조사방법론1(30)+조사방법론2(30)+사회통계(40)

사회통계 [책:사회조사분석사 2급]
제1장 자료의 정리
통계학 일반, 자료정리
제2장 기술통계
대표값, 산포도, 비대칭도
3장 확률과 확률분포
확률과 확률변수, 확률분포
제4장 통계적 추정
추정일반, 점추정, 구간추정
제5장 가설검정
가설검정의 개념 및 용어, 가설검정의 오류, 가설의 요소 및 검정절차, 모평균의 검정
제6장 표본크기
표본크기 일반, 표본크기의 결정
제7장 통계분석
교차분석, 분산분석, 상관분석, 회귀분석, 기타분석

[책:사회조사분석]
제Ⅰ부 경험적 연구의 기초
제1장 과학적 연구
제2장 경험적 연구방법

제Ⅱ부 자료의 수집
제3장 자료 수집 방법
제4장 표집 방법(표본추출방법)
제5장 질문지 작성
제6장 측 정
제7장 지수와 척도

제Ⅲ부 자료의 분석
제8장 기술통계(자료정리, 기술통계)
제9장 추리통계(확률, 확률분포, 통계적추정, 가설검정)
제10장 분할표분석
제11장 집단 간 비교분석
제12장 회귀분석과 경로분석
제13장 군집분석, 다차원척도분석, 판별분석 및 로지스틱회귀분석

8장. 기술통계 [책:사회조사분석]

변수(Variable, 관심의 대상이 되는 특성)에 따른 자료의 분류:

① 범주형 자료(Categorical Data):
특성이 속하는 범주로 표현될 수 있는 자료
- 혈액형자료, 성별자료, 취업상태자료

② 수치형 자료(Numerical-valued Data):
숫자로 표현될 수 있는 자료
ⓐ 이산형자료(Discrete Data) - 월별판매량자료, 교통사고수
ⓑ 연속형자료(Continuous Data) - 키, 생존기간 자료

* 수치형자료는 범주형자료로 변환가능

제1절. 도표

1. 도수분포표: 범주(구간), 도수, 상대도수 정리한 표

도수(빈도수,Frequency): 각 범주(구간)에 속하는 관측값의 수
-Category can be built by surveyor or data analyst.

상대도수(Relative Frequency): 각 범주의 도수의 전체자료의 크기에 대한 비율
= 그 범주의 도수/전체자료수

(1) 자료값의 종류가 적은 경우(범주형,이산형자료)

eg) 종교 조사(인구주택총조사 2005년) & 혈액형조사

종교	도수	상대도수
불교	10,726,463	22.8
개신교	8,616,438	18.3
카톨릭	5,146,147	10.9
원불교	129,907	0.3
기타	351,811	0.7
무교	21,867,055	46.5
무응답	188,104	0.4
계	47,025,925	100%

혈액형	도수	상대도수
A	7	17.5
AB	4	10.0
B	13	32.5
O	16	40.0
계	40	100%

(2) 자료값의 종류가 많은 경우(수치형자료)~ 구간도수분포표

5-15개의 구간으로 나누기- 자료의 최소값, 최대값 이용하여 각 구간 길이 같게

eg) 60명 중간고사 성적, { 88, 80,...,55} → 오름차순 정렬 {36,48,...,99}

자료의 최소값 = 45, 자료의 최대값 = 99 → 각구간길이=10

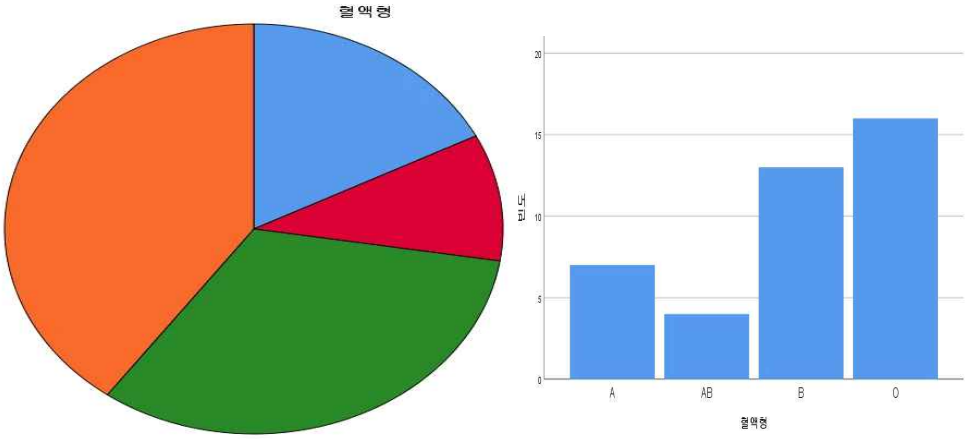
점수	도수	상대도수
30-39	1	1.7
40-49	3	5.0
50-59	8	13.3
60-69	10	16.7
70-79	17	28.3
80-89	13	21.7
90-100	8	13.3
계	60	100%

2. Chart

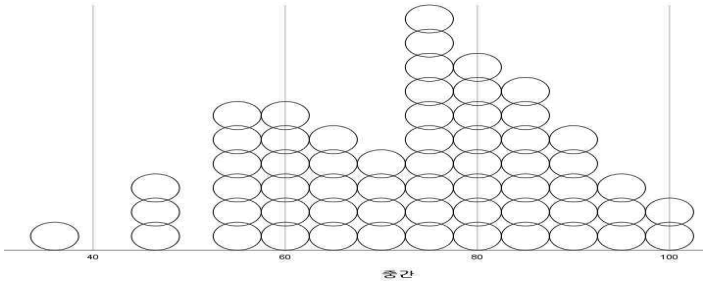
(1)원도표(pie chart): 조각의 각도=상대고수*360°

(2)막대그림(bar chart): 상대도수 or 도수를 y축으로

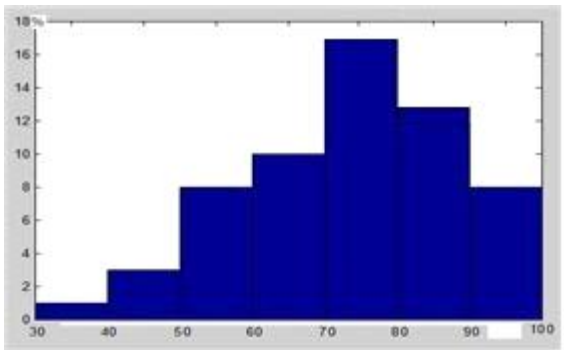
eg)혈액형자료



(1) 점도표(Dot Plot): 자료수 많은 경우, 자료의 퍼진 정도 쉽게 파악



(2)Histogram: 구간도수분포의 상대도수를 y축으로 한 막대그림



구매빈도 Stem & 잎

1.00	3 .	6
3.00	4 .	588
8.00	5 .	34456789
10.00	6 .	2222346778
17.00	7 .	12234445556778889
13.00	8 .	0112344577788
8.00	9 .	11134789

줄기 너비: 10

(3)줄기-잎 그림(Stem-and-Leaf Plot)-Histogram의 변형

-구간도수분포 이용

-줄기에 구간의 끝점의 몇 개 자리수 표시~ 구간시작값/단위(구간길이)

-잎에 관측값에서 몇 개 자리수 뺀 값 표시

㉠ 단위=각구간의 길이

㉡ 줄기: (각 구간의 시작점/단위) 표시 - 줄기마디

㉢ 각 줄기마디의 잎부분: (관측값- 줄기의 마디의 값×단위) 표시 ~오름차순으로

eg) 60명 중간고사 성적

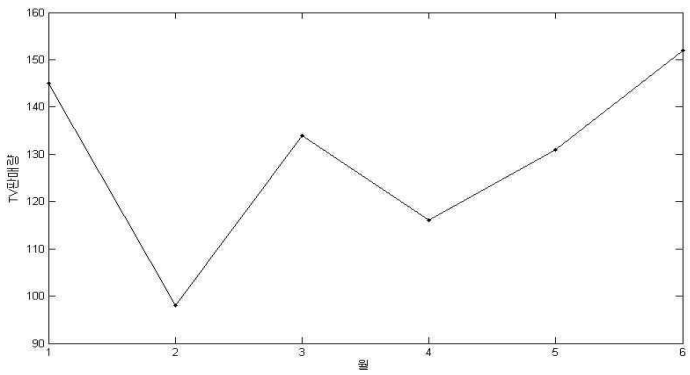
㉠ 단위=10점

㉡ 줄기에 3,4,5,6,7,8,9 표시

㉢ 30−39 의 관측값 = 36 → 줄기마디 3의 잎부분에 (36−3×10)=6,
40−49 의 관측값 = 45,48,48 → 줄기마디 4의 잎부분에 (45−4×10)=5,
(45−4×10)=5, (48−4×10)=8,

* Histogram보다 많은 정보 표시 *특정자료의 상대적 위치파악 용이

(4)꺽은선그림(polygon graph)~ 변수의 변화에 따른 분포의 변화



제2절. 기술통계치

비율(proportion): 상대도수

비(ratio) : 두변수의 관계 ~ 성비(gender ratio)=여성100명당 남성수

율(rate): 증감의 변화 ~ 인구성장률

제3절. 중심집중치 :주어진 자료의 중심위치를 나타내는 값 or 자료를 표현하는 대표값

1.평균(mean): $\frac{1}{n}\sum_{i=1}^{n} x_i$

2.중앙값(median):자료를 오름차순으로 정리한 경우 중앙에 해당되는 값

n=홀수: $x_{(n+1)/2}$, n=짝수: $(x_{(n/2)} + x_{(n/2+1)})/2$

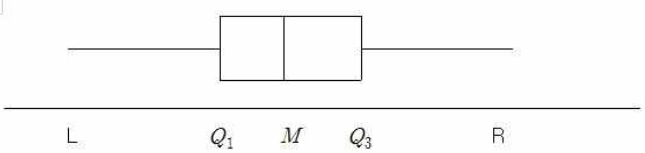
3.최빈값(mode)

제4절. 산포도(dispersion) : 주어진 자료가 흩어져 있는 정도

1.범위(range): $x_{(n)} - x_{(1)}$

2.사분편차: $(Q_3 - Q_1)/2$

Q_3:제3사분위수(자료의 3/4이 이값이하), Q_1:제3사분위수(자료의 1/4이 이값이하)



상자그림(Box plot)

3.분산& 표준편차: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$

4.변동계수(coefficient of variation,$\sigma/\bar{x}$): 측정단위가 서로 다른 자료의 산포도를 비교할 때.

*왜도(Skewness): 분포가 좌우로 기울어진 정도

$$sk = \frac{\sqrt{n(n-1)}}{n-2} \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^{3/2}}$$

sk=0 ~ 좌우 대칭

sk<0 ~ 왼쪽으로 기울어진 형태 (skewed left)

sk>0 ~ 오른쪽으로 기울어진 형태 (skewed right)

*첨도(Kurtosis): 자료가 제일 많이 모인 부분(mode)의 뾰족한 정도

$$g_2 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3, \quad kr = \frac{n-1}{(n-2)(n-3)}((n+1)g_2 + 6)$$

kr=0 ~ 표준정규분포와 같은 모양

kr<0 ~ 표준정규분포보다 더 밋밋한 모양(flat)

kr>0 ~ 표준정규분포보다 더 뾰족한 모양 (peaked)

SPSS

(1) 혈액형자료(범주형자료)

메뉴바의 분석(A) – 기술통계량(E) – 빈도분석(F) 창에서 변수부분으로 이동

도표(C) ––막대도표, 원도표, 히스토그램 중에서 고르기, 창에서 확인 √

(2) 60명 성적자료(수치형자료)의 점도표 & 기술통계량

① 메뉴바의 그래프(G) >>레거시대화상자 >>산점도/점도표

– 단순점도표: x축변수로 변수(점수)이동

메뉴바의 분석(A) – 기술통계량(E) – 데이터 탐색(E) 창에서 점수를 종속변수로 이동 (통계량(S)에서 기술통계 & 백분위수 √)

도표에서 기술통계의 줄기와 잎그림 √(자동으로 상자그림 생성됨)

② 메뉴바의 분석(A) – 기술통계량(E) –기술통계(D) 창에서 점수를 종속변수로 이동 – 옵션에서 모두 선택

③ 메뉴바의 변환(T)–다른변수로 코딩변경 (다른변수명=성공여부)

성적을 오른쪽으로 이동(이용)

출력변수 이름(N) : "성공" – 바꾸기 ↵

기존 값 및 새로운 값(o) ↵

최저값에서 다음 값까지 범위(G) "59"

출력변수가 문자열임(B) √

새로운 값–기준값(A): "실패"

추가 ↵

기타 모든값

새로운 값–기준값(A): "성공"

추가 ↵ -계속 ↵ – 확인 ↵

	중간	성공여부
1	88	성공
2	80	성공
3	75	성공
4	74	성공
5	67	성공
6	54	실패
7	66	성공
8	77	성공
9	62	성공
10	99	성공

메뉴바의 분석(A) >> 기술통계량>> 빈도분석(F)

		빈도	퍼센트	유효 퍼센트	누적 퍼센트
유효	성공	48	80.0	80.0	80.0
	실패	12	20.0	20.0	100.0
	전체	60	100.0	100.0	

메뉴바의 분석(A) >> 평균비교>> 평균분석 ~ 종속변수=중간, 레이어1=성공여부
옵션에서 필요한 통계량 이동

성공여부	평균	표준편차
성공	78.58	10.273
실패	51.92	6.612
전체	73.25	14.419

2변량 자료(Bivariate data, 교차분류된 자료(cross-classified data))
: 각 개체로부터 2가지 변수의 값을 구하는 경우 얻어진 자료
(1) 분할표(Contingency Table)-범주형자료
eg) 성별과 커피선호도에 대한 응답결과와 성별~ 안좋아함-좋아함, 남~여

		안좋아함	좋아함	
성별	남	160	344	504
	여	224	445	669
		384	789	1173

임의로 선택된 사람이 커피좋아하는 여자일 확률=445/1173=0.38
임의로 선택된 사람이 커피좋아하는 사람일 확률=789/1173=0.67
임의로 선택된 사람이 남자인 경우 그사람이 커피좋아하는 사람일 확률=344/504=0.68

카이제곱 독립성검정:
H_0: 성별과 커피선호도는 연관성이 없다.
유의확률<0.05이면 유의수준5%로 H_0기각 ~ 연관성이 있다고 결론.

(2) 산점도(scatter diagram)-수치형자료
자료 = $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$
관측값 (x_1, y_1) - x_1 : 개체1의 변수1의 관측값 , y_1 : 개체1의 변수2의 관측값
산점도 : 2변수의 관계를 개략적으로 파악 가능

2. 상관계수(Correlation coefficient, r)

: 2변수의 선형관계의 측도

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}},\ S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2,\ S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

① $-1 < r < 1$

② r의 절대값이 클수록 강한 선형관계 나타낸다.

r ≈0.9

r ≈0.5

r ≈−0.9

r ≈−0.5

r ≈0

r ≈0

r=0.9~ 강한 양의 상관관계 ~ not indicate a causal relationship btwn 2 variables