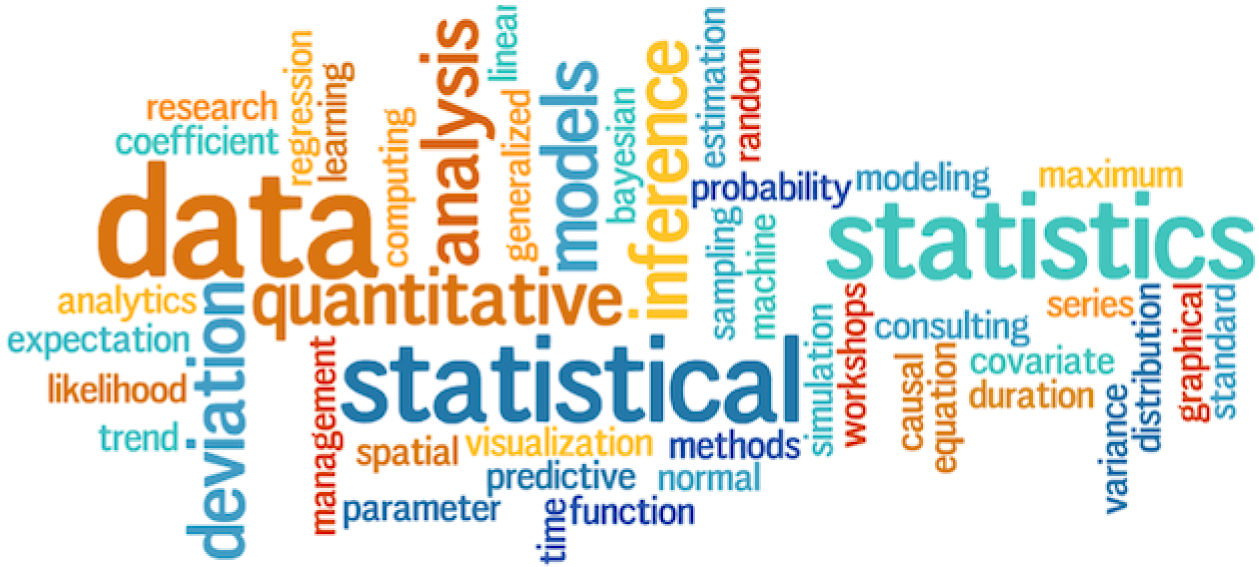


통계기반 데이터 분석 (Ch 2)

데이터사이언스 & A.I 전공
정화민 교수

기술통계학(Descriptive Statistics)

- ## 기술통계학

  수집한 데이터를 정리하여 요약하고 데이터가 어떤 특성을 갖고 있는지 해석하는 통계학 분야
  - 데이터의 요약 방법 : 시각화를 통한 자료의 요약, 각종 통계 숫자를 이용한 자료의 요약

- ## 기초통계량

  - 표본평균, 분산, 표준편차, 수치요약, 최빈값 , 중앙값 등

```r
#서강대학교 정보통신 대학원 (통계기반 데이터 분석)
#  R studio에서 평균, 표본 분산, 표본 표준편차는 각각 mean(), var(), sd(), summary로 구한다.
mean (1:4)
var (1:4)
sd(1:4)
summary(1:4)
```

```r
#서강대학교 정보통신 대학원 (통계기반 데이터 분석)
#  R studio에서 평균, 표본 분산, 표본 표준편차는 각각 mean(), var(), sd(), summary로 구한다.
mean (1:4)
```

```
## [1] 2.5
```

```r
var (1:4)
```

```
## [1] 1.666667
```

```r
sd(1:4)
```

```
## [1] 1.290994
```

```r
summary(1:4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.75    2.50    2.50    3.25    4.00
```

기술통계학(Descriptive Statistics)

- 기술통계학

  예) R 데이터 EuStockMarkets 의 기술통계 코딩 예

```
#서강대학교 정보통신 대학원 (통계기반 데이터 분석)
# (1) 데이터 내용구조 파악
data(EuStockMarkets) # 데이터 세트 사용
#데이트 세트의 구조 파악 (row ,column)
dim(EuStockMarkets)
#데이터세트 이름을 입력하면 해당 데이터 출력
EuStockMarkets
#전체데이터 중 'DAX' 변수에 해당하는 데이터만 출력하기 위함. Germany DAX (Ibis), Switzerland SMI, France CAC, and UK FTSE
EuStockMarkets[,'DAX']
# (2)요약데이터확인
summary(EuStockMarkets)
mean(EuStockMarkets[,'DAX']) # 'DAX'데이터의 평균
#분석하고자 하는 데이터를 그래프 등을 통한 시각화 기법을 활용하여 데이터에 대한이해도를 높인다.
median(EuStockMarkets[,'DAX']) # 'DAX'데이터의 중앙값
range(EuStockMarkets[,'DAX']) # 'DAX'데이터의 범위
summary(EuStockMarkets[,'DAX']) # 중심화 경향 및 분포 파악
var(EuStockMarkets[,'DAX']) # 'DAX'데이터의 분산 계산
sd(EuStockMarkets[,'DAX']) # 'DAX'데이터의 표준편차 계산
```

기술통계학(Descriptive Statistics)

- 기술통계학
  예) R 데이터 EuStockMarkets 의 기술통계 코딩 및 분석결과 예

```
#서강대학교 정보통신 대학원 (통계기반 데이터 분석)
# (1) 데이터 내용구조 파악
data(EuStockMarkets) # 데이터 세트 사용
#데이트 세트의 구조 파악 (row ,column)
dim(EuStockMarkets)
```

```
## [1] 1860    4
```

```
#데이터세트 이름을 입력하면 해당 데이터 출력
EuStockMarkets
```

```
## Time Series:
## Start = c(1991, 130)
## End = c(1998, 169)
## Frequency = 260
##              DAX    SMI    CAC   FTSE
## 1991.496 1628.75 1678.1 1772.8 2443.6
## 1991.500 1613.63 1688.5 1750.5 2460.2
## 1991.504 1606.51 1678.6 1718.0 2448.2
## 1991.508 1621.04 1684.1 1708.1 2470.4
## 1991.512 1618.16 1686.6 1723.1 2484.7
## 1991.515 1610.61 1671.6 1714.3 2466.8
```

```
mean(EuStockMarkets[,'DAX']) # 'DAX' 데이터의 평균
```

```
## [1] 2530.657
```

```
#분석하고자 하는 데이터를 그래프 등을 통한 시각화 기법을 활용하여 데이터에 대한이해
median(EuStockMarkets[,'DAX']) # 'DAX' 데이터의 중앙값
```

```
## [1] 2140.565
```

```
range(EuStockMarkets[,'DAX']) # 'DAX' 데이터의 범위
```

```
## [1] 1402.34 6186.09
```

```
summary(EuStockMarkets[,'DAX']) # 중심화 경향 및 분포 파악
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1402    1744    2141    2531    2722    6186
```

```
var(EuStockMarkets[,'DAX']) # 'DAX' 데이터의 분산 계산
```

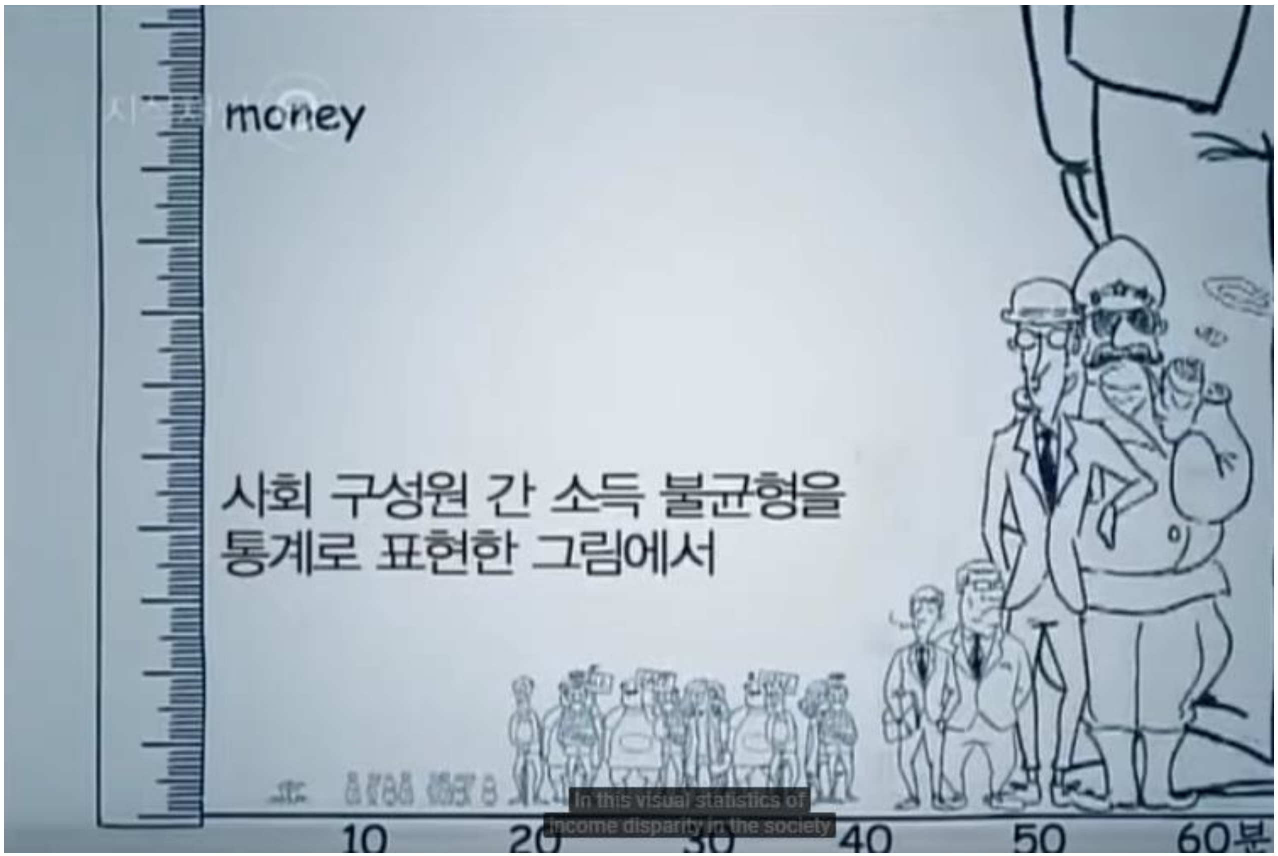
```
## [1] 1176775
```

```
sd(EuStockMarkets[,'DAX']) # 'DAX' 데이터의 표준편차 계산
```

```
## [1] 1084.793
```

기술통계학(Descriptive Statistics)

- 평균의 함정 (동영상)
  - 평균에 들어 가기는 어렵다 (평균 mean , 중앙값 median , 최빈값 mode)



Source https://www.youtube.com/watch?v=Pp_Pd6GZLOE

확률 , 확률분포

- 확률
  - 確率(확률) : 굳을 확, 비율 률

    '(어떤 결정 등을) 굳힐 비율'
  - probability

    probable : (명) '(어떤 일이) 있을 것 같은', '개연성 있는'

    '개연성' 혹은 '개연성 있는 일' , 개연성의 사전적 의미 : 어떤 일이 일어날 수 있는 확실성의 정도

- 고전적 확률 (수학적 확률)
  - 임의의 사건 A가 발생할 수학적 확률은 표본공간의 원소의 개수(O) 중 사건 A에 해당하는 근원사건의 개수(n)입니다 ($\frac{n}{O}$).
  - 예) 주사위를 굴려 홀수가 나올 확률
    - 주사위의 각 눈이 나올 확률은 전체 6개의 눈으로 구성되어 있으며 각각이 나올 확률은 동일하다고 가정하면 확률은 1/6.
    - 홀수인 사건을 구성하는 근원사건의 수는 {1, 3, 5}으로 세 개.
    - 전체 눈의 개수는 6이고 이로부터 홀수눈의 확률은 3/6 = ½ .

Source : 제대로 알고 쓰는 R 통계분석

확률 , 확률분포

▪ 통계적 확률

1) 동일한 조건하에서 같은 실험을 N번 반복

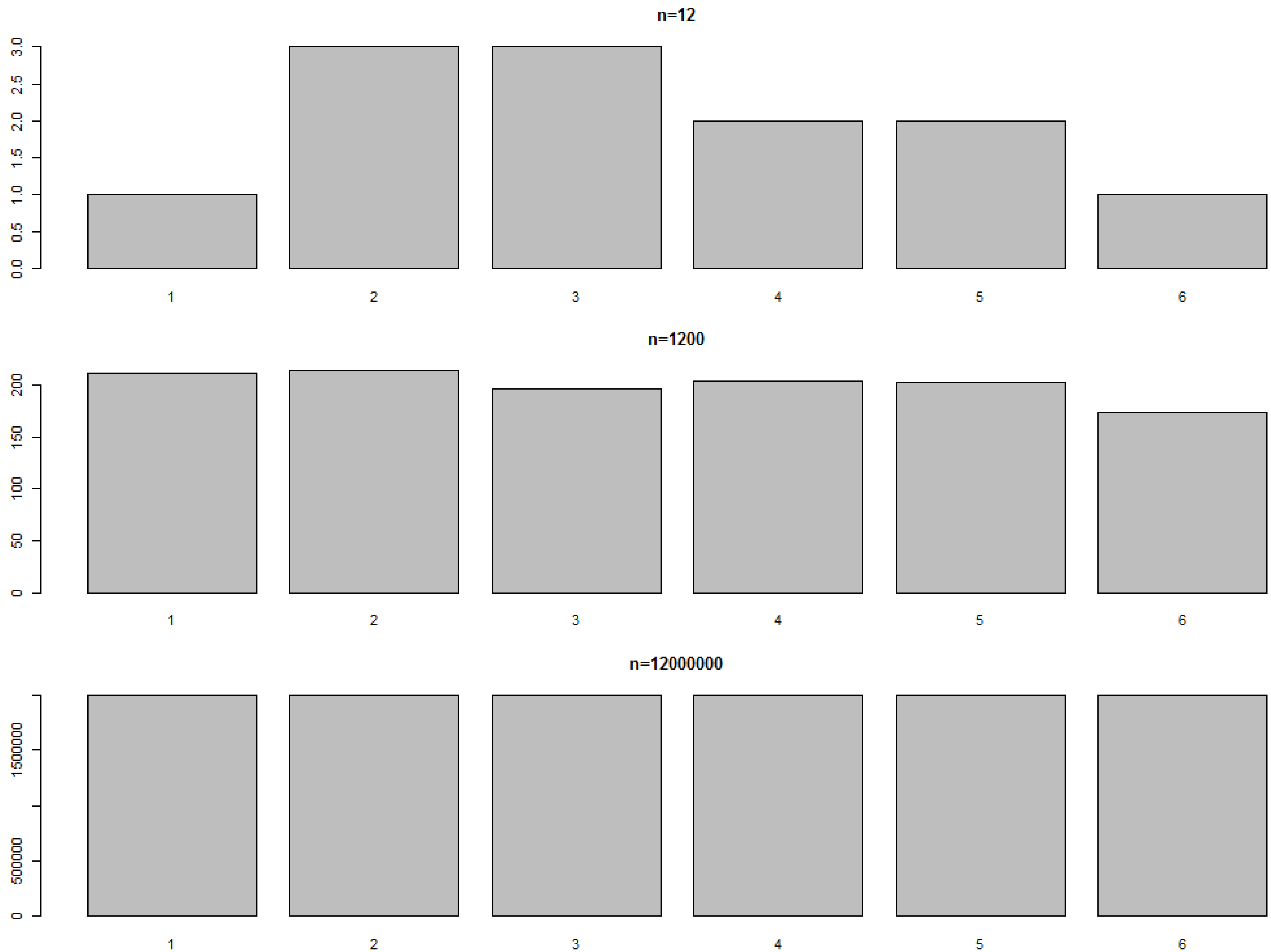
2) 사건 A가 모두 몇 번 발생했는지를 조사 : n

3) 사건 A가 발생할 확률 : $\mathrm{P(A)} = \frac{n}{N}$

- 실험의 반복횟수 N은 매우 커야 그 값을 받아들일 수 있으며, 반복횟수가 커짐에 따라 사건 A의 상대도수($\frac{n}{N}$)가 상수 P(A)로 접근해가는 경향을 보입니다.
- 예) 주사위를 여러 번 굴려 나온 눈을 관찰

시행횟수	1의 눈	2의 눈	3의 눈	4의 눈	5의 눈	6의 눈
12	1	3	3	2	2	1
1200	211	214	196	204	202	173
12000000	2002632	1999749	2000328	1999958	1996037	2001296

확률 , 확률분포



Source : 제대로 알고 쓰는 R 통계분석

확률 , 확률분포

- 공학단위로 많이 사용되는 그리스 문자
  - 뮤 (Mu) 통계학에서 모평균으로 사용
  - 씨그마(Sigma) 주로 모두 더하기
  - 입실론(Epsilon) " 집합원소" 또는 적다의 개념

그리스 문자 기호 및 발음 표기

문자 표기	발음 표기	문자 표기	발음 표기
Α α	Alpha (알파)	Ν ν	Nu (누)
Β β	Beta (베타)	Ξ ξ	Xi / Ksi (크사이)
Γ γ	Gamma (감마)	Ο ο	Omicron (오미크론)
Δ δ	Delta (델타)	Π π	Pi (파이)
Ε ε	Epsilon (입실론)	Ρ ρ	Rho (로오)
Ζ ζ	Zeta (제타)	Σ σ	Sigma (씨그마)
Η η	Eta (에타)	Τ τ	Tau (타우)
Θ θ	Theta (쎄타)	Υ υ	Upsilon (업실론)
Ι ι	Lota (이오타)	Φ φ	Phi (파이)
Κ κ	Kappa (카파)	Χ χ	Chi (카이)
Λ λ	Lambda (람다)	Ψ ψ	Psi (프사이)
Μ μ	Mu (뮤)	Ω ω	Omega (오메가)

※ 참고

평균을 μ라 쓰는 이유는 영어에서 평균은 mean, m에 해당하는 그리스어 소문자는 μ(mu, 뮤)이기 때문.

표준편차를 σ라 쓰는 이유는 영어에서 표준편차는 standard deviation을 의미하고, s에 해당하는 그리스어 소문자는 σ(sigma, 시그마)이기 때문.

확률 , 확률분포

- 확률 공리 (공리적 확률)

  - 표본공간 Ω상의 임의의 사건 A에 대한 실수치 함수에 대해

1) P(A)는 0과 1사이의 값을 갖고(0 ≤ P(A) ≤ 1),

2) 반드시 일어나는 사건(표본공간 전체)의 값은 1이며(P(Ω) = 1),

3) 서로 배반인 사건 A_1, A_2, ..., A_n, ... 의 합집합에 대해 다음을 만족하면,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} A_i$$

함숫값 P(A)를 사건 A의 확률이라 합니다.

  - 확률 공리는 확률이 만족해야 하는 기본 성질이며 이를 통해 확률 계산을 시행함

확률 , 확률분포

- **확률분포**

- 확률변수가 취할 수 있는 값과 발생할 확률을 대응한 관계

- 확률변수가 X 가질 수 있는 임의의 실측값 x에 대해

$$F(x) = P(X \leq x)$$

와 같이 정의된 함수 F를 확률변수 X의 누적분포함수, 또는 간략히 분포함수라고 함.

- 분포의 특성인 모수에 따라 분포의 모양이 결정됨.

- 확률질량함수, 확률밀도함수

  - 확률변수 X가 실측값 x를 가질 확률($P(X = x)$)에 대한 함수를 $f(x)$로 나타냅니다.

$$f(x) = P(X = x)$$

  - 확률변수가 취하는 값이 이산형일 경우에는 확률질량함수, 연속형일 경우에는 확률밀도함수라 부릅니다.

확률 , 확률분포

- **베르누이 시행**
  - p의 확률로 원하는 결과가 나타났을 때 '성공'으로, 1-p의 확률로 그렇지 않은 결과가 나타났을 때 '실패'로 하는 두 가지 결과가 나타나는 확률실험.
  - 성공 확률 p가 베르누이 시행의 모수.
  - 확률변수 X가 베르누이 시행에 따라 성공일 때 1, 실패일 때 0을 가질 경우 확률질량함수는 다음과 같음 .

$$f(x) = p^x \cdot (1-p)^{1-x}, \qquad x = \begin{cases} \text{성공} & 1 \\ \text{실패} & 0 \end{cases}$$

  - 예) 주사위를 던져 3의 배수의 눈이 나오면 상금 얻는 게임
    - 성공 : 3의 눈, 6의 눈이 나오는 경우, X=1

$$P(X=1) = p^{x=1} \cdot (1-p)^{1-(x=1)} = p$$

    - 실패 : 성공의 경우가 아닌 눈이 나오는 경우, X=0

$$P(X=0) = p^{x=0} \cdot (1-p)^{1-(x=0)} = 1-p$$

확률 , 확률분포

▪ 이항분포 개요

- 성공 확률이 p로 동일한 베르누이 시행을 n번 반복해서 실험하는 경우
    - 실험이 n번 반복되더라도 성공 확률 p는 변하지 않고 동일
    - 각 실험이 서로 독립적으로 시행
- n번 반복 실험에서 성공의 횟수가 따르는 분포를 이항분포라고 함.
- 이항분포의 모수
    - n : 시행의 횟수
    - p : 성공의 확률
- 이항분포의 표기 : 위의 두 모수를 이용하여 $B(n,p)$
    - 확률변수 X가 이항분포를 따를 때 $X \sim B(n,p)$와 같이 나타냄.

확률 , 확률분포

- 정규분포

 - 이항분포에서 시행 횟수 n이 커지면, 그에 따라 이를 따르는 확률변수 X가 갖는 확률 ($P(X = x)$) 계산은 복잡해 짐.

- 프랑스 태생의 수학자 드무아브르(1667~1754) 가 성공 확률이 0.5이고 시행 횟수 n이 아주 큰 이항분포가 어떤 함수와 비슷해지는 것을 발견

 - 좌우가 대칭인 종모양(확률분포의 확률값이 x축에 가까이 다가가나 확률이 0이되지 않는)의 형태와 유사.

 - n 이 충분히 크다면 이산형이 아닌 연속형처럼 다루는 것이 가능.

 - 이런 형태를 갖는 분포는, 이항분포가 아닌 다른 분포에서도 이와 닮아감을 밝힘. (라플라스(1749~1827))

 - 관측 오차가 이러한 분포를 따른다는 점이 발견되어 폭넓게 사용.(가우스(1777~1855)

확률 , 확률분포

- ## 정규분포

① 종모양의 형태를 가짐.

양 끝이 아주 느린 속도로 감소하지만, 축에 닿지 않고 -∞와 ∞까지 계속됩니다.

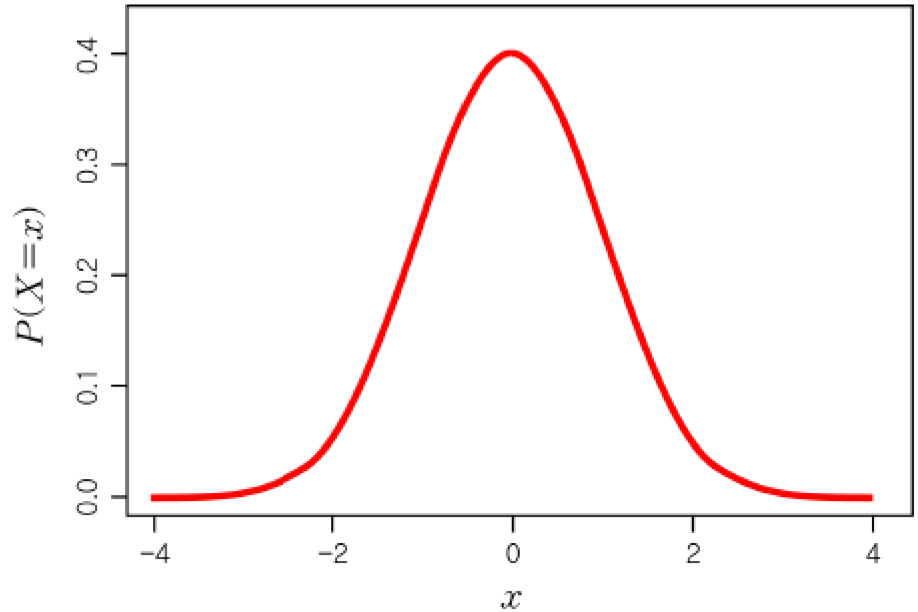
② 평균을 중심으로 좌우대칭

③ 평균 주변에 많이 몰려 있으며 양 끝으로 갈수록 줄어듬

④ 평균과 표준편차로 분포의 모양을 결정합니다.

- 정규분포의 모수는 평균 μ와 표준편차 σ(분산 σ^2)로, $N(\mu, \sigma^2)$으로 나타냅니다.

평균이 0이고 표준편차가 1인 정규분포



- ## 표준 정규분포

- 평균이 0이고 표준편차가 1인 정규분포($N(0,1^2)$)를 **표준정규분포**라하고 대문자 Z로 표시

- 모든 정규분포는 표준정규분포로 변환할 수 있음.

확률 , 확률분포

- 표준 정규분포 예

예 : 어느 대학교 남학생들 키의 평균은 170cm, 표준편차는 6cm입니다. 이 대학교에서 남학생의 키가 182cm 이상일 확률은 다음과 같이 구합니다.(남학생의 키는 정규분포를 따로 있는 것으로 가정

$$P(X \geq 182) = 1 - P(X \leq 182) = 1 - \int_{-\infty}^{182} \frac{1}{6\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-170}{6}\right)^2} dt$$

- 표준화 변환을 통한 표준정규분포로 계산

$$z = \frac{x-\mu}{\sigma} = \frac{182-170}{6} = \frac{12}{6} = 2$$

이를 이용하여 표준정규분포에서 구하면 다음과 같음

- $P(Z \geq 2) = 1 - P(Z \leq 2)$
- 표준정규분포표에서 z값이 2가 되는 값, 즉 행에서 2.0을 찾고 열에서 0.00을 찾은 값은 0.977(유효숫자 소숫점 세째자리)

다음페이지 표로부터 표준정규분포에서 2보다 작을 확률은 0.977이고, z가 2보다 클 확률은 1-0.977≈0.023임.

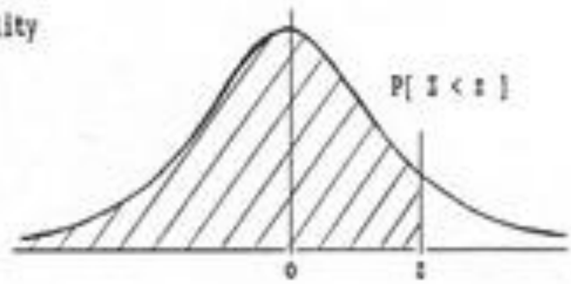
- 이제 다시 원래의 정규분포로 돌아가서 z 값으로 변환하여 2가 된 원래의 값을 구해보면 182. 이를 통해 182cm보다 클 확률은 0.023이 됨을 알 수 있음.

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z
i.e.

$$P[\, Z < z \,] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000