Ⅰ. 산업보건의 통계(통계학의 기본지식)

오 은 하 Ph.D.

• 자료의 요약

- 대표값 : 평균, 중앙값(중위수), 최빈값 등
- 변동 : 범위, 분산, 표준편차 등
- 변동계수

• 산술평균(평균, mean)

- 평균은 대개 산술평균을 의미함. 즉 측정값들의 합을 측정값의 개수로 나눈 것을 의미

- $\mathrm{X} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{X}_i$

•중앙값(median)

- 관찰값의 가운데수를 의미
- 관찰값이 홀수개인 경우 : 순서대로 되어 있는 관찰값의 $\frac{n+1}{2}$ 번째 값
- 관찰값이 짝수개인 경우 :

$$\frac{1}{2}\left[\left(\frac{n}{2}\right)\text{번째 값} + \left(\frac{n}{2}+1\right)\text{번째 값}\right]$$

• 최빈값(Mode)

- 빈도가 가장 많이 나타나는 값
- 여러 개일 수도, 없을 수도 있다.

• 분포의 형태와 평균의 위치

- 대칭분포 : 평균, 중앙값, 최빈값이 모두 같은 위치
- 치우친 분포 : 평균이 가장 큰 영향을 받음. 중앙값은 평균값과 최빈값사이

1) 좌우 대칭형태
2) 오른쪽으로 치우친 형태
3) 왼쪽으로 치우친 형태

• 범위(range, R)

- R = 최대값 - 최소값
- 자료의 두 값만을 사용
- 너무 크게 되고, 극단값에 영향을 많이 받음.

• **분산(variance, var)**

- 관찰값들이 평균으로부터 얼마나 퍼져 있는가를 측정하는 것
- 표본집단의 경우, 제곱한 값들의 평균을 계산할 때 n보다는 n-1의 값으로 나누어 주게 됨. 이것은 전체 모집단의 분산을 추정하는데 있어서 좀더 표율적인 값이 계산됨
- $s^2 = \frac{1}{n-1}\sum(X-\bar{X})^2$
- 분산의 단위는 관찰값 단위의 제곱이 됨

• **자유도(degree of freedom)**

= 표본의 개수 - 1

= n - 1

- 자유도의 정의 : 주어진 통계량의 특정 체계내에서 임의로 결정될 수 있는 자료의 수

• **표준편차(standard deviation. S.D)**

- $s=\sqrt{\frac{1}{n-1}\sum(X-\bar{X})^2}$ 또는
- $s=\sqrt{\frac{1}{n-1}\left\{\sum X^2-\frac{1}{n}(\sum X)^2\right\}}$
- 자료와 같은 단위를 가짐
- 정규분포에 따르는 자료의 경우 관찰값의 약 70%가 평균을 중심으로 좌우 1 표준편차 내에 속하게 되고, 약 95%가 좌우 2 표준편차 내에 속하게 됨

• **변동계수(coefficient of variation, C.V)**

- $c.v=\frac{s}{\bar{X}}\times 100(\%)$
- 자료의 단위에 무관함
- 다른 단위로 측정된 자료의 비교에 유용

• **표본추출의 변동 및 표준오차(standard error, SE)**

- 표본은 모집단을 추정하기 위한 것
- 추출된 표본에는 변동이 존재
- 모집단의 평균을 추정하기 위해 서로 독립적인 표본집단을 추출해서 각각의 표본집단들의 평균을 구하고, 이 표본평균들의 분포를 만든다고 가정하면, 이 표본평균 분포의 평균값은 모집단의 평균과 같게 될 것이다. 또한 이 분포의 표준편차

를 실제 얻은 표본에서 구한 평균을 표준오차라 함.

• **정규분포(가우스 분포)의 특성**

- 종모양, 평균을 중심으로 좌우대칭
- 표준편차가 클수록(평균에서 멀어질수록) 높이는 낮아지고 완만해짐

• **표준정규분포**

- 평균은 0, 표준편차는 1인 정규분포, 즉(0,1)
- 변수가 정규적으로 분포되어 있다면 단위를 변화시켜도 분포모양에는 영향을 주지 않음
- 표준정규분포의 변환이란 각 측정값에서 평균을 빼고 표준편차로 나누어주는 것을 말함

$$SND: \ z=\frac{X-\mu}{\sigma} \quad N(0,1)$$

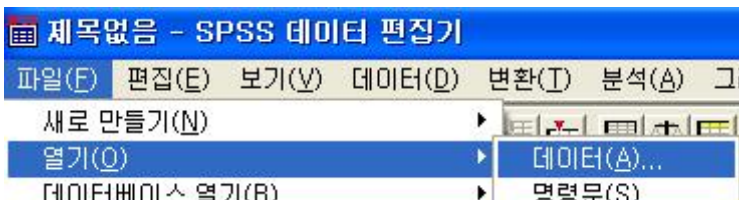
Ⅱ. SPSS를 이용한 통계 분석 방법

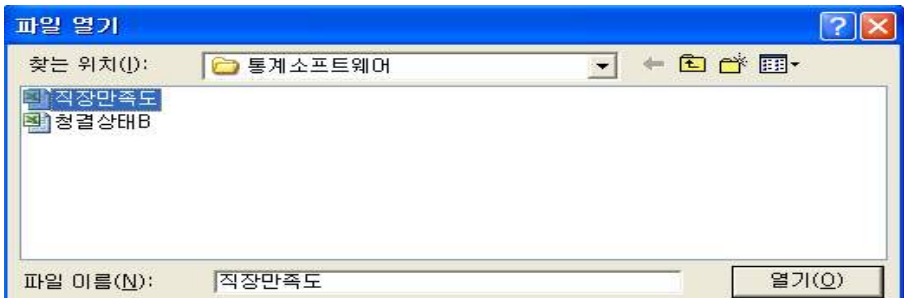
1. 엑셀파일 SPSS로 파일 불러오기.

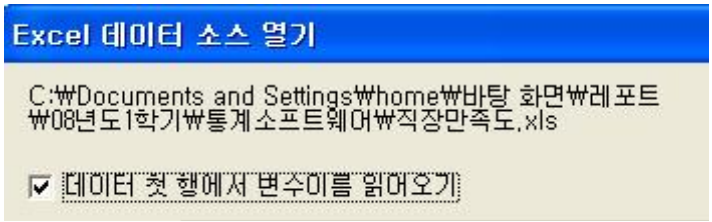
(엑셀파일)

	A	B	C	D	E	F	G	H	I	J	K
1	ID	SEX	AGE	EDU	STA	YEAR	S1	S2	S3	S4	S5
2	1	1	2	1	1	3	3	1	3	4	5
3	2	2	1	3	2	3	5	2	5	1	3
4	3	2	4	4	3	4	3	3	2	2	1
5	4	2	4	1	3	3	5	1	5	3	4

SPSS에서 파일 불러오기







파일 불러오기 할때 유의점은
엑셀파일에 변수값 유무 에 따라 SPSS에서 데이터 첫 행에서 변수이름 읽어오기
를 꼭 확인 해야 한다. 엑셀 파일에서 변수가 없을 경우에 위의 상자에 체크를 하게 되면 데이터의 자료가 완전히 바뀌므로 꼭 유의해야 하며 파일 불러 오기할 때 가장 중요한 작업이다.

	ID	SEX	AGE	EDU	STA	YEAR	S1	S2	S3	S4	S5	M1	M2	M3
1	1	1	2	1	1	3	3	1	3	4	5	3	4	3

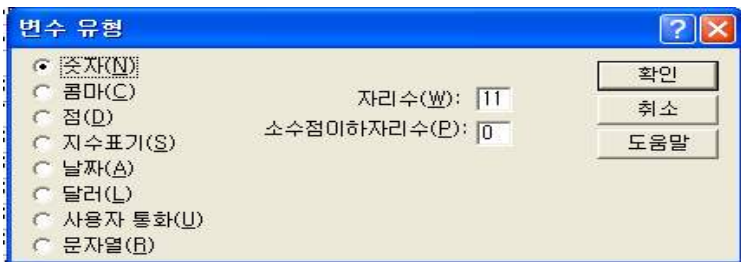
위의 그림을 보면 변수가 입력된 것을 확인 할 수 있다. 만약 엑셀파일에서 변수가 입력되지 않았을 경우 데이터 첫 행에서 변수이름 읽어오기를 체크 한다면 위의 데이터 들이 모두 변수가 되어서 데이터가 모두 바뀌게 된다. 그리고 또 한 가지 중요한 점은 자주 저장을 해주는 것이다. 시스템 오류나 개인의 실수로 인해 데이터가 모두 지워 질수도 있기 때문에 저장은 꼭 자주 해 주어야 한다.

2. 변수보기

엑셀에서 불러온 파일의 변수 보기 창.

	이름	유형	자리수	소수점이하자리	설명	값	결측값	열	맞춤	측도
1	ID	숫자	11	0		없음	없음	8	오른쪽	척도
2	SEX	숫자	11	0		없음	없음	8	오른쪽	척도
3	AGE	숫자	11	0		없음	없음	8	오른쪽	척도
4	EDU	숫자	11	0		없음	없음	8	오른쪽	척도
5	STA	숫자	11	0		없음	없음	8	오른쪽	척도

1) 이름 : 데이터에 있는 모든 변수들이 보이며 수정 이 가능하다.
2) 유형 : 변수 유형을 달리 지정하지 않으면 SPSS는 새로운 변수를 숫자변수로 간주 한다. 변수 유형을 바꾸려면 관련부분을 누르면 변수 유형 대상상자가 열린다.



- 숫자 : 변수값이 숫자인 경우
- 콤마 : 세자리마다 콤마를 표기
- 문자 : 문자변수를 표기하는 경우

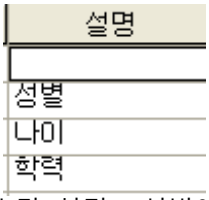
※ 변수의 값이 숫자인지 문자열인지 확인을 한 후 변수 유형에서 수정 해줘야지 변수

값을 입력 할 수 있다.

3) 자리수 : 입력할 자료의 자리 수 설정.

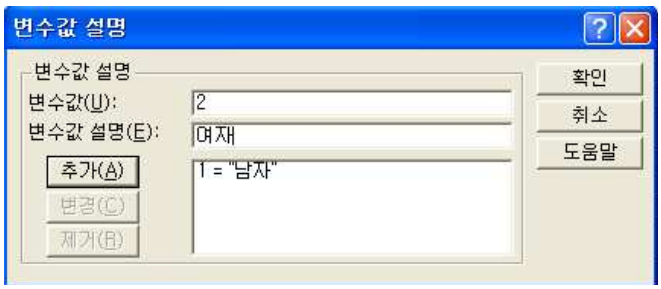
4) 소수점이하의 자리 : 입력할 자료의 소수점이하자리

5) 설명 : 변수에 대한 부연적인 설명을 입력한다.

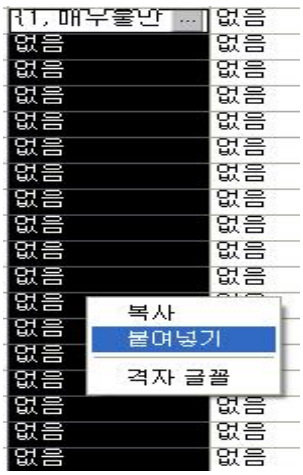


6) 변수값 설명 : 성별에 대한 변수값 설명을 입력 하는 방법이다.

성별에는 남자, 여자 가 있기 때문에 남자는 보통 1 여자는 2 로 입력을 해준다. 설문지나 조사지에서 간혹 여자를 1로 한 경우가 있을수도 있으므로 확인해보고 입력해주어야 한다.



그리고 변수값 설명중 반복되는 변수의 설명값을 입력할 경우 복사해서 붙여넣기 해주면 된다



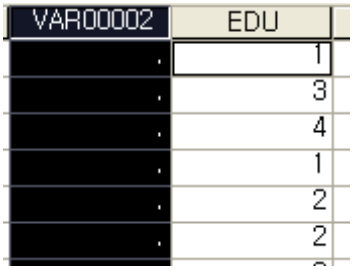
7) 결측값 : ① 시스템 결측값 : 수치변수에 공백문자가 할당된 경우 이를 시스템 결측값이라고 한다. 시스템 결측값은 점(.) 으로 표기된다.

② 시용자-결측값 : 데이터가 어떻게 누락되었는가를 나타내기 위해 사용되는 결측값으로 설문조사에서 '잘 모르겠음', '해당 없음', '응답 거부' 등의 사용자 결측값으로 지정한다.

3. 변수삽입

1) 기존의 입력된 변수들 사이에 새로운 변수 추가

변수 삽입 -> 삽입후 (삽입하는 변수의 왼쪽에 생긴다.)



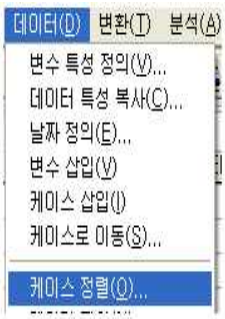
2) 케이스 삽입

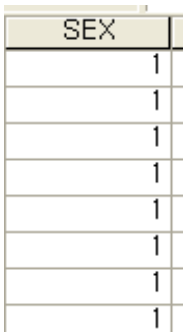
케이스 삽입 -> 삽입후 (삽입하는 변수의 위쪽에 생긴다.)

새롭게 생긴 변수 및 케이스는 모두 결측값으로 입력된다.

4. 케이스 정렬

성별 변수의 변수값 을 오름차순 으로 나타내는 방법





위의 그림을 보면 성별 1 은 남성을 나태낸다.

오름차순을 하여 변수값 1이 성별변수의 상위로 올라왔다.

내림차순을 할 경우에도 위의 케이스 정렬에서 내림차순을 선택해서 하면 된다.

5. 빈도 분석

빈도분석은 질적 자료를 대상으로 빈도와 비율을 계산할 때 쓰인다. 그리고 데이터에 질적 자료와 양적 자료가 많을 때 질적 자료를 대상으로 오류가 있는지 확인 할 수 있다.

데이터파일의 입력내용

문항	변수명	내용
일련번호	ID	세 자리 숫자
성별	SEX	1:남자, 2:여자
연령	AGE	1:29세 이하, 2:30~39 3:40~49세, 4:50세 이상
학력	EDU	1:고졸, 2:전문대졸, 3:대졸, 4:대학원졸
직위	STA	1:사원급, 2:주임급, 3:과장이상
근무년수	YEAR	1:5년이하, 2:6~20년, 3:11~15년, 4:16년이상
보수관련문항 5개	S1-S5	1:매우불만, 2:약간불만, 3:보통, 4:약간만족, 5:매우만족
상사관련문항 4개	M1-M4	1:매우불만, 2:약간불만, 3:보통, 4:약간만족, 5:매우만족
대인관계문항 5개	R1-R5	1:매우불만, 2:약간불만, 3:보통, 4:약간만족, 5:매우만족
업무관련문항 5개	J1-J5	1:매우불만, 2:약간불만, 3:보통, 4:약간만족, 5:매우만족
복지관련문항 4개	W1-W4	1:매우불만, 2:약간불만, 3:보통, 4:약간만족, 5:매우만족

위의 표는 직장만족도 데이터의 대한 사항이다. 위에서 질적 자료에 대한 빈도 분석과 부적절한 응답, 오류가 있는지 확인 해 보겠다.

위의 빈도분석 상자를 보면 변수에는 SEX EDU 등이 들어 가 있다 이들은 질적 자료를 대상으로 빈도분석 하는 것이기 때문에 질적 자료를 변수에 입력해 주어야 한다.

SEX

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	1	126	43.2	43.2	43.2
	2	166	56.8	56.8	100.0
	합계	292	100.0	100.0	

빈도 분석 결과 성별에 의 빈도 분석 결과 이다.
남성(1) 의 빈도가 292 명중 126 명이 남자인 것을 말하며 여성(2) 의 비도가 292 명 중 166 명 인 것을 말한다. 퍼센트는 292 명중 각 남성과 여성의 퍼센트를 나타낸 것이다. 유효 퍼센트와 펴센트가 같은 것으로 보아서 결측치는 없는 것 으로 보인다.

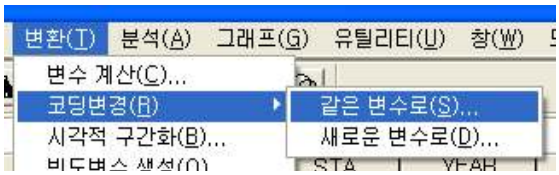
S5

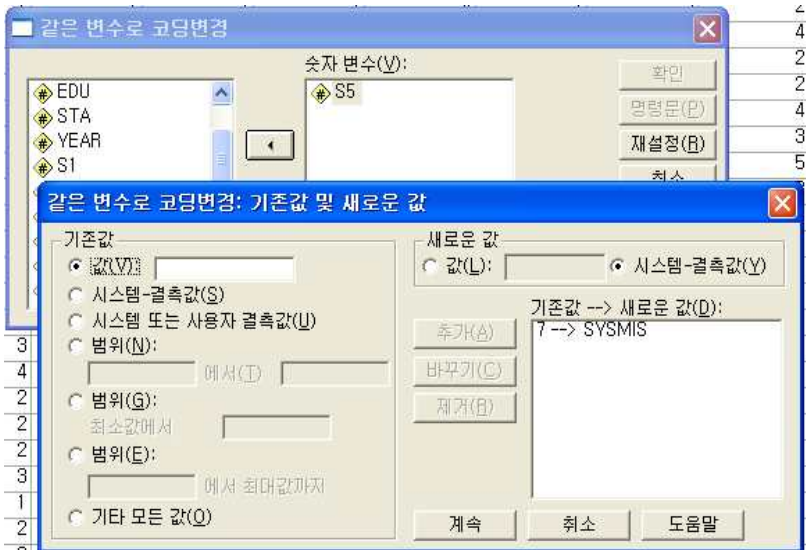
		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	1	40	13.7	13.7	13.7
	2	71	24.3	24.3	38.0
	3	72	24.7	24.7	62.7
	4	70	24.0	24.0	86.6
	5	38	13.0	13.0	99.7
	7	1	.3	.3	100.0
	합계	292	100.0	100.0	

위의 결과를 보면 부적절한 응답이 있는 것을 볼 수 있다. 문항은 5개 밖에 없는데 7으로 응답한 것을 확인 할 수 있었다. 입력 오류값을 시스템 결측값으로 바꿔주려면 S5 문항에서 내림차순으로 하여 결측값으로 바꿔주는 방법과 코딩변경을 사용하여

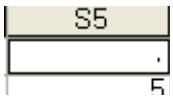
결측값으로 바꿔주는 방법이 있다 .

6. 코딩변경





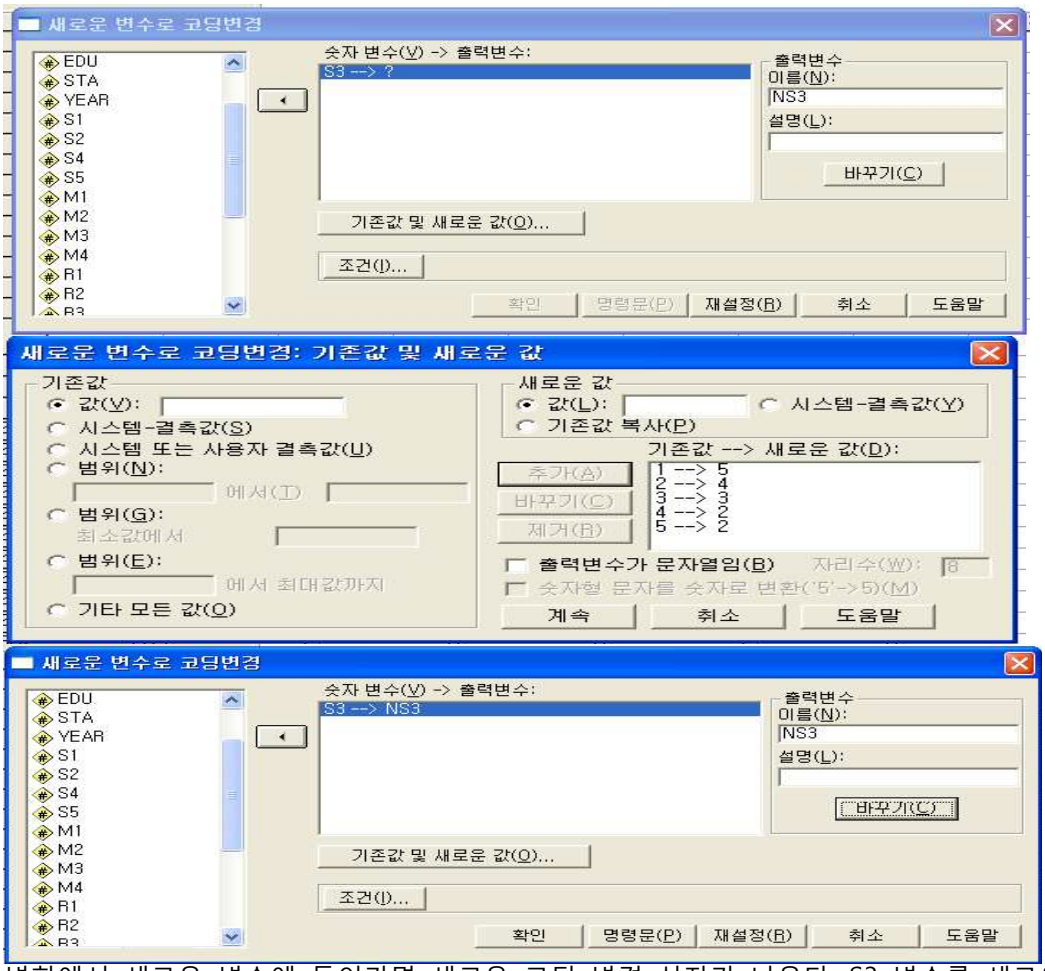
변환 -> 코딩변경 -> 같은 변수로 에 들어가서 S5 변수를 숫자 변수로 옮긴다. 이유는 S5 변수에서 입력 오류 값인 7 이 있기 때문이다. S5를 선택한 후 같은 기존 값 및 새로운 값 단추를 누르면 위의 상자가 나타난다. 우리가 지금 하려고 하는 것은 S5 의 입력 오류 값인 7을 결측 값으로 바꿔야 하기 때문에 기존 값 : 7을 입력하고 새로운 값에는 시스템 결측 값 을 누른 후 추가 한다. S5변수에 7 이 정말로 결측 값으로 변했는지 확인해 보자.



S5 변수에 결측 값이 생긴 것을 확인 할 수 있다.

• 새로운 변수로 코딩 변경

S3에서만 매우불만이 5, 약간불만은 4, 보통은 3, 약간만족은 2, 매우 만족이 1로 처리 되었다고 한다. 이 응답값을 다른 문항의 응답값과 같은 형태로 변환시키기 위해서 새로운 변수명 생성 해야 한다.



변환에서 새로운 변수에 들어가면 새로운 코딩 변경 상자가 나온다. S3 변수를 새로운 변수로 바꿔 주어야 한다. 출력변수(새로운 변수) 에는 자신이 알아 볼 수 있는 변수 NS3 로 지정한 후 기존값 및 새로운 값 단추를 입력한다. S3 변수에서 내용 값이 다른 문항과 반대 로 입력 되어 있기 때문에 기존값에는 1 새로운 값에는 5 를 입력한후 추라 버튼을 누른다. 이와 같이 1->5, 2->4, 3->3, 4->2, 5->,1 을 추가로 입력한 후 확인을 누른 후 바꾸기 버튼을 누르면 된다.

• 역문항

위와 같이 변수값을 반대로 지정해 주는 이유는 역문항 이기 때문이다. 역문항은 설문자가 설문조사 응답을 올바르게 했는지 알아보기 위해 문항 중간에 역문항을 넣어 두는 것이다. 역문항의 문제 예를 들어 보면 아래와 같다.

다음은 청결상태에 대한 질문입니다.

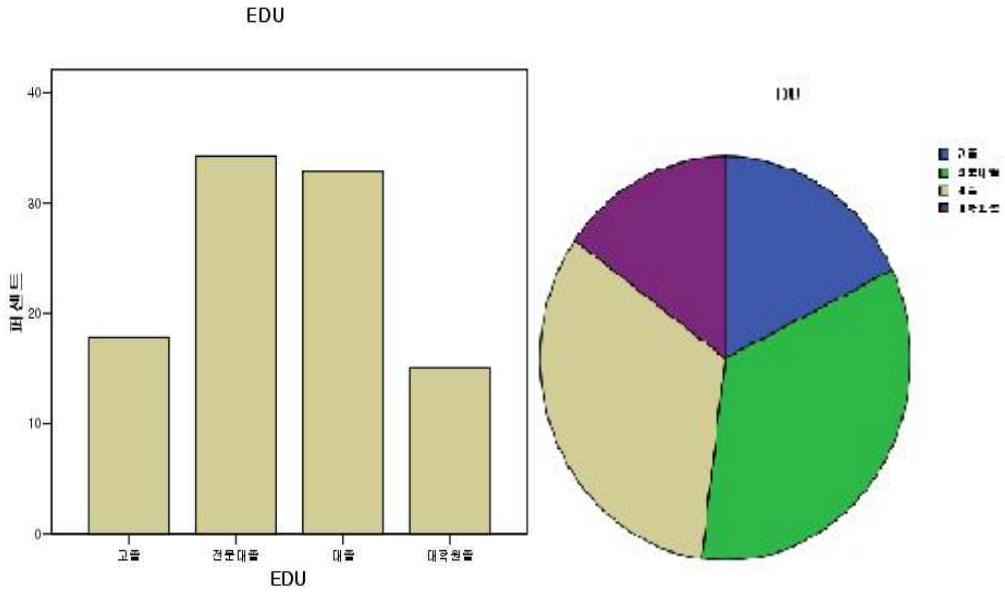
변수명	내용	매우 아니다	아니다	그렇다	매우 그렇다
Q4	선생님은 식사 전에 비누로 손을 씻습니까?	①	②	③	④
Q5	손톱과 발톱을 항상 짧게 깎지 않습니까?	①	②	③	④
Q6	목욕은 규칙적으로 합니까?	①	②	③	④
Q7	양말이나 속옷은 하루에 한번씩 갈아입습니까?	①	②	③	④
Q8	외출해서 돌아오면 손발을 깨끗이 씻습니까?	①	②	③	④
Q9	식사 후와 잠자기 전에는 반드시 양치질을 하십니까?	①	②	③	④

위의 문항을 보면 Q4~ Q9 까지의 문항이 있다 문항을 읽어보면 Q5 문항을 제외 하고 모두 청결을 깨끗이 유지 하는 경우에는 "매우 그렇다"를 표기 하게 되어있다. 하지만 Q5의 문항은 청결을 깨끗이 유지 해야 할 경우에는 매우 아니다 로 표기 하게끔 되어있다. 이렇게 응답자가 올바르게 응답을 했는지 알아 보기위해 역문항을 넣어 주는 것이다.

7. 빈도분석 에서의 그래프

직장 만족도 데이터에서 학력에 대한 그래프를 그려보자 그래프에는 아래의 표에서 볼 수 있듯이 막대도표 원도표 히스토그램 이 있다. 기본으로 도표는 지정되지 않았으며 특별히 도표를 그려 알아보고 싶을 때 빈도분석 상자에서 도표 단추를 입력하여 원하는 도표를 나타내어 볼 수 있다.

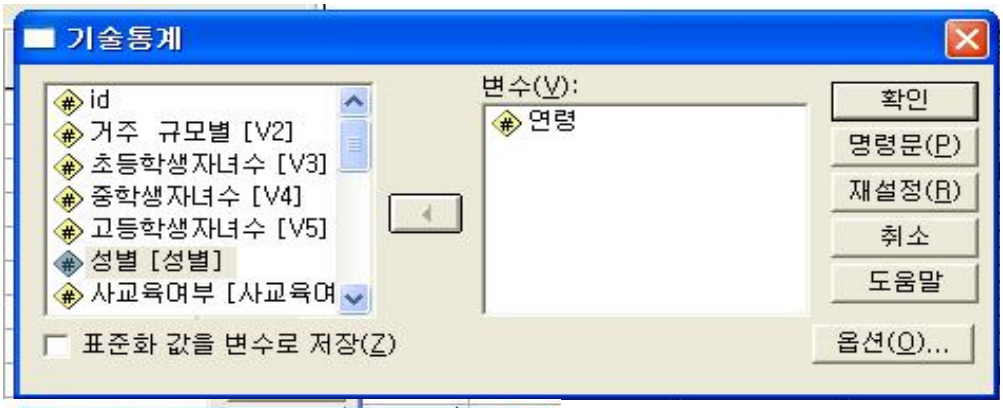
학력이 어떻게 분포 되어있는지 각각의 도표를 통해서 알아보자

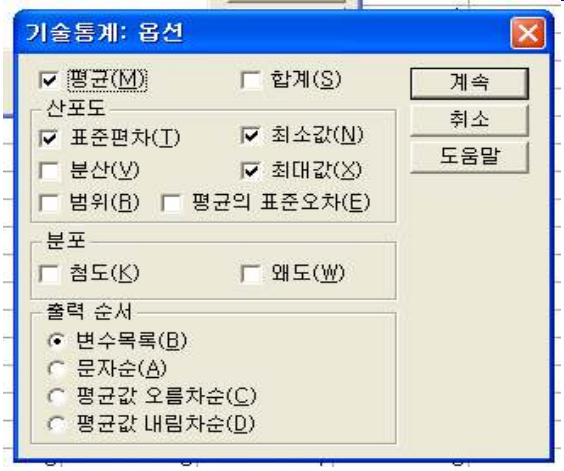


위의 두 개의 도표를 보면 전문대 졸 과 대졸의 인원이 가장 많은 분포를 보이며 고졸과 대학원 졸 인원이 적은 분포를 보인다.

8. 기술통계

빈도분석에서는 질적 자료를 대상으로 하였지만 기술통계에서는 양적 자료를 태상으로 통계 통계량 값을 구할 때 사용 한다.





기술통계량

	N	최소값	최대값	평균	표준편차
연령	295	29	229	40.75	12.015
유효수 (목록별)	295				

학부모 여론조사 데이터를 가지고 기술통계량을 구해보았다 학부모 여론조사 데이터에는 많은 양적 자료가 있지만 그중 연령을 대상으로 기술 통계량 값을 구 해보면 위와 같다. 위의 상자를 보면 295명중 29세 나이가 최소값을 나타내고 229세 나이가 최대값을 보이고 있다 나이가 229세는 없으므로 입력 오류값이라고 판단 결측치로 제거 해준 후 다시 기술통계량 값을 구해보면

기술통계량

	N	최소값	최대값	평균	표준편차
연령	294	29	63	40.11	4.848
유효수 (목록별)	294				

오류값을 결측치로 제거 후 기술통계량 값은 위와 같다. 오류값을 수정하기전의 N 값은 295 이었으나 오류 값을 결측 값으로 고쳤기 때문에 N 값은 294 로 줄었으며 최대 값은 63세 이며 평균 나이는 40로 나왔다. 표준편차는 관측 값들이 평균을 중심으로 얼마나 퍼져 있는가를 측정하는 통계량 값을 뜻한다. 이 외에 다른 통계량 값을 구해보자.

범위는 최대값에서 최소값을 뺀 값으로 표준편차나 분산의 보조통계량으로 사용되며 분산은 관측값들이 평균으로부터 얼마나 떨어져 있는가를 측정하는 통계량으로, 표준편차를 제곱한

기술통계량

	N	범위	최소값	최대값	합계	평균	표준편차	분산	왜도		첨도	
	통계량	통계량	통계량	통계량	통계량	통계량	통계량	통계량	통계량	표준오차	통계량	표준오차
연령	294	34	29	63	11793	40.11	4.848	23.499	.705	.142	1.915	.283
유효수 (목록별)	294											

값이다. 첨도는 분포의 모양이 중심점에서 뾰족한가를 나타내는 통계량이다. 왜도는 분포의 모양이 얼마나 좌우대칭인지를 나타내는 통계량이다.

9. 변수계산

변수값들을 계산할 때 사용 된다. 직장 만족도 데이터를 가지고 변수 계산을 해보자.
먼저 변수 계산을 하기 전에 주의할 점은 함수식과 수식 이다. 함수식은 결측치가 있어도 결측치가 제외하고 계산 하지만 수식 으로 계산할 경우 결측 값이 있으면 결측값으로 계산 된다. 간단히 살펴보자

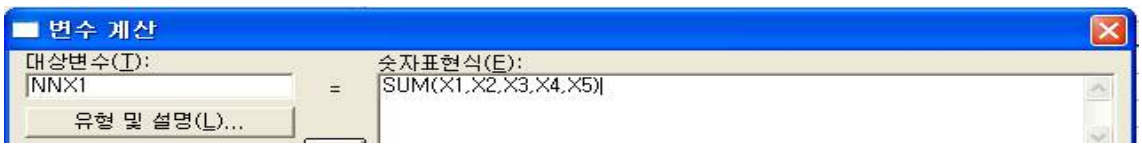
	X1	X2	X3	X4	X5
1	3	1.00	5.00	33.00	49.00
2	10	2.00	43.00	75.00	3.00
3	.	3.00	2.00	32.00	26.00
4	1	3.00	32.00	26.00	21.00
5	4	2.00	.	78.00	11.00

위와 같은 데이터가 있다 데이터를 보면 X1 와 X3 변수에는 결측값이 있는 것을 확인 할 수 있다. 위의 데이터를 가지고 변수 계산을 해보자

• 수식을 이용한 변수 계산

	NX1
1	91.00
2	133.00
3	.
4	83.00
5	.

• 함수식을 이용한 변수 계산



NNX1
91.00
133.00
63.00
83.00
95.00

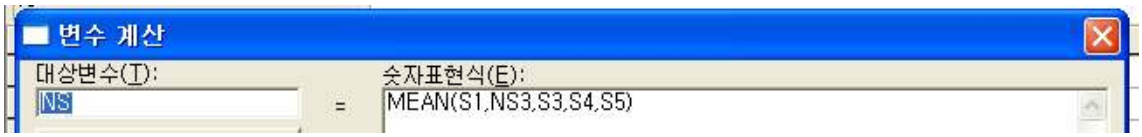
위의 표에서 볼 수 있듯이 수식으로 변수 계산을 하였을 경우에는 변수에 결측값이 하나라도 있으면 결측값으로 나오는 것을 확인 할 수 있다. 반면에 함수식을 이용하였을 때에는 결측 값이 있더라도 결측값이 제외된 상태에서 값이 나오는 것을 알 수 있다.

직장 만족도 데이터로 문제를 풀어 보자

5개의 새로운 변수를 생성하시오. 새롭게 생성되는 변수는 해당 변수들이 가지는 값 중에 결측치가 있더라도 결측치가 제외된 값이 나오도록 한다(즉, 새롭게 생성된 변수에는 결측치가 없도록 한다).

① 보수관련문항 5개의 평균을 계산한『보수』
② 상상관련문항 5개의 평균을 계산한『상사』
③ 대인관계문항 5개의 평균을 계산한『대인관계』
④ 업무관련문항 5개의 평균을 계산한『업무』

문제를 보면 결측치가 없도록 한다 라고 나와 있다 이는 수식으로 계산하는것이 아닌 함수식으로 계산하라는 것을 뜻한다.



보수 관련 문항 5개의 평균을 계산한 함수식 이다.

NS
3.00
3.60
3.60
3.60
3.60
3.60
3.60
3.40
3.80

보수 관련 문항 5개의 평균을 낸 값 위와 같은 방법으로 모두 평균을 낸다.

NS	NM	NR	NJ
3.00	3.00	2.40	3.25
3.60	2.75	2.60	2.33
3.60	3.00	3.40	3.00
3.60	3.00	3.20	3.00

위의 표는 각 변수의 데이터를 평균한 값이다.

10. 데이터 탐색

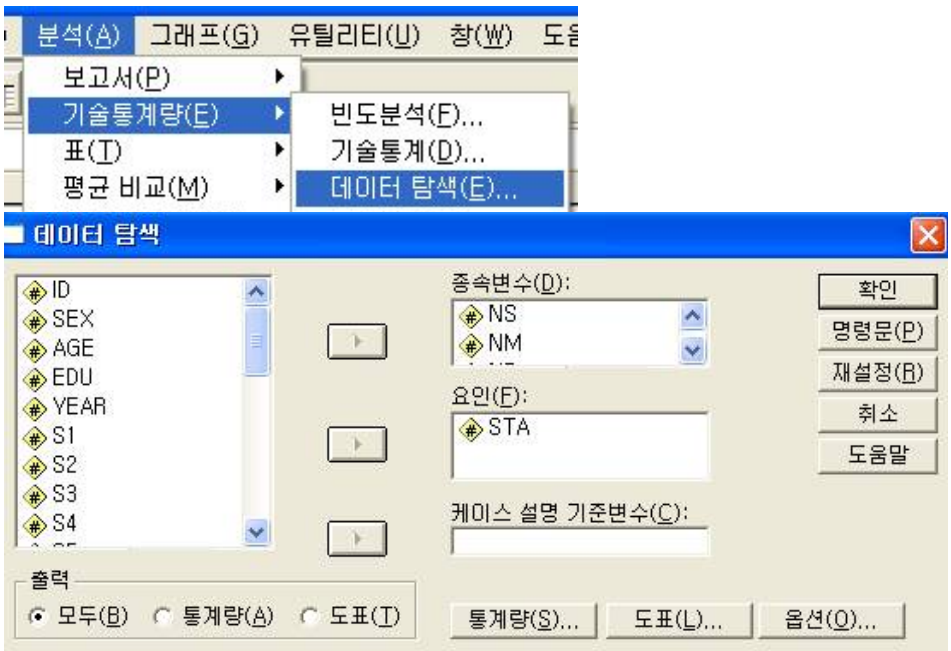
데이터 탐색은 데이터의 구조와 특징을 알아내기 위한 분석 방법으로서 다른 요인들과 같이 볼 때 사용 된다.

아래의 문제를 데이터 탐색을 이용해 풀어 보자.

다음과 같이 기술통계량을 구하시오.

	구분	N	평균	표준편차	최소값	최대값
사원급	보수					
	상사					
	대인관계					
	업무					
	복지					

위의 문제를 보면 사원급은 직위를 나타내며 보수, 상사, 대인관계, 업무 는 변수 계산을 이용하여 새로운 변수를 만들었던 변수를 나타낸다. 즉 직위에 따른 각각의 문항의 기술 통계량 구하는 것이다. 직위와 다른 요인과 같이 봐야 하기 때문에 데이터 탐색을 이용한다.



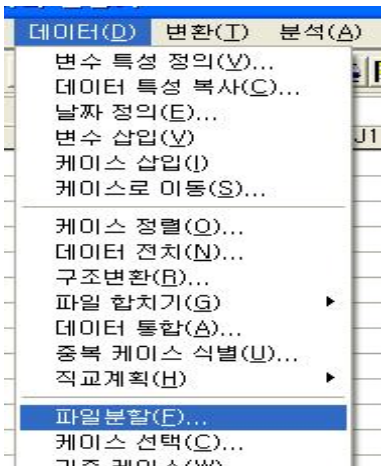
위와 같은 방법으로 종속변수에는 변수계산 한 변수를 넣어주고 요인에는 같이 보고자 하는 변수 즉 직위를 넣어준다.
아래의 표는 결과창의 일부 이다.

	STA			통계량	표준오차
NS	사원급	평균		2.9714	.05819
		평균의 95% 신뢰구간	하한	2.8555	
			상한	3.0873	
		5% 절삭평균		2.9740	
		중위수		3.0000	
		분산		.261	
		표준편차		.51064	
		최소값		1.80	
		최대값		4.20	
		범위		2.40	
		사분위수 범위		.80	
		왜도		-.036	.274
		첨도		-.110	.541

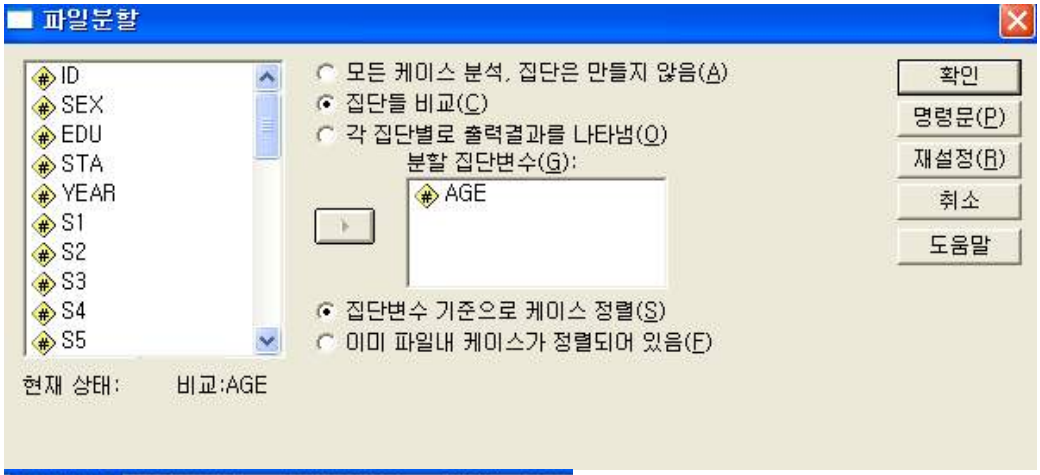
11. 파일분할

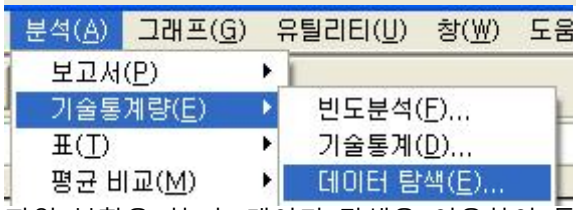
데이터 탐색은 두 가지의 요인을 함께 알아보고자 할 때 사용 하지만 여기서 한 가지 요인을 더 추가해서 알아보고자 할 때 사용된다. 문제를 보면서 살펴보자

연령이 30대이면서 직위가 주임급인 사람들의 특징을 알아보고자 할 때 직장 만족도의 연령은 4개의 값으로 분류되어 있다. 여기서 문제는 각 연령대 별로 나누어서 알아보아야 하기 때문에 연령을 파일 분할하여 알아 볼 수 있다.

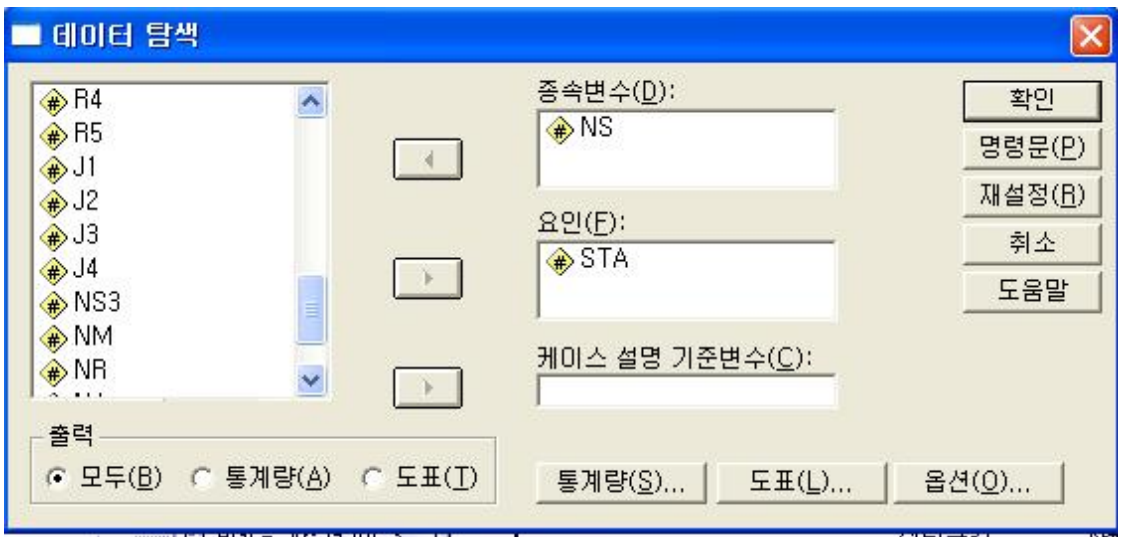


위의 파일분할 상자를 보면 AGE 변수가 분할 집단 변수에 들어 가 있는 것을 알 수 있다. 문제가 연령별에 따f라 알아보아야 하기 때문에 집단들 비교에 선택을 해 준다.





파일 분할을 한 후 데이터 탐색을 이용하여 통계량 값을 구한다.



위의 데이터 탐색 상자를 보면 요인에는 직위가 들어 가있으며 종속변수에는 위에서 변수 계산한 값들이 들어 가있다. 즉 파일분할을 이용하여 연령대 별로 나누어 준 후 데이터 탐색을 이용하여 직위에 따른 각 변수의 통계량을 구하는 것이다.

문제 30대 이면서 직위가 주임급인 사람의 특징

위의 표는 30대 이면서 주임급인 사람의 보수 관련 문항에 대한 특징이다.
다른 관련 문항에 대한 특징도 알아보고 싶다면 위의 종속변수에 다른 문항을 추가 적으로 입력 한 후에 알아보면 된다.

주임급	평균		3.0133	.06828
	평균의 95% 신뢰구간	하한	2.8757	
		상한	3.1509	
	5% 절삭평균		3.0086	
	중위수		3.0000	
	분산		.210	
	표준편차		.45806	
	최소값		2.20	
	최대값		3.80	
	범위		1.60	
	사분위수 범위		.80	
	왜도		.069	.354
	첨도		-1.137	.695

여기서 주의 할 점은 파일분할을 한 후 파일분할 한 데이터를 다시 복귀 시켜주지 않으면 파일분할 된 데이터로 계속 결과가 나오기 때문에 주의해야 한다.

12.교차분석

교차분석은 한 변수의 빈도분석표를 작성한 것과는 달리 두 개 이상의 행과 열을 갖는 교차표를 작성하게 할 때 유용하다.

<자료> 학부모 여론 조사

문제

1. 당신은 자녀에게 학교 이외의 사교육을 시키고 있습니까?

① 있다 ② 없다

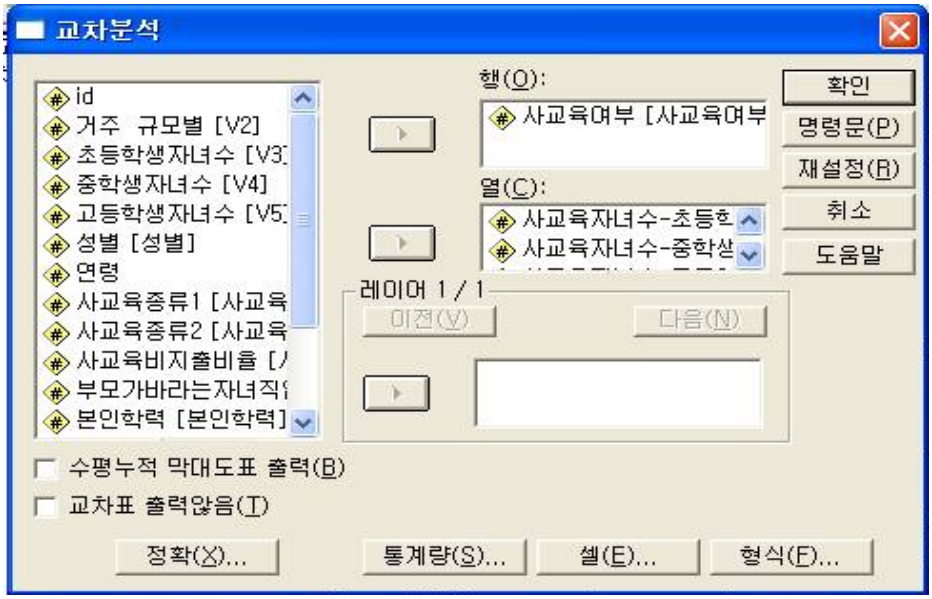
2. 사교육을 시키는 자녀는 몇 명입니까?

초등학교 자녀 _____명, 중학교 자녀 _____명, 고등학교 자녀 _____명

위의 문제를 1번을 보면 사교육 여부에 관한 문항이다. 응답자가 "없다"라고 응답할 경우에는 2번 문항에는 사교육을 시키는 자녀의 수를 입력 하면 안 된다. 이렇듯 교차 분석은 서로 상관 있는 문항 끼리의 관계를 알아 보기 위해 사용 한다.

위의 교차분석 상자를 보면 행에는 사교육 여부 변수가 들어 가 있으며 열에는 사교육 자녀수 입력해 준다. 결과창을 살펴보면

위의 결과창에서 케이스 처리 요약을 보면 295명을 대상으로 설문 조사를 한 결과 결측값



케이스 처리 요약

	케이스					
	유효		결측		전체	
	N	퍼센트	N	퍼센트	N	퍼센트
사교육여부 * 사교육자녀수-초등학생	295	100.0%	0	.0%	295	100.0%
사교육여부 * 사교육자녀수-중학생	295	100.0%	0	.0%	295	100.0%
사교육여부 * 사교육자녀수-고등학생	295	100.0%	0	.0%	295	100.0%

사교육여부 * 사교육자녀수-초등학생 교차표

빈도

		사교육자녀수-초등학생				
		없음	1명	2명	3명	전체
사교육여부	있음	64	78	26	2	170
	없음	124	1	0	0	125
전체		188	79	26	2	295

사교육여부 * 사교육자녀수-중학생 교차표

빈도

		사교육자녀수-중학생			
		없음	1명	2명	전체
사교육여부	있음	105	59	6	170
	없음	124	0	1	125
전체		229	59	7	295

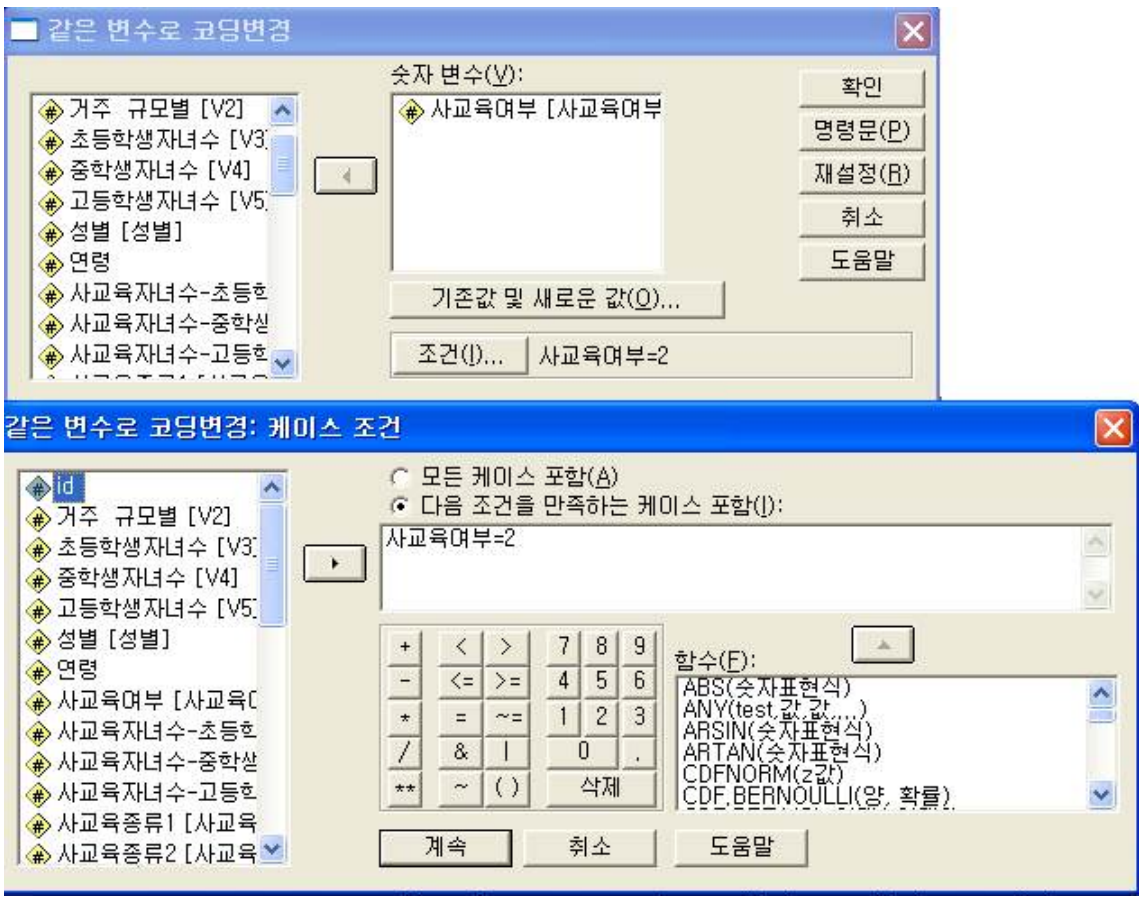
사교육여부 * 사교육자녀수-고등학생 교차표

빈도

		사교육자녀수-고등학생				
		없음	1명	2명	3명	전체
사교육여부	있음	133	30	7	0	170
	없음	124	0	0	1	125
전체		257	30	7	1	295

은 없어 보인다. 그리고 사교육 여부와 사교육 자녀수의 교자표를 보면 첫 번째 표를 보면 사교육 여부에는 "없음"을 입력 하였는데 사교육을 시키고 있는 자녀수에는 1명을 입력 한 것으로 나와 있다. 또한 중학생 교차표에도 사교육 여부를 "없음"에 입력 하였는데 2명 입

력 하였으며 고등학생 또한 입력 오류 값이 나와 있는 것을 알 수 있다. 입력 오류값을 결측값으로 입력 해 주기 위해 코딩변경을 사용하여 입력 오류 값을 결측값으로 바꿔준다. 코딩변경을 사용하여 바꿔주려면 위에서 설명한 것과 같이 해야 하지만 여기서는 사교육 여부를 "없음" 으로 입력한 문항에 한에서 결측값을 으로 입력해 줘야 하기 때문에 조건을 입력하여야 한다.



위의 같은 변수로 코딩변경 창에서 조건 단추를 클릭한다. 사교육 여부를 "없음" 이라고 응답한 문항에 대해서만 결측값 으로 입력해 줘야 하기 때문에 위의 창 과 같이 입력 해 준다.

조건을 입력한 후 기존값및 새로운 값 단추를 누른후 위와 같이 오류값을 결측값으로 입력해 준다. 오류값을 시스템 결측값으로 입력 후에 다시 교차분석을 해본다.

케이스 처리 요약

	케이스					
	유효		결측		전체	
	N	퍼센트	N	퍼센트	N	퍼센트
사교육여부 * 사교육자녀수-초등학생	294	99.7%	1	.3%	295	100.0%
사교육여부 * 사교육자녀수-중학생	294	99.7%	1	.3%	295	100.0%
사교육여부 * 사교육자녀수-고등학생	294	99.7%	1	.3%	295	100.0%

사교육여부 * 사교육자녀수-초등학생 교차표

빈도

		사교육자녀수-초등학생				
		없음	1명	2명	3명	전체
사교육여부	있음	64	78	26	2	170
	없음	124	0	0	0	124
전체		188	78	26	2	294

사교육여부 * 사교육자녀수-중학생 교차표

빈도

		사교육자녀수-중학생			
		없음	1명	2명	전체
사교육여부	있음	105	59	6	170
	없음	124	0	0	124
전체		229	59	6	294

사교육여부 * 사교육자녀수-고등학생 교차표

빈도

		사교육자녀수-고등학생			
		없음	1명	2명	전체
사교육여부	있음	133	30	7	170
	없음	124	0	0	124
전체		257	30	7	294

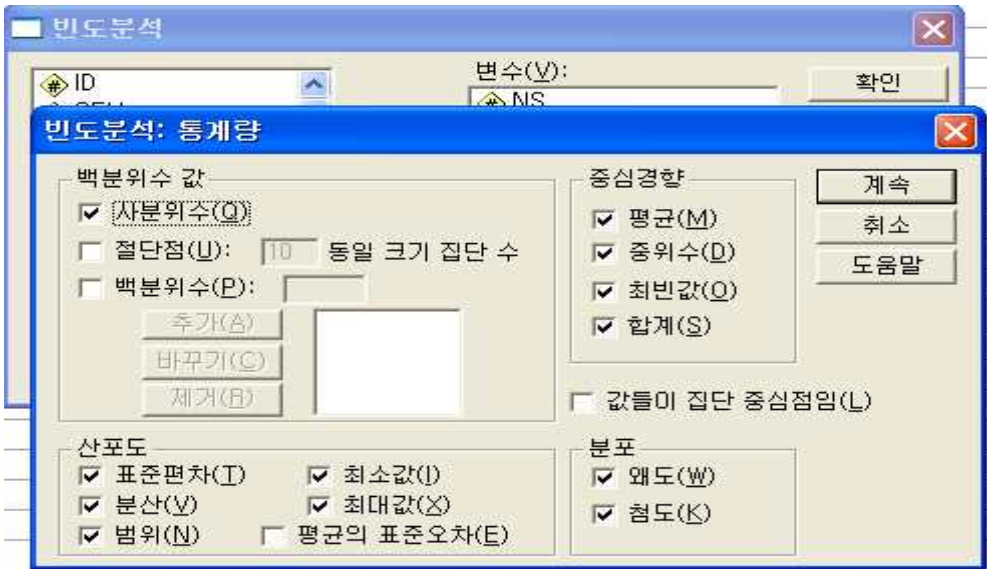
위의 결과창을 보면 사교육 여부에 "없음"을 입력 했을 때 사교육 자녀수에 입력한 값이 모두 없어 진 것을 확인 할 수 있다. 위의 결과창을 해석하면 사교육를 시키는 초등학생의 수는 78명의 응답자가 1명의 초등학생을 교육시키고 26명의 응답자가 2명, 2명의 응답자가 총 3명의 초등학생을 사교육 시킨다. 중학생을 사교육을 시키는 중학생의 수는 59명의 응답자가 1명의 중학생 사교육 시키고 6명의 응답자가 2명을 사교육 시킨다고 응답하였다. 그리고 사교육 시키는 고등학생의 수는 30명의 응답자가 1명, 7명의 응답자가 2명 사교육 시킨다.

13. 빈도 분석에서의 통계량 값

위의 빈도분석에서 입력 오류값을 코딩변경으로 이용하여 결측값으로 바꿔주는 작업을 해서 빈도분석에서의 통계량 값을 구하는 작업을 못했다. 직장만족도 자료 이용하여 빈도 분석의 통계량 값을 구해보자.

구분	N	평균	표준편차	최소값	최대값
보수					
상사					
대인관계					
업무					

참고) 보수 상사 대인관계 업무 는 각 문항을 평균한 값을 뜻한다.



위의 문제에서 평균 표준편차 최소값 최대값 의외에 빈도분석창의 통계량 단추 를 눌러 들어가면 더 많은 통계량 값을 구할 수 있다.

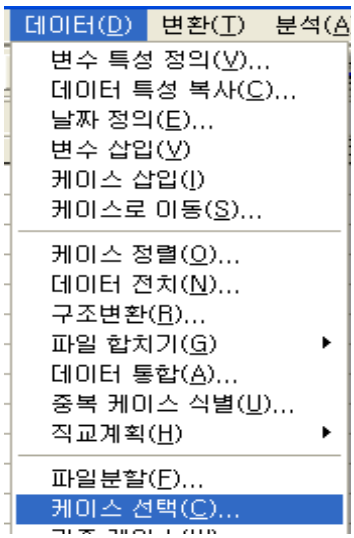
평균 : 각 문항 값의 합계에서 총 N 의 수를 나누어 준 값.
중위수 : 중앙이라고도 하며, 크기순서로 가장 작은 값 가장큰 값으로 나열하였을 경우에 정 중앙에 위치하는 관측값, 관측값의 수가 짝수 일 경우에는 중앙에 위치하는 두 관측값의 산술평균이 중위수가 된다.
최빈값 : 가장 많은 빈도를 가지고 있는 관측값
합계 : 모든 관측값을 합산한 값
왜도 : 분포의 비대칭도 분포의 모양이 얼마나 좌우대칭인지를 나타내는 값
첨도 : 분포의 뾰족한 정도 분포의 모양이 중심점에서 얼마나 뾰족한가를 나타내는 통계량
사분위수 : 가장 작은 값에서 큰값의 순서로 나열했을 때 아래서부터 25% 50% 75%에 해당 되는 값이 출력 된다.
표준편차 : 관측값들이 평균을 중심으로 퍼진 정도

분산 : 표준편차의 제곱값
범위 : 최대값에서 최소값을 뺀값
최소값 : 관측값중 가장 작은 값
최대값 : 관측값중에 가장 큰값

14. 케이스 선택

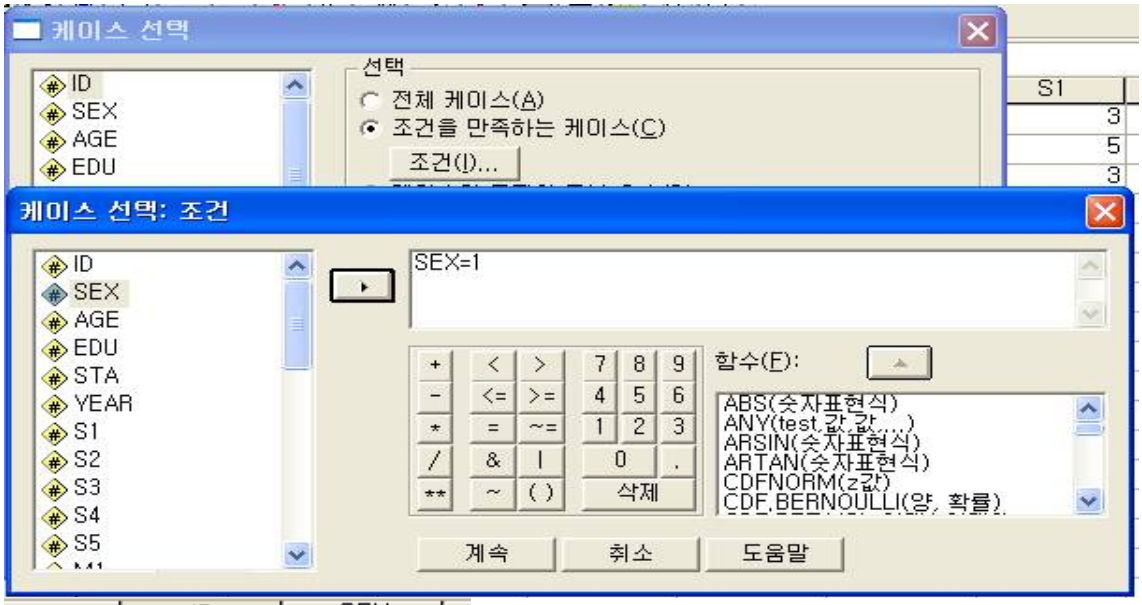
케이스 선택 기능은 조건에 맞는 케이스만 분석하고 나머지 케이스는 분석을 하지 않을 목적에 적합한 기능이다 만약 특정 조건에 맞는 케이스만을 선택 하는 방법은 데이터에서 케이스 선택에 가서 사용 하면 된다.

예를 들어 데이터 직장만족도 데이터를 에서 성별이 남자를 대상으로 보수, 상사, 대인관계, 업무에 대한 통계량 값을 알아보고자 할 때 남자를 대상으로 알아 봐야 하기 때문에 남자를 케이스 지정 하는 법을 살펴보면 아래와 같다.



케이스 선택에 들어가서 조건을 만족하는 케이스를 선택후 조건을 지정해 준다. 조건을 만족하는 케이스로 간 이유는 성별 남자를 대상으로 알아봐야 하기 때문이다. 조건에서 성별 SEX = 1 입력 해 준다.

옆의 표와 같이 성별이 여성인 케이스는 모두
제거 된 것으로 확인 할 수 있다.



	ID	SEX
1	1	1
2	2	2
3	3	2
4	4	2
5	5	1
6	6	1
7	7	2
8	8	1
9	9	2
10	10	1

• 위의 케이스 조건에 계속 이어서 남성만을 대상으로 각 항목별 나무-줄기 그림과 상자그림 을 그려보겠다.

데이터 탐색에서 종속변수는 각 항목의 평균값과 요인

에는 성별을 입력한다. 그리고 데이터 탐색의 도포 창에서는 상자도표 를 살펴 보자.

요인수준들과 함께 : 특정한 종속변수에 대하여 각 집단별로 상자도표를 나란히 출력, 집단간

의 비교를 할 경우에 사용된다.

종속변수들과 함께 : 특정한 집단에 대하여 각 종속변수별로 상자도표를 나란히 출력, 집단의 여러 종속변수들을 비교할 경우에 사용한다.

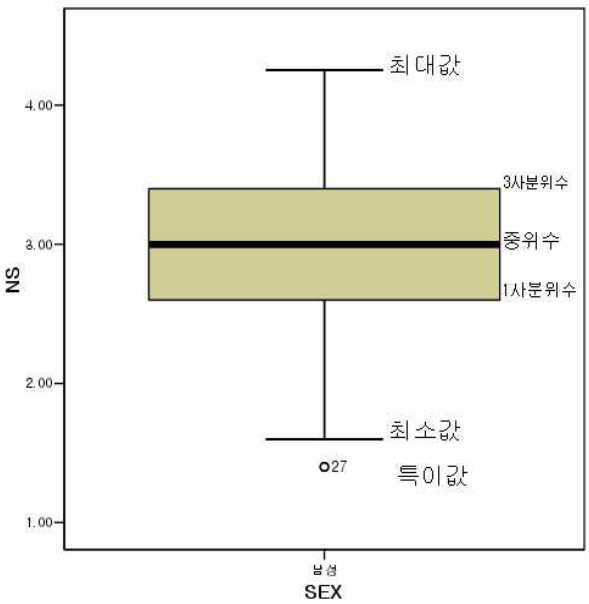
기술통계에서의 각 항목을 보면 나무줄기 그림이 있으며 히스토그램이 있다. 자신이 원하는선택하여 볼 수 있다. 둘 다 같이 선택 할 수 있다.

SEX			통계량	표준오차
NS	남성	평균	2,9611	,04885
		평균의 95% 신뢰구간 하한	2,8644	
		상한	3,0578	
		5% 절삭평균	2,9665	
		중위수	3,0000	
		분산	,301	
		표준편차	,54830	
		최소값	1,40	
		최대값	4,25	
		범위	2,85	
		사분위수 범위	,80	
		왜도	-,144	,216
		첨도	,309	,428

```
Frequency    Stem &  Leaf

     2.00 Extremes    (=<1.4)
      .00         1 .
     1.00         1 .  6
      .00         1 .
     4.00         2 .  0000
     4.00         2 .  2222
    13.00         2 .  4444444444455
    19.00         2 .  6666666666666667777
    16.00         2 .  8888888888888888
    17.00         3 .  00000000000000000
    15.00         3 .  222222222222222
    17.00         3 .  44444444444444445
     8.00         3 .  66666666
     3.00         3 .  888
     5.00         4 .  00000
     2.00         4 .  22

Stem width:      1.00
Each leaf:       1 case(s)
```



각항목별로 그림을 안 알아보고 대표적으로 보수에 관련된 그림만 알아보겠다. 왼쪽은 나무-줄기 그림이고 오른쪽 그림은 상자 그림 이다. 여기서 왜도와 첨도에 대해서 알아보자 왜도는 분포의 비대칭 측도이다. 정규분포는 대칭이므로 왜도값이 0이다. 양의 왜도를 가지는 분포에는 오른쪽으로 긴 꼬리가 나타나고 유의한 음의 왜도를 가지는 분포에는 왼쪽으로 긴 꼬리가 나타난다. 첨도는 관측값이 중심점 주위에 군집하는 범위에 대한 측도 이다. 정규분포의 경우 첨도 통계량 값은 0 이다. 양의 첨도는 관측값이 정규분포에서의 관측값보다 더 많이

군집되어 있으며 꼬리가 더 길다는 의미이고, 음의 첨도는 관측값이 더 적게 군집되어 있고 꼬리도 짧다는 의미이다. 위의 결과 창에서 비교하자면 왜값이 -1.44 나왔으므로 분포형태가 왼쪽으로 긴 꼬리 형태가 나타나며 첨도값은 0.309 이므로 양의 값이다. 즉 관측값이 중앙으로 조금 몰려 있다는 것을 뜻한다.

15. 다중응답

다중응답은 한 문항에서 두 가지 이상의 항목을 선택하도록 구성된 문항을 다중응답문항이라 한다.

이분법

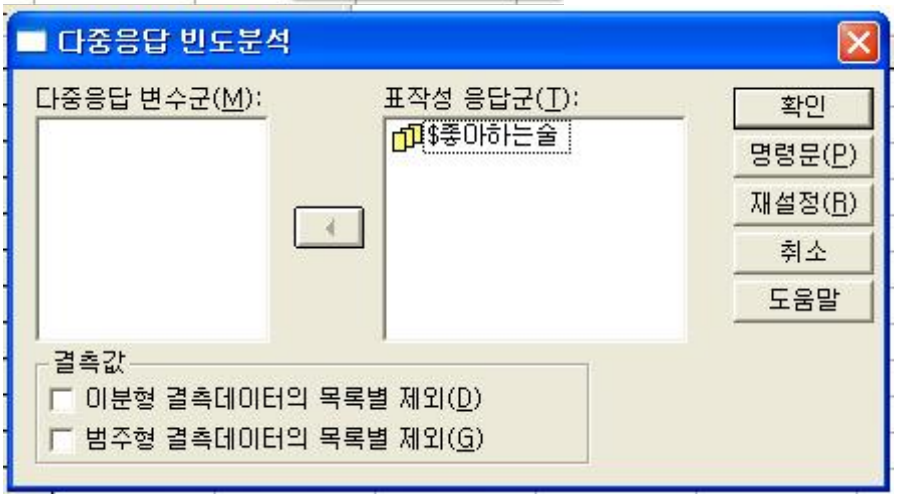
	VAR00001	X1	X2	X3	X4	X5
1	소주	.	.	1.00	1.00	.
2	맥주	.	1.00	1.00	.	.
3	양주	1.00	.	1.00	1.00	.
4	고량주	1.00	.	.	.	1.00

위의 데이터는 좋아하는 술을 개수에 상관없이 체크 한 데이터 이다. 한 개의 문항에 여러개의 술을 선택하였으므로 다중응답 이다.

분석의 다중응답에서 변수군 정의에 들어간다. 그럼 위와 같은 창이 나오는데 각 문항을 변수군에 포함된 변수로 지정해주고 변수들의 코딩형식에는 이분형의 빈도화 값 = 1 로 입력 해 준다. 이분형은 응답이 두 개 밖에 없는 것으로 (0 : 아니오 1 : 예) 이다. 그리고 이름에는 "좋아하는 술" 에 대해서 조사를 하였으므로 "좋아하는 술" 이라고 입력해 준다.
그리고 추가 후 다시 분석 다중응답으로 간다.

변수군 정의를 하기 전에는 밑에 빈도 분석이 안 나왔었는데 변수군 정의를 해준 후 빈도 분석이 나온 것을 확인 할 수 있다. 빈도 분석에 들어간다.

빈도분석에 들어가서 좋아하는 술을 표작성 응답군 으로 이동 한다.



결과창.

위의 표는 다중응답에 대한 표 이다. 표를 보면 X1 은 좋아하는 술이 2개 있음을 나타내며 퍼센트로는 22.2 % , X2 는 좋아하는 술이 1개 이며 11.1%, X3 는 좋아하는 술이 3 개, 33%.3, X5 는 좋아하는 술이 2개 이고 22.2% 이다. X5 는 좋아하는술은 1개 이며 11.1% 를 분포를 나타내고 있다.

Group $좋아하
(Value tabulated = 1)

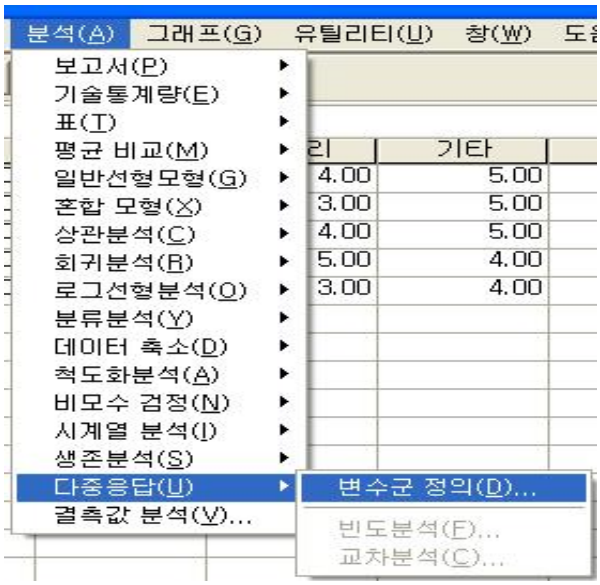
Dichotomy label	Name	Count	Pct of Responses	Pct of Cases
	X1	2	22.2	50.0
	X2	1	11.1	25.0
	X3	3	33.3	75.0
	X4	2	22.2	50.0
	X5	1	11.1	25.0
Total responses		9	100.0	225.0

중복법

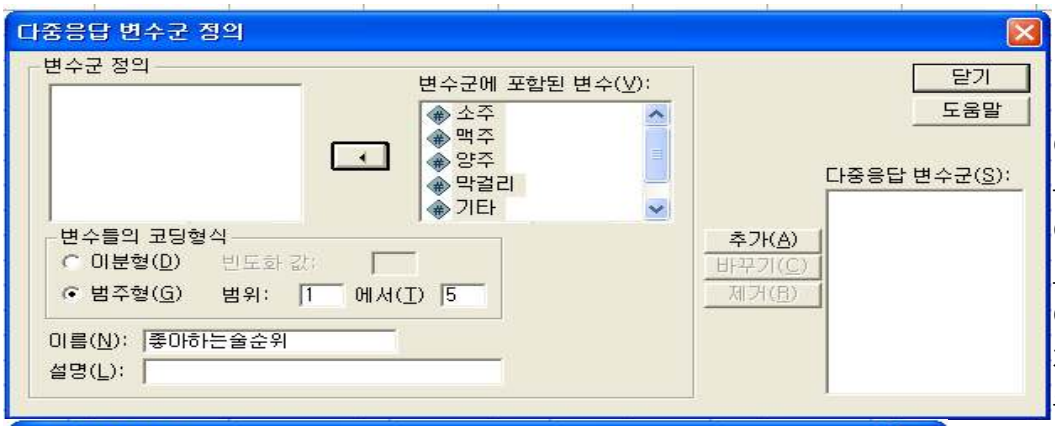
이분법은 0 , 1 두 개의 값을 입력하였지만 중복법은 모두다 선택 하는 것을 말한다. 데이터를 보고 살펴보면 아래와 같다.

	ID	소주	맥주	양주	막걸리	기타
1	조승현	1.00	3.00	2.00	4.00	5.00
2	강효진	2.00	1.00	4.00	3.00	5.00
3	김용한	2.00	1.00	3.00	4.00	5.00
4	이창수	3.00	2.00	1.00	5.00	4.00
5	김민정	2.00	1.00	5.00	3.00	4.00

위이 데이터는 개인이 좋아하는 술에 대해서 나타난 데이터 이며, 좋아하는 술을 순서 대로 나열 한 것이다.



이분형과 동일하게 변수군 정의로 들어 간다.



이분형은 값이 0,1로 되어있었지만 위의 데이터는 1~5까지 있기 때문에 범주형에서 범위를 지정해 준다. 다음으로 좋아하는 술에 대해서 순위를 나타내는 방법은 아래와 같다.

분석에서의 기술통에서 각 변수의 평균값으로도 순위를 낼 수 있기 때문에 기술통계에 가서 평균값을 산출해 순위를 알아 보면 다음과 같다.

기술통계량

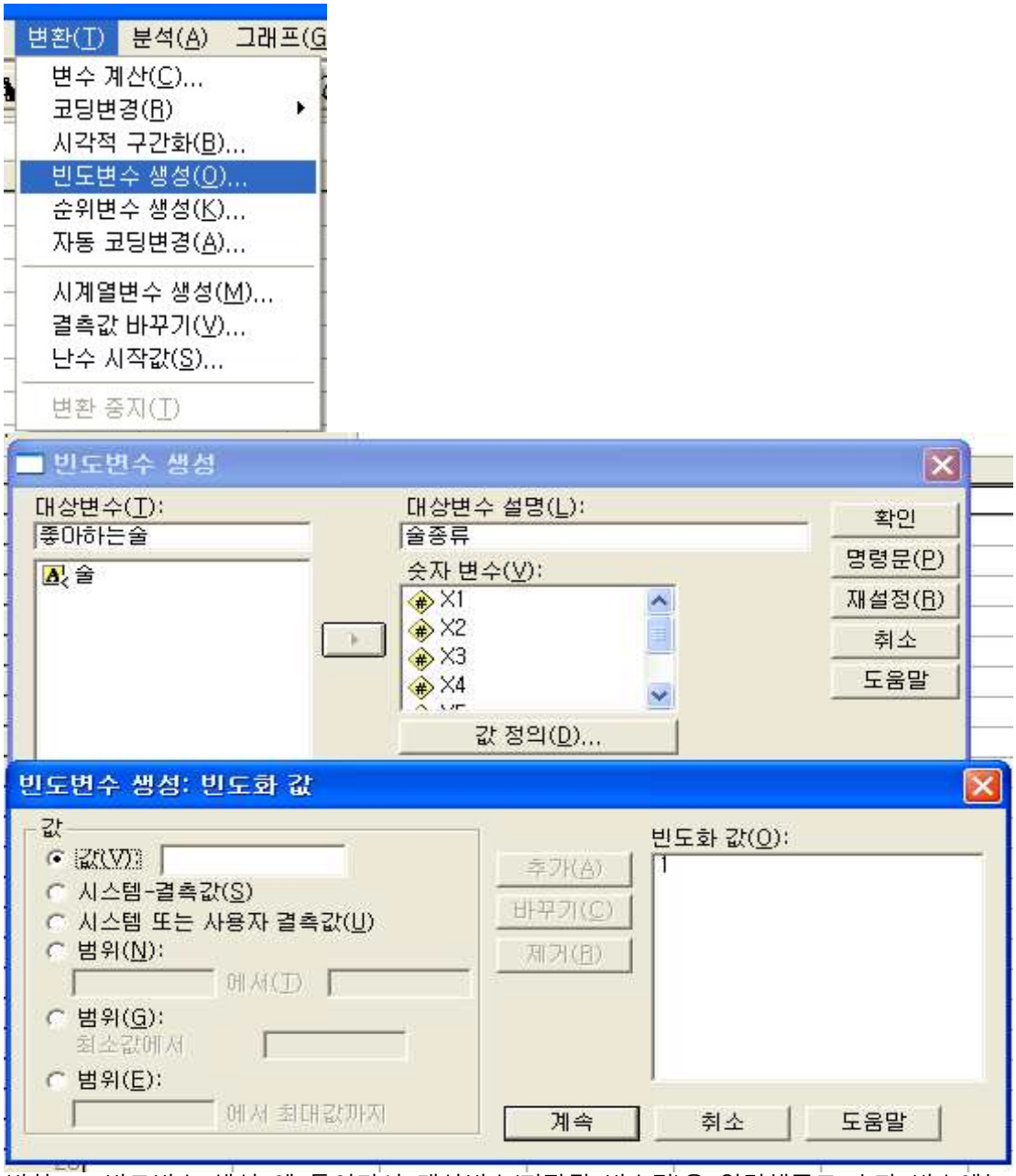
	N	평균
소주	5	2.0000
맥주	5	1.6000
양주	5	3.0000
막걸리	5	3.8000
기타	5	4.6000
유효수 (목록별)	5	

위의 평균값을 보면 맥주가 1.6 으로 가장 작으므로 맥주를 좋아하는 사람이 많았으며 그 다음은 소주 양주 막걸리 기타 순이 되겠다.

16. 빈도 변수 생성

특정변수들 내에 임의 값이 몇 번 나타나는지를 파악하기 위하여 빈도를 변수값으로 하는 변수를 생성하기 위한 것이다.

	술	X1	X2	X3	X4	X5
1	소주	1.00	1.00	2.00	2.00	4.00
2	맥주	2.00	3.00	1.00	1.00	3.00
3	양주	3.00	2.00	3.00	3.00	2.00
4	막걸리	4.00	4.00	5.00	4.00	1.00
5	기타	5.00	5.00	4.00	5.00	5.00



변환 -> 빈도변수 생성 에 들어가서 대상변수(저장될 변수명)을 입력해주고 숫자 변수에는 각 변수들을 이동시킨다. 그리고 값 정의 에 들어가서 자신이 알아보고자 하는 값을 입력 후 추가 버튼을 누른다. 나는 가장 좋아하는 술의 빈도를 알아보기 위해서 1 을 입력 하였다.

	술	X1	X2	X3	X4	X5	좋아하는술
1	소주	1.00	1.00	2.00	2.00	4.00	2.00
2	맥주	2.00	3.00	1.00	1.00	3.00	2.00
3	양주	3.00	2.00	3.00	3.00	2.00	.00
4	막걸리	4.00	4.00	5.00	4.00	1.00	1.00
5	기타	5.00	5.00	4.00	5.00	5.00	.00

위의 데이터를 보자 좋아하는 술 이라는 변수 값이 새로 생긴 것을 확인 할 수 있으며 변수의 값들이 뜻하는 것은 소주를 가장 좋아한다고 응답한 사람이 2명 맥주를 가장 좋아한다고 응답

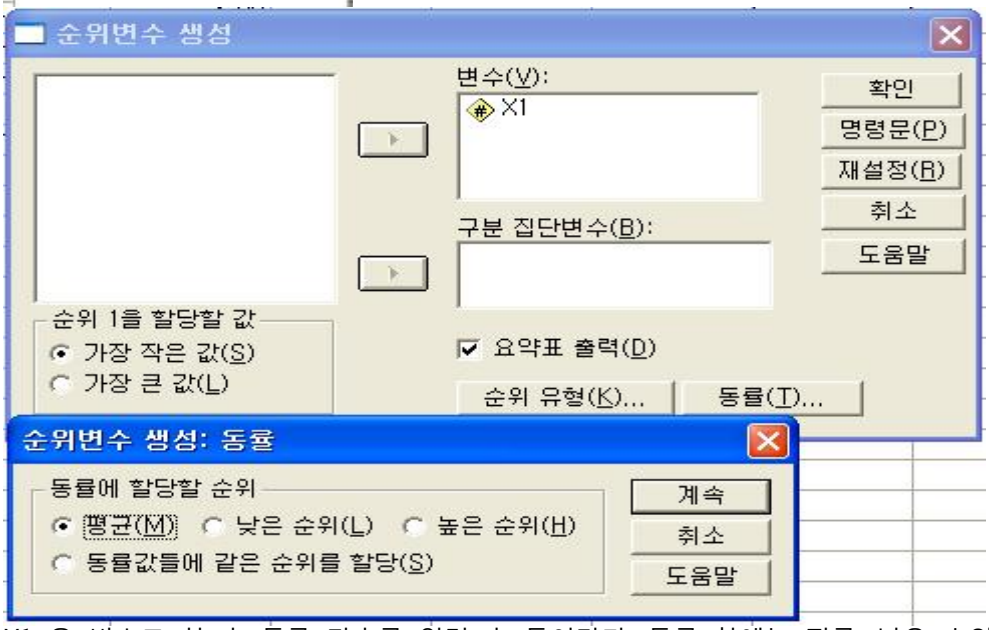
한사람이 2명 양주와 기타는 각각 없으며 막걸리를 가장 좋아한다고 응답한 사람은 1명이다.

17. 순위변수 생성

	술	X1
1	소주	1.00
2	맥주	1.00
3	양주	3.00
4	막걸리	4.00
5	기타	5.00

옆의 데이터와 같이 좋아하는 술을 순서대로 나열한 데이터가 있다 그런데 소주와 맥주의 순위 같을 경위 순위를 어떻게 정할까? 이러한 경우에 순위변수 생성을 이용하여 순위를 정할 수 있다.

변환 -> 순위 변수 생성에 들어 간다.



X1 을 변수로 한 후 동률 단추를 입력 후 들어간다. 동률 창에는 평균, 낮은 순위, 높은 순위, 동률값들에 같은 순위를 할당 이 있다. 각각 비교해서 입력하면 아래와 같다.

	술	X1	RX1	RAN001	RAN002	RAN003
1	소주	1.00	1.500	1.000	2.000	1.000
2	맥주	1.00	1.500	1.000	2.000	1.000
3	양주	3.00	3.000	3.000	3.000	2.000
4	막걸리	4.00	4.000	4.000	4.000	3.000
5	기타	5.00	5.000	5.000	5.000	4.000

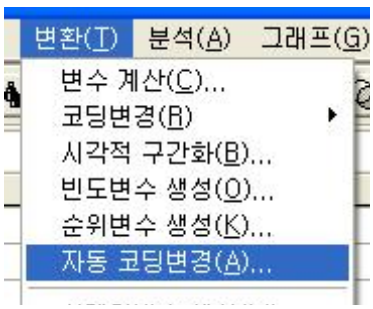
RX1 은 평균으로 순위를 나타낸 것이며 RAN001 은 낮은 순위 RAN002 는 높은 순위 RAN003

은 동률값들에 같은 순위를 할당 한 것이다. 평균부터 살펴보면 소주와 맥주는 각각 1.5 등으로 나왔으며 나머지의 술들은 이전과 동일하게 나왔다. 낮은 순위의 경우에는 소주와 맥주가 1위로 나왔으며 높은 순위의 경우에는 소주와 맥주가 2위로 나왔다. 동률 값들에 같은 순위로 할당 한 경우에는 소주와 맥주는 1위로 나왔지만 양주 막걸리 기타 는 1 단계씩 내려 간 것을 확인 할 수 있다.

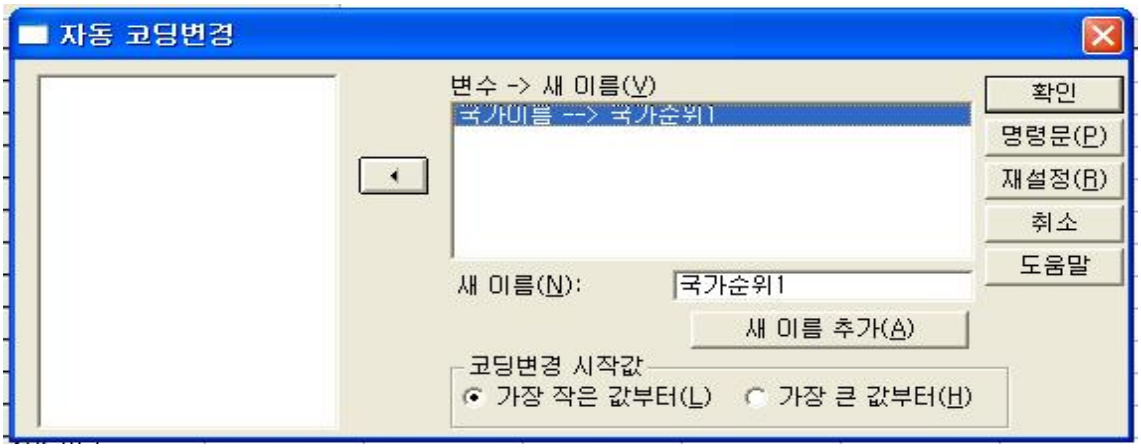
18. 자동코딩변경

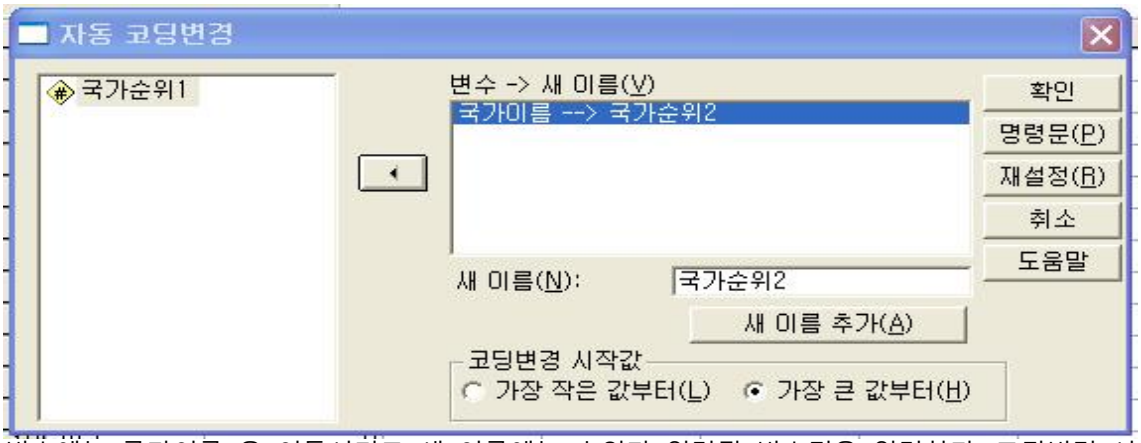
	국가이름
1	한국
2	일본
3	중국
4	러시아
5	미국
6	영국
7	독일
8	네덜란드
9	프랑스
10	체코
11	사우디아
12	네팔
13	가나
14	홍콩
15	브라질
16	멕시코
17	벨기에
18	스위스
19	스웨덴
20	인도

옆의 표를 보면 국가 이름이 나열 되어 있다. 국가 이름이 있는 경우 순위는 어떻게 정할까? 이렇게 문자변수는 통계분석시 불편한 경우가 많다 이러한 경우에 자동코딩 변경을 이용하여 순위를 정할 수 있다.



변환 -> 자동코딩변경





변수에는 국가이름 을 이동시키고 새 이름에는 순위가 입력될 변수명을 입력한다. 코딩변경 시작값은 가장 작은 값부터 와 가장 큰 값부터 가 있는데 이들을 각각 살펴보면 아래와 같다.

	국가이름	국가순위1	국가순위2
1	한국	20	2
2	일본	16	6
3	중국	17	5
4	러시아	6	16
5	미국	8	14
6	영국	14	8
7	독일	5	17
8	네덜란드	3	19
9	프랑스	19	3
10	체코	18	4
11	사우디아	11	11
12	네팔	4	18
13	가나	2	20
14	홍콩	21	1
15	브라질	10	12
16	멕시코	7	15
17	벨기에	9	13
18	스위스	13	9
19	스웨덴	12	10
20	인도	15	7

결과를 살펴보자 가장 작은 값으로 순위를 정한 국가 순위1 변수를 보면 가나의 등수 가 가장 높고 홍콩이 가장 낮은 순위 이다. 이는 오름차순 순으로 순서를 정한 것과 같다.

국가순위2 변수는 가장 큰 값부터 순위를 정한 것이다. 이때 가장 작은 값부터의 경우와 반대 이다. 홍콩이 가장 높은 순위 이며 가나가 가장 낮은 순의이다. 이것은 내림차순으로 정렬한 것과 같은 경우이다.

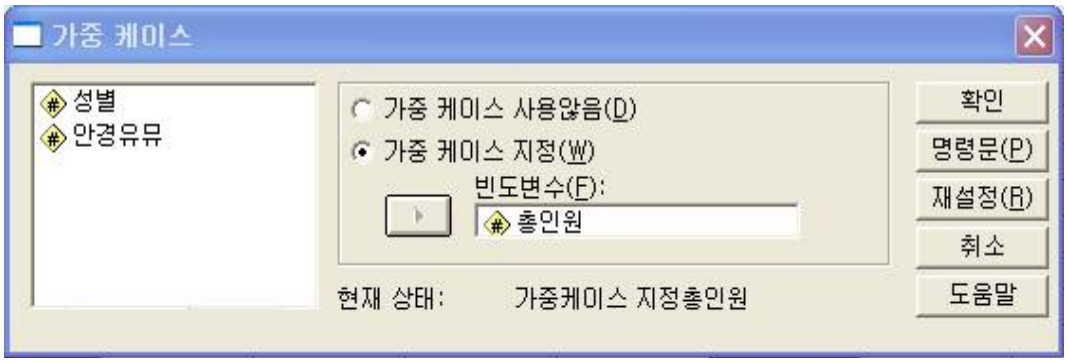
19.가중케이스

가중케이스는 특정한 변수의 관측값을 가중케이스로 지정하여 해당되는 값만큼 빈도수로 비중을 두어 처리 하는 것 이다.

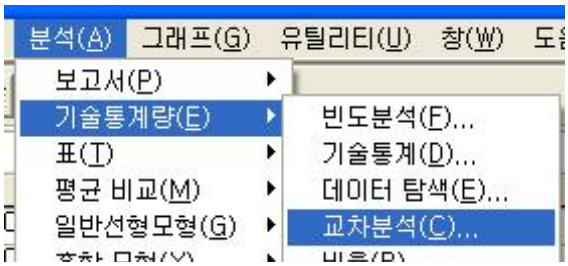
	성별	안경유무	총인원
1	1.00	1.00	20.00
2	1.00	2.00	10.00
3	2.00	1.00	15.00
4	2.00	2.00	15.00

위의 데이터가 있을 때 성별 1 = 남자 , 2 여자 이고 안경을 쓰는 사람을 1 안쓰는 사람을 2

라고 하자. 위의 데이터 같은 경우는 변수의 관측값을 가중케이스로 지정하여야 한다. 가중케이스로 지정하지 않고 빈도분석으로 통계량을 구하면 관측값 자체의 평균과 합만 나오기 때문이다. 가중케이스로 지정하는 방법은 아래와 같다.

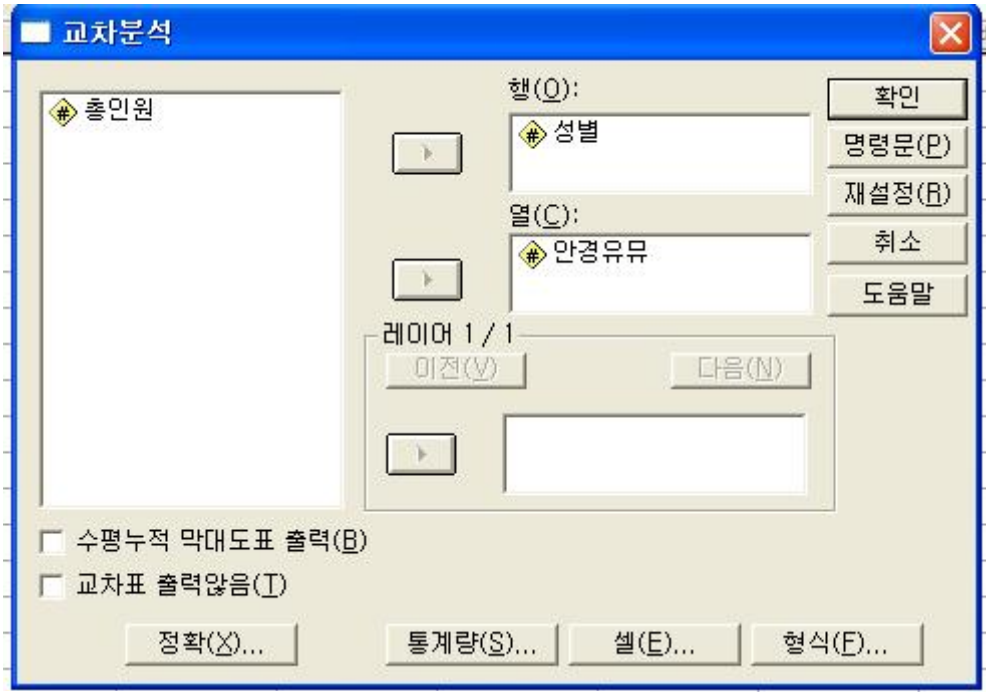


위의 그림과 같이 데이터 -> 가중 케이스 에 들어가면 가중케이스 창이 뜬다. 가중케이스 지정은 총인원으로 지정해 준다.



분석 -> 교차분석 으로 들어간다.

교차분석에서 행에는 변수 성별과 열에는 안경유무 변수를 지정해 준다.



성별 * 안경유무 교차표

빈도

		안경유무		전체
		안경유	안경무	
성별	남자	20	10	30
	여자	15	15	30
전체		35	25	60

결과창 해석 -> 총인원을 가중케이스로 지정해준 결과 관측값들이 인원수로 바뀌었다.

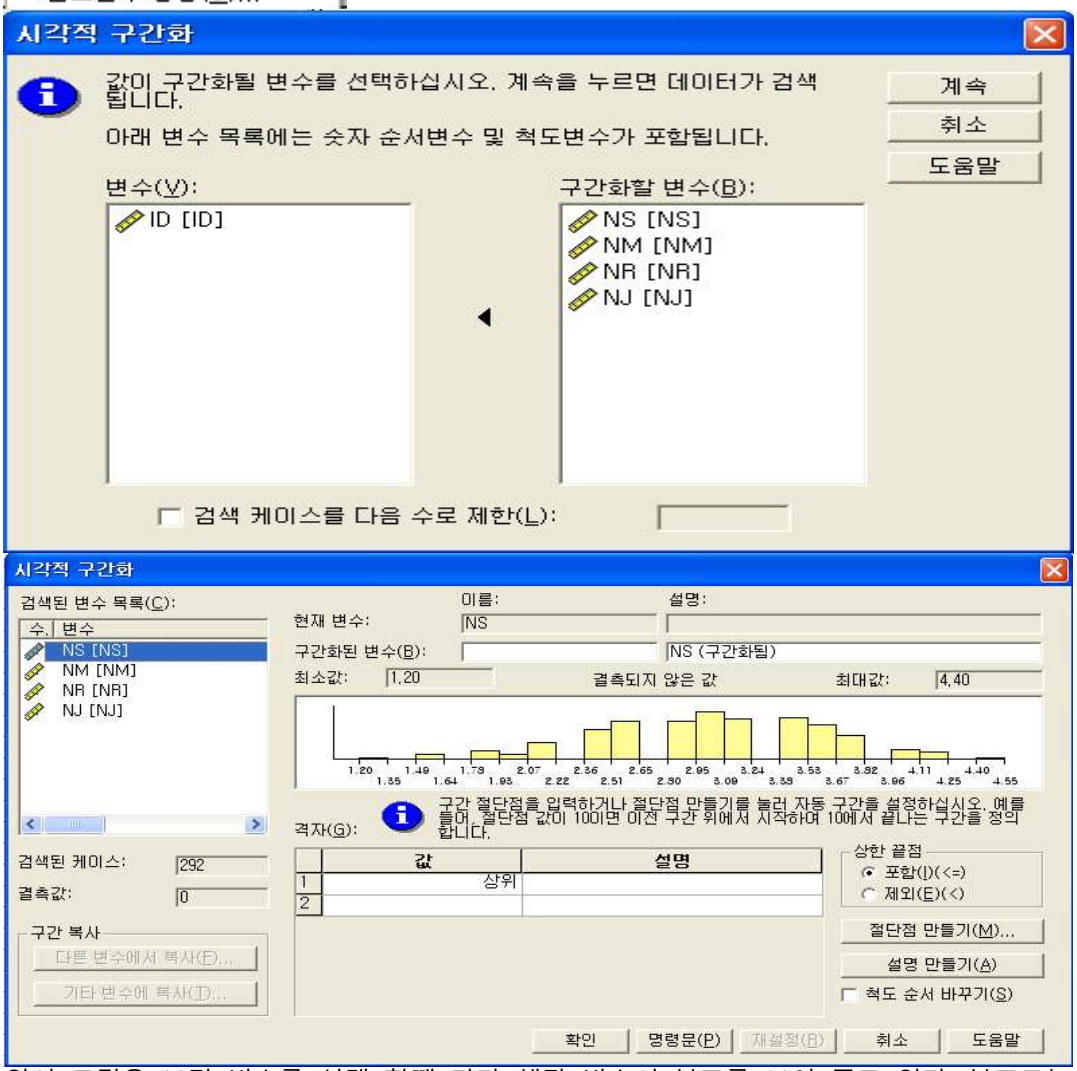
총 60명중에 안경을 착용한 사람은 35명이며 착용을 안한사람은 25 명이며, 남자 30명 중에 안경을 창용한 사람은 20명 착용하지 않은 사람은 10명 이 나왔으며 여자 30명중 안경을 착용한 사람은 15명, 착용하지 않은 사람도 15명 이다.

20. 시각적 구간화

자신의 데이터를 시각적으로 볼 수 있게 원하는 구간을 설정해줘서 분포형태를 한눈에 볼 수 있게 해주는 기능이다.

<자료> 직장만족도
변환 -> 시각적 구간회에 들어간다.

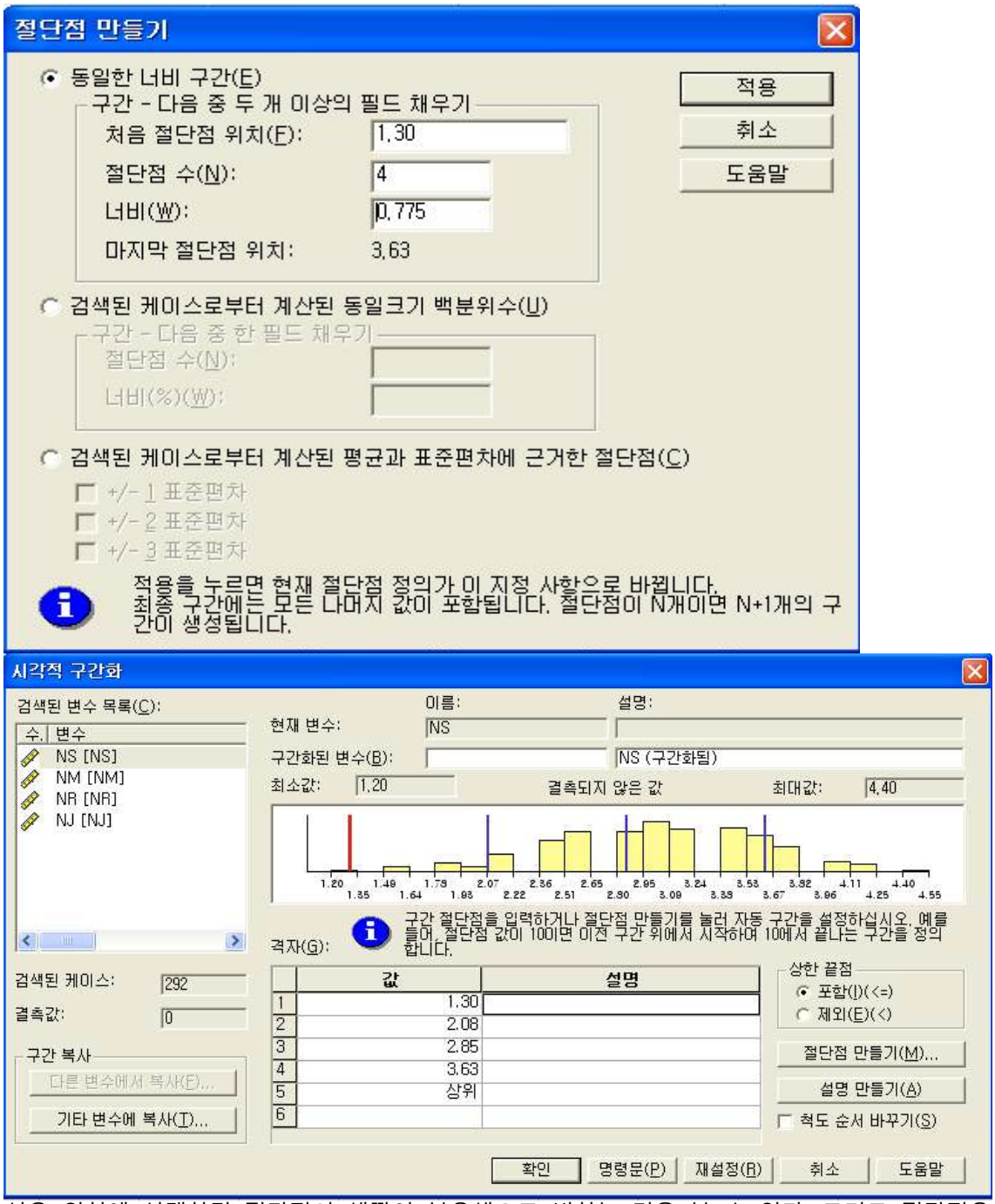
시각적 구간화에 들어가면 위와 같은 창이 뜬다. 구간화 할 변수는 직장만족도 데이터에서 각 항목의 평균값 이다.



위의 그림을 보면 변수를 선택 할때 마다 해당 변수의 분포를 보여 주고 있다. 분포도는 마치 히스토그램과 비슷하게 생겼다. 변수의 분포를 구간을 나누어서 보려면 위의 절단점 만들기 단추를 눌러 들어간다.

위의 창을 보자 처음 절단점 위치는 구간을 나누어서 볼 경우 처음 구간이라고 볼 수 있으며 절단점의 수는 몇 개의 구간으로 볼 것인지 선택 하는 것이다. 처음 절담점 위치와 절단점 수를 입력하여 주면 너비는 자동적으로 계산된다.

절단점의 위치와 절단점 수를 입력할 결과 이다. 처음 절단점의 위치는 내가 입력한 1.3 이며 절단점 수는 위의 그림에서 볼 수 있듯이 4개 이다. 그리고 위의 값, 설명 창에서 자신이 보고

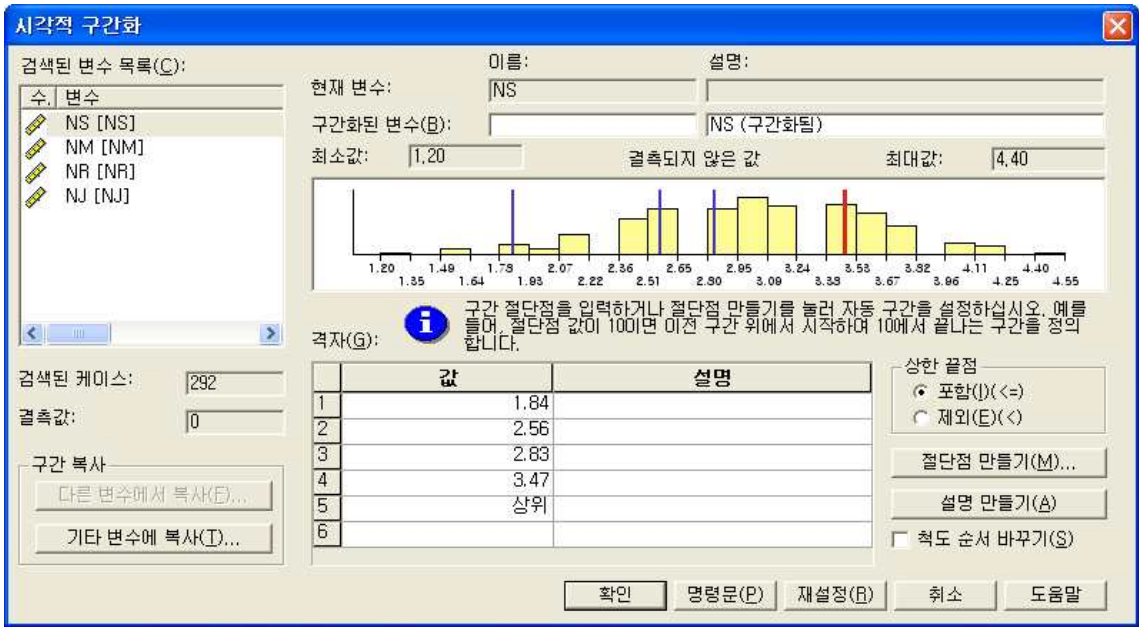


싶은 위치에 선택하면 절단점의 색깔이 붉은색으로 변하는 것을 볼 수 있다. 그리고 절담점은 마우스로 임의적으로 이동 시킬 수 있다.

위의 절단점을 마우스로 움직여 보았다. 위의 값이 바뀐 것을 알 수 있다.

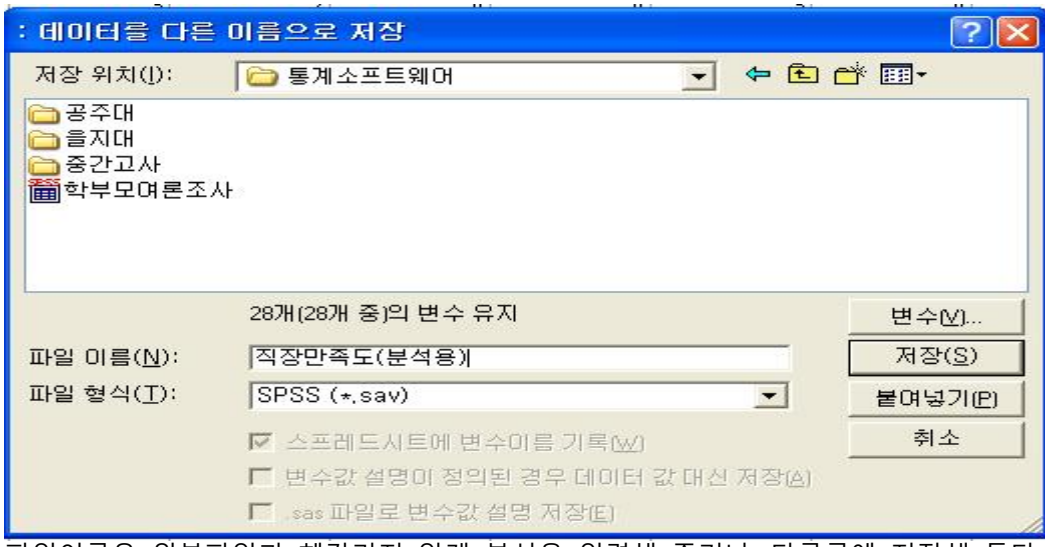
21. 저장하기

데이터를 입력하고 출력하는 것은 중요하지만 이보다 더 중요한 것은 자신이 한 작업을 저장 하는 것이다. 작업을 완벽히 했다 하여도 저장을 하지 않으면 작업을 다시 해야 하기 때문이다.



•SPSS데이터창 저장하기

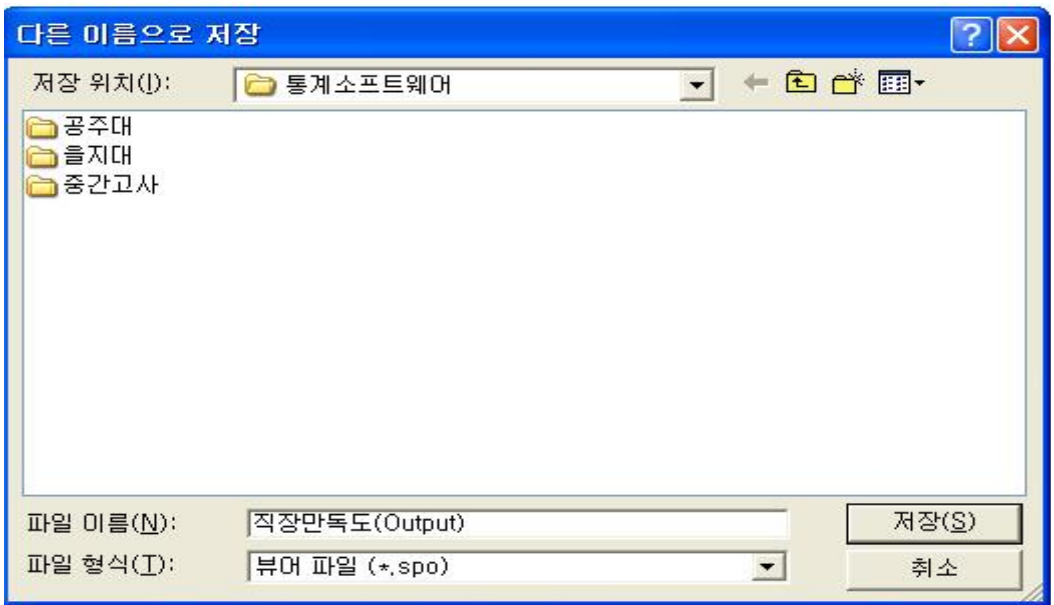
원본 파일과 겹쳐서 저장될 수 있으므로 다른 이름으로 저장해 둔다.



파일이름은 원본파일과 헷갈리지 않게 분석용 입력해 주거나, 다른곳에 저장해 둔다.

•Output 창 저장하기

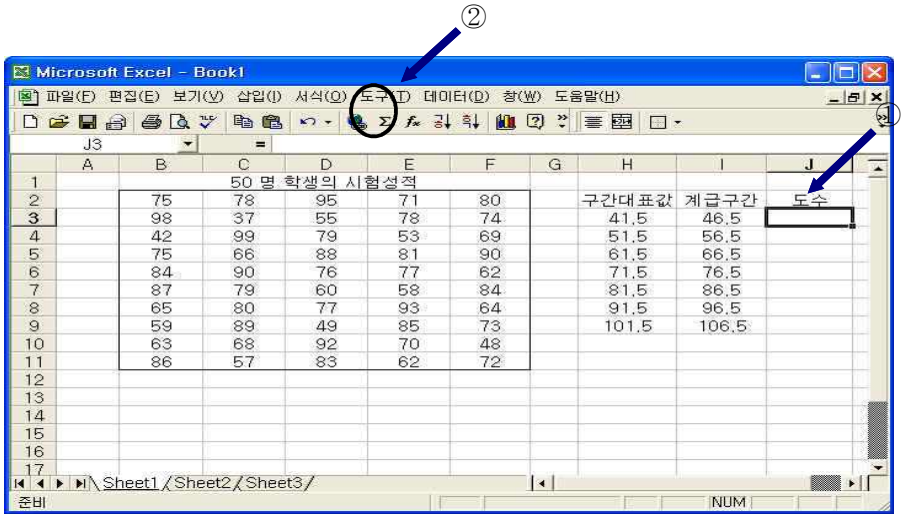
Output 창에서 다른 이름으로 저장하기에 들어간다.



Output 을 저장하는 이유는 결과창을 저장해 두지 않으면 데이터로 다시 작업을 수행 해야 하기 때문이다. 그리고 파일이름은 알아보기 쉽게 직장만족도(Output)라고 저장한다.

Ⅲ. Excel을 이용한 통계 분석 방법

1. Excel을 이용한 50명 학생의 시험성적 자료에 대한 도수분포표와 히스토그램 작성법

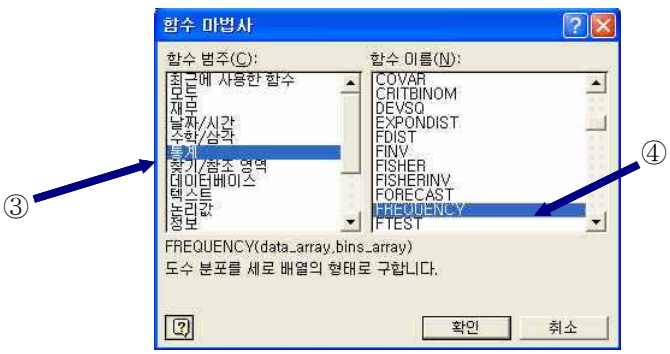


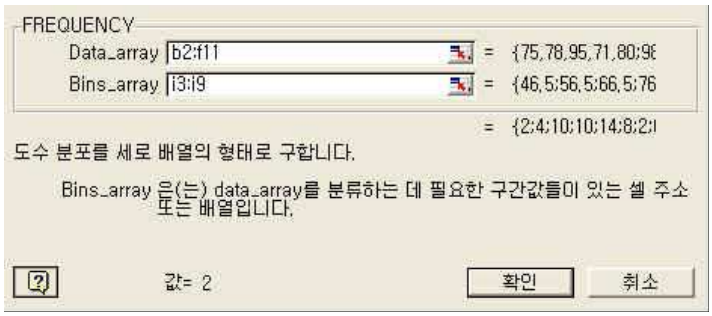
· 도수분포표의 계급값 입력

주어진 자료는 (B2:F11) 셀에 정리되어 있다. 이 자료의 히스토그램을 얻기 위해서는 일단 도수분포표를 구하여야 하는데, 여기서는 7개의 계급구간과 대표값을 정해 H열과 I열에 입력하였다. I열의 계급구간 46.5은 (46.5이하)로, I4의 56.5는 (46.5 < 값 ≤ 56.5) 등으로 인식한다. 다음은 각 계급구간에 속하는 값들의 빈도수를 구한다. (도수분포표)

· 도수의 계산

도수분포표(Frequency table)를 구하려면 우선 J3 셀을 선택하고(위의 그림에서 ①) '함수 마법사' 버튼(②)을 클릭한다. 아래 대화상자에서 함수 종류는 통계를 선택하고, 함수이름은 FREQUENCY를 선택한 후 확인 단추를 누른다.

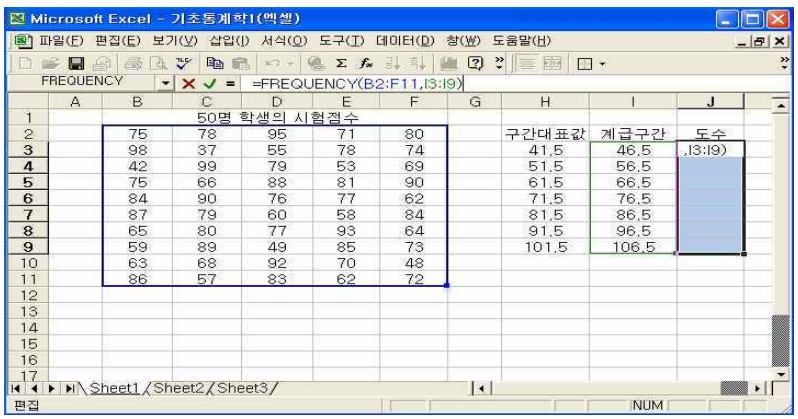




위와 같은 FREQUNCY 대화상자에서 Data_array에 B2:F11을 입력하고 Bins_array에는 집단으로 분류하는 계급구간이 적혀 잇는 셀의 주소를 적는다. 위의 대화 상자에 {2; 4; 10; 10; 14; 8; 2} 등과 각 계급구간에 따른 빈도수가 미리 보여 진다.

· 도수가 출력될 영역 선택

확인을 누르면 수식 입력줄에 =FREQUENCY(B2:F11,I3:I9)이 나타나는데, 이때 Enter 키를 치던가, 수식 입력줄 좌측의 체크버튼을 클릭하면 J3셀에만 0이 출력되므로 주의해야 한다. 도수분포는 J3~J9 셀에 출력되어야 하므로, 다음 그림과 같이 커스를 드래그해서 도수가 출력되어야 할 영역인 J3에서 J9까지 선택하고 수식입력줄을 클릭 한다.



· 출력결과의 수정

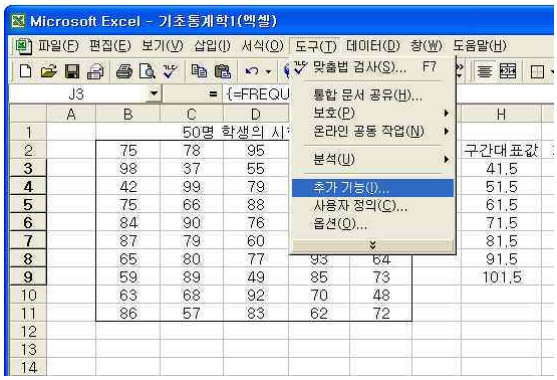
다수의 영역에 이 수식을 입력하기 위해 Ctrl + Shift + Enter 키를 누르면 오른편의 그림과 같이 J3~J9 영역에 각 계급 별 도수가 출력된다. 이때 수식 입력줄의 =FREQUENCY (B2:F11,I3:I9) 였던 수식이 {=FREQUENCY (B2:F11,I3:I9)} 로 바뀐 것을 확인할 수 있다.

<참고> 분석도구에 의한 히스토그램 출력

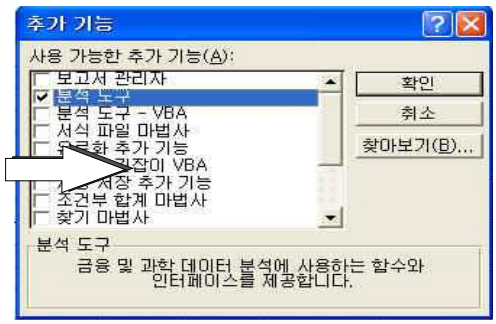
엑셀에서는 여러 가지 종류의 통계적 데이터 분석을 대화형으로 입력하여 매우 편리하게 그 출력을 얻는 '데이타 분석' 도구 기능이 있다. 그러나 통계함수를 이용하면 다른 셀에 함수를 복사할 수도 있고, 입력된 데이터를 고쳤을 때 자동으로 결과도 바뀌지만, 통계 데이터 분석은 간편하게 결과를 얻는 반면 입력값의 변화가 바로 반영되지 않는 불편함도 있다.

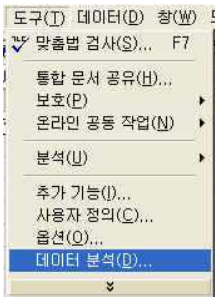
· 통계 데이타 분석도구의 설치

이 분석도구는 옵션항목으로 처음 사용자라면 먼저 이 기능을 설치하여야 사용할 수 있다. 데이터분석도구를 본인의 컴퓨터에 설치하려면 우선 '도구'메뉴에서 추가기능을 설치한다.



추가기능 대화상자에서 분석도구 버튼에 체크하고 확인을 누르면, '도구'메뉴에 '데이터분석' 항목이 추가 설치된다.

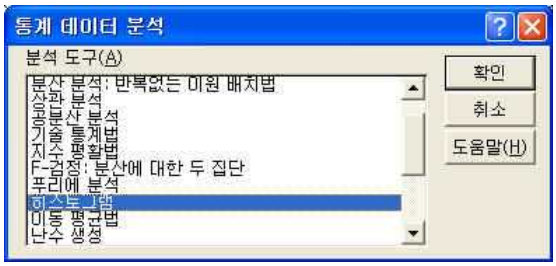




H	I	J
구간대표값	계급구간	도수
41.5	46.5	2
51.5	56.5	4
61.5	66.5	10
71.5	76.5	10
81.5	86.5	14
91.5	96.5	8
101.5	106.5	2

· 통계데이타분석 대화상자

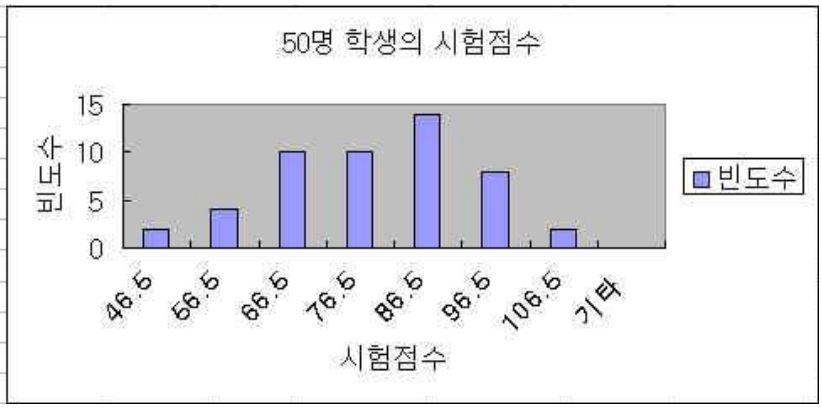
도구메뉴에서 '데이타 분석'을 클릭하면 다음과 같은 통계 데이타 분석 대화상자가 나온다. 여기서 히스토그램을 선택하고 확인단추를 누른다.



· 히스토그램 출력

히스토그램 대화상자에서 자료의 입력범위와 계급구간을 입력하고, 출력범위와 출력사항을 선택하면 된다. (여기서도 마찬가지로 우선 계급구간이 시트에 입력되어 있어야 한다) 차트출력박스에 체크하여 출력범위 K2에 계급과 빈도수(도수분포표), 히스토그램을 얻는다.

계급	빈도수
46.5	2
56.5	4
66.5	10
76.5	10
86.5	14
96.5	8
106.5	2
기타	0

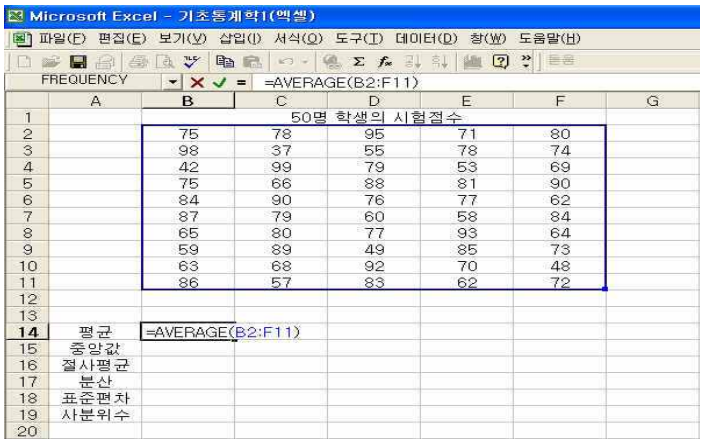


2. Excel을 이용한 50명 학생의 시험성적(p. 33) 자료에 대한 기초통계량 구하기.

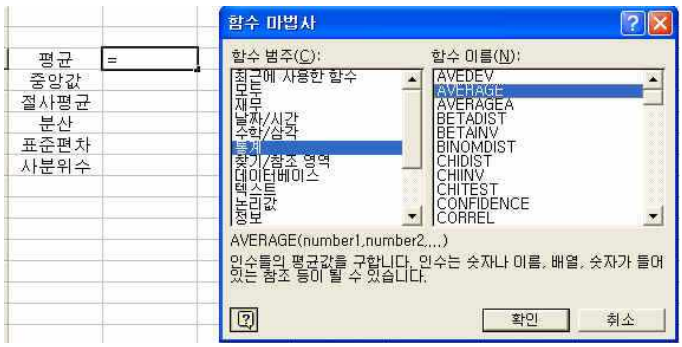
단, 구하는 방법은 수식입력창에 명령어와 데이터의 범위를 입력하는 방법과 함수마법사를 이용하는 방법이 있다.

(a) 평균(Mean)

값을 구해서 넣을 셀(B14)에 직접 '=AVERAVE(B2:F11)'을 입력하든지, 수식입력창에 입력해서 Enter키를 치면 된다.



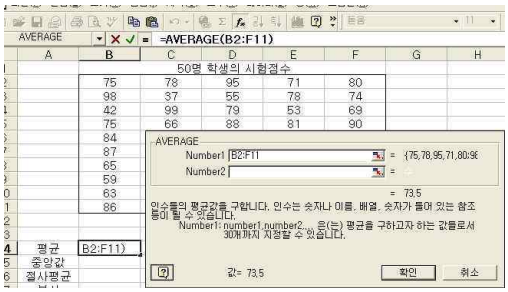
또는 함수 마법사에서 AVERAGE 함수를 불러 데이터의 범위를 입력하면 평균값 73.5가 구해진 것을 볼 수 있으며 확인 버튼을 누르면 B14 셀에 값이 입력된다.

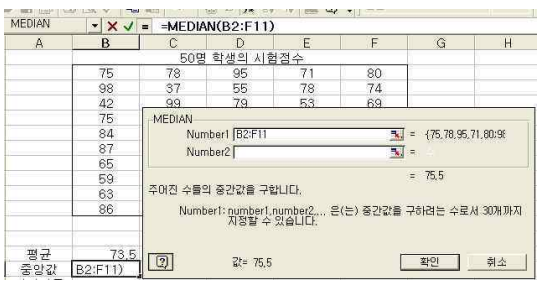


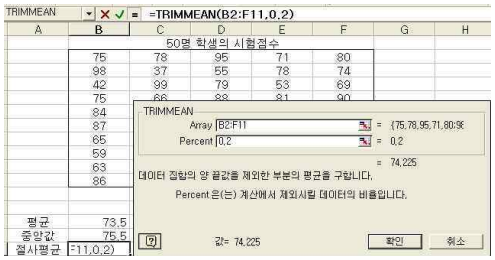
같은 방법으로, 중앙값(=MEDIAN(B2:F11)), 75.5도 구하면 된다.

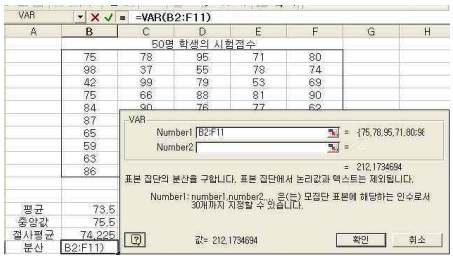
(b) 절사평균(Trimmed mean)

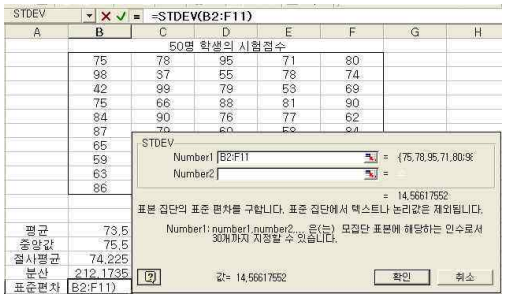
절사평균(Trimmed mean)은 자료의 양극단에서 일부(100 × α%)를 제외한 나머지 값들의 평균을 구한다. 여기서 α=0.1인 10% 절사평균을 구하려면 양극단에서 10%의 자료 5개씩을 제외한 40개 자료의 평균을 구한다. 수식입력창에 '=TRIMMEAN(B2:F11), 0.2'을 입력해서 구하든지 함수 마법사를 이용해서 구하면 74.225의 절사평균을 구한다.











(c) 분산과 표준편차

모분산과 표본분산의 분모가 다르기 때문에 Excel에서는 Population(모집단)의 P를 덧붙여 모분산과 표본분산을 구별하여 준다. 즉, 모분산은 '=VARP(데이타범위)'이고 표본분산은 '=VAR (데이타범위)'이다. 모표준편차도 '=STDEVP(데이타범위)'이고 표본표준편차는 '=STDEV(데이타범위)'이다. 위에서 구한 방법과 같이 표본분산과 표본표준편차를 구하면 다음과 같이 나온다.

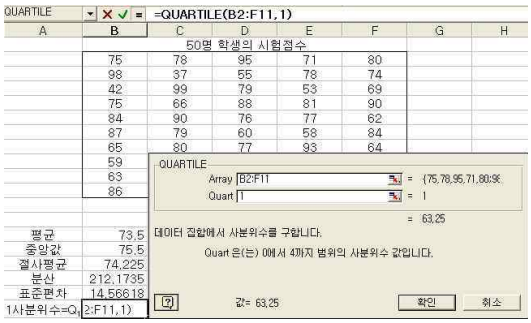
(d) 사분위수(Quartile), 백분위수(Percentile)

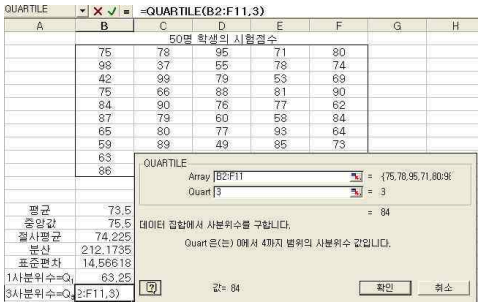
자료를 순서대로 늘어놓고, 작은 값부터 시작해 25% 위치에 있는 값을 제 1사분위수(Q_1, lower quatile) 혹은 제 25백분위수(25-th percentile), 50% 위치에 있는 값을 제 2사분위수 (Q_2), 또는 중앙값(Median), 제 50백분위수(50-th percentile), 75% 위치에 있는 값을 제 3사분위수(Q_3, upper quartile), 제 75백분위수(75-th percentile)라 한다. 구하는 방법은 수식입력창 에 다음과 같이 입력하든지, 함수마법사를 이용한다.

제 1 사분위수 : '=QUARTILE(데이타 범위, 1)' 또는 '=PERCENTILE(데이타범위, 0.25)'
제 2 사분위수 : '=QUARTILE(데이타 범위, 2)' 또는 '=PERCENTILE(데이타범위, 0.50)'
제 3 사분위수 : '=QUARTILE(데이타 범위, 3)' 또는 '=PERCENTILE(데이타범위, 0.75)'

* 사분위수 범위(Interquartile range), $IQR = Q_3 - Q_1$

사분위수 범위(IQR) : '=QUARTILE(데이타 범위, 3)-QUARTILE(데이타 범위, 1)'

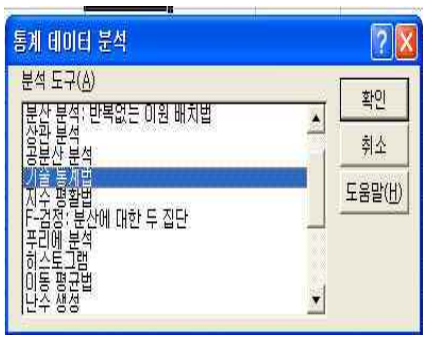


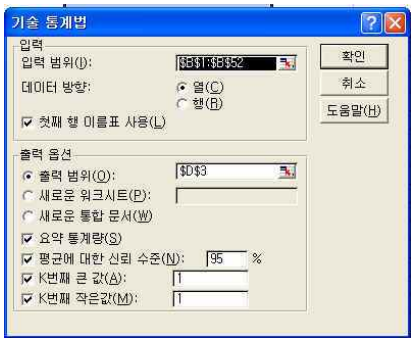


사분위수 범위(IQR)는 오른쪽 그림과 같이 입력하고자 하는 셀에 다음과 같이 입력하고 Enter 키를 치면 구해진다.

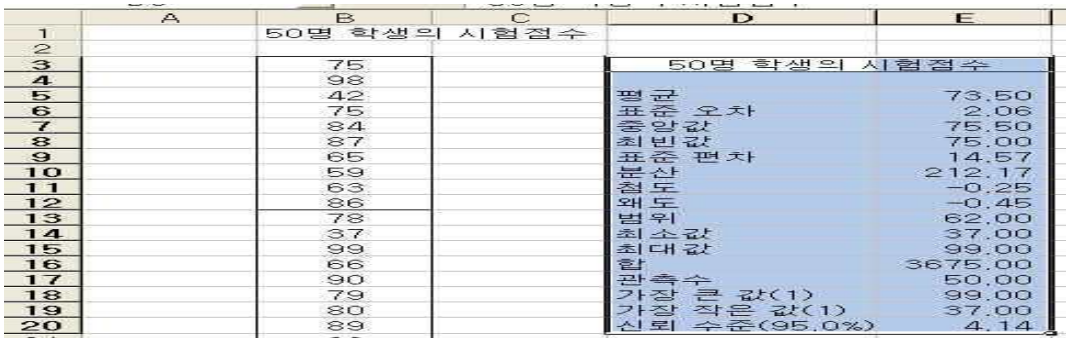
3. '도구'에서 '데이터분석'을 이용한 50명 학생의 시험성적자료에 대한 기초통계량 구하기.

Excel에 내장되어 있는 데이터분석 도구로 분석하는 방법이다. 주의할 것은 자료가 한 열이나 한 행으로 배열되어 있어야 한다는 점이다. 여기서는 자료를 B1에서 B52 셀까지 재배치하였다. 우선 '도구' 메뉴에서 '데이터분석'을 선택하면 다음의 대화상자가 나오는데, 여기서 기술통계법을 선택 후 확인 버튼을 누른다.





자료가 배열된 B1:B52(제목 포함) 셀을 입력범위로 입력한다. 데이터 방향은 열이고, 첫째 행은 자료수치가 아닌 자료의 이름(50명 학생의 시험점수)이므로 첫째 행 이름표 사용 박스에 체크한다. 결과는 같은 시트 내에 D3셀에 출력한다.



	A	B	C	D	E
1		50명 학생의 시험점수			
2					
3		75		50명 학생의 시험점수	
4		98			
5		42		평균	73.50
6		75		표준 오차	2.06
7		84		중앙값	75.50
8		87		최빈값	75.00
9		65		표준 편차	14.57
10		59		분산	212.17
11		63		첨도	-0.25
12		86		왜도	-0.45
13		78		범위	62.00
14		37		최소값	37.00
15		99		최대값	99.00
16		66		합	3675.00
17		90		관측수	50.00
18		79		가장 큰 값(1)	99.00
19		80		가장 작은 값(1)	37.00
20		89		신뢰 수준(95.0%)	4.14

여기서, 출력결과 중에서 설명하지 않은 통계량을 간략히 소개한다.

ⓐ 표준오차(standard error)는 $\bar{x}$의 표준편차를 의미한다. 즉, x의 표준편차를 $\sqrt{n}$으로 나누어준 값(=14.57/$\sqrt{50}$)이다.(8장의 표본분포에서 다시 설명)

ⓑ 왜도(skewness)는 분포의 좌우대칭성에 관한 측도로 대칭이면 0, 왼쪽으로 긴 꼬리를 가지면 음수, 그 반대면 양수 값을 가진다.

ⓒ 첨도(Kurtosis)는 분포모양이 최빈값 부근에서 얼마나 뽀족한가를 표현하는 측도이다. 뒤에서 배울 표준정규분포의 경우 3의 값을 가지며, 이 보다 크면 더 날카로우며, 작으면 뭉툭함을 의미한다.

ⓓ 신뢰수준(95%)은 9장의 구간추정에서 다룰 내용으로, 간략히 표현하면 실제적인 평균(모평균)이 (표본평균±4.14) 구간 안에 있을 거라고 95%정도 신뢰한다는 의미이다.