인공지능 수학: 통계학
1. 기본 개념

윤상민

E-mail : smyoon@kookmin.ac.kr
Office : 02-910-4645

개념 정의

- 통계학(statistics)
  - 데이터의 수집 (collection), 구성 (organization), 분석 (analysis), 해석 (interpretation), 표현 (presentation)에 관한 학문
  - 기술통계학 (descriptive statistics)
  - 추측통계학 (inferential statistics)

개념 정의

- 모집단 (population)
  - 어떤 질문이나 실험을 위해 관심의 대상이 되는 개체나 사건의 집합
  - 전교 남학생의 키
- 모수(parameter)
  - 모집단의 수치적인 특성
  - 키의 평균
- 표본(sample)
  - 모집단에서 선택된 개체나 사건의 집합

도수 (Frequency)

- 정의
  - 어떤 사건이 실험이나 관찰로부터 발생한 횟수
- 표현 방법
  - 도수분포표 (Frequency Distribution Table)
  - 막대그래프 (Bar graph)
    - 질적 자료
  - 히스토그램 (Histogram)
    - 양적 자료

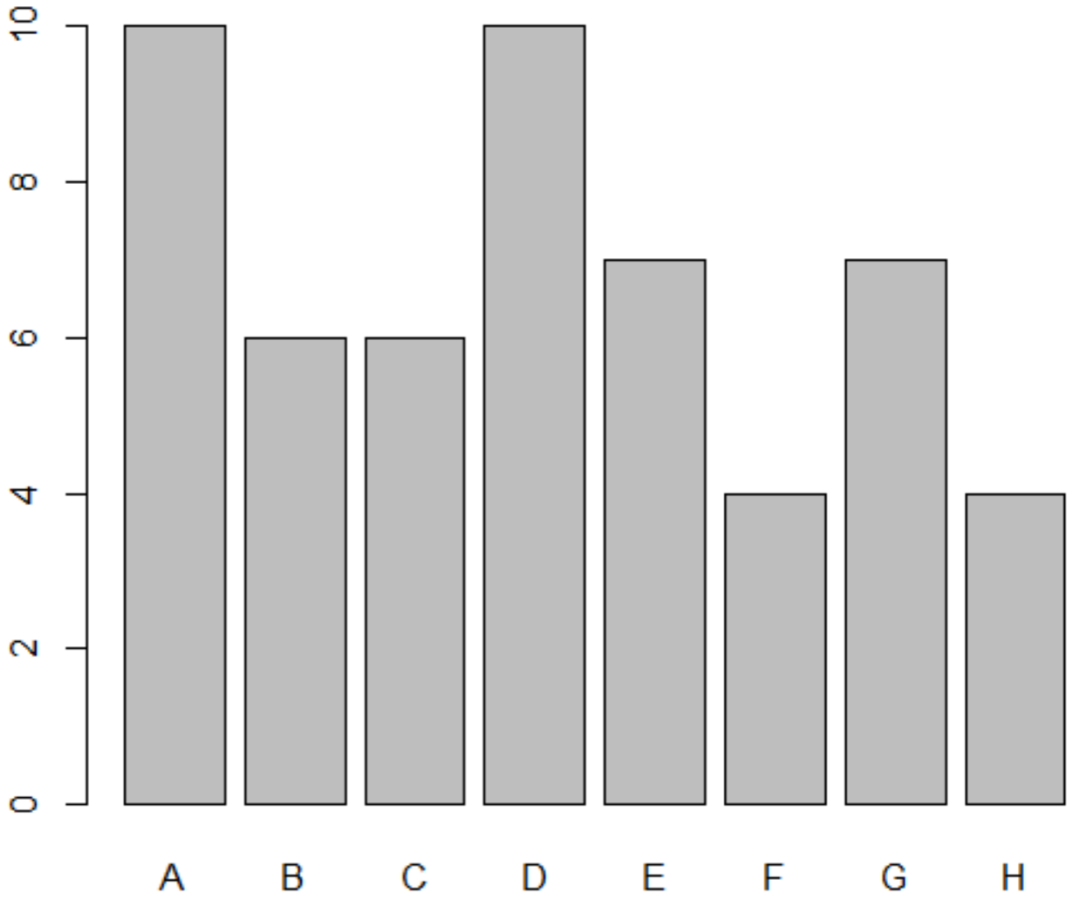
도수

- 질적 데이터
  - A A A A A A A A A A B B B B B B C C C C C C D D D D D D D D D E E E E E E E F F F F G G G G G G G H H H H H
- 도수분포표

A	B	C	D	E	F	G	H
10	6	6	10	7	4	7	4

도수

- 막대그래프

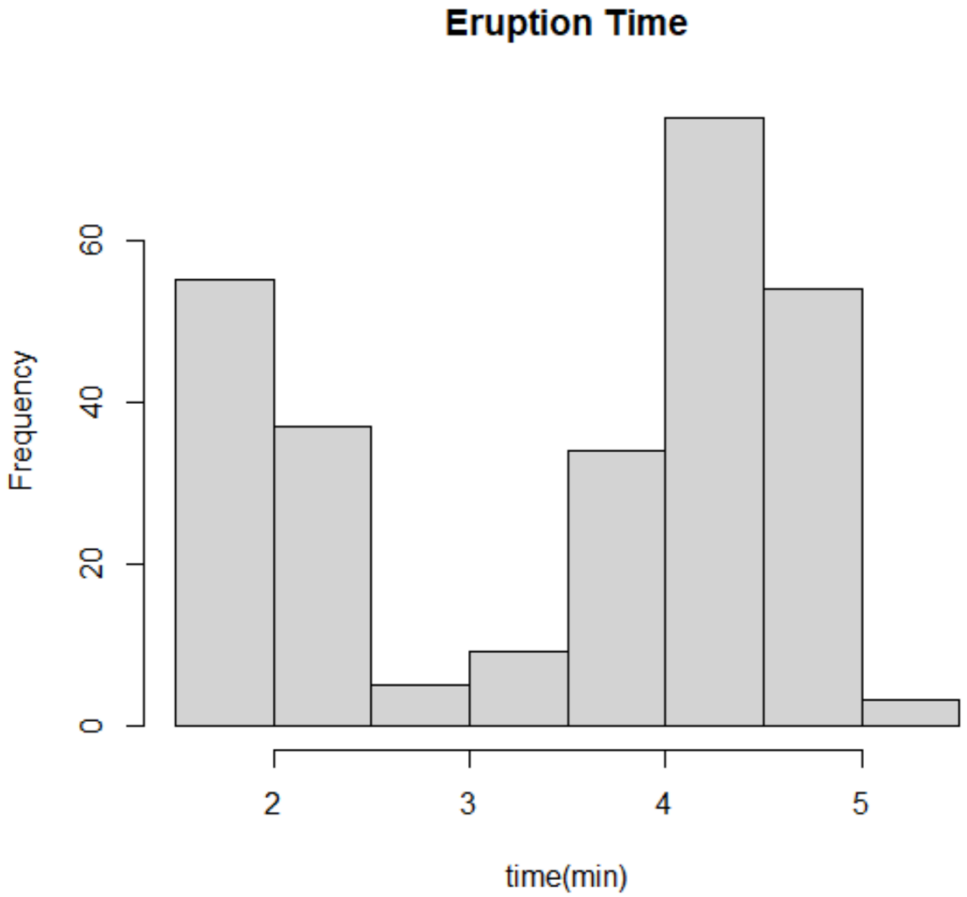


도수

- 양적 데이터
  - 3.600 1.800 3.333 2.283 4.533 2.883 4.700 3.600 1.950 4.350 1.833 3.917 4.200 1.750 4.700 2.167 1.750 4.800 1.600 4.250 1.800 … 4.317

도수

- 히스토그램



줄기-잎 그림

- Stem and Leaf Diagram
  - 양적 자료를 줄기와 잎으로 구분
    - The decimal point is 1 digit(s) to the left of the |

```
16 | 070355555588
18 | 000022233333335577777777888822335777888
20 | 000022233788000357 78
22 | 0002335578023578
24 | 00228
26 | 23
28 | 080
30 | 7
32 | 2337
34 | 250077
36 | 0000823577
38 | 2333335582225577
40 | 000000335778888800223355557 7778
42 | 033355557788002333333555577778
44 | 022223355577800000000233333577788 88
46 | 0000233357700000023578
48 | 00000022335800333
50 | 0370
```

상대도수

- 도수를 전체 원소의 수로 나눈 것

값	도수	상대도수
A	10	0.185
B	6	0.111
C	6	0.111
D	10	0.185
E	7	0.13
F	4	0.074
G	7	0.13
H	4	0.074

54개

scipy 모듈

$ sudo apt-get install python3-numpy python3-scipy python3-matplotlib python3-pandas python3-sympy python3-nose

평균

- mean

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

```
>>> a=
[79, 54, 74, 62, 85, 55, 88, 85, 51,
85, 54, 84, 78, 47]
>>> len(a)
14
>>> import statistics
>>> statistics.mean(a)
70.07142857142857
```

평균

- 모평균 μ
  - 모집단 전체 자료일 경우
- 표본 평균 $\bar{x}$
  - 모집단에서 추출한 표본일 경우

중앙값 (Median)

- 평균의 경우 극단 값의 영향을 많이 받음

```
b=[79, 54, 74, 62, 85, 55, 88, 85, 51,
85, 54, 84, 78, 47, 1000]
```

```
[6] import statistics

[10] b = [79, 54, 74, 62, 85, 55, 88, 85, 51, 85, 54, 84, 78, 47, 1000]

mean_value = statistics.mean(b)
print(mean_value)

132.06666666666666
```

중앙값 (Median)

- Median
  - 주어진 자료를 높은 쪽 절반과 낮은 쪽 절반으로 나누는 값을 의미
  - 자료를 순서대로 나열했을 때 가운데 있는 값
  - 자료의 수: n
    - n 이 홀수: $\frac{(n+1)}{2}$번째 자료값
    - n 이 짝수: $\frac{n}{2}$번째와 $\frac{n}{2}+1$번째 자료값의 평균

중앙값 (Median)

- Median

```
>>> a = [79, 54, 74, 62, 85, 55, 88, 85,
51, 85, 54, 84, 78, 47]
>>> b = [79, 54, 74, 62, 85, 55, 88, 85,
51, 85, 54, 84, 78, 47, 1000]
>>> a = sorted(a)
>>> a
[47, 51, 54, 54, 55, 62, 74, 78, 79, 84,
85, 85, 85, 88]
>>> b = sorted(b)
>>> b
[47, 51, 54, 54, 55, 62, 74, 78, 79, 84,
85, 85, 85, 88, 1000]
>>> statistics.median(a)
76.0
>>> statistics.median(b)
78
```

분산 (Variance)

- 편차의 제곱의 합을 자료의 수로 나눈 값
  - 편차: 값과 평균의 차이
- 자료가 모집단일 경우: 모분산

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- 자료가 표본일 경우: 표본분산

$$s^2 = \frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

분산 (Variance)

```
>>> statistics.variance(a)
234.37912087912088
>>> statistics.variance(b)
57868.78095238096
```

표본분산

```
>>> import scipy
>>> import scipy.stats
>>> scipy.stats.tvar(a)
234.37912087912088
```

표준편차 (Standard Deviation)

- 분산의 양의 제곱근
- 모표준편차 (population standard deviation)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- 표본표준편차 (sample standard deviation)

$$s = \sqrt{\frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

표준편차 (Standard Deviation)

- Standard Deviation

```
>>> statistics.stdev(a)
15.30944547915178
>>> statistics.stdev(b)
240.55930859640614
```

- 모분산, 모표준편차

```
>>> statistics.pvariance(a)
217.6377551020408
>>> statistics.pstdev(a)
14.752550799846134
```

표준편차 (Standard Deviation)

```
>>> import numpy
>>> numpy.var(a)
217.6377551020408
>>> numpy.std(a)
14.752550799846134

>>> numpy.var(a, ddof=1)
234.37912087912085
>>> numpy.std(a, ddof=1)
15.309445479151778
```

ddof:
Delta Degrees
of Freedom

범위 (Range)

- 자료를 정렬하였을 때 가장 큰 값과 가장 작은 값의 차이

```
>>> max(a) - min(a)
41
>>> max(b) - min(b)
953
>>> numpy.max(a)- numpy.min(a)
41
```

사분위수 (Quartile)

- 전체 자료를 정렬했을 때 ¼, ½, ¾ 위치에 있는 숫자
  - Q1: 제 1사분위수
  - Q3: 제 3 사분위수

```
>>> numpy.quantile(a, .25)
54.25
>>> numpy.quantile(a, .5)
76.0
>>> numpy.quantile(a, .75)
84.75

>>> numpy.quantile(a, .60)
78.8
```

사분위수 (Quartile)

- 사분위범위 (IQR, interquartile range)
  - Q3 – Q1

```
>>> numpy.quantile(a, .75) -
numpy.quantile(a, .25)
30.5
>>> numpy.quantile(b, .75) -
numpy.quantile(b, .25)
30.5
```

사분위수 (Quartile)

- 참고
  - numpy.quantile(a, q, axis=None, out=None, overwrite_input=False, **method='linear',** keepdims=False, *, weights=None, interpolation=None)
  - **method**: str, optional
    - This parameter specifies the method to use for estimating the quantile. There are many different methods, some unique to NumPy. The recommended options, numbered as they appear in [1], are:
    - 'inverted_cdf'
    - 'averaged_inverted_cdf'
    - 'closest_observation'
    - 'interpolated_inverted_cdf'
    - 'hazen'
    - 'weibull'
    - **'linear' (default)**
    - 'median_unbiased'
    - 'normal_unbiased'

z-score

- 어떤 값이 평균으로부터 몇 표준편차 떨어져 있는지를 의미하는 값
  - 모집단의 경우

$$z = \frac{x - \mu}{\sigma}$$

  - 표본의 경우

$$z = \frac{x - \bar{x}}{s}$$

정리

- 통계학의 정의
- 개념 정리
- 다양한 측도