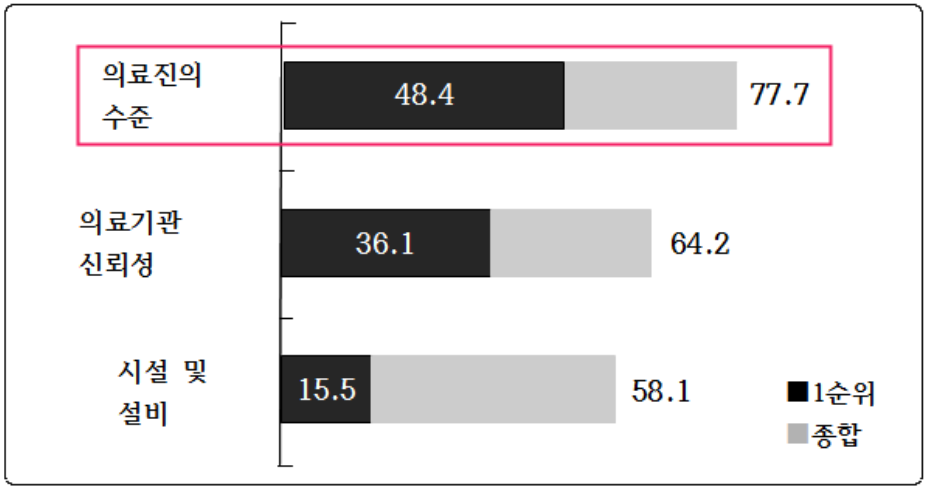
제2장 기술통계와 확률

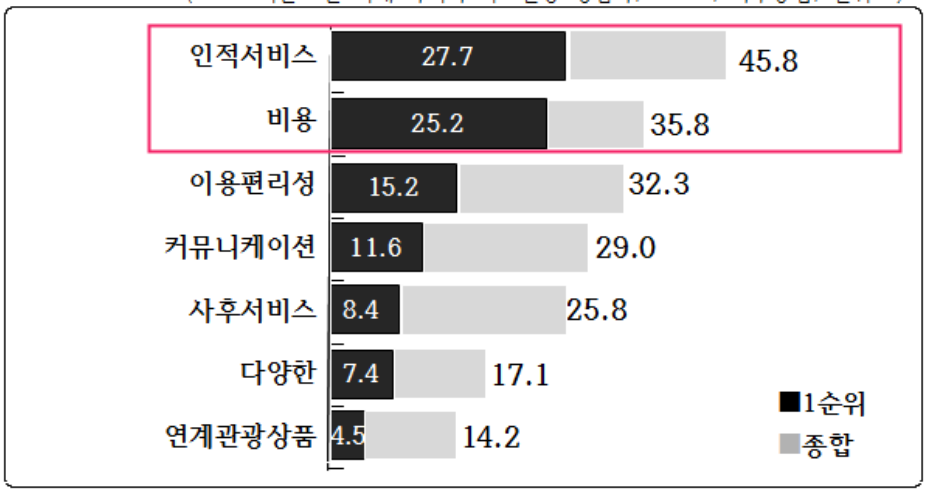
사례:해외의료관광 의향자 빈도분석

- 의료관광 결정시 중요 서비스 요인

(Base: 최근 2년 이내 아시아 의료관광 경험자, n=310, 복수응답, 단위:%)



(Base: 최근 2년 이내 아시아 의료관광 경험자, n=310, 복수응답, 단위:%)



	전체	국가별			월평균 가구소득			
		일본	중국	미국	1500 USD 이하	1501~3500 USD 이하	3501~7000 USD 이하	7001 USD 이상
사 례 수	(310)	(100)	(108)	(102)	(71)	(86)	(77)	(59)
의료진의 수준	48.4	40.0	42.6	62.7	36.6	48.8	49.4	57.6
의료기관 신뢰성	36.1	37.0	43.5	27.5	42.3	37.2	33.8	28.8
시설 및 설비	15.5	23.0	13.9	9.8	21.1	14.0	16.9	13.6

	전체	국가별			월평균 가구소득			
		일본	중국	미국	1500 USD 이하	1501~3500 USD 이하	3501~7000 USD 이하	7001 USD 이상
사 례 수	(310)	(100)	(108)	(102)	(71)	(86)	(77)	(59)
인적 서비스	27.0	27.0	33.3	22.5	35.2	26.7	18.2	27.1
비용	25.2	27.0	10.2	39.2	16.9	31.4	35.1	13.6
이용편리성	15.2	19.0	12.0	14.7	14.1	14.0	16.9	18.6
커뮤니케이션	11.6	16.0	8.3	10.8	5.6	8.1	14.3	23.7
사후서비스	8.4	5.0	9.3	10.8	5.6	12.8	6.5	5.1
다양한 연계관광상품	7.4	4.0	16.7	1.0	14.1	4.7	5.2	8.5
보험사연계	4.5	2.0	10.2	1.0	8.5	2.3	3.9	3.4

3-1 자료의 요약

- 자료의 요약
  - 가장 기초적인 데이터 이해 방법
- 다음과 같은 상황에서 분석방법은?
  - 매년 신입생이 1만 명 입학한다고 가정
  - 3년 동안의 신입생 성적을 비교해야 함
- 자료의 요약방법
  - 도수를 이용한 방법
  - 중심화 경향을 통한 방법
  - 산포경향을 통한 방법
  - 기타 요약방법
    - 줄기-잎사귀 도표(Stem-and-Leaf plot)
    - 백분위수(Percentile)
    - 상자 도표(Box plot)
- 표준화를 통한 자료의 변환

도수를 이용한 요약(1/2)

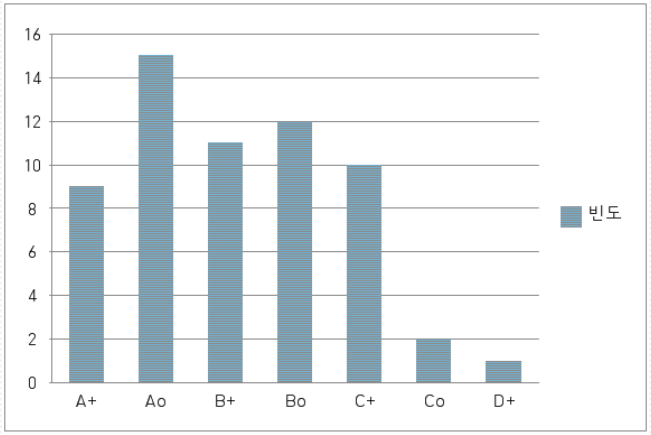
- **도수(빈도수: Frequency) : 특정한 구간(계급: Class)이나 조건을 만족하는 자료의 개수**
- **도수분포표: 모든 구간과 구간에 따른 도수를 정리한 표**
- **전체적인 자료의 분포를 파악하는데 도움을 줌**
- **그래프(히스토그램, 파이차트)와 같이 많이 사용됨**

B+ Ao B+ Ao B+ Co Bo Ao C+ Bo B+ Bo C+ B+ Bo
B+ Ao B+ D+ A+ A+ Ao C+ Bo Ao B+ A+ Ao C+ Bo
C+ Co Bo Ao C+ Bo A+ Ao C+ B+ Ao B+ Bo C+ Ao
A+ A+ Ao C+ Bo C+ Ao A+ A+ Ao B+ Bo A+ Ao Bo

성적	(빈)도수
A+	9
Ao	15
B+	11
Bo	12
C+	10
Co	2
D+	1
계	60

도수를 이용한 요약(2/2)

- 히스토그램

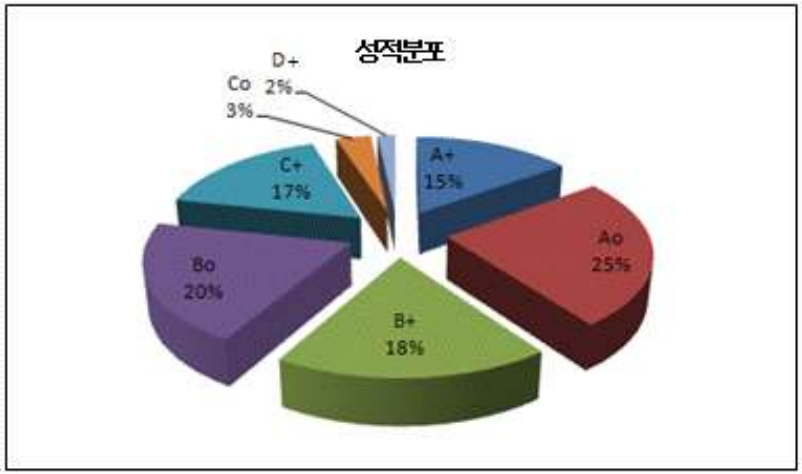


- 구간 폭의 결정
  - 구간 수 = (자료의 최대값-자료의 최소값)/(구간 폭의 넓이)
  - 구간 수는 7±2

- 상대도수, 누적도수, 누적상대도수

성적	빈도	상대도수	누적도수	누적상대도수
A+	9	0.15(=9/60)	9	0.15
Ao	15	0.25	24(=9+15)	0.40
B+	11	0.18	35	0.58 (=35/60)
Bo	12	0.20	47	0.78
C+	10	0.17	57	0.95
Co	2	0.03	59	0.98
D+	1	0.02	60	1.00
계	60	1.00		

- 파이차트



중심화 경향을 통한 요약

- **중심화 경향 (Central tendency) : 여러 개의 자료의 가운데, 또는 중심 위치에 해당하는 값**
- **예제 데이터 (12개)**

15	19	12	10	15	18	16	21	11	15	25	75

- **중앙값(Median): 자료를 크기 순으로 나열한 후 가운데에 위치한 값**
  - **중앙위치 = (*N*+1)/2 , *N*은 자료의 갯수**
- **최빈값(Mode) : 자료에서 가장 빈번하게 나타나는 값**
- **산술평균: 모든 자료의 합을 자료의 수로 나눈 값**
- **기하평균: 모든 자료를 곱한 결과의 N 제곱근, N은 자료의 개수**
- **위 데이터의 중심화 경향을 나타내는 값은 각각 얼마인가?**

- **이상치(Outlier)와 중심화 경향을 나타내는 값의 관계는?**

산포경향을 통한 요약

- **산포: 데이터가 서로 흩어진(떨어진) 정도**
  - **두 데이터 set A={49, 50, 51}, B={0, 50, 100}의 산술평균은?**
- **편차=관찰치-평균**
  - **두 데이터 set A, B의 편차는?**
  - **두 데이터 set A, B 편차의 합은?**
- **절대편차 = |편차|**
  - **두 데이터 set A, B의 절대편차는?**
- **편차 제곱의 평균 (분산, σ^2)**
  - **두 데이터 set A, B의 편차제곱의 합은?**
  - **두 데이터 set A, B의 편차제곱의 평균(분산)은?**
- **편차 제곱의 평균의 제곱근 (표준편차, σ)**
  - **두 데이터 set A, B의 표준편차는?**
  - **표준편차의 의미는? 데이터는 평균에 표준편차정도 떨어져 있음**
- **분산=표준편차^2**

기타 요약 방법(1/2)

- **줄기-잎사귀 도표(Stem-and-Leaf plot)**
  - **앞의 데이터를 줄기-잎사귀 도표로 나타내면**

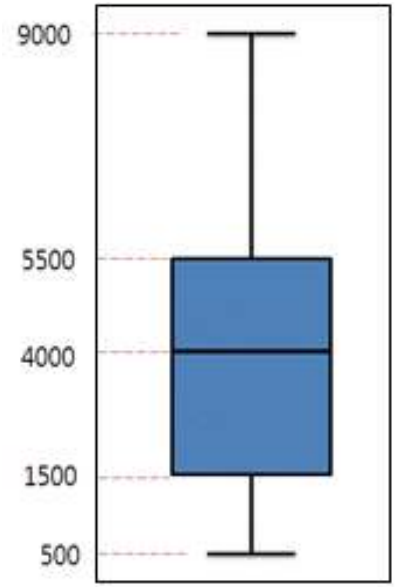
Stem	Leaf
1	1 5 5 6
2	2 4 7 8
3	1 4 6
4	2

- **백분위수(Percentile)**
  - **순서상 오름차순 정렬 후 그 순위를 100등분한 것**
  - **만일 한국의 연간 가구 소득의 백분위수를 계산했는데 제2백분위수가 1100만원 이라면 '1100만원 이하는 버는 가정이 전체의 2%라는 의미'**
- **사분위수(Quartile): 제1사분위수(제25백분위수), 제2사분위수(제50백분위수),제3사분위수(제75백분위수)**

기타 요약 방법(2/2)

- **상자도표(Box plot): 전체자료의 모습을 요약하기 위해 최소값, 제1사분위수, 제2사분위수, 제3사분위수를 이용하여 그린 도표**
  - **최대값이나 최소값이 너무 크거나 작으면 제90백분위수와 제10백분위수를 사용하기도 함**
  - **한국의 연간 가구소득의 백분위수가 다음과 같다면 상자도표는?**

백분위 (percentile)	해당 백분위 수 (만원/년)
10	500
25	1500
50	4000
75	5500
90	9000



자료의 표준화(Standardization)

- 여러 종류의 자료를 비교해야 하는데 평균이나 분산 등이 다른 경우?
  - 자료의 표준화가 필요
  - 예) 장학금을 주는 기준이 '직전학기 평점평균', '보호자의 월소득'인 경우에 평점은 0~4.5이고 월소득은 0~수 천만원
  - 가능한 방법: 평점과 월소득의 합으로? 등간이 아님
  - 대안: 평균과 표준편차를 이용한 표준화
    - 공식: $Z_A = \frac{X_A - \mu_A}{\sigma_A}$
- 연구문제1

초등학교 6학년 학생들의 평균 키는 150㎝, 표준편차는 15㎝이며, 평균 몸무게는 40㎏, 표준편차는 10㎏이라고 한다. 그런데 순이의 키는 160㎝인데 반해, 몸무게는 48㎏이다. 순이의 몸무게가 키에 비해 많이 나간다고 할 수 있는가?

확률과 확률분포

- **확률의 개념**
  - $\mathbf{Pr(A)} = \frac{\text{사건 } A\text{가 발생하는 가짓 수}}{\text{발생 가능한 모든 가짓 수}}, \mathbf{0} \leq \boldsymbol{Pr(A)} \leq \mathbf{1}$
  - **동전 2개를 던져 모두 같은 면이 나올 확률은?**
- **확률의 종류**
  1. **고전적 확률(논리적 확률): 논리적으로 계산이나 추론이 가능한 경우**
     - **예) 주사위를 한 번 던져서 3이 나올 확률은?**
  2. **상대빈도의 확률: 전체의 사건 중에 특정 사건이 발생한 비율**
     - **예) 공장에서 생산된 제품 만개 중 100개 가 불량일 때 제품이 불량인 확률은?**
  3. **주관적 확률(경험적 확률): 논리적 계산이나 추론이 불가능하고 일상생활에서 여러 차례 경험을 통해 주관적으로 판단된 확률**
     - **예) 친구 A에게 전화를 걸었을 때 받을 확률?**

결합확률과 주변확률

- **결합확률(Joint probability)**
  - **두 개 이상의 사건이 동시에 일어나는 확률**
  - **연습문제**
    - **다음의 표에서 임의의 환자 1명이 당뇨이면서 고혈압일 확률, Pr(당뇨∩고혈압)은?**

당뇨 \ 고혈압	있음	없음	합계
있음	250	50	300
없음	100	100	200
합계	350	150	500

- **주변확률(Marginal probability)**
  - **두 가지 이상의 사건이 동시에 일어나는 경우 하나의 사건이 특정한 값을 가지고 다른 사건은 발생 가능한 여러 경우를 모두 다 합친 확률**
    - **연습문제**
      - **500명 중 임의로 선택한 한 명이 고혈압을 보유할 확률은?**

조건부 확률

- **조건부 확률(Conditional probability)**
  - **특정한 사건이 발생하였다는 전제하에 다른 사건이 발생할 확률**
  - **표현법 Pr(B|A): A라는 사건이 발생했다는 전제하에 B사건이 발생할 확률**
    - **공식 Pr(B|A)=** $\frac{Pr(B \cap A)}{Pr(A)}$
  - **연습문제2**
    - **한림전자에서는 소비자들이 TV 광고 기억여부와 냉장고 구매여부에 관해 800명을 대상으로 조사한 결과 다음과 같은 표를 얻었다.**

	광고를 기억함	광고를 기억하지 못함	합계
구입함	150	90	240
구입하지 않음	250	110	360
합계	400	200	600

    - **Q1) 광고를 기억하지 못한 고객이 냉장고를 구입할 확률은?**
    - **Q2) 광고를 기억했다는 조건하에 구입을 하지 않을 확률은?**

확률변수와 확률분포

- 변수 : 미지의 수, 정해지는 수
  - 예) $2x+6=10 \rightarrow x = 2$
- 확률변수: 여러 개의 값이 가능한 변수와 변수가 특정한 값을 가질 확률이 있는 경우
  - 예) 주사위 1개를 던져 나온 눈의 수 = x
  - 변수+확률을 가지고 있음
- 확률분포표: 확률변수의 값과 그에 따른 확률을 모은 표
- 연습문제
    - 동전 두 개를 던져 앞면이 나온 개수의 확률분포표?

x	Pr($X=x$)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

확률변수의 기대값과 분산

- 확률변수의 기대값의 계산
  - 기대값은 확률변수의 값과 그 때의 확률을 곱한 총합
    - $E(x)= \Sigma x_i Pr(X=x_i)$
  - 예) 주사위 1개를 던져 나온 눈의 수의 기대값
- 기대값과 평균의 관계
  - 평균=기대값, 즉 $\mu= E(x)$
- 연구문제 3
  - 오른쪽 그림과 같은 다트판에 다트를 던져서 표시된 금액을 상금으로 가져가는 게임을 하기로 한다. 던질 때 마다 500원을 내고 다트가 다트판을 벗어나면 새로 던질 수 있는 기회를 무상으로 제공한다. 두 가지 질문에 답해보자
    - Q1) 한번 던질 때 상금의 기대값은 얼마인가?
    - Q2) 이 게임을 계속하는 것이 유리한가? 불리한가?