[개념 통계] 도수 분포표와 히스토그램

https://arnongdatanote.tistory.com/24

[개념 통계] 도수 분포표와 히스토그램

기술통계는우리가수집한데이터가어떻게생겼는지(대표값은무엇인지? 어떻게분포하고있는지?)를파악하는데사용하는통계기법이라고할수있고, 추리통계는그수집한데이터를이용해서우리가예측하고싶어하는것을확률적으로판단하는통계기법들이라고할수있습니다.

- [통계 노트/통계 개념 정리] - [개념 통계] 기술 통계와 추리 통계란 무엇인가?

이번포스팅에는기술통계기법중에서하나인도수분포표(frequency table)와히스토그램(histogram)에대해서알아보겠습니다. 도수분포표와히스토그램은중고등학교과정을무사히마치셨다면매우익숙한용어일거라고생각합니다. 하지만그것을왜사용하고또그것을어디에쓰는지는잘모르고있습니다. 그래서이번포스팅에서는도수분포표와히스토그램을데이터를파악하고이해하는데어떻게쓰일수있는지알아보겠습니다.

도수 분포표 (Frequency table): 특정 구간에 속하는 자료의 개수를 나타내는 표

도수분포표는영어로 Frequency table입니다. 직역하자면빈도표입니다. 즉도수분포표란자료의 분표를몇개의구간으로나누고, 나누어진각 구간에속하는자료가 몇 개인지정리한표입니다. 그런데왜구간을나눌까요? 그 이유는 구간을나누면개별적인데이터를보는것보다데이터의전체적인분포 즉 모양을요약해서볼수있기때문입니다. 예컨대우리가어느 학년55명의 수학 점수데이터를가지고있다고합시다. 개별55개 데이터가 숫자로만 주욱~ 있다고 생각해봅시다. 수학 점수분포가한 눈에들어올까요? 당연히안들어오겠죠. 그래서데이터가어떻게분포하고있는지 점수 범위를 0-10점, 10-20점, 20-30 점… 90-100점이렇게구간을만

들고이범주에들어간인원수를세면간단하게 55명의 수학 점수 분포를파악할수있겠죠. 이게바로도수분포표입니다.

그렇다면도수분포표는어떻게만들까요? 바로다음과같은 절차와 방법으로만듭니다.

(1) **자료의갯수를센다.**

(2) **자료 내에서 최대 / 최소값을찾는다.**

(3) **몇 개구간(급의 수)으로나눌지결정한다.**

- ▶ 자료의개수나분포에따라달라져야한다.
- ▶ 각구간에 5개이상의숫자가들어가도록하는것이좋다.
- ▶ 너무많은구간을나누지않도록한다. (일반적으로 5-15구간)

(4) **구간의 폭 (급의폭)을구한다.**

- ▶ 구간폭 = (최대값-최소값)/구간수
- ▶ 되도록이면정수, 짝수, 5의배수를사용하는것이좋다.

(5) **구간의 경계값 (급의경계값)을구한다.**

(6) **구간별 자료의 갯수 (도수)를 적는다.**

실제로 도수 분포표를 작성해봅시다. 아래 자료는 대학에서 임의로 선정한 남학생 55명의 신장(단위 cm)을 기록한 것입니다. 이제 이것에 대한 도수 분포표를 작성해 봅시다.

170	178	171	168	173	178	171	174	170	170	175
170	169	166	162	170	171	175	175	171	171	170
172	179	164	170	181	178	180	177	166	169	168
165	163	175	166	178	165	168	167	177	168	177
174	174	176	179	169	173	167	170	173	170	162

▲R과 함께하는 통계학의 이해 (최용석, BigBook) 자료 인용

(1) 자료의 갯 수를 센다.

▶ 55개

(2) 자료 내에서 최대 / 최소값을찾는다.

▶ 최대 180, 최소 162

(3) 몇 개구간(급의 수)으로나눌지결정한다.

▶ 구간 수는 5개로 설정

(4) 구간의 폭 (급의폭)을구한다.

▶ 구간폭 = (최대값-최소값)/구간수

▶ (180-162) / 5 = 3.6

▶ 3.8은 정수가 아니므로 구간 폭은 4로 정한다.

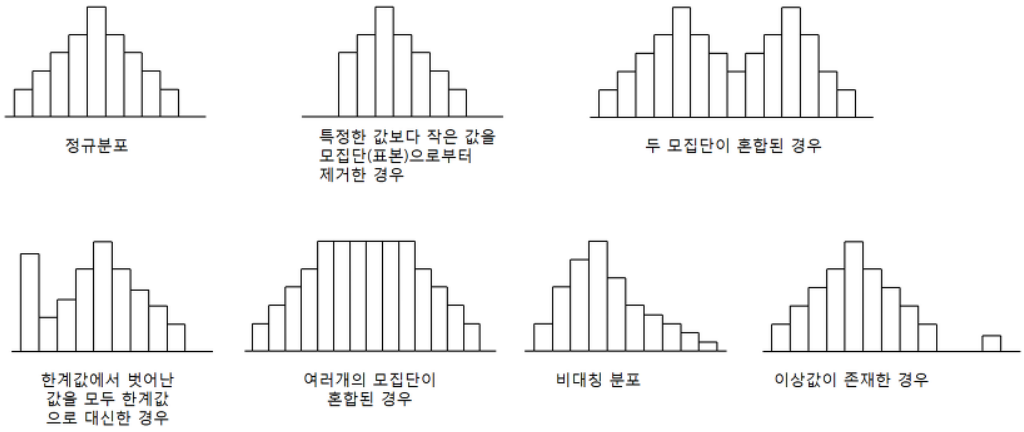
(5) 구간의 경계값 (급의경계값)을구한다.

(6) 구간별 자료의 갯수 (도수)를 적는다.

계급구간(cm)	도수
161.5 이상 165.5 미만	6
165.5 이상 169.5 미만	12
169.5 이상 173.5 미만	18
173.5 이상 177.5 미만	11
177.5 이상 181.5 미만	8
합계	55

히스토그램 (Historgram): 도수 분포표를 시각적으로 표현한 막대 그래프

위의 표처럼 도수분포표가완성되었습니다. 위 도수 분포표는 구간 수가 4개 밖에 없어서 그나마 쉽게 데이터 분포가 어떻게 생겼는지 파악할 수 있습니다. 그런데 구간이 10개 이상이라면? 그렇다면 도수분포표에 숫자가너무많습니다.따라서 구간별데이터가가시적으로들어오지않게 될 것입니다. 이러한 상황에서 도수분포표를눈에확!들어오게 만들어주는것이바로히스토그램입니다. 히스토그램의 x축(가로축)은 구간을나타내고, y축(세로축)은각구간별빈도수를나타냅니다. 그렇다면 히스토그램으로 우리는 무엇을 알 수 있을까요? 당연히 구간별 빈도수 겠지요? 하지만 그것보다 더 중요한 것. 바로 그 빈도수가 무엇을 결정할까요? 바로 히스토그램의 모양입니다. 그리고 그 히스토그램의 모양으로 우리는 우리가 수집한 데이터가 어떻게 생겼는지 한눈에 볼 수 있습니다. 바로 아래 그림처럼 말이죠.



즉, 히스토그램을 그리면, 수집한 데이터가 종모양(정규분포)를 모양을 이루고 있는지, 아니변 두 집단이 혼합된 경우처럼 생겼는지, 특정 구간에 빈도가 몰려 있는 비대칭적인 분포를 이루고 있는지, 또는 유독 튀는 이상한 값이 들어 있는지를 한 눈에 파악할 수 있습니다. 다시 요약하자면, 도수 분포표와 히스토그램을 이용하면 우리가 수집한 데이터를 요약하여 전반적인 생김새(분포)를 파악할 수 있는 것입니다.