Lecture Notes for
Introduction to Probability and Statistics
© 2013, Sungkuk Lim
Hanyang University, Ansan, 425-791, Korea
calltimy@hanyang.ac.kr

Probability and Statistics / 확률과 통계

강의노트 07

통계 - 이산확률분포 1

52. 확률변수(random variable)
: 확률변수는 시행의 수치 결과

- 확률과 통계를 듣는 클래스의 학생집단에서 한 학생을 추출한다고 할 때
- 그 학생의 { 키, 몸무게, 가족수입, 수능성적 } 등은 그 학생의 특징을 나타내는 수치로 된 변수들이고 이는 모두 확률변수가 될 수 있다.
- 2개의 동전을 던지는 시행의 결과, 앞면(F)이 나오는 횟수

결과	BB	BF	FB	FF
앞면횟수	0	1	1	2

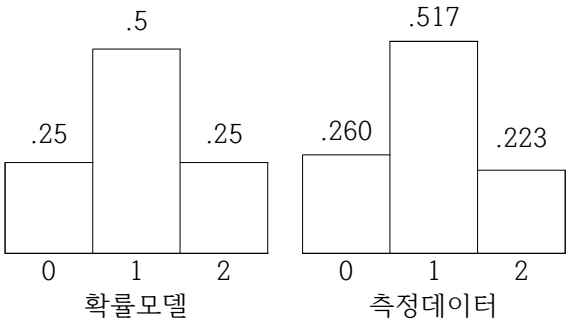
- 확률변수는 대문자를 사용, 소문자는 확률변수(random variable)의 값(value)을 의미

53. 동전 던지기

동전 두 개를 던진다. 그때 발생하는 경우의 수는 {(앞면, 앞면), (앞면, 뒷면), (뒷면, 앞면), (뒷면, 뒷면)} 이 된다.
앞면이 나온 수를 기록한다.

확률모델		측정데이터	
P[X]	x	n_x (발생횟수)	n_x/n (상대도수)
.25	0	260	.260
.50	1	517	.517
.25	2	223	.223

54. 확률 모델과 측정 데이터로 그린 히스토그램



55. P(X=x) 의 의미

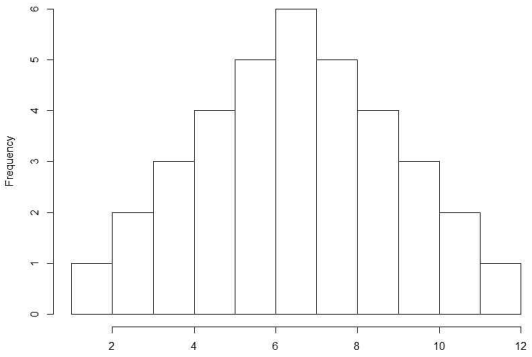
- P(X=x) : 확률변수 X 가 x 의 값을 가질 확률

x	0	1	2
P(X=0)	1/4	1/2	1/4

- P(Y=y), 두 주사위의 숫자의 합을 확률변수 Y로 놓은 경우 : 확률변수는 Y, Y가 취할 수 있는 값은 y이고, y는 2부터 12까지의 정수에 해당
- 확률은 장기적인 관점에서 본 어떤 사건의 상대도수
- 즉, 어떤 시행을 무수히 반복하면 그 결과의 상대도수 히스토그램은 그 확률변수의 확률히스토그램과 유사한 형태가 될 것

y	P(Y=y)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36

```
8       5/36
9       4/36
10      3/36
11      2/36
12       1/36
====================
```



56. 평균과 표준편차

▷ 평균값 : 모든 데이터의 값을 더한 후 데이터의 개수로 나눈값

▷ 자료의 총합을 자료의 총수로 나눈 값.

▷ $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} \dfrac{x_i}{n}$

▷ 중앙값 : 데이터의 중점

▷ 편차(deviation) : 관측값과 평균의 차이

▷ $d = x - \overline{x}$

▷ $\sum_{i=1}^{n} (x - \overline{x}) = 0$

▷ 제곱합(sum of squares, 자승합) : 편차를 제곱한 값의 총합

▷ $SS = \sum_{i=1}^{n} (y - \overline{y})^2$

▷ 분산 : 편차의 제곱의 합을 총 변량의 개수로 나눈 값

▷ $VarX = \sigma^2 = E[(X - \mu)^2]$

▷ $\sigma^2 = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$

▷ 표본분산 : 분모에 n-1을 쓴 값

▷ $s^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$

▷ 표준편차 : 데이터가 평균 $\overline{x}$ 에서 떨어져 있는 평균거리

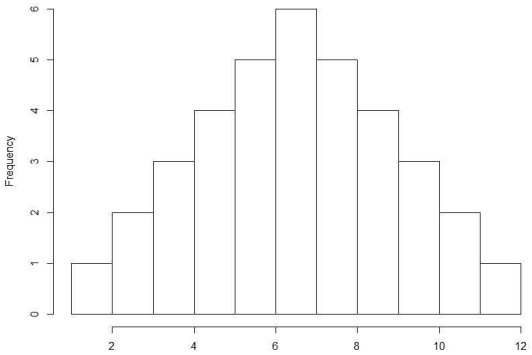
▷ $\sigma = \sqrt{VarX} = \sqrt{\sigma^2}$

▷ 표본의 표준편차 $s = \sqrt{s^2}$

▷ 산포도 : 자료들이 대표 값 주위에 흩어진 정도

57. z-점수, 평균과 표준편차의 특성

● 대칭적인 히스토그램에서 주로 사용



● z-점수 : 표준점수, 평균으로부터의 표준편차만큼의 거리에 대한 정의

● $z_i = \dfrac{x_i - \overline{x}}{s}$

58. z = 2 의 의미

--> 평균보다 표준편차의 2배만큼 크다.

59. 일반적인 데이터의 모양 (경험법칙)

평균에서 종모양에 가까운 데이터들의 경우

● 평균에서 표준편차의 한배 이내에 약 68%

● 평균에서 표준편차의 두배 이내에 약 97%

의 데이터가 들어 있다.