 편집기

3. 데이터 보기 / 변수 보기



# SPSS Statistics의 기본 구성 (2)

## ❖ SPSS Statistics의 기본 화면 구성 : 변수 편집기

The screenshot shows the SPSS Statistics Data Editor window with the Variable View tab selected. The window title is "disabled.sav [데이터집합1] - PASW Statistics Data Editor". The menu bar includes File (F), Edit (E), View (V), Data (D), Transform (T), Analyze (A), Direct Marketing (M), Graphs (G), Utilities (U), Window (W), and Help (H). The toolbar contains icons for file operations, editing, viewing, data management, analysis, and help. The variable list table is as follows:

	이름	유형	너비	소수점아...	설명	값	측정값	일	맞춤	측도	역할
1	id	숫자	4	0	Serial ID...	없음	없음	8	오른쪽	알 수 없음	입력
2	marital	숫자	1	0	결혼상태	{1, 미혼}...	없음	8	오른쪽	알 수 없음	입력
3	susfamno	숫자	2	0	부양가족수(증...	없음	99	8	오른쪽	알 수 없음	입력
4	jobsat1	숫자	1	0	직무만족도1(R)	{1, 매우그렇...	9	8	오른쪽	알 수 없음	입력
5	gender	숫자	1	0	성별	{1, 남성}...	9	8	오른쪽	알 수 없음	입력
6	age	숫자	2	0	(만)연령	없음	99	8	오른쪽	알 수 없음	입력
7	educ	숫자	1	0	최종학력	{1, 무학}...	9	8	오른쪽	알 수 없음	입력
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											

변수 편집/입력기



# SPSS Statistics의 기본 구성 (3)

## ❖ SPSS Statistics의 기본 화면 구성 : SPSS Statistics Viewer

The screenshot shows the SPSS Statistics Viewer window with the following content:

- Left Panel:** A tree view showing the output structure: '출력결과' (Output) > '로그' (Log) > '기술통계' (Descriptives) > '제목' (Title) > '노트' (Note) > '활성 데이터 집합' (Active Data Set) > '기술통계량' (Descriptive Statistics).
- Main Panel:**
  - GET**  
FILE='C:\Documents and Settings\hcahn\My Documents\강의1-경영정보처리\bipl4\disabled.sav'.  
DATASET NAME 데이터집합1 WINDOW=FRONT.  
DESCRIPTIVES VARIABLES=marital gender  
/STATISTICS=MEAN STDDEV MIN MAX.
  - 기술통계**  
[데이터집합1] C:\Documents and Settings\hcahn\My Documents\강의1-경영정보처리\bipl4\disabled.sav
  - 기술통계량**

	N	최소값	최대값	평균	표준편차
결혼상태	100	1	3	1.56	.538
성별	100	1	2	1.66	.476
유효수 (목록별)					

PASW Statistics 프로세서 준비 완료

**SPSS Statistics는 데이터와 결과를 별도의 파일로 관리**

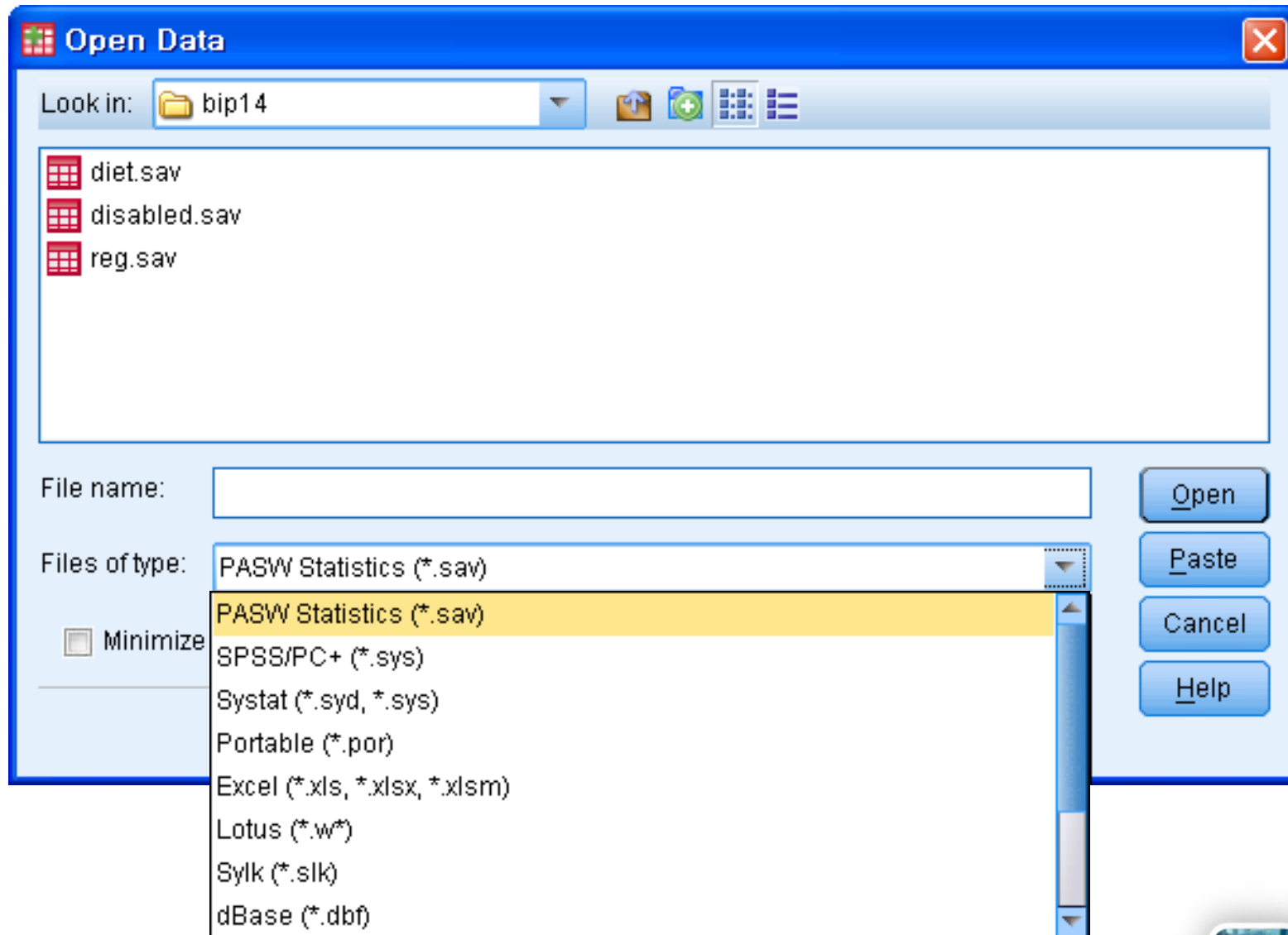
데이터 파일 : \*.sav

결과 파일 : \*.spv 또는 \*.spo (구버전)



# 자료의 입력 (1)

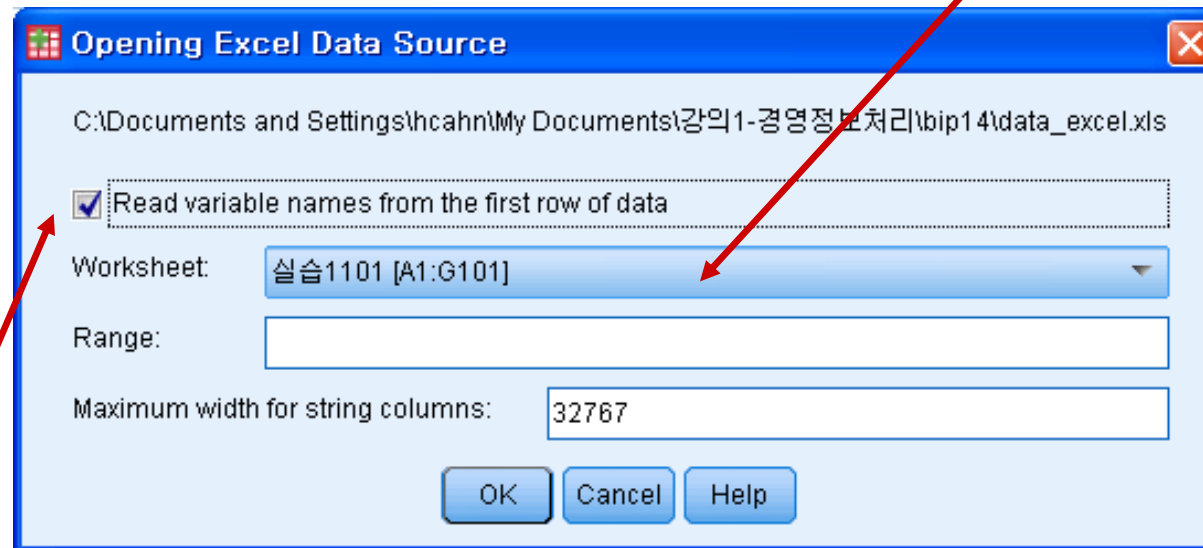
## ❖ 다른 유형의 자료를 입력 받기



# 자료의 입력 (2)

## ❖ 다른 유형의 자료를 입력 받기 : EXCEL 파일

**EXCEL의 어떤 워크시트를 읽어들이는 것인가?**



**EXCEL 워크시트의 맨 앞줄을 변수명으로 입력받을 것인가?**



# 자료의 입력 (3)

## ❖ 다른 유형의 자료를 입력 받기 : Text 파일 (1)

**Text Import Wizard - Step 1 of 6**

Welcome to the text import wizard!

This wizard will help you read data from your text file and specify information about the variables.

Does your text file match a predefined format?

☐ Yes ☒ No

Browse...

Text file: C:\Documents and Settings\hcahn\My Documents\강의1-경영정보처리\bip14\data\_text

	0	10	20	30	40	50	60	70
1	X만족도	Y만족도	Z만족도	W만족도	GENDER	AGE	INCOME	
2	4	3	3	4	1	20	12	
3	3	3	3	3	1	27	15	
4	3	3	3	3	1	27	15	
5	3	3		5	2	26	10	
6				3	2	24	20	

< Back Next > Finish Cancel Help



# 자료의 입력 (4)

## ❖ 다른 유형의 자료를 입력 받기 : Text 파일 (2)

**Text Import Wizard - Step 2 of 6**

How are your variables arranged?

☒ **D**elimited - Variables are delimited by a specific character (i.e., comma, tab).

☐ **F**ixed width - Variables are aligned in fixed width columns.

Are variable names included at the top of your file?

☒ **Y**es

☐ **N**o

Text file: C:\Documents and Settings\hcahn\My Documents\강의1-경영정보처리\bip14\data\_text

0 10 20 30 40 50 60 70

1	4	3	3	4	1	20	12
2	3	3	3	3	1	27	15
3	3	3	3	3	1	27	15
4	3	3		5	2	26	10
5				3	2	24	20

< Back Next > Finish Cancel Help



# 자료의 입력 (5)

## ❖ 다른 유형의 자료를 입력 받기 : Text 파일 (3)

**Text Import Wizard - Delimited Step 3 of 6**

The first case of data begins on which line number?

How are your cases represented?

☒ Each line represents a case

☐ A specific number of variables represents a case:

How many cases do you want to import?

☒ All of the cases

☐ The first  cases.

☐ A random percentage of the cases (approximate):  %

Data preview

	0	10	20	30	40	50	60	70
1	4	3	3	4	1	20	12	
2	3	3	3	3	1	27	15	
3	3	3	3	3	1	27	15	
4	3	3		5	2	26	10	

< Back Next > Finish Cancel Help





# 자료의 입력 (6)

## ❖ 다른 유형의 자료를 입력 받기 : Text 파일 (4)

Text Import Wizard - Delimited Step 4 of 6

Which delimiters appear between variables?

☒ Tab ☐ Space ☐ Comma ☐ Semicolon ☐ Other:

What is the text qualifier?

☒ None ☐ Single quote ☐ Double quote ☐ Other:

Data preview

X만족도	Y만족도	Z만족도	W만족도	GENDER	AGE	INCC
4	3	3	4	1	20	12
3	3	3	3	1	27	15
3	3	3	3	1	27	15
3	3		5	2	26	10
			3	2	24	20
3	3	4	4	2	29	25
3	1	1	4	2	22	10
			3	2	26	50
2	2	2	4	2	25	10

< Back Next > Finish Cancel Help



# 자료의 입력 (7)

## ❖ 다른 유형의 자료를 입력 받기 : Text 파일 (5)

Text Import Wizard - Step 5 of 6

Specifications for variable(s) selected in the data preview

Variable name:  Original Name: X만족도

Data format:

Data preview

X만족도	Y만족도	Z만족도	W만족도	GENDER	AGE	INCC
4	3	3	4	1	20	12
3	3	3	3	1	27	15
3	3	3	3	1	27	15
3	3		5	2	26	10
			3	2	24	20

< Back Next > Finish Cancel Help



# 자료의 입력 (8)

## ❖ 다른 유형의 자료를 입력 받기 : Text 파일 (6)

**Text Import Wizard - Step 6 of 6**

You have successfully defined the format of your text file.

Would you like to save this file format for future use?

☐ Yes ☐ No [Save As...](#)

Would you like to paste the syntax?

☐ Yes ☐ No ☒ [Cache data locally](#)

Press the Finish button to complete the text import wizard.

**Data preview**

X만족도	Y만족도	Z만족도	W만족도	GENDER	AGE	INCC
4	3	3	4	1	20	12
3	3	3	3	1	27	15
3	3	3	3	1	27	15
3	3		5	2	26	10
			3	2	24	20
3	3	4	4	2	29	25
3	1	1	4	2	22	10

< Back Next > Finish Cancel Help

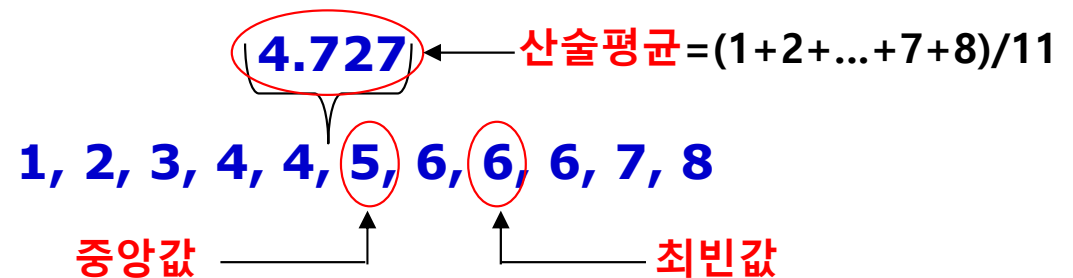


# 기술통계 실습 (1)

## ❖ 빈도분석

- 남자가 더 많을까? 여자가 더 많을까?
- **Analysis > Descriptive Statistics > Frequencies**
- 'Statistics' 버튼을 누르면, 다음과 같은 추가적인 통계정보 파악 가능
  - Percentile Values (백분위수) : 상위 x%의 기준이 되는 값이 어디인가?
  - Quartile (사분위수) : 상위 25%, 50%, 75%의 기준이 되는 값이 어디인가?

- Mean (산술평균)
- Median (중앙값)
- Mode (최빈값)



- Variance (분산) / Std. deviation (표준편차) / Kurtosis (첨도) / Skewness (왜도)  
: 다음 슬라이드에 상세 설명
- 'Charts' 버튼을 누르면, Histogram 등을 그릴 수 있음

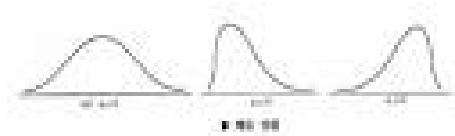


# 기술통계 실습 (2)

## ❖ 왜도와 첨도

- 왜도(Skewness) : 분포의 편중(치우침)을 나타내는 지표

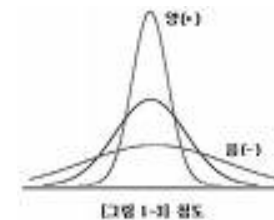
$$skewness = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^3}{(N-1)s^3}$$



- A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of zero. A distribution with a significant **positive skewness** has a **long right tail**. A distribution with a significant **negative skewness** has a **long left tail**. As a rough guide, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

- 첨도(Kurtosis) : 분포의 뾰족한 정도를 나타내는 지표

$$kurtosis = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^4}{(N-1)s^4}$$



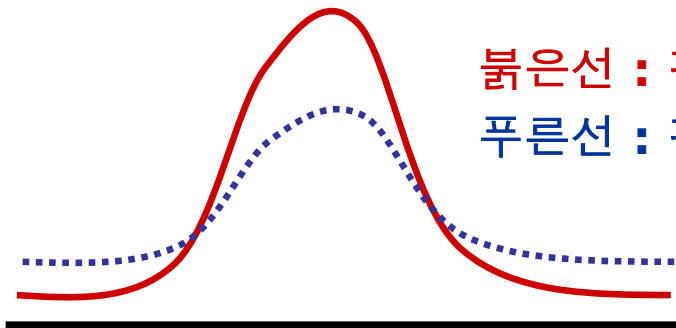
- A measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is 0. **Negative kurtosis** indicates that the observations cluster more and have **longer tails** than those in the normal distribution and **positive kurtosis** indicates the observations cluster less and have **shorter tails**.



# 기술통계 실습 (3)

## ❖ 기술통계량

- 응답자 중 용돈이 가장 높은(혹은 낮은) 사람의 소득은 얼마일까?
- 응답자 나이의 평균은 몇 살일까?
- X상품 선호도에 대한 관측치들은 평균을 중심으로 얼마나 퍼져있을까?  
→ 분산 이나 표준편차를 통해 확인 가능

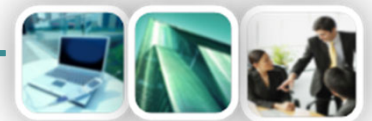


붉은선 : 평균 근처에 몰려있음 → 분산, 표준편차 작음  
푸른선 : 평균으로부터 떨어져 있음 → 분산, 표준편차 큼

$$S = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n-1}}$$

$$V = S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$$

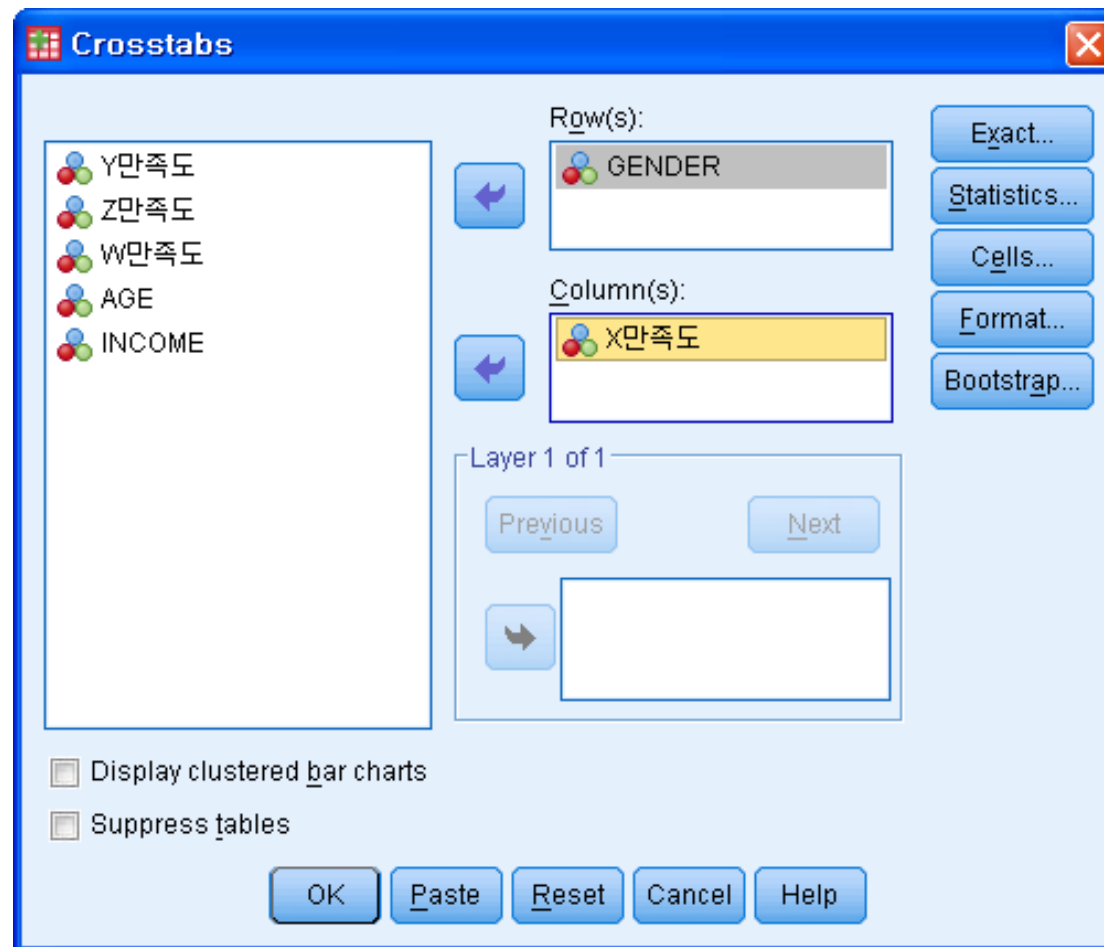
- Analysis > Descriptive Statistics > Descriptives



# 기술통계 실습 (4)

## ❖ 교차표

- 응답자의 성별에 따른 제품X에 대한 평가결과는 어떻게 달라지는가?
- **Analysis > Descriptive Statistics > Crosstabs**



# 카이제곱 검정: 교차분석 (1)

## ❖ 언제 쓰는가?

- 다음의 예를 보자.

(**disabled.sav** : 장애인복지 종사자들의 근무실태에 대한 설문조사)

결혼상태 \* 성별 Crosstabulation

			성별		Total
			남성	여성	
결혼 상태	미혼	Count	6	40	46
		% within 결혼상태	13.0%	87.0%	100.0%
		% within 성별	17.6%	60.6%	46.0%
		% of Total	6.0%	40.0%	46.0%
	기혼	Count	28	24	52
		% within 결혼상태	53.8%	46.2%	100.0%
		% within 성별	82.4%	36.4%	52.0%
		% of Total	28.0%	24.0%	52.0%
	기타	Count		2	2
		% within 결혼상태		100.0%	100.0%
		% within 성별		3.0%	2.0%
		% of Total		2.0%	2.0%
Total	Count	34	66	100	
	% within 결혼상태	34.0%	66.0%	100.0%	
	% within 성별	100.0%	100.0%	100.0%	
	% of Total	34.0%	66.0%	100.0%	

- “결혼 상태가 성별에 따라 차이가 있을까?”  
→ 차이가 있다고 보여질 경우, 과연 그 차이가 통계적으로 유의한가?





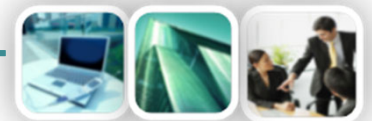
# 카이제곱 검정: 교차분석 (2)

## ❖ 언제 쓰는가?

이산형 변수로 구성된 2개의 변수가 서로 독립인지 아닌지를  
확인하고자 할 때 사용한다.

즉, 두 이산형 변수가 서로 상관관계가 있는지, 없는지를 통계적으로 검정할  
때 사용되는 기법이 카이제곱 검정(교차분석)이라고 할 수 있다.

- **이산형(discrete) 변수** : 비계량적인 변수로서, 관측대상을 범주로 나누어 분류한 후, 그에 따라 기호나 숫자를 부여한 변수. 여기서는 “평균”의 의미가 없다!  
(예) 성별, 선호정당, 출신국가, 직업구분, 주택보유 여부 등
- **연속형(continuous) 변수** : 계량적인 변수로서, 일반적인 수치적 방법에 의해 측정 가능한 변수. 여기서는 “평균”이 의미가 있다!  
(예) 금액, 거리, 무게, 시간 등



# 카이제곱 검정: 교차분석 (3)

## ❖ 척도의 4가지 종류

### ● 명목척도 (Nominal Scales)

- 정의 : 대상의 특성을 분류하거나 확인할 목적으로 사용하는 척도
- 예시 : 축구선수 등번호, 입시에 사용되는 출신학교 코드
- 분석방법 : 최빈값(Mode), 교차분석(Cross Tabulation) 등

### ● 서열척도 (Ordinal Scales)

- 정의 : 측정대상간의 순서를 밝히기 위해 사용하는 척도 (양적 비교 불가)
- 예시 : 수학시험 석차, 선호도 우선순위
- 분석방법 : 중간값(Median), 서열상관관계, 비모수통계기법 등

### ● 등간척도 (Interval Scales)

- 정의 : 부여된 순위 사이의 간격이 동일한 척도
- 예시 : 온도
- 분석방법 : 평균값(Mean), 상관계수, 표준편차, 모수통계기법 등

### ● 비율척도 (Ratio Scales)

- 정의 : 순위 사이의 간격이 동일하고, 비율계산이 가능한 척도
- 예시 : 무게, 투표율
- 분석방법 : 모든 모수통계기법

“이산형 변수”  
로 통칭

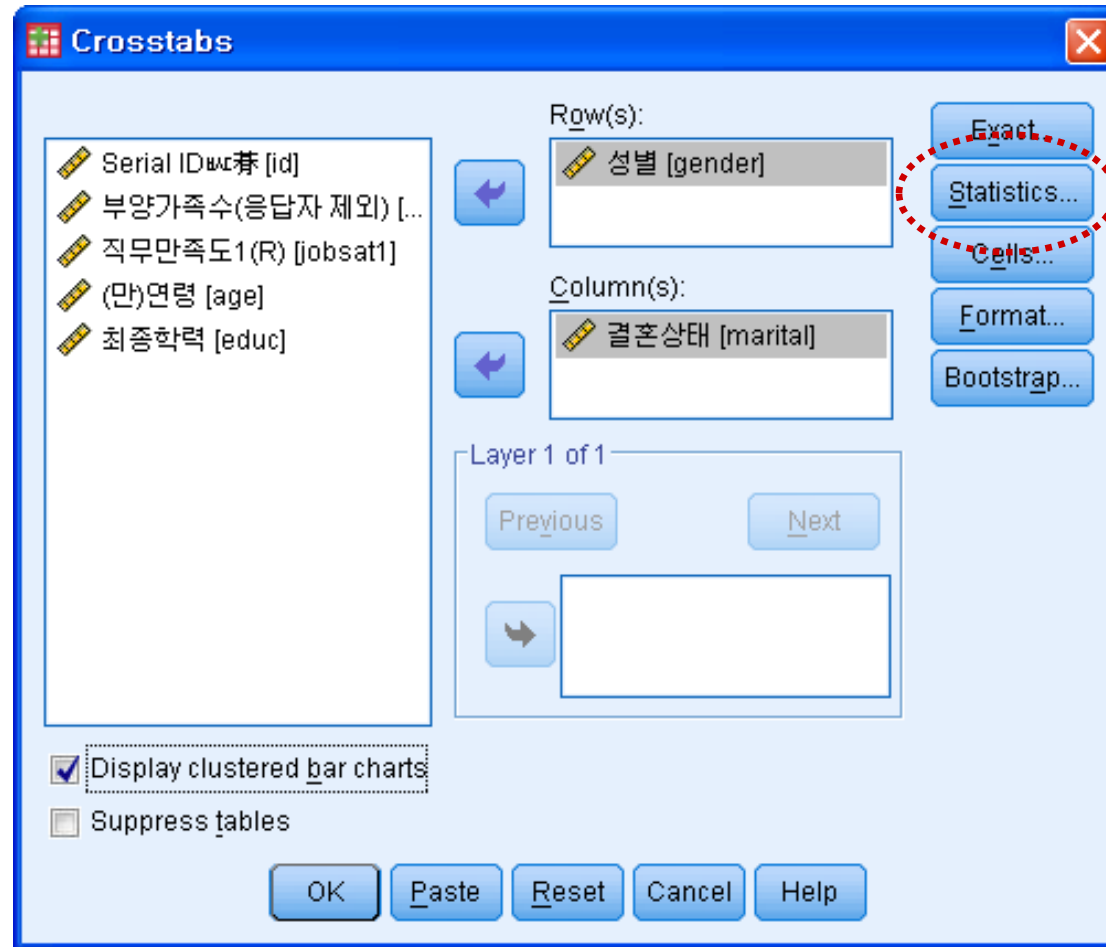
“연속형 변수”  
로 통칭



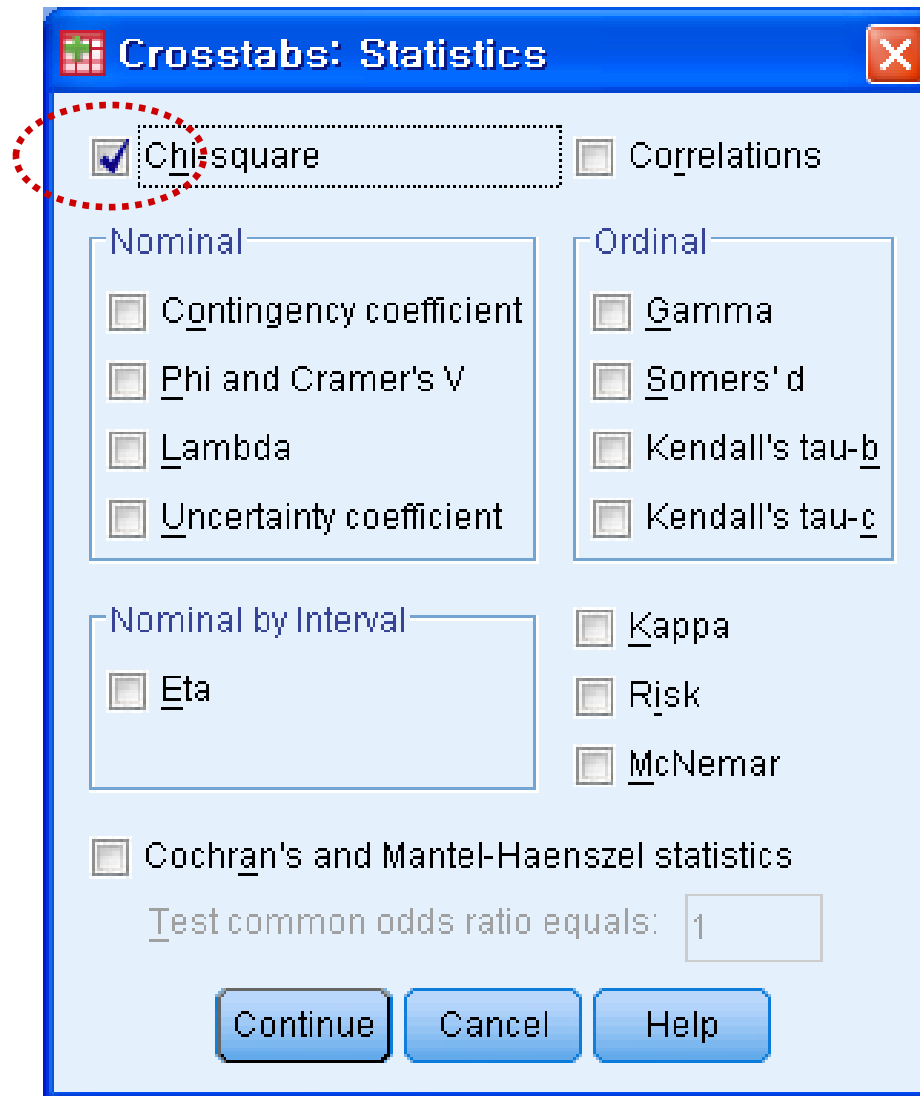
# 카이제곱 검정: 교차분석 (4)

## ❖ 적용 방법

- Analysis > Descriptive Statistics > Crosstabs



# 카이제곱 검정: 교차분석 (5)



# 카이제곱 검정: 교차분석 (6)

## ❖ 결과 해석

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	19.160 <sup>a</sup>	2	.000
Likelihood Ratio	20.804	2	.000
Linear-by-Linear Association	12.367	1	.000
N of Valid Cases	100		

a. 2 cells (33.3%) have expected count less than 5. The minimum expected count is .68.

- 우리는 **Asy. Sig.** (점근 유의확률, **p-value**라고도 함)을 토대로 평가한다.
- “95% 신뢰수준 하에서” 로 해석하고 싶으면, 유의확률 < 0.05 인지 확인
- “90% 신뢰수준 하에서”로 해석하고 싶으면, 유의확률 < 0.1 인지 확인  
(즉, 유의확률은 작으면 작을수록 좋다!)



# 가설검정의 기초

## ❖ 과학분야에서의 증명: 반증법에 의거해 증명

- “모든 사람은 정직하다”라는 명제가 있을 때, 이 명제가 참인지 거짓인지를 확인하는 접근법에는 2가지가 존재
  - (1) 모든 사람을 일일이 뒷조사해서, 정직한지 확인하는 방법
  - (2) 정직하지 못한 사람(사례)을 하나 찾아내, 명제가 거짓임을 입증하는 방법
- 위 2가지 방법 중 어느 것이 효율적일까? (현실적으로 가능한 방법일까?)
  - 1,000명이 정직함을 확인했을 때, 1,001명째 사람이 정직하다고 확언할 수 있을까?
- 귀무가설( $H_0$ )과 대립가설( $H_a$ )
  - 우리가 알고 싶은 명제(주로 A와 B는 다르다, A와 B는 차이가 있다 등)는 대립가설로 두고, 그 반대인 귀무가설(주로 A와 B는 같다, A와 B는 차이가 없다 등)이 기각됨을 보임으로서 내가 입증해 보이고 싶은 명제를 증명

## ● 대표적인 귀무가설과 대립가설

$$H_0 : \rho_{AB} = 0$$

$$H_a : \rho_{AB} \neq 0$$

[카이제곱검정]

$$H_0 : \mu_A = \mu_B$$

$$H_a : \mu_A \neq \mu_B$$

[t-검정]

$$H_0 : \sigma_A = \sigma_B$$

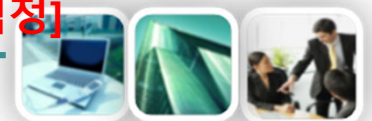
$$H_a : \sigma_A \neq \sigma_B$$

[F-검정(등분산검정)]

$$H_0 : \mu_A = \mu_B = \mu_C$$

$$H_a : \mu_A \neq \mu_B \neq \mu_C$$

[ANOVA 검정]



# 독립표본 t-검정 (1)

## ❖ 언제 쓰는가?

- 전체 응답자 중 남자와 여자 사이의 연령은 차이가 있는가?

Group Statistics

		N	Mean	Std. Deviation	Std. Error Mean
(만)연령	성별 남성	34	37.68	9.929	1.703
	여성	66	28.61	6.346	.781

서로 독립된 두 집단간의 평균의 차이가 통계적으로 유의미한지  
비교하고자 할 때 사용한다.

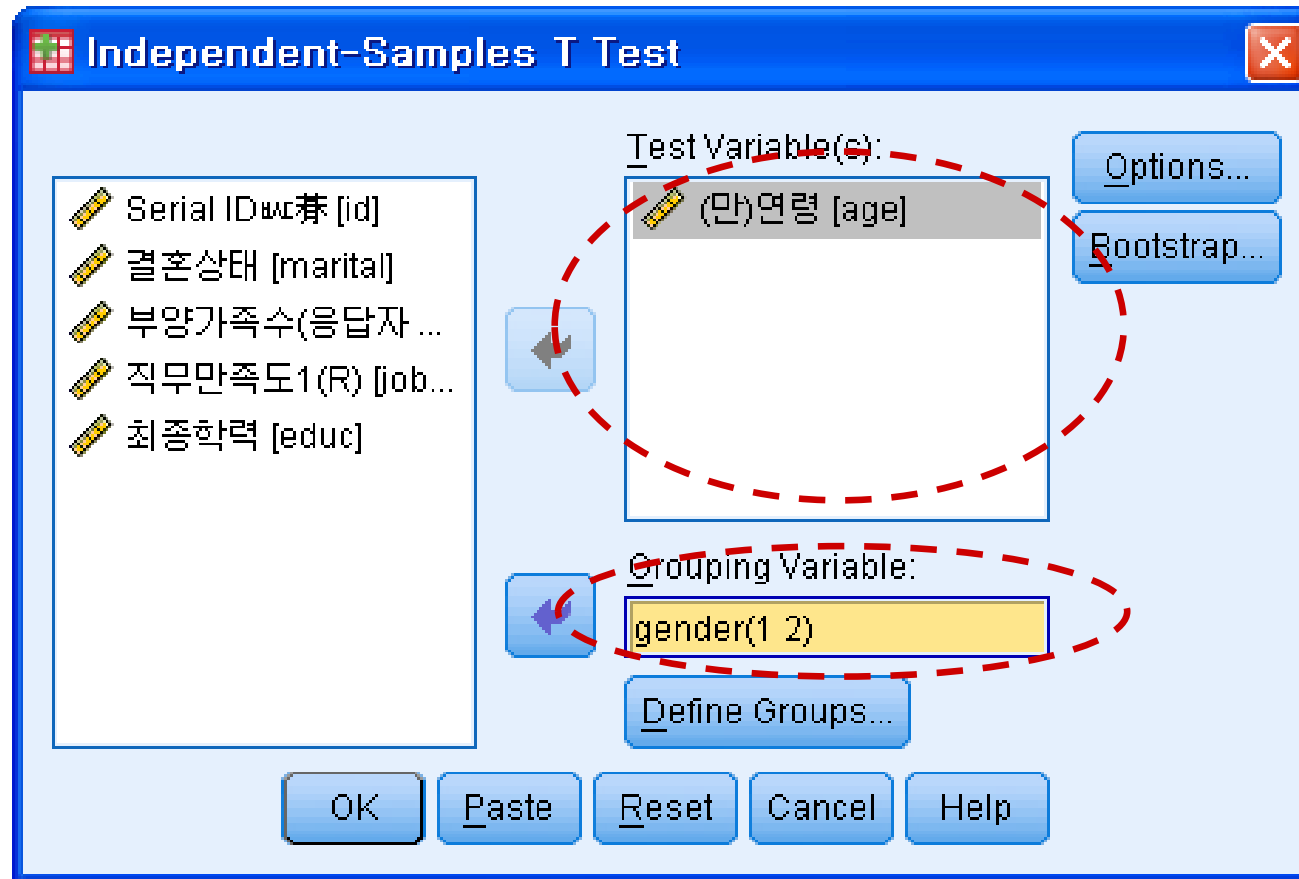
즉, 서로 독립된 **두 집단**에 대해 각 집단별 **특정 연속형 변수** 평균값이  
서로 차이가 있는지, 없는지를 통계적으로 검정할 때 사용되는 기법이  
독립표본 t-검정이다.



# 독립표본 t-검정 (2)

## ❖ 적용 방법

- Analysis > Compare Means > Independent-Samples t-test





# 독립표본 t-검정 (3)

## ❖ 결과 해석

- 1단계. “분산이 같다”는 가정이 제대로 성립되는지 확인
- 2단계. “두 집단간 평균이 같다”는 가설이 맞는지 확인

Independent Samples Test									
	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
(만)연령	7.049	.009	5.551	98	.000	9.07	1.634	5.828	12.313
Equal variances assumed									
Equal variances not assumed			4.842	47.287	.000	9.07	1.873	5.302	12.839

- 위와 같은 경우, 분산은 같다는 가정이 깨지게 되며, 따라서 아래의 값들을 보고 판단해야 함
- 결과적으로 두 집단간 연령의 평균은 통계적으로 유의하게 차이가 난다고 할 수 있음 (95% 신뢰수준 하에서)



# 대응표본 t-검정 (1)

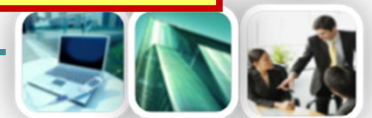
## ❖ 언제 쓰는가?

- 앞의 독립표본 t-검정의 경우, “남성”집단과 “여성”집단은 서로 공통요소가 없는 ‘독립된 집단’임
- 하지만, 다음과 같은 경우는 어떤가? (**diet.sav**)
  - 한 회사에서 자사가 개발한 한 달간의 식이요법 프로그램이 효과가 있는지 여부를 분석하고자 함
  - 식이요법에 참가한 10명의 몸무게가 Before / After에서 다음과 같이 변화하였음

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
요법전	10	54	82	66.50	9.801
요법 후	10	50	76	63.50	9.312
Valid N (listwise)	10				

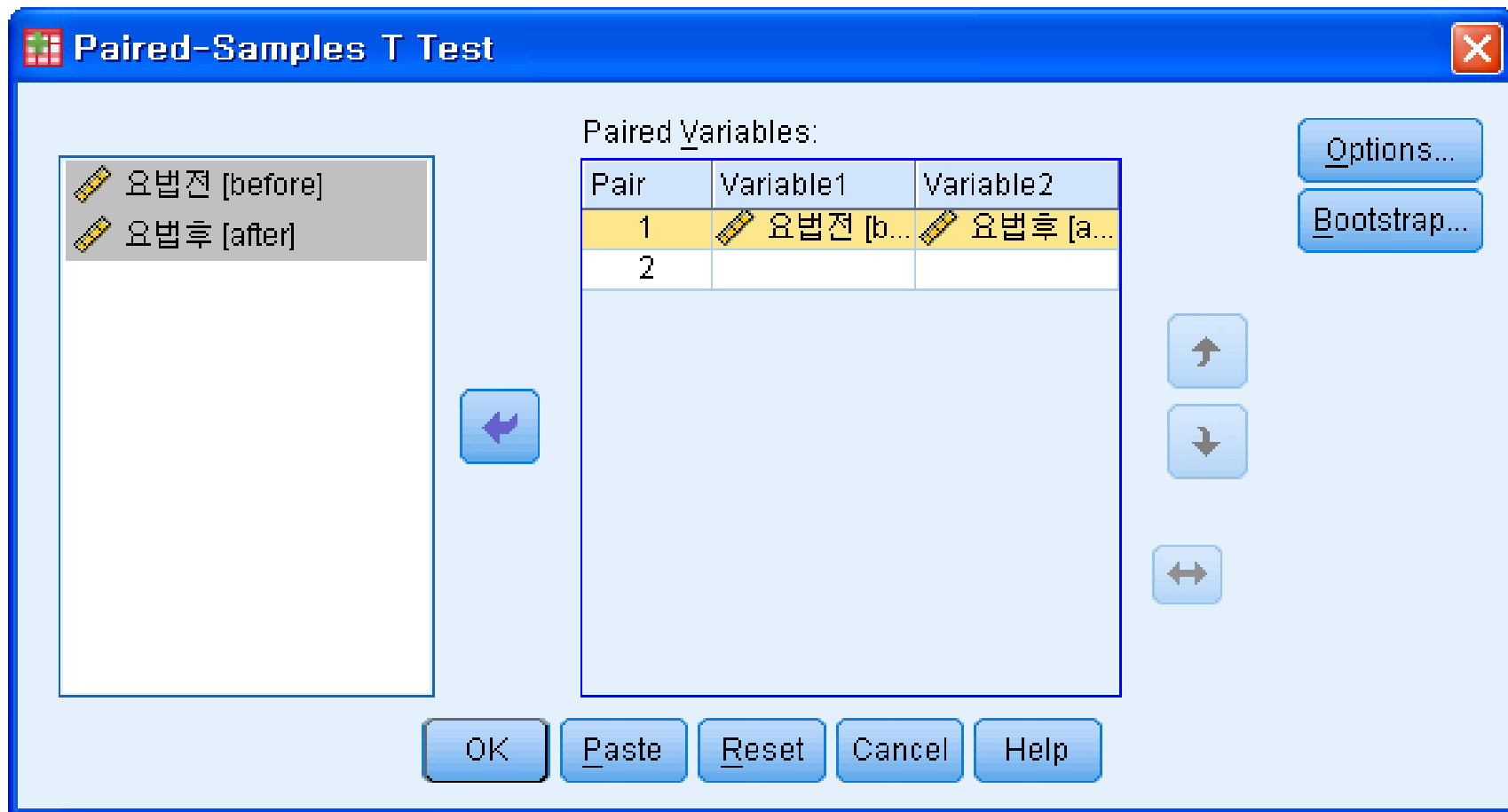
위 예와 같이, 서로 동일한 모집단에서 추출된 **두 표본**에 대해  
**특정 연속형 변수** 평균값이 서로 차이가 있는지, 없는지를 통계적으로  
검정할 때 사용되는 기법이 **대응표본 t-검정(Paired-samples t-test)**이다.



# 대응표본 t-검정 (2)

## ❖ 적용 방법

- Analysis > Compare Means > Paired-Samples t-test



# 대응표본 t-검정 (3)

## ❖ 결과 해석

- “두 집단간 평균이 같다”는 가설이 맞는지 확인

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1    요법전 - 요법후	3.000	3.232	1.022	.688	5.312	2.935	9	.017

- p-value가 0.017이므로, 95% 신뢰수준 하에서 두 집단간 평균은 통계적으로 유의하게 차이가 난다고 할 수 있음
  - 따라서, 통계적으로 볼 때, 이 회사의 식이요법 프로그램은 효과가 있다고 잠정적으로 결론지을 수 있음



# 일원배치 분산분석: One-way ANOVA (1)

## ❖ 언제 쓰는가?

- 학력수준에 따라 직무만족도의 수준은 차이가 있는가?  
(현재 학력수준은 8개로 나누어져 있음)

Descriptives

직무만족도1(R)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
고등학교 이하	34	3.38	.985	.169	3.04	3.73	1	5
초급/전문/대학교	55	3.36	.950	.128	3.11	3.62	2	5
대학원이상	11	3.45	1.214	.366	2.64	4.27	1	5
Total	100	3.38	.982	.098	3.19	3.57	1	5

세 개 이상의 집단간의 평균의 차이가 통계적으로 유의미한지  
비교하고자 할 때 사용한다.

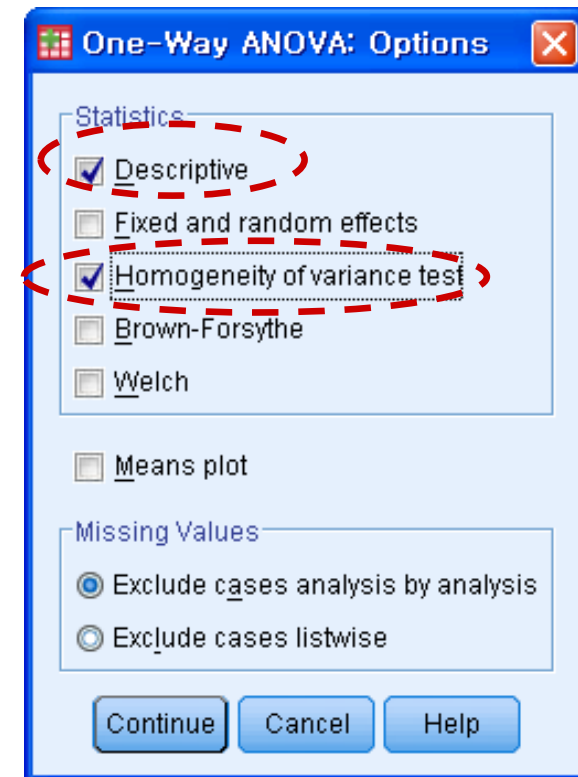
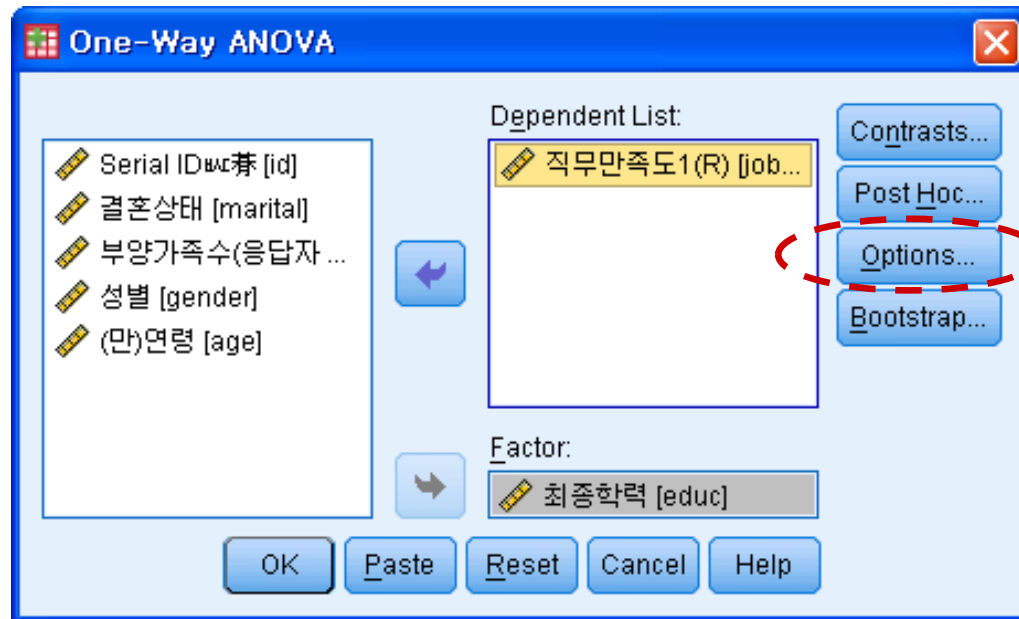
즉, 셋 이상의 집단에 대해 각 집단별 특정 비율변수 평균값이 서로 차이가 있는지,  
없는지를 통계적으로 검정할 때 사용되는 기법이 ANOVA다.



# 일원배치 분산분석: One-way ANOVA (2)

## ❖ 적용 방법

- Analysis > Compare Means > One-way ANOVA



# 일원배치 분산분석: One-way ANOVA (3)

## ❖ 결과 해석

### Test of Homogeneity of Variances

직무만족도1(R)

Levene Statistic	df1	df2	Sig.
.327	2	97	.722

### ANOVA

직무만족도1(R)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.076	2	.038	.039	.962
Within Groups	95.484	97	.984		
Total	95.560	99			

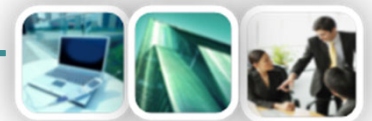
- 분산이 같다는 가정에 대해서는 유의확률 $>0.05$ 이면, 신뢰수준 95%하에서 각 집단의 분산은 같음
- 유의확률 $>0.05$  이면, 신뢰수준 95%하에서 각 집단간 직무만족도는 같음 (평균 간 통계적으로 유의한 차이가 없음)



# 가설검정의 정리

- ❖ 분석의 대상이 되는 두 변수의 특징에 따라, 다음과 같이 다른 검정기법을 적용해야 함

기법	대상변수A	대상변수B	적용 예
카이제곱검정	이산형	이산형	성별과 결혼유무 사이에 유의한 관계가 있는가?
독립표본t검정	이산형 (2그룹/독립)	연속형	성별에 따른 평균 취업율의 차이가 있는가?
대응표본t검정	이산형 (2그룹/Pair)	연속형	보충수업 후, 3학년 1반 수학성적의 향상이 있는가?
일원배치 분산분석	이산형 (3그룹 이상)	연속형	거주지역에 따른 평균 소득액의 차이가 있는가?





## 개인과제 01:



다이어트 인터넷 쇼핑몰 데이터셋에 대해 전처리한 뒤,  
종속변수(특정상품 구매여부)와 95% 신뢰수준 하에서  
통계적으로 유의한 독립변수들의 리스트를 산출하시오.

제출물: 전처리 완료된 데이터 파일 및 통계분석 결과(변수 리스트)

제출기한: Week04 강의일 밤(00:00)까지