CHAPTER 6

기초 통계량 분석

분류형(범주형) 변수 데이터에 대한 정리 방법으로는 숫자 요약인 빈도 분석과 그래프 요약인 파이 차트, 바 차트가 이용된다. 측정형 변수에 대한 숫자 요약은 일반적으로 자료의 중앙 위치와 자료의 흩어진 정도를 나타내는 두 개의 값으로 축약된다. 즉, 크기 n 개의 데이터의 가진 정보가 2 개 숫자 요약으로 축약(data reduction) 된다. 데이터의 중앙 위치에 대한 통계량 평균(mean)이나 중앙값(median)이 있고 흩어진 정도를 측정하는 통계량으로는 표준 편차(standard deviation), 범위(range), IQR 등이 있다. 평균은 표준 편차와 중앙값은 범위나 IQR 과 함께 발표되며 이들을 기초 통계량(elementary statistic)이라 한다. 측정형 변수에 대한 그래프 방법으로는 히스토그램(histogram), 줄기-잎 그림(stem and leaf plot), 상자-수염 그림(box-whisker plot) 등이 있는데 가장 유용한 것은 상자 수염 그림이다.

설문 조사 데이터의 경우 측정형 변수는 (1)측정 가능한 것에 대한 개방형 문항(예: 소득, 연령)이나 (2)리커드 척도 문항이다. 일반적으로 설문 조사 데이터 분석의 경우 줄기-잎 그림, 상자 수염 그림과 같은 그래프 요약은 하지 않으므로 본 책에서도 측정형 변수에 대한 그래프 요약은 생략하기로 한다. 리커드 척도 문항의 경우에도 1-5 점으로 계량화 하여 측정형 변수 분석 방법을 사용하기도 하지만 빈도 분석(만족 비율 혹은 불만족 비율)만으로 보고서를 작성하기도 한다.

6.1. 기초 통계량

6.1.1. 중앙 위치

(1)평균(mean)

평균은 관측치의 절대 크기의 중앙이므로 모든 관측치를 더한 값을 관측치 수(n)로 나눈 값이므로 산술 평균(arithmetic average)과 동일한 개념이다. n 개의 관측치 $(x_1, x_2, \ldots, x_n)$ 의 평균은

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

이다.

(예제)(1, 4, 6, 5, 6, 2)의 평균은 (1+4+6+5+6+2)/6=4 이다.

(2)중앙값 (median)

자료의 크기 중심인 평균과는 달리 중앙값은 자료의 순서의 중심이다. 자료의 중앙값을 계산하기 위하여 자료의 순서 통계량(order statistics)을 먼저 구해야 한다. 순서 통계량이란 관측치를 크기 순으로 정렬한 후 제일 작은 값부터 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 으로 표기하고 이를 순서 통계량(order statistics) 이라 한다. 만약 표본의 크기 n 이 홀수이면 $M = x_{(\{n+1\}/2)}$, 짝수이면 $M = [x_{(n/2)} + x_{(\{n+2\}/2)}]/2$ 이 중앙값이 된다.

①순서 통계량 (order statistics)

크기가 n 인 표본 자료 관측치(observation) $(x_1, x_2, \ldots, x_n)$ 을 크기 순으로 정렬한 후 가장 작은 관측치를 $x_{(1)}$, 가장 큰 관측치를 $x_{(n)}$ 이라 표현하고 $x_{(1)}$, $x_{(2)}$, .., $x_{(n)}$ 을 순서 통계량이라 한다. 순서 통계량에 대해 다음이 성립한다.

ⅰ) $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$

ⅱ)최소값(min): $x_{(1)}$

ⅲ)최대값(max): $x_{(n)}$

ⅳ)범위(range): $x_{(n)} - x_{(1)}$

(예제)크기 6 인 표본 관측치 (1, 4, 6, 5, 6, 2)의 순서 통계량은 (1, 2, 4, 5, 6, 6)이다. 최소값은 $x_{(1)}=1$, 최대값은 $x_{(6)}=6$ 이고 범위는 $x_{(6)}-x_{(1)}=5$ 이다.

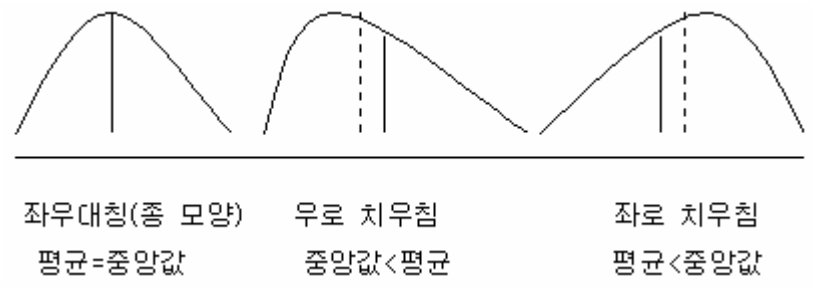
②깊이(depth)

사분위 값을 구하려면 자료의 깊이 (depth) 개념을 이용하면 편리하다. (Tukey 제안) 관측치를 크기 순으로 정렬한 후 각 양쪽 끝에서 1 부터 번호를 매겨 그 번호를 자료의 깊이라 정의한다. 즉 최대값, 최소값의 깊이는 각 1 이다. Depth(중앙값=M)= $(n+1)/2$ 이고 사분위 깊이는 Depth(Q1)=Depth(Q3)=([Depth(M)]+1)/2 이다. (기호)[x]=x 를 넘지 않는 최대 정수.

(예제)크기 6 인 표본 관측치 (1, 4, 6, 5, 6, 2)의 중앙값 길이는 (6+1)/2=3.5 이고, 사분위 길이는 ([3.5]+1)/2=(3+1)/2=2 이다.

(3)평균과 중앙값 비교

자료의 측정치 중 다른 측정치에 비해 아주 크거나 아주 작은 측정치(극단치)가 존재하는 경우 순서의 중심인 중앙값과는 달리 크기의 중심인 평균은 극단치가 존재하는 쪽으로 치우치게 된다. 극단치 중 수집 자료에 포함하여 분석하기에는 부적절하게 크거나 작은 측정치를 이상치(outlier)라 한다. 다음은 확률 분포 함수 (히스토그램) 형태에 따른 중앙값과 평균의 관계이다.



(예제)크기 10 인 자료 (1, 2, 3, 4, 5, 6, 7, 8, 9, 55)의 경우 평균과 중앙값을 구하면 평균은 $\bar{x}=10$ 이고 중앙값은 $M=[x_{(5)}+x_{(6)}]/2=(5+6)/2=5.5$ 이다. 중앙값 길이는 (10+1)/2=5.5 이고 사분위 길이는 ([5.5+1])/2=3 이다. 그러므로 제 일사분위 Q_1 은 $x_{(3)}=4$ 이다.

위의 자료에서 중앙 위치를 나타내는 숫자 요약으로는 중앙값인 5.5 가 평균인 10 보다 더 합리적이다. 이와 같이 극단치가 존재하는 경우 자료의 중앙 위치를 나타내는 통계량

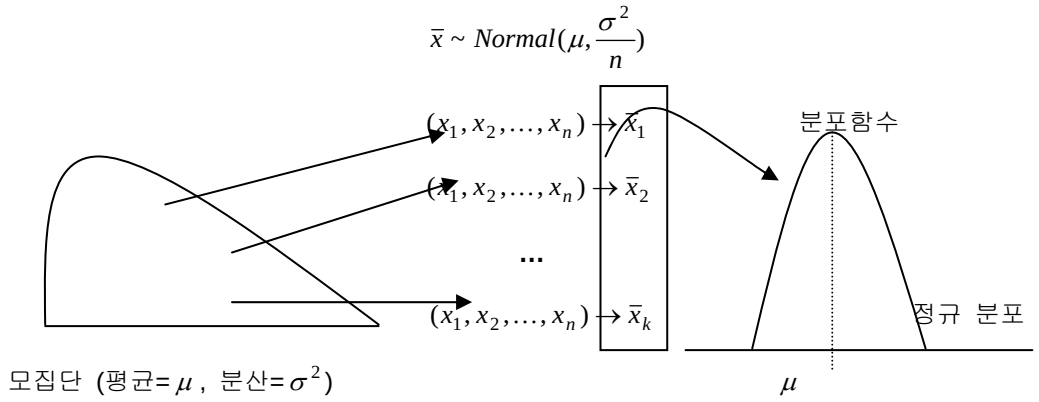
으로 중앙값이 평균보다 더 이상적인 값이다. 그리고 자료에 극단치가 존재하지 않으면 중앙값과 평균이 거의 일치하므로 자료의 중앙 위치에 대한 요약 값이라는 측면에서는 중앙값이 평균보다 더 합리적이다.

(4)평균 사용 이유

데이터의 분포가 좌우 대칭이 아닌 경우 중앙값이 평균에 비해 더 합리적인 중앙 위치 값임에도 불구하고 평균을 사용하는 이유는 다음과 같다.

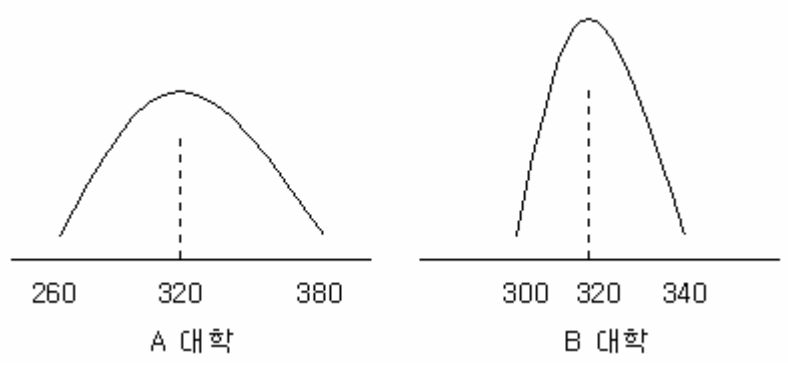
평균에 대한 수학적 전개가 용이하고 중앙값과는 달리 평균의 분포 함수를 쉽게 구할 수 있기 때문이다. 통계학에서 대수의 법칙(Law of Large Number)과 함께 가장 널리 사용되는 중심극한정리(Central Limit Theorem)에 의하면 표본의 크기 n 인 큰 경우(대표본) 모집단의 분포에 상관없이 표본 평균의 분포 함수는 정규분포를 따른다. 통계량의 분포함수를 구할 수 있는 것이 장점이 되는 이유는 통계량의 분포를 알아야 모수에 대한 신뢰구간을 구하거나 모수에 대한 가설 검정이 가능하다**.**

⏭ 중심 극한 정리 (Central Limit Theorem): 표본의 크기 n 이 큰 경우 (20~30 이상) 모집단의 분포 형태에 관계없이 표본 평균의 분포는 정규분포를 따른다.



6.1.2. 흩어진 정도 (산포도: spread)

측정형 변수에 대한 분석에서 중앙 위치만을 알고 있으면 자료 전체에 대한 정보를 얻는데 한계를 느낀다. 왜냐하면 자료의 측정치들의 흩어진 정도나 측정치들이 가질 수 있는 값의 범위는 얼마인지는 자료의 중앙 위치만으로 알 수 없다. 같은 평균을 갖더라도 흩어진 정도에 차이가 있으면 자료의 특성은 다르다.



위의 그림은 두 대학의 수능 성적 자료 분포 함수이다. 자료의 중앙 위치 값(320 점)만으로 보면 두 대학 신입생들의 성적은 동일하다고 말할 수 있을 것이다. 그러나 확률 분포 함수를 살펴보면 두 대학 신입생들의 수능 성적은 전혀 다른 정보를 가지고 있음을 쉽게 알 수 있다. A 대학의 신입생은 매우 우수한 학생이 있고 상대적으로 성적이 낮은 학생들도 있으나 B 대학 신입생들의 성적은 큰 차이를 보이지 않고 평균 점수에 집중되어 있다.

(1)범위와 IQR

측정 자료의 최대값과 최소값의 차이를 범위(range)라 한다. 범위는 계산이 편리하다는 장점이 있으나 범위 계산할 때는 두 측정치(최대값, 최소값)만 사용되므로 다른 측정치들의 정보가 전혀 고려되지 않고 이상치나 극단치가 존재하는 경우 범위가 커지는 단점이 있다. 크기가 10 인 자료 (1, 2, 3, 4, 5, 6, 7, 8, 9, 55)와 같이 극단치가 존재하면 자료의 범위는 54 로 커지게 된다.

극단치가 존재하는 경우 산포도의 계산 값인 범위가 커지는 단점을 보완하기 위하여 삼사분위 값과 일 사분위 값의 차이인 IQR 값을 산포도로 사용하기도 하지만 이것 역시 다른 측정치의 정보는 무시되는 단점을 가지고 있다.

p%-percentile(p 백분위 값) 데이터 관측치 중 p%가 그 값보다 작고 (1-p)%가 그 값보다 클 때 그 값을 p% 백분위 값이라 한다. 일사분위(First Quartile, Low Quartile) Q1은 관측치 중 25%가 그 값보다 작고 75%가 그 값보다 클 때 그 값을 일사분위라 정의한다. 이사분위(Second Quartile, Median) Q2은 관측치 중 50%가 그 값보다 작고 자료의 50%가 그 값보다 클 때 그 값을 이사분위라 정의하고 이를 특히 중앙값이라 한다. 삼사분위(Third Quartile, Upper Quartile) Q3는 관측치 중 75%가 그 값보다 작고 자료의 25%가 그 값보다 클 때 그 값을 삼사분위라 정의한다. 그리고 (Q3-Q1)을 자료의 IQR(Inter-Quartile Range)라 한다.

(예제)크기 6 인 표본 관측치 (1, 4, 6, 5, 6, 2)의 순서 통계량은 (1, 2, 4, 5, 6, 6)이다. 중앙값의 깊이는 $(6+1)/2 = 3.5$ 이고 일사분위와 삼사분위 깊이는 $([3.5]+1)/2 = 2$ 이다. 그러므로 일사분위 $Q_1 = x_{(2)} = 2$, 이산분위 중앙값 $Q_2 = (4+5) = 4.5$, 삼사분위 $Q_3 = x_{(5)} = 6$ 이다. 그러므로 $IQR = 6-2 = 4$ 이다.

자료의 분포가 한쪽으로 치우쳐 있지 않은 종 모양(bell shaped)인 경우 자료의 대략적인 산포도를 측정치로 범위나 IQR 가 계산되기도 하지만 IQR 은 나무 상자 그림에 주로 사용되고 자료의 범위보다는 최대값과 최소값을 주로 사용하므로 그 값 자체만으로는 자료 정리에 거의 이용되지 않는다.

(2)표준편차 및 분산

자료의 산포도 측정치로 가장 많이 사용되는 분산(variance)은 각 측정치(x_i)들이 평균($\bar{x}$)으로부터 떨어진 정도(차이)를 제곱한 값들을 합한 후 자료의 수로 나눈 값이고 표준편차(standard deviation)는 분산의 양의 제곱근 값으로 정의된다. 모집단 전체 자료의 분산을 모집단 분산(σ^2), 표준 편차를 모집단 표준 편차(σ : sigma)라 하고 표본 자료의 경우는 표본 분산(s^2), 표본 표준 편차(s)라 한다.

표준편차나 분포는 순서 통계량에 의해 자료의 흩어진 정도를 나타내는 범위나 IQR 과는 달리 측정치들이 평균으로부터 떨어진 정도를 숫자로 나타낸다. 측정치와 평균의 차이를 제곱함으로써 멀리 떨어질수록 자료의 흩어진 정도에 더 많은 영향을 미치게 된다. 즉 표준 편차나 분산은 자료들이 평균으로부터 평균적으로 얼마나 떨어져 있는지를 나타내는 수치이다.

⏭ 모집단 분산, 표준 편차 계산식: $\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - u)^2}{N}$, $\sigma = \sqrt{\sigma^2}$

⏭ 표본 분산, 표준 편차 계산식: $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2]$, $s = \sqrt{s^2}$

(예제)크기가 6 인 자료 (1, 4, 6, 5, 6, 2)의 분산과 표준 편차를 구해보자.

$$s^2 = \frac{1}{6-1}\sum_{i=1}^{6}(x_i - 4)^2 = \frac{(1-4)^2 + (4-4)^2 + \cdots + (2-4)^2}{5} = 4.4$$

$$s = \sqrt{4.4} = 2.098$$

※(n-1)로 나누는 이유?

①통계학에서는 모집단 분산의 추정치로 표본 분산을 사용하는데 (n-1)로 나누어 계산한 표본 분산이 불편성을 (unbiased)을 갖게 된다. 만약 표본 분산을 계산할 때 표본 크기 n 으로 나누면 편기(biased) 추정치가 된다.

②자유도(degree of freedom)에 관한 문제이다. 자유도는 자료로부터 통계량을 계산하는 경우 이 통계량에 대해 독립적인 정보를 갖는 자료 측정치의 개수이다. 독립적인 정보란 의미는 자료의 측정치 중 그 통계량을 계산하는데 필요란 측정치가 몇 개인가 하는 것이다. 그러므로 자료의 수와 자유도는 다를 수 있다. 예를 들어보자. 표본 자료 (7, 8, 9, 10, 11)에서 통계량 평균을 구하는 경우 5 개 측정치가 모두 필요하나 표준 편차를 구하는 경우는 5 개 측정치 값이 모두 필요하지는 않다. 표준 편차를 구하기 위해서는 우선 평균을 먼저 구하므로 5 개 측정치 중 하나가 없어도 그 값을 알 수 있다. 위 표본 자료의 평균은 9 이므로 5 개의 측정치 중 10 이 없어져도 (7, 8, 9, 11), 4 개의 측정치와 평균 9 만 있으면 없어진 측정치가 10 인지 알 수 있다. 그러므로 크기가 n 인 표본 자료의 분산의 자유도는 (n-1)이 된다.

(3)평균과 표준 편차

측정형 자료에 대한 기초 통계량을 정리할 때 평균과 표준 편차를 사용하는 것이 좋다. 분산은 측정치들을 제곱한 값들을 합한 것이므로 측정치들을 단순히 합한 평균과는 단위가 다르지만 제곱근을 구한 표준 편차는 평균과 단위가 일치하기 때문이다.

(4)표준 편차 해석

표준 편차는 자료의 흩어진 정도를 나타내는 수치이므로 이를 이용하여 측정 자료를 해석하는데 이용할 수 있다. 예를 들어, 수능 성적이 동일한 학생 100 명을 대상으로 새로운 학습 방법을 적용한 후 100 문제를 풀게 하여 평균이 80 점이고 표준 편차 1 을 얻었다고 하자. 성적의 분포가 종 모양이라면 100 명 학생 성적은 77~83 점에 대부분 분포할 것이다. 그러므로 학생들의 학습 능력은 동일하다고(homogeneous) 생각할 수 있다. 만약 표준 편차가 10 이라면 학생들의 점수는 40~100 점 사이에 있으므로 학습 능력은 다소 차이가 있음을 알 수 있다.

(예제)두 모집단의 분포를 비교하는 경우 표준 편차가 더 유용하게 이용된다. 능력이 동일한 두 의사에게 진료를 받기 위해 기다리는 시간을 조사하였더니 다음과 같았다 (단위: 분). 어느 의사에게 가서 진료를 받는 것이 유리할까? 경험적 법칙(Empirical Rule)에 의하면 의사 A 의 경우 기다리는 시간은 8±3*0.26, 즉 7.22~8.78 분 사이이

고 의사 B는 4.14~11.87 분이다. 의사 A의 경우 대부분 8분 정도 기다리나 의사 B의 경우 운이 좋으면 5 분 이내 진찰을 받을 수 있으나, 운이 나쁘면 11 분 이상 기다려야 한다. 여러분이면 어느 의사에게 진찰을 받으러 가겠는가?

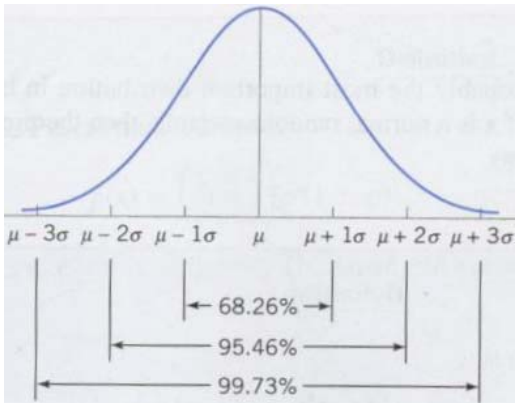
	데이터										평균	표준편차
의사 A	**9**	**8**	**7**	**7**	**8**	**9**	**9**	**7**	**8**	**8**	**8**	**0.26**
의사 B	**14**	**7**	**8**	**11**	**15**	**4**	**3**	**6**	**7**	**5**	**8**	**1.29**

참고 Empirical Rule(경험적 규칙): 자료의 분포가 좌우 대칭이면 다음이 성립한다.

(1)전체 관측치 중 적어도 68.3%는 평균±표준편차 범위에 있다

(2)전체 관측치 중 적어도 95.5%는 평균±2*표준편차 범위에 있다

(3)전체 관측치 중 적어도 99.7%(대부분)는 평균±3*표준편차 범위에 있다



(5)변동 계수 (CV: Coefficient of Variation)

측정 단위에 따라 표준 편차의 값의 크기가 달라지므로 단위가 다른 두 집단을 비교하는 경우 두 표준 편차의 단위를 같게 할 필요가 있다. 이를 위하여 표준편차를 평균으로 나눈 값에 100 을 곱한 값을 변동 계수(CV: Coefficient of Variation)라 하고 상대 변동(분산) 개념으로 정의하고 있다.

⏭ 변동 계수: 표본 자료의 평균이 $\bar{x}$, 표준 편차가 s인 경우 $CV = \frac{s}{\bar{x}} \times 100(\%)$

(예제)고등학교 3 학년인 A, B 학생의 공부 습관을 조사하여 한 달간 조사하여 A 학생은 평균 3 시간, 표준 편차는 0.5, B 학생은 6 시간 표준 편차 0.8 인 결과를 얻었다. 어느 학생이 더 꾸준히 공부하는 습관을 가지고 있을까? 이에 대한 답을 위해 변동 계수를 계산하면 된다. 위의 계산 결과 B 학생이 더 꾸준히 공부하는 습관을 가지고 있다고 결론 지을 수 있다.

A 학생 공부시간에 대한 변동 계수 $= 0.5/3 \times 100(\%) = 16.7\ (\%)$
B 학생 공부시간에 대한 변동 계수 $= 0.8/6 \times 100(\%) = 13.3\ (\%)$

(6)표준 편차 와 표준 오차

표준 편차 (standard deviation)는 자료의 표준 편차이고 표준 오차 (sampling error)는 표본 평균의 표준 편차이므로 표준 오차는 $\sigma/\sqrt{n}$ (모집단 표준 편차를 알 경우) 혹은 $s/\sqrt{n}$ (모집단 표준 편차 모를 경우) 이다.

6.2. 설문 분석에서 기초 통계량 사용

리커드 척도나 열린 문항(open item: 측정형 변수인 소득, 판매액, 키 등을 주관식 형태로 물어보는 경우) 형태로 조사된 문항을 분석하는 방법으로 기초 통계량 분석을 할 수 있다. 이는 문항의 응답 척도가 측정형이기 때문이다. 설문 데이터 기초 통계량은 중앙값이나 범위가 아니라 **평균과 표준 편차를 사용**한다. 설문 조사에 기초 통계량 분석이란 각 문항에 대해 평균과 표준편차를 구하여 문항 자료(응답 결과)를 정리하는 방법이다. 응답자들의 생각의 평균이 어디이고 그들의 생각이 얼마나 흩어져 있는지를 측정하는 것이다.

설문 분석에서 기초 통계량을 구하는 것은 응답자들의 평균 리커드 점수의 평균(5 점 척도의 경우 5 점 만점에 몇 점 정도)이나 응답자들의 선택의 흩어짐(산포도; 표준편차)을 구하여 응답자들이 어느 정도 만족하고 있고 그들의 의견이 얼마나 흩어져 있는지 알 수 있다. **표준편차가 크다는 것은 사람들의 의견이 많이 상이하다는 것**이다.

6.2.1. 리커드 척도 문항에 대한 기초 통계량 허점

(1)리커드 척도 문항의 경우 평균의 의미는 무엇인가? 경상대 건물 안 공간(Q4) 만족도 점수가 2.91 이었다(7 점 만점). 강의실 공간(Q6) 만족도는 2.76 이었다. 각각 어떻게 해석할 수 있는가? 약간 답답하다. 약간보다 조금 더 많이 답답하다. 해석하기 어렵다. 사실 리커드 척도 문항은 점수로 계량화(quantify) 하였지만 해석에 문제가 발생한다. 그러므로 이런 경우 앞에서 지적하였듯이 만족(만족 이상)하는 사람들의 빈도(비율)나, 불만족(불만족 이하) 느끼는 **사람들의 빈도(비율)를 표로 정리(수정된 빈도표)**하는 것이 유용하다.

(2)리커드 척도의 또 다른 문제점은 서로 다른 응답자 집단의 리커드 점수를 비교하는 경우 발생한다. 예를 들어 대전 서구청과 동구청의 민원인 만족도를 비교한다고 하자. 대전 서구 주민의 구청 만족도는 3.2(이를 100 점 만점으로 환산하면 55 점)였고 동구 주민의 구청 만족도는 3.5(62.5)였다. 과연 동구청 만족도가 더 높은가? 리커드 문항 척도 응답점수=(응답자 만족 점수)+(성향)으로 구성되어 있어 비판적인 응답 성향을 가진 응답자 집단의 경우에는 동일 만족 수준이라도 리커드 점수가 낮을 수 있다. 그러므로 같은 만족을 느껴도 까다로운 집단의 경우 응답 점수가 낮게 된다. 이런데 그 리커드 척도 점수 평균만으로 단순히 비교하는 것은 잘못된 것이다. 이런 이유로 몇 년 전 서울 지역 구청 만족도 조사에서 꼴찌를 한 서초 구청이 반발한 것도 이런 이유에서이다. OO 조사 은행 만족도 1 위?라고 선전하는 은행이 있다. 믿을만한가? 제대로 조사였을까? [예: 1996 년 설문조사 수강생들이 대전 지역 5 개 지역 대학생 만족도를 조사한 적이 있다. ○○대학생의 만족도가 제일 낮았다. 그럼 정말 ○○대 학생들의 만족도는 다른 학교에 비해서 낮은가? 이것은 2 등이 느끼는 피해 의식이 아닐까? 2 등은 늘 1 등과 비교하므로 불만족을 많이 느낀다.

(3)리커드 척도 문항은 집단간 비교를 위하여 조사하면 문제가 발생한다. 그러므로 집단간 분석보다는 설문 조사 내의 다른 문항(주로 인구학적 변인 문항)과의 관계를 분석하는 것이 보다 적절한 분석 방법이다. 예를 들면, 시설 만족도 10 개 문항 중 만족도가 가장 낮은 곳은 어디인지 조사하거나 (만족도 문항간 평균 비교 ➔ 이를 이용하여 만족도 가장 맞은 부분에 집중적으로 투자), 성별 혹은 출신 지역별 시설 만족도의 차이는 있는가? (t-검정, 분산 분석) 조사하는 것이다.

6.2.2. T-score (T-점수)

연봉 책정을 위하여 각 직원들의 고과 점수를 매겨야 할 필요가 있다. 이 때 팀장의 평가 성향에 따라 점수 부여 정도가 달라질 수 있으므로 팀 장의 점수(1-100 점)를 그대로 사용하면 점수에 인색하거나 평가 점수 폭이 큰 팀장에 속한 직원은 불이익을 받게 된다. 그렇다고 각 직원의 평가 점수를 부여하지 않을 수 없는 경우 T-점수를 사용하여 점수를 변환하여 그 점수를 이용하여 각 직원의 업무 능력 점수로 사용하면 된다. $\overline{Q}$: 전체 직원 총 평가 점수 평균, S_Q : 전체 직원 평가 점수 표준 편차 Q_i : i 번째 직원 평가 점수, $\overline{Q_i}$: i 번째 응답자 속한 팀 평균, S_{Qi} : i 번째 응답자 속한 팀 표준 편차라 하면 각 개인의 T-점수는 $T = \overline{Q} + S_Q * (\frac{Q_i - \overline{Q_i}}{S_{Qi}})$ 이다.

6.3. 통계소프트웨어 사용 방법

예제 설문 CODING.txt 를 SPSS 데이터로 만들어 SURVEY.sav 에 저장했고, SAS 경우에는 프로그램을 SURVEY.sas 로 저장하였다 예제 설문에서 적도 문항 4 개(SAS: Q14-Q17, SPSS: V14-V17)에 대하여 각각 기초 통계량 분석을 실시해 보자.

6.3.1. SAS

도구(T) ▶ 옵션(O) ▶ 설정(P)... ▶ ☑ HTML 생성 설정하고 프로그램을 실행하자. (페이지 89-90 참고)

```
PROC MEANS DATA=SURVEY MEAN STD;
    VAR Q22-Q25;
RUN;
```

MEAN, STD 통계 Key-word 를 써 주면 평균과 표준 편차만 출력된다.

변수	평균값	표준편차
Q14	3.7307692	1.2186437
Q15	3.3923077	1.2909714
Q16	2.3384615	1.1244557
Q17	3.4769231	1.7127673

다음은 SAS 웹 문서 출력 결과를 엑셀로 가져가 정리한 내용이다.

문항	평균값	표준편차	100 점 환산
교수 강의	3.73	1.22	45.5
질의 응답	3.39	1.29	39.8
상담 기회	2.34	1.12	22.3
조교	3.48	1.71	41.3

참고 1-5 점 리커드 점수를 100 점 만점으로 환산하는 방법: $100점점수 = (리커드점수-1)*25$

<리커드 점수를 100 점 만점으로 환산하는 예>

5점 척도	3.12 ▶(3.12-1)*25 = 53점
	3 ▶(3-1)*25 = 50점
7점 척도	3.12 ▶(3.12-1)*100/6 = 35.3점
	4 ▶(3-1)*100/6 = 50점

교수에 대한 만족도 조사 중 강의에 대한 만족도가 가장 높고 상담 기회에 대한 만족도가 가장 낮으므로 교수들은 학생들에게 시간을 할애하여 상담 기회를 제공한다면 교수에 대한 만족도를 높일 수 있을 것이다. 조교에 대한 학생들의 만족도는 3.48 로 다른 항목에 비해 높으므로 평균만 이용한다면 문제가 없어 보인다. 그러나 표준 편차가 1.71 로 가장 높다는 것에 유념해야 한다. 표준 편차가 높다는 것은 매우 만족하는 학생도 많은 반면 매우 불만족 하는 학생의 비율도 높다는 것을 의미한다. 만족을 느끼는 사람이 만족 정도를 말하는 것보다 불만족 하는 학생이 불만족을 말하는 빈도가 높고 (역)효과를 더 많이 내므로 비록 만족도 점수는 높으나 조교에 대해 불만족 느끼는 항목을 조사하여 긴급히 시정해야 할 것이다.

TABULATE procedure 를 사용하면 더 다루기 쉬운 표를 적성할 수 있다. FORMAT=5.2 옵션은 출력 결과를 5 자리이고 소수점 2 자리로 출력하라는 의미이다.

```
PROC TABULATE DATA=SURVEY FORMAT=5.2;
    VAR Q14-Q17;
    TABLE (Q14 Q15 Q16 Q17),(MEAN STD);
RUN;
```

TABLE 문의, 앞은 행 뒤는 열의 형식 지정

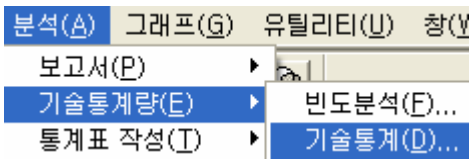
	Mean	Std
Q14	3.73	1.22
Q15	3.39	1.29
Q16	2.34	1.12
Q17	3.48	1.71

만약 성별(Q1)로 Q14-Q17 의 평균과 표준 편차를 구하려면 다음 프로그램을 실행한다.

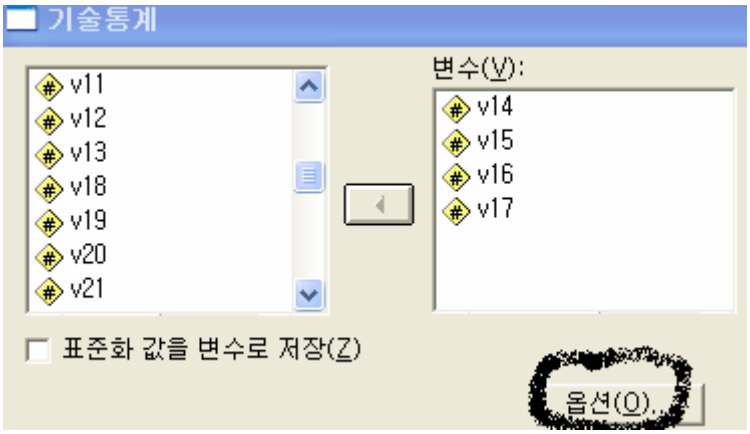
```
PROC TABULATE DATA=SURVEY FORMAT=5.2;
    CLASS Q1
    VAR Q14-Q17;
    TABLE Q1, (Q14 Q15 Q16 Q17)*(MEAN STD);
RUN;
```

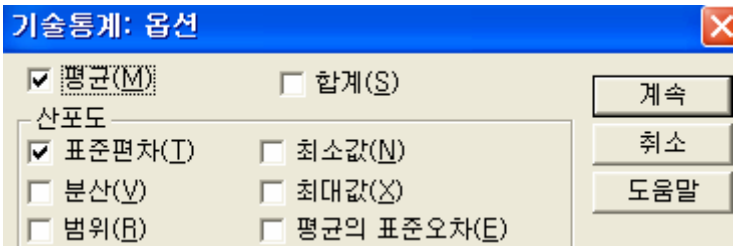
6.3.2. SPSS

메뉴에서 다음과 같이 기술 통계를 선택한다.



기술 통계 창에서 옵션(O)를 선택하여 원하는 기술 통계를 선택할 수 있다.





결과는 SAS 출력 결과와 동일하다. 표에서 오른쪽 마우스 버튼을 눌러 “복사”를 선택한 후 엑셀 문서로 가져가거나 “개체 복사”를 눌러 워드 문서로 가져가 이용한다.

기술통계량

	N	평균	표준편차
V14	130	3,73	1,22
V15	130	3,39	1,29
V16	130	2,34	1,12
V17	130	3,48	1,71
유효수 (목록별)	130		

6.4. 리커드 척도 문항 비교

6.4.1. 개념

리커드 척도 문항 만족도간 평균 비교는 가능한가? 학생들이 느끼는 강의(3.73), 질의응답(3.39), 상담 기회(2.34) 만족도의 차이가 있는지 통계 검정이 가능한가? 결론부터 말하자면 불가능하다. 마치 세 집단(범주) 평균 차이 검정처럼 보여 분산 분석(9장 참고)이 적절한 것처럼 보이지만 분산 분석을 실시할 수 있는 자료가 아니다. 각 분야 만족도 점수는 서로 독립이 아니므로 분산분석을 실시하여 범주의 차이를 볼 수 없다.

설문 조사에서는 범주간 평균 차이 비교를 하지 않는 것이 옳다. 굳이 해야 한다면 범주 2개씩 짝진 t-검정을 여러 번(kC_2 , k : 범주 수) 실시하면 된다. 예를 들어 (강의, 질의응답), (질의 응답, 상담 기회), (상담 기회, 강의) 각각에 대해 짝진 표본 평균 차이 t-검정을 실시해야 한다. 세 개의 가설을 각각 검정한 후 하나로 합치려 한다면 유의 수준을 다음 공식에 의해 고쳐야 한다. $1-(\alpha^*)^k = 0.05$ k 개의 귀무가설들을 하나로 합쳤을 때 유의수준이 0.05가 되게 하려면 이 식을 만족하는 α^* 를 유의 수준으로 사용해야 한다. 범주가 3개인 경우 각 귀무가설의 유의수준은 0.017로 해야 한다.

(예제)생산 기계가 두 대 A, B 들어 왔다. 어느 기계 성능이 좋은지 알아보기 위하여 8명의 전문가를 선정하여 각 기계의 성능 점수를 부여하였다. 기계 성능의 차이가 있는지 적절한 검정하시오. 점수는 정규 분포를 따른다고 가정하자. (유의수준=0.05)

전문가	기계 A	기계 B	차이
1	74	78	-4
2	76	79	-3
3	74	75	-1
4	69	66	3
5	58	63	-5
6	71	70	1
7	66	66	0
8	65	67	-2

(1)가설

①귀무가설: $H_0 : \mu_a = \mu_b \Leftrightarrow \mu_d = 0$ (A, B 기계 성능에 차이가 없다)

②대립가설: $H_0 : \mu_a \neq \mu_b \Leftrightarrow \mu_d \neq 0$ (A, B 기계 성능에 차이가 있다)

필요한 통계량 계산: 측정치의 차이 평균($\bar{d}$)과 표준 편차(s_d)를 계산한다.

(2)검정통계량: $T = \frac{\bar{d} - \mu_d}{s_{\bar{d}}} = \frac{\bar{d} - 0}{s_d / \sqrt{n}} = \frac{1.375 - 0}{2.67 / \sqrt{8}} = -1.46$

(3)결론

임계치 검정 통계량의 절대값이 임계치 $t(1-\alpha/2;7) = 2.365$ 보다 작으므로 귀무가설은 채택되고 두 기계의 성능 차이는 없다고 할 수 있다. P-값을 계산하면

```
P=CDF('T',-1.46,7)*2;
```

0.18766이므로 유의수준 0.05에서는 귀무가설 채택.

```
DATA ONE;
    INPUT M1 M2 @@;
    CARDS;
    74 78 76 79 74 75 69 66
    58 63 71 70 66 66 65 67
RUN;
PROC TTEST DATA=ONE;
    PAIRED M1*M2;
RUN;
```

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
d	8	-3.607	-1.375	0.8566	1.7649	2.6693	5.4327	0.9437

T-Tests

Variable	DF	t Value	Pr > |t|
d	7	-1.46	0.1885

6.4.2. 설문 분석에 이용

예제 설문에서 (강의, 질의 응답), (질의 응답, 상담 기회), (상담 기회, 강의) 각각에 대한 짝진 표본 평균 차이 t-검정을 실시해 보자.

```
PROC TTEST DATA=SURVEY;
    PAIRED Q14*Q15;
RUN;
```

T-Tests

Difference	DF	t Value	Pr > |t|
Q14 - Q15	129	3.33	0.0011

```
PROC TTEST DATA=SURVEY;
    PAIRED Q15*Q16;
RUN;
```

T-Tests

Difference	DF	t Value	Pr > |t|
Q15 - Q16	129	9.19	<.0001

```
PROC TTEST DATA=SURVEY;
    PAIRED Q16*Q14;
RUN;
```

T-Tests

Difference	DF	t Value	Pr > |t|
Q16 - Q14	129	-13.12	<.0001

(강의, 질의 응답), (질의 응답, 상담 기회), (상담 기회, 강의) 각각이 유의 수준 0.017에서 유의적인 차이가 있다는 결론을 얻었다. 그러므로 학생들이 교수 (강의, 질의 응답, 상담 기회)에 느끼는 만족도에는 유의적인 차이가 있고 강의(3.73)>질의 응답(3.39)>상담 기회(2.34) 순이다.

6.5. 리커드 척도 문항 빈도 분석

리커드 척도 문항도 보기 문항이므로 빈도 분석으로 정리할 수 있다. 각 보기에 대한 빈도, 상대 빈도 (퍼센트 %) 작성하거나 그래프를 그리지만 일반적으로 “만족” (매우 만족+만족) 혹은 “불만족” (불만족+ 매우 불만족)의 비율을 구하여 정리한다.

예제 설문에서는 7 점 척도로 조사하였으므로 (5+6+7)을 만족하는 그룹 (1+2+3)을 불만족 그룹으로 하자. 5 점 척도일 경우에는 (4+5)=만족, (1+2)=불만족으로 생각한다.

6.5.1. SAS

다음은 7 점 척도인 경우 만족 혹은 불만족 비율을 구하는 프로그램이다. **5 점 척도인 경우에는 만족에 관심이 있다면(표 왼쪽 셀) 5 대신 4 를, 불만족에 관심이 있다면(표 오른쪽) 3 대신 2 를 사용하면 된다.**

만족에 관심이 있는 경우

```
DATA SURVEY1;
  SET SURVEY;
   Q14_G="보통";
   IF (Q14>=5) THEN Q14_G="만족";
RUN;

PROC FREQ DATA=SURVEY1;
    TABLE Q14_G;
RUN;
```

Q14_G	도수	백분율
만족	35	26.92
보통	95	73.08

불만족에 관심이 있는 경우

```
DATA SURVEY1;
  SET SURVEY;
  Q14_G="보통";
  IF (Q14<=3) THEN Q14_G="불만족";
RUN;

PROC FREQ DATA=SURVEY1;
    TABLE Q14_G;
RUN;
```

Q14_G	도수	백분율
보통	78	60.00
불만	52	40.00

만족을 느끼는 사항보다는 불만족 느끼는 사항에 의해 만족 정도가 결정되므로 교수, 조교에 대한 만족도 문항 Q14-Q17 에서 불만족 비율의 차이를 보고 싶다면 다음과 같이 하면 된다.

```
DATA SURVEY1;
  SET SURVEY;
    Q14_G="보통";
    IF (Q14<=3) THEN Q14_G="불만족";
    Q15_G="보통";
    IF (Q15<=3) THEN Q15_G="불만족";
    Q16_G="보통";
    IF (Q16<=3) THEN Q16_G="불만족";
    Q17_G="보통";
    IF (Q17<=3) THEN Q17_G="불만족";
RUN;

PROC FREQ DATA=SURVEY1;
    TABLE Q14_G Q15_G Q16_G Q17_G /NOCUM;
RUN;
```

Q14_G	도수	백분율
보통	78	60.00
불만	52	40.00

Q15_G	도수	백분율
보통	61	46.92
불만	69	53.08

Q16_G	도수	백분율
보통	16	12.31
불만	114	87.69

Q17_G	도수	백분율
보통	65	50.00
불만	65	50.00

교수가 “상담 기회”를 제공하는가에 대한 불만족 비율이 가장 높으므로 학생들의 만족도를 높이기 위해서는 교수는 상담 시간을 정하여 운영할 필요가 있다. 위의 결과는 평균을 이용하여 얻은 결과와 유사하다. NOCUM 옵션은 누적(cumulative) 통계량을 출력하지 말라는 옵션이다.

그럼 만족도 문항에 대해 빈도 분석이나 기초 통계량 분석 중 어는 것이 옳은가? 빈도와 평균 분산을 함께 적어 보고서를 적성한다.

```
DATA SURVEY1;
    SET SURVEY;
    V=Q14;G="Q14";OUTPUT;
    V=Q15;G="Q15";OUTPUT;
    V=Q16;G="Q16";OUTPUT;
    V=Q17;G="Q17";OUTPUT;
RUN;

PROC FREQ DATA=SURVEY1;
    TABLE G*V/NOPERCENT NOCOL;
RUN;
```

변수 2 개를 만든다. V 는 각 문항의 응답을 G 는 문항 번호에 대한 변수이다.

TABLE G*V: G 를 행, V 를 열로 하여 교차표 작성.
NOPERCENT: 샐 퍼센트(%) 출력하지 않기.
NOCOL: 열 퍼센트 출력하지 않기.

빈도
행 백분율

NOPERCENT

NOCOL

G * V 교차표

G	V 1	2	3	4	5	6	7	총합
Q14	7 5.38	11 8.46	34 26.15	43 33.08	30 23.08	3 2.31	2 1.54	130
Q15	11 8.46	17 13.08	41 31.54	42 32.31	12 9.23	4 3.08	3 2.31	130
Q16	35 26.92	38 29.23	41 31.54	13 10.00	1 0.77	1 0.77	1 0.77	130
Q17	22 16.92	17 13.08	26 20.00	31 23.85	18 13.85	8 6.15	8 6.15	130
총합	75	83	142	129	61	16	14	520

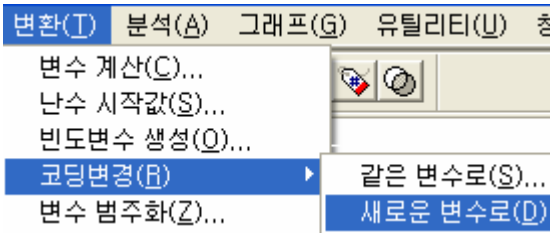
엑셀에서 빈도표 오른쪽에 평균과 분산을 출력 결과를 붙인 것이다. 이런 표를 만들 때는 굳이 만족 혹은 불만족 비율로 빈도표를 만들 필요는 없다. 엑셀에서 같은 셀에 두 줄을 사용하려면 한 줄을 완성한 후 ATL+ENTER 를 치면 줄이 바뀌게 되므로 글을 입력하면 된다.

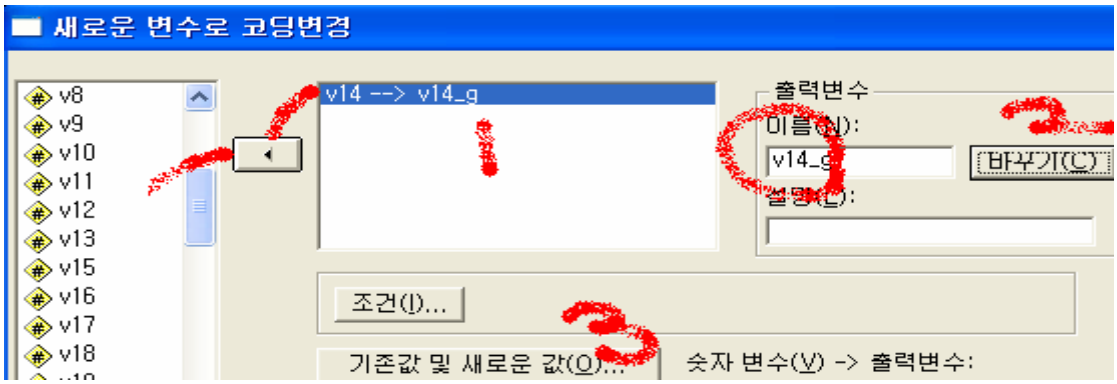
문항	1	2	3	4	5	6	7	평균	표준편차
교수 강의	7	11	34	43	30	3	2	3.73	1.22
(%)	5.38	8.46	26.15	33.08	23.08	2.31	1.54		
질의응답	11	17	41	42	12	4	3	3.39	1.29
(%)	8.46	13.08	31.54	32.31	9.23	3.08	2.31		
상담기회	35	38	41	13	1	1	1	2.34	1.12
(%)	26.92	29.23	31.54	10	0.77	0.77	0.77		
조교	22	17	26	31	18	8	8	3.48	1.71
(%)	16.92	13.08	20	23.85	13.85	6.15	6.15		

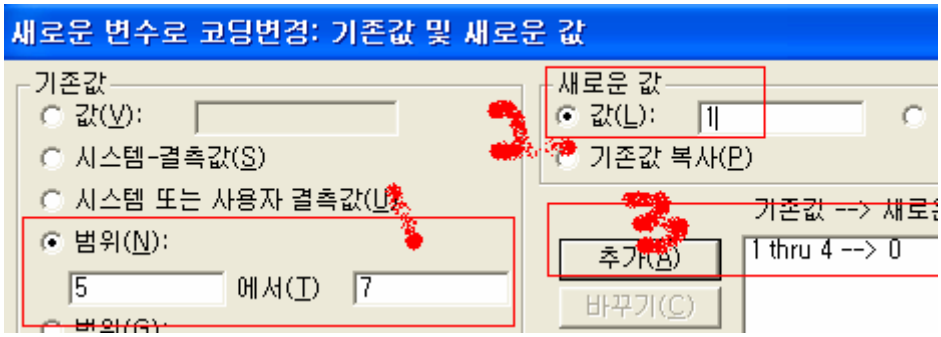
5 점 척도를 사용한 경우에는 1, 2, 3, 4, 5 의 빈도만 있으므로 표가 더 간결하다. 리커드 척도 문항은 위와 같이 정리하거나 만족(혹은 불만족)하는 사람의 비율을 표로 적성하여 발표하면 된다.

6.5.2. SPSS

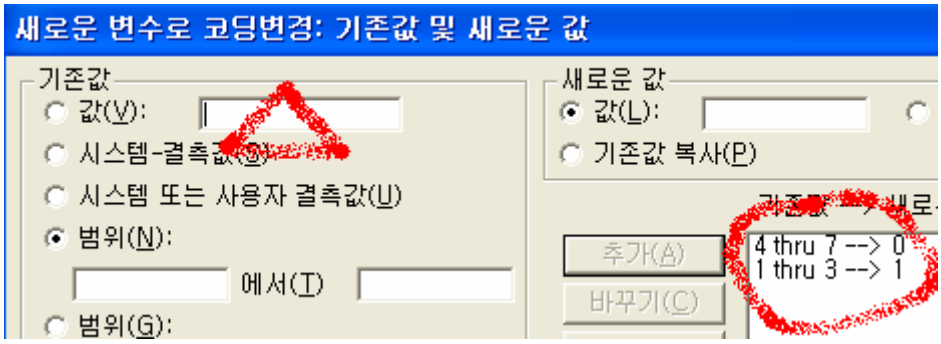
우선 V14-V17 변수에 대해 데이터 변환을 해야 한다. 각 변수에 대해 따로 해 주어야 하므로 V14 를 예를 들어 살펴보기로 하자.







만약 5 점 척도라면 1 thru 3-->0 으로 변환하고 4 thru 5--> 1 로 변환하여야 한다. 그리고 범위로 하지 않고 개별 값으로 하려면 아래 △ 부분을 사용하면 된다.

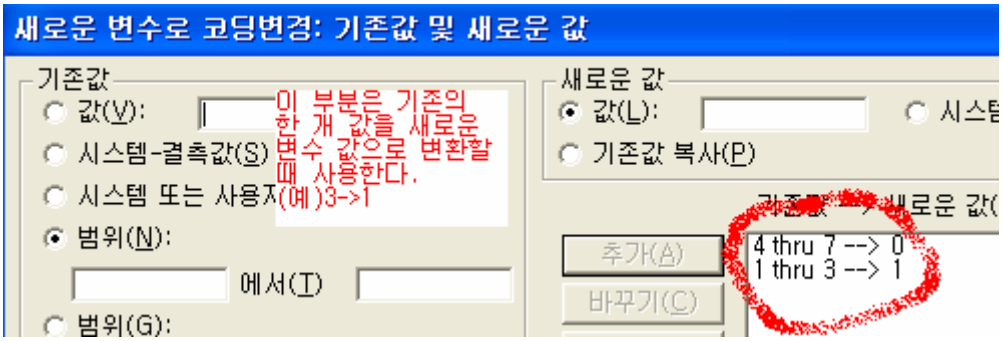


마지막 열에 V14_G 이라는 새로운 변수가 생긴다. V14_G 변수에 대해 빈도 분석을 실시하면 만족하는 응답자의 비율을 얻을 수 있다.

	v34	v35	v14_g
1	2	5	.00
2	3	.	.00
3	.	.	.00
4	5	6	.00
5	4	5	1.00

default 로 소수점 2 자리 숫자가 만들어지므로 시트 아래 변수 보기 폴더를 눌러 자리 수를 지정해 주면 된다.

불만족 비율을 얻고 싶다면 "새로운 변수 코딩 변경" 화면 창에서 다음과 같이 지정해 주면 된다.



6.6. 만족도 문항 역 코딩

다음과 같이 리커드 척도 문항에서 보기를 좋은 것부터 열거한 경우 예를 들어 살펴 보자.

1. 정부가 추진하고 있는 일자리 창출 노력에 대해 당신은 얼마나 만족하고 있습니까?
①매우 만족 ②만족 ③보통 ④불만족 ⑤매우 불만족

리커드 척도 문항에 대해 빈도 분석만 할 경우에는 매우 만족을 1, 만족을 2, ... 순으로 코딩 해도 문제가 없으나 (빈도만 보는 것이므로) 기초 통계량을 구하는 경우 매우 만족은 5 점, 만족은 4 점, .. 이런 식으로 사용하므로 문제가 발생한다.

그럼 코딩 할 때 미리 기초 통계량 계산할 것을 대비하여 ①번 선택하면 5 로, ②는 4, ③은 3, ④는 2, ⑤는 1 로 코딩 해야 하는가? 이렇게 코딩을 하다 보면 코딩 오류가 많이 발생하게 되므로 코딩 할 때는 그냥 번호 순으로 입력하고 나중 프로그램에서 정정해 주면 된다. Q99 번 문항이 5 점 척도이고 역으로 코딩 되어 있다고 하자.

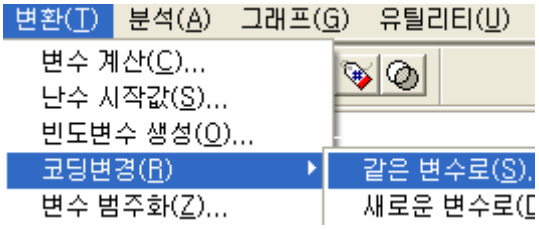
6.6.1. SAS

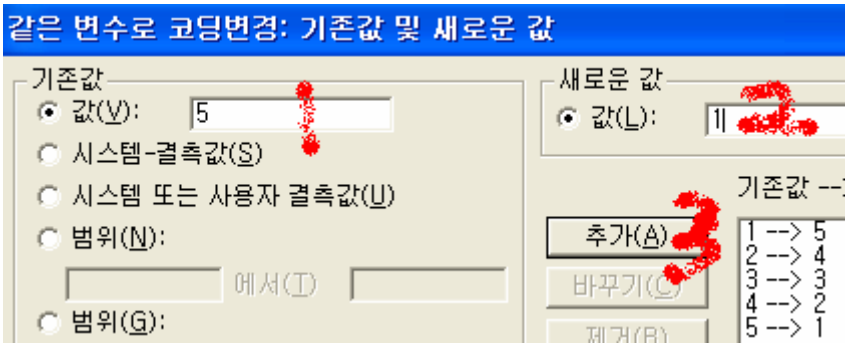
Q66 의 원래 값이 1 이면 5 가 되고 5 이면 1 이 되므로 적절하게 변환된다. 만약 예제 설문처럼 7 점 척도였으면 6 대신 8 을 사용해 주면 된다.

```
DATA SURVEY1;
    SET SURVEY;
    Q99=6-Q99;
RUN;
```

6.6.2. SPSS

같은 변수 메뉴를 선택하는 것을 제외하고는 6.5.2. 절의 방법과 동일하다.





6.7. 우선 순위 문항

5 장에서는 우선 순위 문항에 대한 빈도 분석을 실시하였으나 순위를 값으로 생각하여 순위에 대한 기초 통계량으로 구하여 표로 정리할 수 있다. 예제 설문에서 Q26_1-Q26_5 번 문항(SPSS 는 V26-V29 문항)에 대한 각 보기의 순위 점수 평균과 표준 편차를 구하여 보자.

6.7.1. SAS

(1)PROC MEANS 방법

```
PROC MEANS DATA=SURVEY MEAN STD;
    VAR Q26_1--Q26_5;
RUN;
```

변수	평균값	표준편차
Q26_1	1.5042017	0.6873426
Q26_2	3.3303571	0.9142357
Q26_3	1.6422764	0.7143733
Q26_4	3.8660714	0.7533174
Q26_5	4.5625000	0.7075048

(2)PROC TABULATE 방법

```
PROC TABULATE DATA=SURVEY FORMAT=3.2;
    VAR Q26_1--Q26_5;
    TABLE (Q26_1 Q26_2 Q26_3 Q26_4 Q26_5), (MEAN STD);
RUN;
```

	Mean	Std
Q26_1	1.50	0.69
Q26_2	3.33	0.91
Q26_3	1.64	0.71
Q26_4	3.87	0.75
Q26_5	4.56	0.71

순위는 1~5 까지 있으므로 순위 평균이 낮을수록 학생들이 학과를 선택할 때 중요하게 생각한다. 취업 ▶ 적성 ▶ 학문적 우월성 ▶ 교수 질 ▶ 선후배 관계 순이다. 빈도 분석 결과와 일치한다. 우선 순위의 평균을 사용하는 경우는 ①하나의 값으로 요약할 수 있을 때 ②각 문항의 순위 차이(분산 분석)를 검정할 때이다.

(3)우선 순위 개수와 문항 수가 같지 않다면...

만약 예제 설문과 같이 조사되지 않고 우선 순위를 적으라고 한 설문 조사는 어떻게 분석할 것인가? 다음의 예를 보자.

다음 중 전공을 선택할 때 중요하다고 생각되는 3개를 순서대로 적으시오.

Q26_1 Q26_2 Q26_3

1순위:(3) 2순위:(2) 3순위:(4),

①취업 전망 ②학문적 우월성 ③나의 적성 ④전공 교수의 질 ⑤선후배 관계

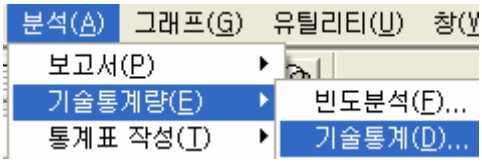
```
DATA SURVEY1;
    SET SURVEY;
        Q26=Q26_1;RANK=1;OUTPUT;
        Q26=Q26_2;RANK=2;OUTPUT;
        Q26=Q26_3;RANK=3;OUTPUT;
RUN;

PROC TABULATE DATA=SURVEY1 FORMAT=3.2;
    CLASS Q26;
    VAR RANK;
    TABLE (Q26), RANK*(MEAN STD);
RUN;
```

	RANK	
	Mean	Std
Q26		
1	1.92	0.99
2	2.07	0.94
3	2.10	0.45
4	1.96	0.20
5	2.00	0.35

Q26 의 1 은 ①취업 전망, 2 는 ②학문적 우월성, 3 은 ③적성, 4 는 ④교수의 질, 5 는 ⑤ 선후배 관계를 의미한다. 그리고 평균은 순위의 평균이므로 이 값이 낮은 문항이 우선 순위가 높다.

6.7.2. SPSS



기술통계량

	N	평균	표준편차
V26	119	1.50	.69
V27	112	3.33	.91
V28	123	1.64	.71
V29	112	3.87	.75
V30	112	4.56	.71
유효수 (목록별)	112		

6.8. 보고서 작성

6.8.1. 리커드 척도 문항

리커드 척도 문항에 대해서는 빈도 분석과 함께 평균과 표준 편차를 함께 작성하는 것이 바람직하다. 다음은 예제 설문에서 1-2 점을 1 점, 3 점을 2 점, 4 점을 3 점, 5 점을 3 점, 6- 7 점을 5 점으로 하여 Q14-Q17 문항을 정리한 것이다. **주로 5 점 척도를 사용하기 때문에 이렇게 예제를 재구성 하였다.**

각 항목에 대해 행 퍼센트가 가장 높은 셀에 파란 색 채우기를 이용하여 보기 쉽게 하였고 만족도 점수가 가장 낮은 항목과 표준 편차가 가장 곳에 빨간 색으로 강조하였다. 평균 점수가 낮다는 것은 만족도가 낮다는 것을 의미하며 표준 편차가 크다는 것은 응답자의 만족 점수의 변동이 크므로 응답 점수가 높은 사람은 물론 낮은 사람도 많다는 것을 의미한다.

항목	1	2	3	4	5	평균	표준편차
교수 강의	18	34	43	30	5	2.77	1.08
	13.85	26.15	33.08	23.08	3.85		
질의 응답	28	41	42	12	7	2.45	1.09
	21.54	31.54	32.31	9.23	5.38		
상담 기회	73	41	13	1	2	1.60	0.82
	56.15	31.54	10	0.77	1.54		
조교	39	26	31	18	16	2.58	1.37
	30	20	23.85	13.85	12.31		

교수들의 상담 기회 제공에 대한 만족도가 가장 낮으므로 교수들은 일주일에 2 시간 정도 시간을 할애하여 학생에게 면담을 제공한다면 학생 만족도를 높일 수 있을 것이다. 조교에 대한 만족도 점수는 다른 항목에 비해 낮지 않지만 표준 편차가 크므로 불만족을 느끼는 학생들의 비율이 교수 강의나 질의 응답에 비해 크다는 것을 의미하므로 학생들이 불만족을 느끼는 사항에 대해 시정할 필요가 있다.(표준 편차 해석: 6.3.1 절 참고)

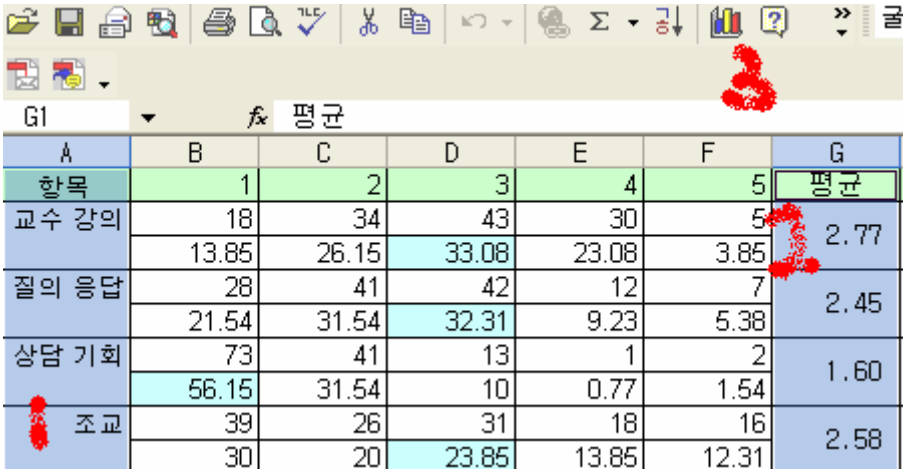
알고 가자 각 항목의 평균이 다른지 (분산 분석) 표준 편차가 다른지 (F-검정, Bartlett 검정, Hartley 검정) 통계적 검정을 해야 한다(6.4 절 참고). 그럼 통계적 유의성 검정은 꼭

필요한가? 그렇지 않다. 설문 조사에서는 항목간 차이가 있다는 것에 집중한다. 위 표는 학교가 학생들의 만족도를 높이기 위해서는 어떤 항목에 우선적으로 집중해야 하느냐를 결정할 때 참고 자료로 사용할 수 있다. 사실 설문 조사에서 통계적 유의성 검정은 학문 연구의 가치 밖에는 없다.

리커드 척도 문항을 개별적으로 그래프화 할 때는 각 문항 보기의 퍼센트에 대한 Bar chart 를 이용하면 된다. 평균, 표준 편차는 각 하나이므로 그래프 그린다는 의미가 없다. 그러나 위의 예제처럼 유사한 리커드 척도 항목을 비교하고자 할 때는 퍼센트에 대한 바 차트나 평균에 대한 바 차트를 그리는 것이 좋다.

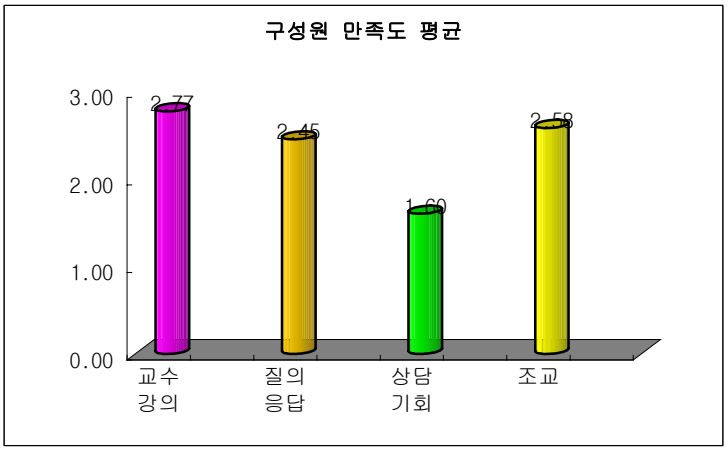
(1)평균 Bar chart (바 차트)

CTRL 키를 누른 상태에서 마우스를 이용하여 다음과 같이 항목과 평균을 선택하고  아이콘을 눌러 그래프를 그리면 된다.



G1 fx 평균

A	B	C	D	E	F	G
항목	1	2	3	4	5	평균
교수 강의	18	34	43	30	5	2.77
	13.85	26.15	33.08	23.08	3.85	
질의 응답	28	41	42	12	7	2.45
	21.54	31.54	32.31	9.23	5.38	
상담 기회	73	41	13	1	2	1.60
	56.15	31.54	10	0.77	1.54	
조교	39	26	31	18	16	2.58
	30	20	23.85	13.85	12.31	

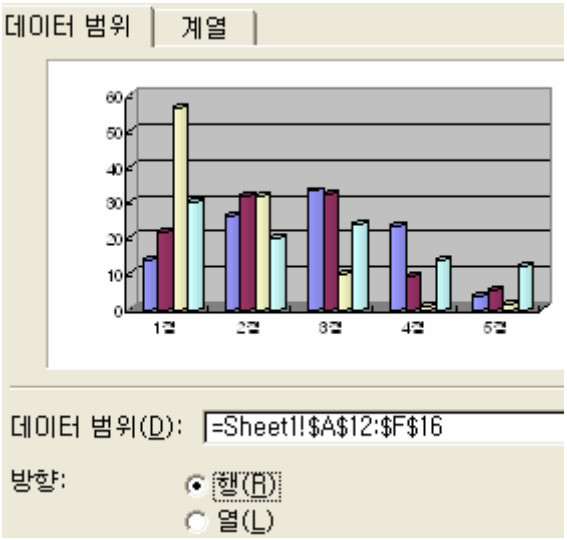


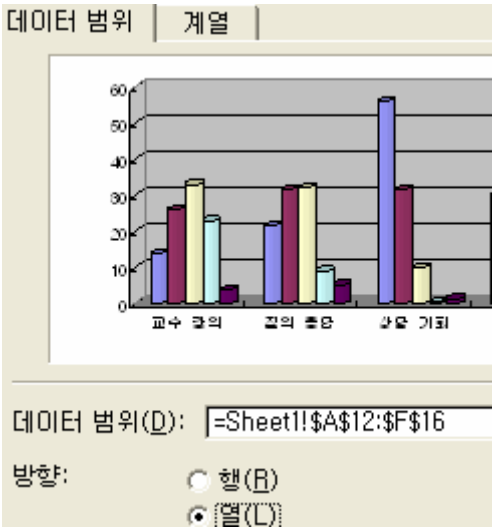
(2)퍼센트 Bar chart (바 차트)

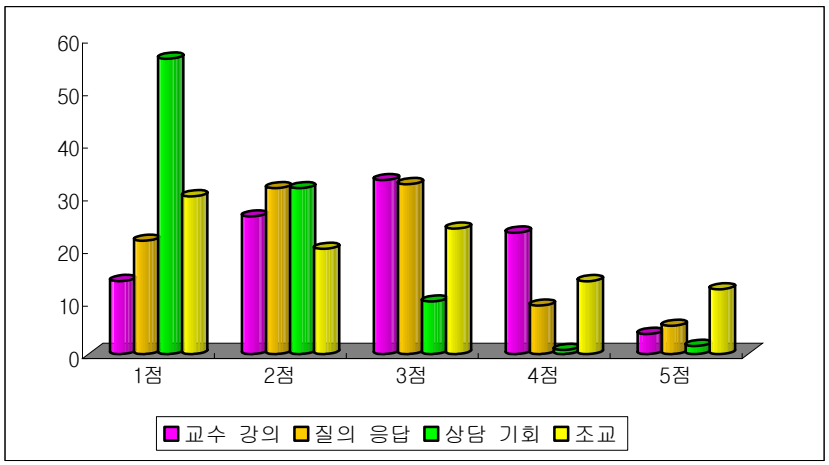
엑셀에서 표를 복사한 후 빈도를 지우고 퍼센트만 남긴 후 마우스로 그래프 그릴 셀들을 다음과 같이 선택하고  아이콘을 눌러 그래프를 그리면 된다.

12	항목	1점	2점	3점	4점	5점
13	교수 강의	13.85	26.15	33.08	23.08	3.85
14	질의 응답	21.54	31.54	32.31	9.23	5.38
15	상담 기회	56.15	31.54	10	0.77	1.54
16	조교	30	20	23.85	13.85	12.31

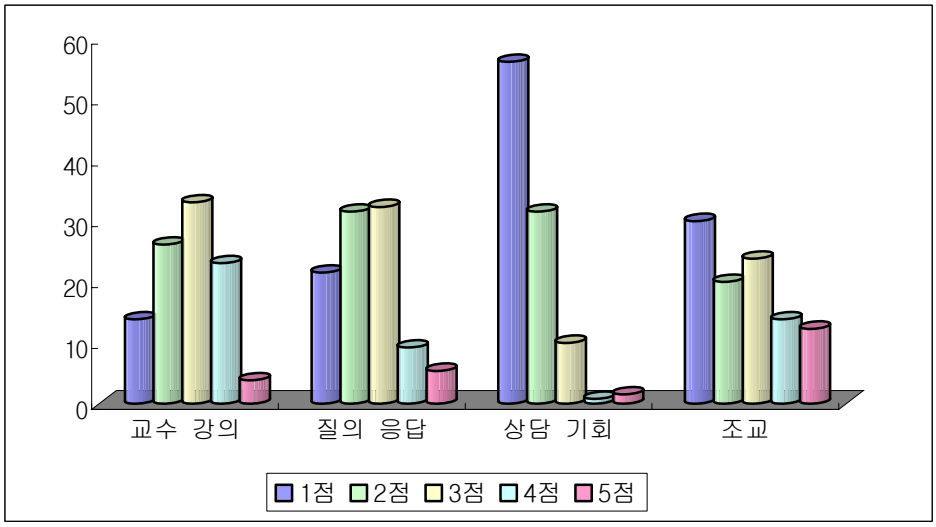
항목별 퍼센트에 대한 그래프를 그릴 때 방법은 다음과 같이 두 가지 방법이 있다.







분석 결과에 대한 해석은 평균 바 차트를 이용한 경우와 유사하다. 위의 그래프를 이용하면 리커드 점수 별로 항목의 퍼센트를 비교할 수 있고 아래 그래프는 항목별로 리커드 점수 분포를 알 수 있다.



조교에 대한 만족도를 보면 매우 만족하는(5 점) 학생이 있는 반면 매우 불만족 비율도 높으므로 이를 시정할 방안을 강구해야 한다. 교수의 상담 기회에 대한 불만족 비율이 높으므로 상담 시간을 갖는 방안을 우선적으로 시행할 필요가 있다.

6.8.2. 우선 순위 문항

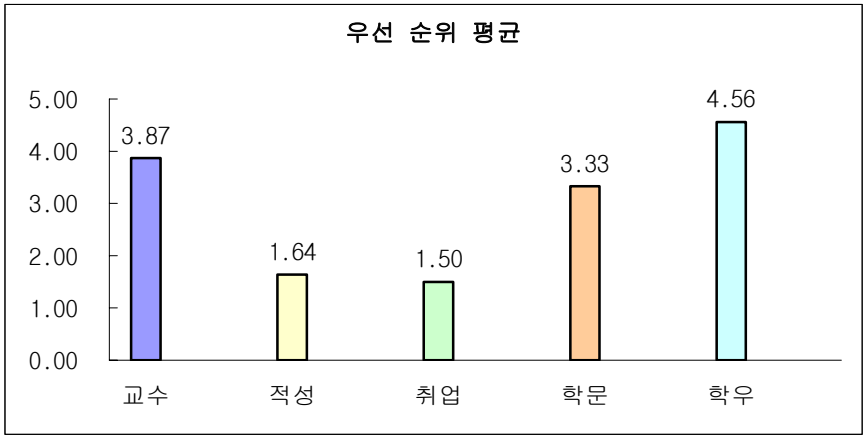
우선 순위 문항도 순위 빈도와 평균, 표준 편차를 표로 작성하거나 그래프를 그리면 된다.

```
DATA SURVEY2;
    SET SURVEY;
    Q26=Q26_1;MAJOR="취업";OUTPUT;
    Q26=Q26_2;MAJOR="학문";OUTPUT;
    Q26=Q26_3;MAJOR="적성";OUTPUT;
    Q26=Q26_4;MAJOR="교수";OUTPUT;
    Q26=Q26_5;MAJOR="학우";OUTPUT;
RUN;
```

```
PROC FREQ DATA=SURVEY2;
    TABLE MAJOR*Q26/NOCOL NOPERCENT;
RUN;

PROC TABULATE DATA=SURVEY2;
    CLASS MAJOR;
    VAR Q26;
    TABLE MAJOR, Q26*(MEAN STD);
RUN;
```

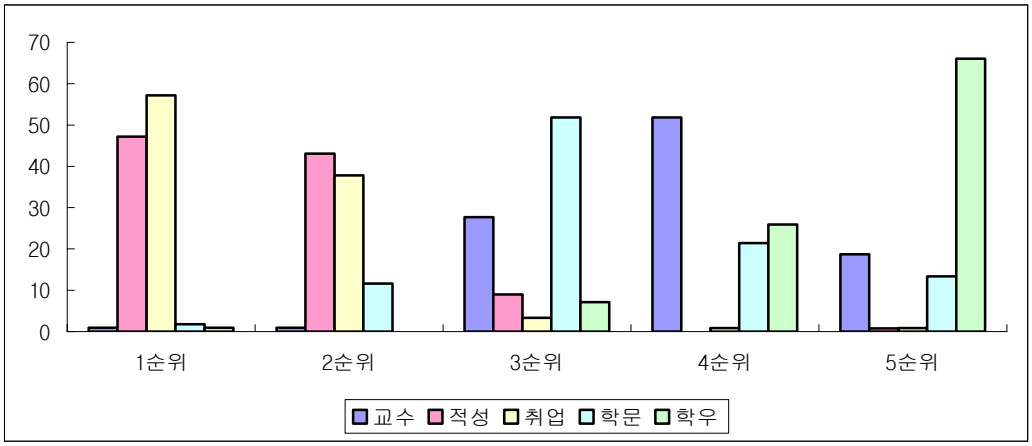
	A	B	C	D	E	F	G	H
1	항목	1순위	2순위	3순위	4순위	5순위	평균	표준편차
2	교수	1	1	31	58	21	3.87	0.75
3		0.89	0.89	27.68	51.79	18.75		
4	적성	58	53	11	0	1	1.64	0.71
5		47.15	43.09	8.94	0	0.81		
6	취업	68	45	4	1	1	1.50	0.69
7		57.14	37.82	3.36	0.84	0.84		
8	학문	2	13	58	24	15	3.33	0.91
9		1.79	11.61	51.79	21.43	13.39		
10	학우	1	0	8	29	74	4.56	0.71
11		0.89	0	7.14	25.89	66.07		



학생들이 전공을 선택할 때 취업, 적성을 우선적으로 고려하고 있었고 학우 관계는 큰 영향을 미치지 않았다. (평균 순위나 아래 퍼센트 바 차트로부터 얻는 정보는 동일하다.) 설문조사가 이루어진 시기가 1 학기 초라 학생들이 대학을 들어올 때 막연히 생각했던 취업 문제, 들어서 알고 있던 적성을 택하였다. 그러나 학년말이 갈수록 MT, 학과 단위의 모임으로 인하여 전공을 선택하는데 선후배(학우) 관계가 주요 항목으로 나타나고 있다. (실제 일들에 대한 논의)

다음은 순위에 대한 빈도표를 항목별로 작성하고 그에 대한 바 차트를 그린 것이다.

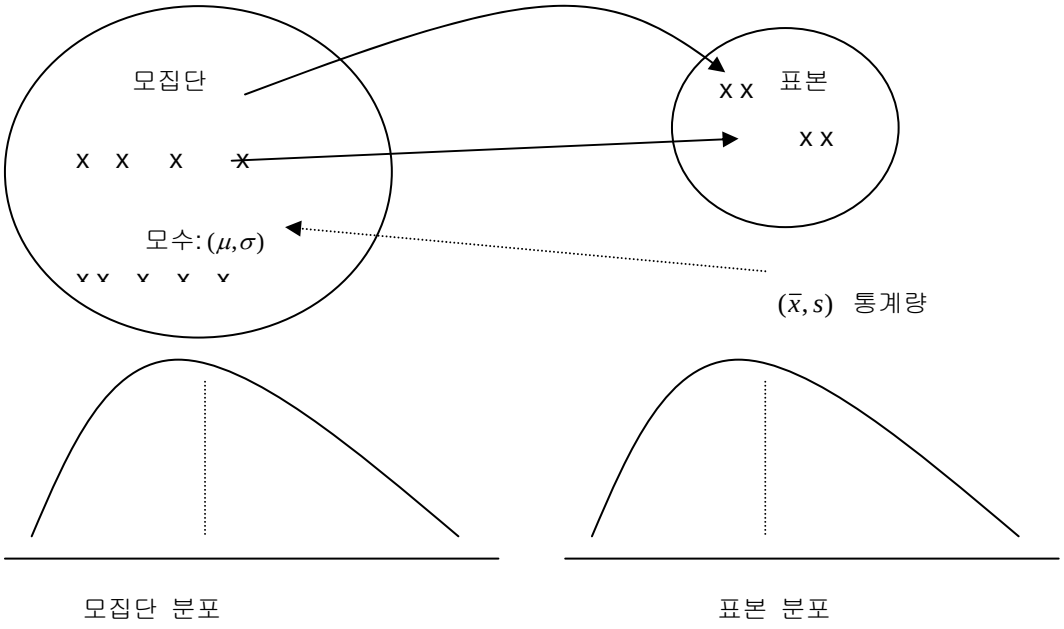
27	항목	1순위	2순위	3순위	4순위	5순위
28	교수	0.89	0.89	27.68	51.79	18.75
29	적성	47.15	43.09	8.94	0	0.81
30	취업	57.14	37.82	3.36	0.84	0.84
31	학문	1.79	11.61	51.79	21.43	13.39
32	학우	0.89	0	7.14	25.89	66.07



6.9. 통계적 가설 검정 (optional)

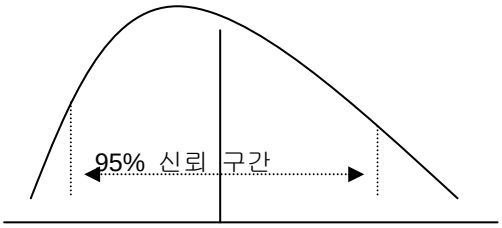
6.9.1. 모수와 통계량

모집단 데이터 특성 값을 모수(parameter)라 허고 그 모수를 추정하기 위하여 표본으로부터 계산된 값을 통계량(statistic)이라 한다. 모집단의 분포와 표본의 분포는 동일하다. 그러나 통계량의 분포는 모집단의 분포와 다를 수 있다. (예: 중심 극한 정리)



6.9.2. 추정

점 추정은 (point estimation) 모집단의 모수를 하나의 값으로 추정하는 방법이다. 구간 추정은 (interval estimation) 모수 값을 구간으로 추정하는 것으로 아래 그림과 같다. 이처럼 구간을 설정하기 위해서는 분포에 대한 정보가 필요하다.



점 추정치

6.9.3. 통계적 가설 (statistical hypothesis)

모수에 대해 알고자 하는 내용을 가설로 설정하여 표본으로부터 구한 통계량을 이용하여 설정한 가설의 진위 여부를 검정하는 것을 가설 검정이라 (hypothesis testing) 한다. 통계학에서 설정되는 통계적 가설은 모수에 대한 하나의 값으로 설정된 귀무가설과 (null hypothesis)과 그에 대립하는 대립가설이 (alternative hypothesis) 있다. 우리의 관심은 대부분 대립 가설에 있으므로 이를 연구 가설이라 (research hypothesis) 한다. 귀무가설은 “=”, “차이가 없다”, “영향을 미치지 않는다” 등으로 표현되므로 아무 내용도 없다는 뜻의 NULL 이라는 이름이 부여되어 있다.

(예제)한남대학교 대학생의 IQ가 120인가를 알고 알아보고자 한다면 다음과 같다.

귀무가설: $\mu = 120$ $\mu =$ 한남대 학생의 평균 IQ vs. 대립가설: $\mu \neq 120$ 양측 검정

만약 IQ가 120 이상인가를 알아보고자 한다면 귀무가설은 그대로이고 대립가설만 $\mu > 120$으로 바꾸면 된다. 이를 단측 가설이라 한다.

통계적 가설 검정은 표본으로부터 구한 통계량을 이용하여 귀무가설을 채택하거나 기각한다. 귀무가설이 기각되면 대립가설이 채택한다. 통계적 가설 검정에서는 다음과 같이 2 가지 오류(error)가 발생한다.

실제 / 검정 결과	귀무가설 참	귀무가설 거짓
귀무가설 기각	1 종 오류(α)	옳은 판단
귀무가설 채택	옳은 판단	2 종 오류(β)

1 종 오류(type I error)와 2 종 오류(type II error)를 동시에 최소화 하는 검정 방법은 존재하지 않으므로 1 종 오류 값을 임의로 설정하고 2 종 오류를 최소화하는 검정 방법을 사용한다. 분석자가 임의로 설정한 1 종 오류를 유의 수준이라 (significant level) 하고 일반적으로 0.1(10% 유의), 0.05(5%), 0.01(1% 매우 유의)을 주로 사용한다.

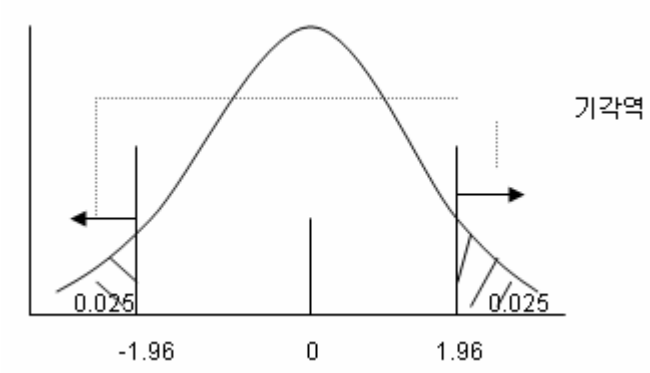
신뢰구간과 가설 검정은 일대일 관계가 있다. 즉 95% 신뢰구간과 5% 유의수준 가설 검정 결과는 동일하다. 예를 들어 한남대학생 IQ 의 95% 신뢰구간을 구하였더니 (100,130) 이었다면 귀무가설에서 $H_0 : \mu = \mu_0$, μ_0 의 값을 100 에서 130 의 값으로 설정하면 귀무가설은 채택된다.

6.9.4. 유의 수준과 유의 확률(p-값)

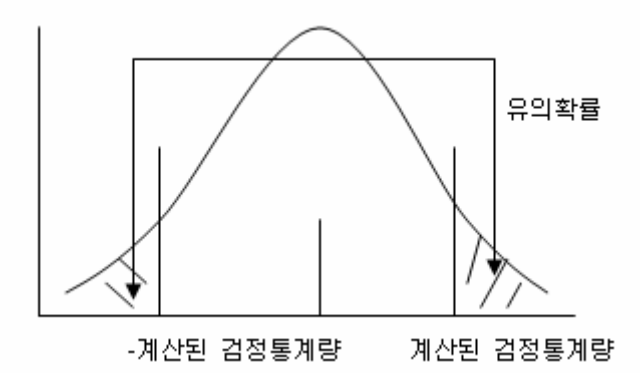
통계적으로 유의 하다는 의미는 통계적으로 차이가 존재한다는 것이다. IQ 에 있어 성별 변인이 미치는 영향이 유의 하다는 것은 성별에 따른 IQ 의 차이가 있다는 것이고 OO 시설 만족도에 학년 변인 문항이 유의 하다는 것은 학년에 따른 만족도 차이가 있다는 것을 의미한다. 통계적 가설을 진위를 검정하기 위하여 계산되는 통계량을 검정 통계량이라 한다. OO 대학 학생들의 IQ 가 대한민국 IQ 평균 120 과 같은가? (귀무가설은 $H_0 : \mu = 120$, 대립가설은 $H_0 : \mu \neq 120$) 알아볼 때 검정 통계량은 다음과 같으며 이 통계량은 중심극한 정리에 의해 정규 분포를 따른다. $T = \frac{\bar{x} - \mu_0 (= 120)}{s/\sqrt{n}} \sim Normal(0,1)$

검정 통계량이 정규 분포를 따르므로 만약 유의 수준을 0.05 라 설정하였다면 표본으로부터 계산된 검정통계량 값의 절대값이 $z_{0.025} = 1.96$ (기각치, 혹은 임계치)값보다 크면 귀무가설을

기각하고(대립가설 채택) 그렇지 않으면 귀무가설을 채택한다. 다음은 유의 수준을 0.05 (5%)로 한 경우 모평균에 대한 가설 검정 기각역을 나타낸 것이다.



표본으로부터 계산된 검정 통계량의 값보다 크거나 작은 부분의 확률을 유의 확률이라 한다. 그러므로 유의 확률이 유의 수준(α)보다 작다면 계산된 검정 통계량이 기각역에 들어감을 의미함으로 귀무가설이 기각된다.



유의 확률을 p-값(p-value)이라 하는데 “귀무가설을 기각할 최소의 유의 수준”이라고 정의하기도 한다. 통계 소프트웨어는 항상 검정 통계량 값과 유의 확률을 함께 출력하므로 유의 확률 개념만 알면 결과 해석에 문제가 없다. 유의 확률(p-값)이 유의 수준보다 작으면 귀무가설을 기각하고 크면 귀무가설을 채택한다.

통계소프트웨어는 양측 검정 기준으로 유의확률을 출력하므로 대립 가설이 단측 가설 검정이라면 유의 확률만 1/2 로 해 주면 된다.

[연습문제]

(1)○○대학교 학생으로 느끼는 점에 대한 4 개 문항(Q22~Q25)에 대해 기초 통계량을 계산하고 빈도표와 함께 정리하시오.

(2)(1)의 결과에 적절한 그래프를 그리고 해석하시오.

(3)팀 프로젝트 설문지에서 리커드 척도 문항에 대한 기초 통계량 분석을 실시하고 적절한 표와 그래프로 요약하시오.

(4)리커드 척도 문항은 빈도 분석과 기초 통계량 분석을 할 수 있다. 이 경우 두 방법의 장단점을 논하시오.