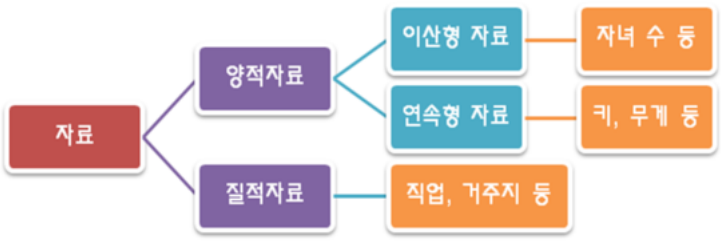
자바쌤~!

https://cafe.naver.com/javassem

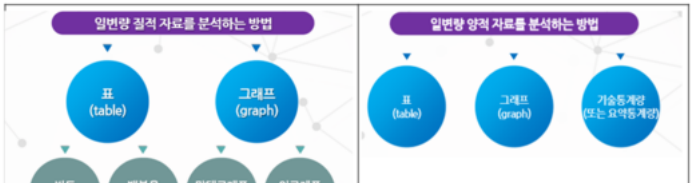
63기-자바 & 빅데이타 >

[파이썬] 데이타종류와 그래프

자바쌤 1:1 채팅
2020.07.06. 08:52 조회 1



(1) 일변량 질적 자료 분석



카페정보 나의활동

매니저 자바쌤
since 2013.03.15.
카페소개

새싹4단계
433 초대하기
즐겨찾는 멤버 65명
게시판 구독수 4회
우리카페앱 수 0회

카페 글쓰기

카페 채팅

즐겨찾는 게시판
전체글보기 4,146

KOSMO
63기-자바 & 빅데이타
팁
심심문제
기타등등
JAVA
Oracle/Mysql
JSP
WebUI
파이썬
Spring
리눅스
하둡+eco
프로젝트

공유해요
알고리즘도전
기술면접 준비하기
실존 면접
기회는 준비된자에게
프로젝트 팁

쉬어가기
청년아카데미
전문가과정
공지사항
카페회칙

카페탈퇴하기

(frequency) (percent) (bar chart) (pie chart)

빈도와 백분율의 개념

빈도	백분율
• 자료가 가지는 각각의 값이 몇 개가 있는지를 구한 수치	• 자료가 가지는 각각의 값이 전체를 100으로 보았을 때에 얼마나 차지하고 있는지를 알려주는 수치

1. 질적 자료

(1) 일변량 질적 자료 분석

1- 막대그래프

- 질적 자료의 특징을 파악하기 위해서 작성하는 그래프
- 막대그래프를 작성할 때에 빈도 또는 백분율을 이용하며 일반적으로 빈도를 더 많이 사용함
- 막대그래프는 세로 막대그래프와 가로 막대그래프가 있음
  - 세로 막대그래프인 경우 : y축제목이 꼭 있도록 해야 함
  - 가로 막대그래프인 경우 : x축제목이 꼭 있도록 해야 함

※ x축, y축의 제목이 막대의 높이(또는 길이)가 무엇을 의미하는지 알 수 있도록 해 줌

- 막대그래프의 최소값은 0부터 시작, 최대값이 포함되도록 y축 눈금을 설정해야 함

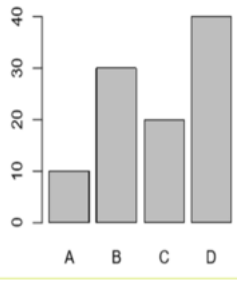
※ 막대그래프에서 절단효과를 넣지 않도록 주의!
(절단효과를 넣으면 의도하지 않게 정보를 왜곡시킬 수 있음)

1. 질적 자료

(1) 일변량 질적 자료 분석

1- 막대그래프

- 주로 범주형 자료를 그림으로 표현하기 위해 사용
  - 막대의 높이가 해당 범주의 빈도수 또는 비율이 됨
  - 각 막대는 일반적으로 분리되도록 그림



1. 질적 자료

(1) 일변량 질적 자료 분석

2- 원그래프

- 범주형 자료의 각 범주별 빈도수(또는 각 범주의 비율)를 원 내부의 각도에 비례하게 그린 그림
- 범주형 변수 중 주로 명목형에 적합함

- 원그래프는 질적 자료가 가지는 항목(값)이 5개 이하인 경우에 적당함
- 각 항목이 전체 중에서 얼마나 차지하고 있는지를 표현하는 데에 유용함
- 강조하고 싶은 항목이 있다면 해당 조각을 분리해서 표현하는 것도 좋음

2. 양적 자료

(1) 일변량 양적 자료 분석

2. 양적 자료

(1) 일변량 양적 자료 분석

- 양적 자료는 질적 자료와 다르게 자료가 가지는 각각의 값에 대한 빈도와 백분율을 구하지 않음

※질적 자료는 데이터가 많다고 하더라도 데이터가 가지는 값의 종류는 몇 개가 되지 않지만,
양적 자료는 데이터가 가지는 값의 종류가 많기 때문

※예 : 5,000명에게 성별과 신장을 조사했고, 5,000명 모두가 성별과 신장에 대해서 응답해 주었을 때
- 성별이라는 질적 자료에는 5,000개의 자료가 있음
- 5,000개가 가지는 값은 두 개의 값인 '남자(또는 1)' 나 '여자(또는 2)'로 구성되어 있음
- 신장이라는 양적 자료에도 5,000개의 자료가 있음
- 5,000개가 가지는 값은 성별처럼 두 개의 값이 아니라 훨씬 더 많은 다양한 값들로 구성되어 있음

2. 양적 자료 - (2) 히스토그램

- 연속인 자료에 대해서 구간을 나누어 해당 구간에 포함된 자료의 빈도(또는 비율)을 기둥의 높이로 하여 막대를 연속으로 그린 그림
- 막대그래프와 히스토그램의 차이

막대그래프	VS	히스토그램
범주형 자료에 적용되고 따라서 각 기둥이 분리되어 그려짐		연속인 자료에 적용되고 따라서 기둥을 분리하지 않음

- 히스토그램은 구간의 빈도나 백분율을 이용하여 작성
- x축은 양적 자료의 구간, y축은 각 구간의 빈도 또는 백분율이 됨
- 히스토그램은 막대그래프와 다르게 막대의 가로가 의미를 가짐
  - 히스토그램에서의 막대의 가로는 각 구간의 너비로 정보를 가지고 있음

2. 양적 자료 - (2) 히스토그램

히스토그램의 활용

- 히스토그램을 통하여 다음과 같은 특징을 파악할 수 있음
  - 각 구간의 현황
  - 빈도가 가장 많은 구간
  - 빈도가 가장 작은 구간
  - 무게 중심
  - 대칭 여부
  - 이상치(outlier) 유무 : 이상하게 크거나 이상하게 작은 값의 유무
  - 단봉 : 봉우리가 한 개
  - 쌍봉 : 봉우리가 두 개

2. 양적 자료 - (3) 박스플롯

사분위에 해당하는 부분은 상자로 표현하고 그 밖의 범위를 선으로 연결하는 그림

상자그림(boxplot)은 이상치 유무, 대칭여부, 자료의 분포에 대한 정보를 얻을 수 있음

[분석]

중앙값 : mean 아닌 median
최소값 : min 아님
최대값 : max 아님

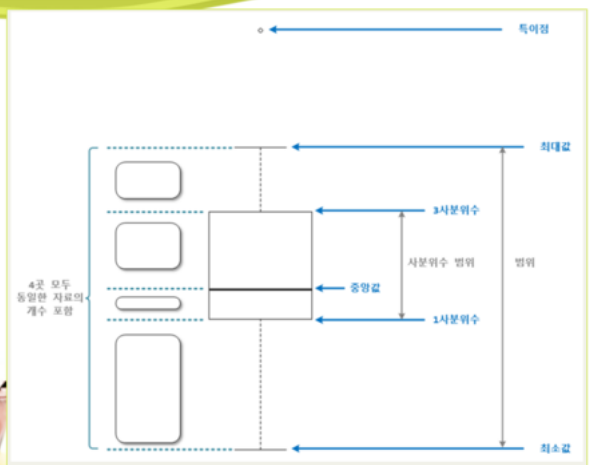
2. 양적 자료 - (3) 박스플롯

상자 그림(boxplot) 해석방법

박스 플롯은 박스와 박스 바깥의 선(whisker)으로 이루어져 있습니다.

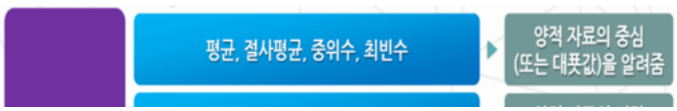
구분	설명
whisker	상자의 좌우 또는 상하로 뻗어나간 선
박스 내부의 가로선	중앙값을 나타냅니다.
lower whisker	최소값 '중앙값 - 1.5 x IQR'보다 큰 데이터 중 가장 작은 값
upper whisker	최대값 '중앙값 + 1.5 x IQR'보다 작은 데이터 중 가장 큰 값
IQR	Inter Quartile Range 제3사분위수 - 제1사분위수 실수 값 분포에서 1사분위수(Q1)와 3사분위수(Q3)를 뜻하고 이 3사분위수와 1사분수의 차이(Q3 - Q1)를 IQR(interquartile range)라고 합니다.
점	이상치(outlier; 아웃라이어) ≒ 특이점 lower whisker보다 작은 데이터 또는 upper whisker보다 큰 데이터가 여기에 해당됩니다.

2. 양적 자료 - (3) 박스플롯



[출처] https://codedragon.tistory.com/

2. 양적 자료 - (4) 기술통계량



2. 양적 자료 - (4) 기술통계량

1- 중심값(대푯값) : 평균, 중위수

평균

반응

- 자료의 합을 자료의 개수로 나눈 값으로 지나치게 큰 값이나 작은 값에 빨리 반응함

중앙값

반응

- 자료를 크기순으로 나열할 때 중앙에 위치한 값으로 지나치게 큰 값이나 작은 값에 늦게 반응함

개념

- 자료의 개수가 홀수이면 중앙에 한 값이 위치하고 개수가 짝수이면 중앙에 위치한 두 값의 평균값
- 중위값, 중위수 등으로도 부름

2. 양적 자료 - (4) 기술통계량

(2) 산포 : 양적자료의 퍼짐 정도

범위(Range) : 최대값(Max) - 최소값(Min)

- 양적 자료의 퍼짐(산포)을 알려주는 값
- 범위가 큰 값을 가지면 자료가 많이 퍼져 있다고(또는 많이 다르다고) 해석함

※범위도 평균과 마찬가지로 아주 작은 값이나 아주 큰 값에 영향을 받기 때문에 해석에 주의가 필요함

사분위 범위(IQR : Inter Quartile Range)

- 제3 사분위수에서 제1 사분위수의 값을 뺀 값

2. 양적 자료 - (4) 기술통계량

(2) 산포 : 양적자료의 퍼짐 정도

분산(Variance)

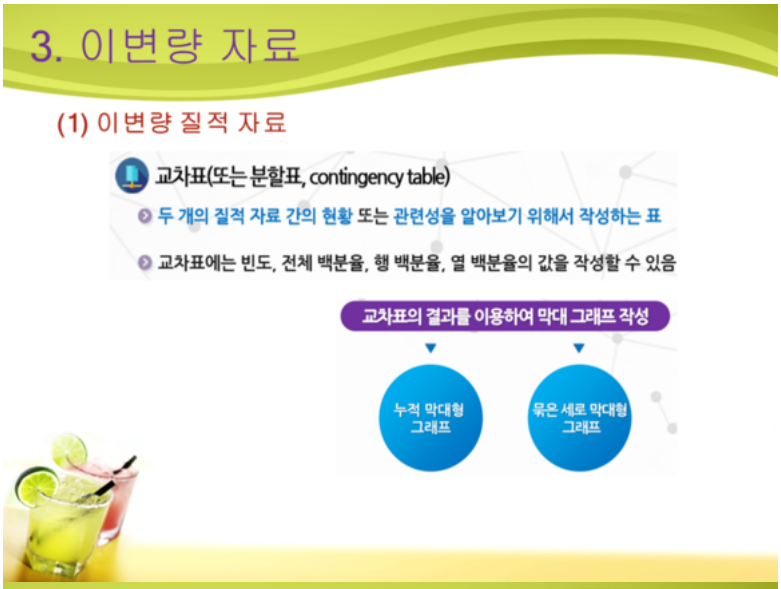
- 각 자료에서 평균을 뺀 값, 즉 편차(deviation)를 제곱한 값을 평균한 것
- 양적 자료의 퍼짐(산포)을 알려주는 값임
  - 분산이 크면 자료가 많이 다르다고 해석함
  - 분산이 작으면 자료가 비슷하다고 해석함

※분산도 아주 작은 값이나 아주 큰 값이 영향을 받는 특징이 있음

표준편차(Standard Deviation)

- 분산의 제곱근(square root)이며, 양적 자료의 퍼짐(산포)을 알려주는 대표적인 값
- 표준편차는 자료들이 평균과 얼마나 다를까를 알려줌
  - 표준편차가 크게 나오면 평균과 차이가 많이 나는 자료들로 구성되어 있다고 해석함
  - 표준편차가 작게 나오면 평균과 비슷한 자료들로 구성되어 있다고 해석함

※ 표준편차도 평균, 범위, 분산과 동일하게 이상한 값에 영향을 많이 받음



댓글 댓글알림

김용선

댓글을 남겨보세요

등록