

# 통계적 분석

황도영

# 통계적 분석

---

1. 확률과 확률변수
2. 확률분포
3. 기술통계
4. 가설검정

# 확률과 확률변수

- 통계학: 여러 사건(event)들을 수학적으로 모델링하고, 이를 분석하는 것이 통계학의 본질
- 사건(event)는 근본적으로 발생하기 전에는 알 수 없으므로 불확실성을 내포하고 있습니다  
이러한 불확실성을 표현할 수 있는 수단이 바로 확률입니다.
- 동전 두 번 던지기를 예시로 들었을 때,

Experiment: 동전을 던지는 행위

Sample: experiment의 결과(동전의 앞/뒤)

Sample space: experiment로 인해 발생하는 모든 Sample의 집합

(가능한 모든 동전 앞/뒤 조합의 집합)

Event: sample space의 부분 집합으로, 어떤 조건을 만족하는 특정한 표본점들의 집합

- 이 때, sample space  $S$ 는 다음과 같습니다.  
 $S = \{(\text{앞}, \text{앞}), (\text{앞}, \text{뒤}), (\text{뒤}, \text{앞}), (\text{뒤}, \text{뒤})\}$

- 확률: 실험을 실시했을 때, 나올 수 있는 모든 경우의 수(sample space) 내에서 특정 사건이 발생하는 비율

$$P(A) = \frac{A\text{사건이일어나는경우의수}}{\text{모든사건이일어나는경우의수}}$$

- 동전 한번 던지기를 예로 들면,  
sample space  $S=\{\text{앞}, \text{뒤}\}$   
event  $A=\text{앞}$  이라고 할 때

동전 앞면이 나올 확률  $P(A) = \frac{A\text{사건이일어나는경우의수}}{\text{모든사건이일어나는경우의수}} = \frac{1}{2}$

- 확률의 성질
  1. 사건  $A$ 가 발생할 확률은  $[0,1]$  사이의 값을 가진다
  2. Sample space내 모든 사건의 확률의 합  $\sum_{A \in \Omega} P(A)$  ,  $\Omega=\text{sample space}$ ) 은 1이다

# 변수

---

- 변수(Variable): 특정 조건에 따라 변하는 값  
→ 확률 변수는 '확률'에 따라 변하는 값
- 독립 변수(x, feature): 다른 변수에 영향을 받지 않는, 오히려 종속 변수에 영향을 주는 변수
- 종속 변수(y, label): 독립변수의 영향을 받아서 변화하는 변수
- 연구자의 목표는 독립변수를 조정할 때 종속 변수가 어떻게 변화하는지를 알아내는 것  
즉, 독립 변수가 원인, 종속 변수가 결과라는 가정이 필요하며  
두 변수간의 관계를 알아내는 것이 중요  
(e.g. correlation, regression, 대부분의 데이터 모델링)

# 확률 변수

---

- 확률 변수(Random Variable)의 정의:  
무작위(Random) 실험을 했을 때,  
특정 확률로 발생하는 각각의 결과를  
수치적 값으로 표현하는 변수
- e.g. (동전 던지기)  
동전을 무작위로 던져서 앞뒤가 나오는 실험(무작위 실험)에서  
앞이 나올 확률  $\frac{1}{2}$ , 뒤가 나올 확률  $\frac{1}{2}$  (일정한 확률) 을 가지고 발생하는 결과를  
앞=1, 뒤=0이라는 실수 값(수치적 값) 으로 표현하는 변수
- 이산 확률 변수(Discrete random variable):  
확률 변수 X가 이산값(정수) 값을 택하는 변수
- 연속 확률 변수(Continuous random variable):  
확률 변수 X가 어떤 구간의 모든 실수값을 택하는 변수

# 확률 분포

- 확률 분포(Probability distribution):  
확률 변수의 모든 값과 그 확률이 어떻게 분포하는지를 의미합니다

e.g.  $X$  = 동전을 두번 던져서 앞이 나오는 확률 변수일때  
 $X$ 가 가질 수 있는 값(sample space) =  $\{0,1,2\}$   
확률 분포는

$x$	0	1	2
$P(x)$	0.25	0.5	0.25

- 확률 함수(Probability function):  
확률 변수  $x$ 를 확률값에 대응시키는(연결시켜주는) 함수  $P(x)$   
e.g.  $P(X=1) = 0.5$



# 확률 분포

- 즉, 확률 변수와 확률 함수를 이용해 sample space내 사건의 확률을 얻을 수 있습니다  
실험의 sample space  $\rightarrow$  [확률 변수  $X$ ]  $\rightarrow$  실수공간  $\rightarrow$  [확률 함수  $f(x)$ ]  $\rightarrow$  확률

e.g. 주사위를 두번 던져서 나온 합이 5 이상 7 이하인 확률은?

$X$  가 가지는 sample space =  $\{2,3,...,12\}$

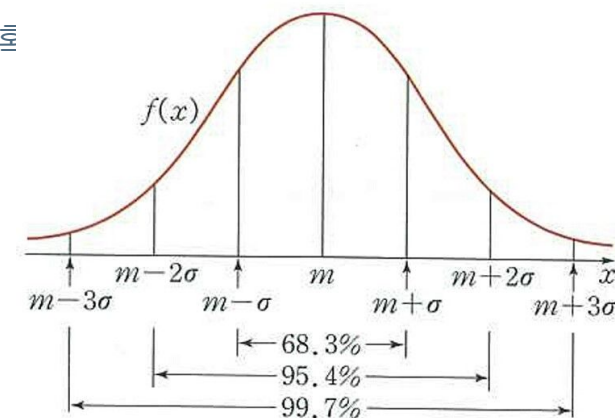
구하고자 하는 확률

$$\begin{aligned} P(5 \leq X \leq 7) &= P(X=5) + P(X=6) + P(X=7) \\ &= 4/36 + 5/36 + 6/36 = 5/12 \end{aligned}$$

- 연속 확률 변수의 경우, 사건이 발생하는 구간의 넓이를 계산하여  
사건이 발생할 확률을 계산할 수 있습니다.

e.g. 정규분포를 따르는 확률 변수  $X$ 가  $(\mu-2\sigma)$  와  $(\mu+2\sigma)$  사이의 값을 가질 확률

$$F(x) = P((\mu-2\sigma) \leq x \leq (\mu+2\sigma)) = \int_{\mu-2\sigma}^{\mu+2\sigma} f(x)dx$$



# 확률질량함수

---

- 확률질량함수(Probability Mass Function, pmf):  
이산확률변수  $X$ 가 취할 수 있는 값  $x_0, x_1, \dots$ 의 각각에 대해  
확률값  $P(X=x_0), P(X=x_1), \dots$ 를 대응시켜주는 확률함수를  $X$ 의 확률질량함수  $f(x)$ 라고 합니다
- 확률질량함수의 성질
  1. 모든  $x$ 에 대해  $f(x) \geq 0$
  2.  $\sum_{i=0}^{\infty} P(x_i) = 1$
  3.  $P(a \leq x \leq b) = \sum_{a \leq x_i \leq b} f(x_i)$
- 누적분포함수  $F(X) = P(X \leq a) = \sum_{x \leq a} f(x)$

# 확률밀도함수

- 확률밀도함수(Probability Density Function, pdf):  
연속확률변수  $X$ 가 취할 수 있는 값의 범위  $[a,b]$ 에 대해서  
확률값  $P(a \leq X \leq b) : \int_a^b f(x)dx$  를 대응시켜주는 확률함수를  $X$ 의 확률질량함수  $f(x)$ 라고 합니다
- 확률질량함수의 성질
  1. 모든  $x$ 에 대해  $f(x) \geq 0$
  2.  $\int_{-\infty}^{\infty} f(x)dx = 1$
  3.  $P(a \leq x \leq b) = \int_a^b f(x)dx$
- 누적분포함수  $F(a) = \int_{-\infty}^a f(x)dx$   
 $P(a \leq x \leq b) = \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx = F(b) - F(a)$
- 연속확률변수의 확률은 '범위의 면적'  
확률밀도함수(pdf)로 곡선을 그리고 '범위 내 곡선 아래의 면적'을 구한다

## 모집단, 모수, 표본

---

- 모집단: 통계학에서 관심의 대상이 되는 모든 개체 값의 집합  
e.g. “대한민국 고등학생의 평균 키를 알고 싶다”  
→ 모집단: 대한민국의 ‘모든’ 고등학생들의 키 값
- 모수(Population parameter, Parameter): 모집단의 특성을 나타내는 통계적인 특성치  
e.g. 모집단이 정규분포를 따른다고 할 때,  
모집단의 분포 특성을 나타내는 모수는  $\mu$ (평균)과  $\sigma^2$ (표준편차)
- 모수는 모집단을 모두 조사(전수검사) 해야 얻을 수 있는 값이지만, 전수검사가 어렵기 때문에  
‘통계적 추론’을 합니다  
통계적 추론: 모집단에서 추출한 표본들의 특성을 분석하여, 모수에 대해 추론하는 과정
- 표본(sample): 전체 모집단에 대해서 샘플링(sampling)을 통해 뽑히는 값으로, 모집단의 부분 집합을 의미  
e.g. 대한민국 고등학생 중 100명을 ‘뽑아’ 키를 조사 → 100개의 sample에서 얻은 통계량을 이용해 모수를 추론

## 모집단, 모수, 표본

---

- 전반적인 프로세스: 모집단이 갖는 분포를 가정(e.g. 정규분포, 포아송분포,...)
  - sample들을 추출
  - 뽑힌 sample들을 통해 얻어진 통계량(e.g. 평균, 분산,..)이 지닌 성질을 이용해 모수를 추정할 수 있습니다
- 모집단의 모수를 잘 추정하기 위해서는 표본을 '잘' 추출하는 것 역시 중요합니다.
  - 모집단에서 sample이 뽑힐 가능성을 모두 '동일' 하게 부여하고, 객관적으로 무작위 추출해야됩니다.
  - sample들을 서로 독립적이며(Independent) 동일한 분포(Identically Distributed), 흔히 말하는 i.i.d 를 따라야 됩니다. 이러한 sample을 random sample이라고 부릅니다.

독립적: sample들이 추출될 때 서로 영향을 미치지 않음  
동일한 분포: sample들이 동일한 모집단으로부터 추출됨
- Sample들의 통계량은 추출할 때 마다 달라지지만, 여러번의 추출을 통해 얻어진 여러 통계량 값의 발생 분포를 그려보면 통계량을 확률 변수로 하는 확률 분포를 얻을 수 있습니다.
  - 결국, 통계량의 확률 함수와 확률 분포를 이용하여 모수를 추정할 수 있게 됩니다.

확률 분포

# 기댓값 & 분산

- 기댓값(Expected value): 어떤 확률적 사건이 평균적으로 가질 수 있는 값.(=평균값,  $E(x)$ ,  $\mu$ )

- 이산확률변수의 기댓값:

$$E(x) = \sum_x x f(x)$$

- 연속확률변수의 기댓값:

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

- 기댓값의 성질:

1.  $E(X + Y) = E(X) + E(Y)$

임의의 실수  $a, b, c$ 에 대해서

2.  $E(aX+b) = aE(X)+b$

3.  $E(aX^2 + bX + c) = aE(X^2) + bE(X) + c$

4.  $E(aX + bY) = aE(X) + bE(Y)$

서로 독립인 두 확률변수  $X, Y$ 에 대해서

5.  $E(XY) = E(X)E(Y)$

## 기댓값 & 분산

---

- 분산(Variance): 분포가 평균값으로부터 얼마나 산포되어있는지 ( $\text{Var}(X)$ ,  $\sigma^2$ )

$$\text{Var}(X) = E[(X-E(X))^2] = E[X^2] - E[X]^2$$

- 이산확률변수의 분산:

$$\text{Var}(x) = \sum_x (x - E(x))^2 f(x)$$

- 연속확률변수의 분산:

$$\text{Var}(x) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x) dx$$

- 표준편차  $\sigma$

$$\sigma = \sqrt{\text{Var}(x)}, \sigma^2 = \text{Var}(x)$$

- 분산의 성질:

서로 독립인 두 확률변수  $X, Y$ 에 대해서

$$1. \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

임의의 실수  $a, b, c$ 에 대해서

$$2. \text{Var}(aX+b) = a^2\text{Var}(X)$$

$$3. \text{Var}(aX+bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$



## 결합확률분포

- 결합확률분포(Joint probability distribution): 두 개의 확률변수  $X, Y$ 에 대해  $P(X=x, Y=y)=f(x,y)$ 를 만족하는  $f(x,y)$ 를  
확률변수  $X, Y$ 의 결합확률분포 혹은 결합확률{질량/밀도}함수(Joint pmf/pdf) 라고 합니다

- 이산  $\sum_{x \in \Omega} \sum_{y \in \Omega} f(x, y) = 1$   $P[(x, y) \in A] = \sum \sum_A f(x, y)$   
이때

- 연속  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dy dx = 1$   $P[(x, y) \in A] = \int \int_A f(x, y) dy dx$   
이때

- (중요) 확률변수  $X$ 와  $Y$ 가 서로 독립이면  $f(x,y) = f_x(x)f_y(y)$

- 공분산(Covariance): 두 개의 확률변수 X와 Y에 대해 X가 변할 때 Y가 변하는 정도를 나타내는 값  
즉, X와 Y가 같이 변하는 정도를 나타내는 값  
 $Cov(X, Y) = E[(X-\mu_x)(Y-\mu_y)]$ , 편차의 곱의 기댓값  
이 때  $(X-\mu_x)$ ,  $(Y-\mu_y)$ 를 편차라 합니다
- 이산확률변수의 공분산  $Cov(X, Y) = \sum_x \sum_y (x - \mu_x)(y - \mu_y)f(x, y)$   
연속확률변수의 공분산  $Cov(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y)f(x, y)dydx$
- 기댓값의 성질 ( $E(XY) = E(X)E(Y)$ ,  $E(aX+b) = aE(X)+b$ ) 를 이용해  
공분산을 유도해보면  
 $Cov(X, Y) = E[(X-\mu_x)(Y-\mu_y)]$   
 $= E[XY - \mu_y X - \mu_x Y + \mu_x \mu_y]$   
 $= (\mu_x, \mu_y \text{는 상수이므로}) E[XY] - \mu_y E[X] - \mu_x E[Y] + \mu_x \mu_y$   
 $= E[XY] - E[X]E[Y]$
- 그런데 '서로 독립인' X, Y에 대해서  $E[XY] = E[X]E[Y]$ 라고 했었습니다  
즉, X와 Y가 서로 독립이면  $Cov(X, Y) = 0$ 입니다.

## 베르누이 분포(Bernoulli distribution)

---

- 베르누이 시행: 어떤 시행의 결과가 1(성공) or 0(실패)인 실험

베르누이 시행에서 확률변수  $X=1$ 일 확률이  $p$ ,  $X=0$ 일 확률이  $q = 1-p$  인 경우  
확률변수  $X$ 는 베르누이 분포를 따른다

- pmf는

$$f(x) = \begin{cases} p, & x = 1 \\ (1 - p), & x = 0 \end{cases}$$

- 기댓값  $E(X) = 1 \times p + 0 \times (1-p) = p$   
 $E[X^2] = 1^2 \times p + 0^2 \times (1-p) = p$   
분산  $\text{Var}(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$
- e.g. 동전을 던져 앞이 나오면 1, 뒤가 나오면 0인 확률변수  $X \sim \text{Bernoulli}(p=0.5)$

## 이항 분포(Binomial distribution)

---

- 베르누이 시행을  $n$ 번 반복했을 때 성공 횟수를 값으로 갖는 확률 변수  $X$ 에 대해  $X$ 는 이항확률변수(Binomial random variable) 이라고 합니다
- 베르누이 시행의 결과가 성공일 확률이  $p$ , 실패할 확률이  $(1-p)$  일때 이 시행을  $n$ 번 반복했을 때 나타나는 확률분포를 이항분포(Binomial distribution)이라고 합니다
- 이항확률변수  $X$ 에 대한 pmf는
$$f(x) = {}_nC_x p^x (1-p)^{n-x}$$
where  $x=0, 1, \dots, n$
- 기댓값과 분산의 경우 이항확률변수  $X$ 는  $n$ 개의 베르누이 확률변수  $B_i$ 의 합이므로
$$E[X] = E[\sum_{i=1}^n B_i] = \sum_{i=1}^n E[B_i] = np$$
$$Var[X] = Var[\sum_{i=1}^n B_i] = \sum_{i=1}^n Var[B_i] = np(1-p)$$
- e.g. 동전을 5번 던져서 2번 앞이 나올 확률
$$= {}_5C_2 (0.5)^2 (1-0.5)^3$$

## 포아송 분포(Poisson distribution)

---

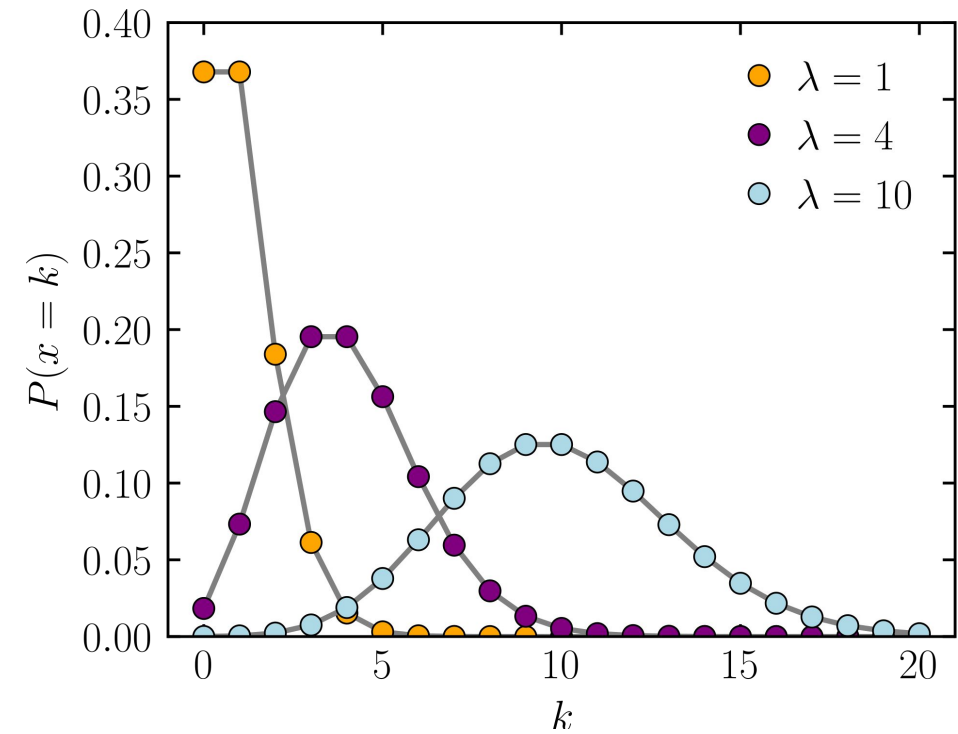
- 포아송 분포에서 모수  $\lambda$ 는 '단위시간/단위공간에서의' 평균 발생횟수  
e.g. 1시간 동안 버스가 정류장에 도착하는 횟수
- 포아송 분포(Poisson distribution):  
단위시간/단위공간에서 어떤 사건이 발생하는 횟수를 확률변수  $X$ 라 할때,  $X$ 는 포아송 분포를 따른다
- 포아송 분포의 전제조건
  1. 독립성: 단위 시간/공간에서 발생한 결과는 중복되지 않은 다른 시간/공간에서 발생한 결과와 독립이다
  2. 일정성: 단위 시간/공간에서 발생한 확률/횟수는 그 시간/공간의 크기에 비례한다.  
즉, 단위 시간/공간에서 발생한 평균발생횟수는 일정하다
  3. 비집락성: 매우 짧은(즉, 같은) 시간/공간에서 두 개 이상의 결과가 동시에 발생할 확률은 0이다

## 포아송 분포(Poisson distribution)

- 포아송 분포의 pmf

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

- 포아송 분포의 기댓값과 분산은 모두  $\lambda$ 입니다



## 균등분포(Uniform distribution)

- 균등분포: 모든 확률변수값에 대해 균일한 확률을 갖는 확률 분포

- 구간  $[a, b]$  내 모든 구간에서 일정한 크기의 확률을 가지는 확률변수  $X$ 의 pdf는

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & elsewhere \end{cases}$$

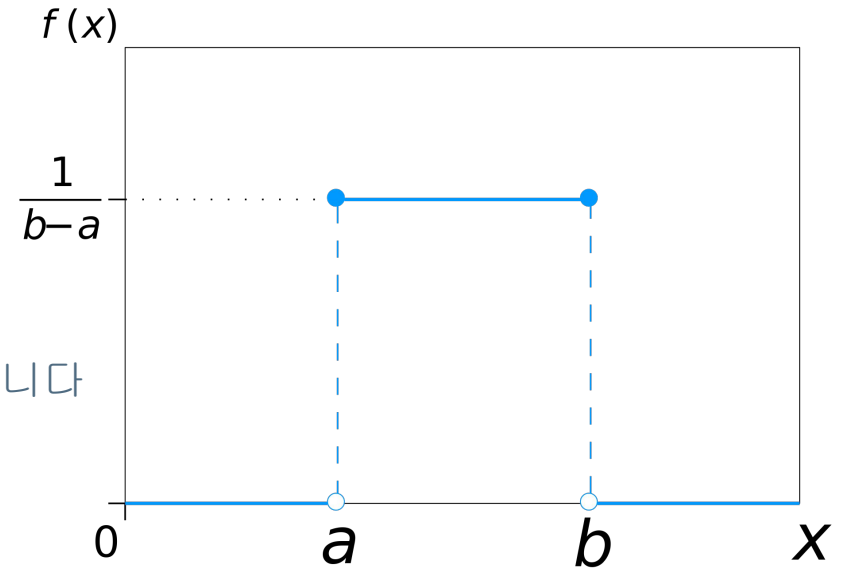
- 모든 확률의 합은 1이므로, 구간  $[a, b]$  사이의 모든 확률의 합은 1입니다

- 균일분포의 누적분포함수  $F(x)$ 는

$$F(x) = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}$$

- 균등분포의 기댓값  $E[X] = \frac{a+b}{2}$

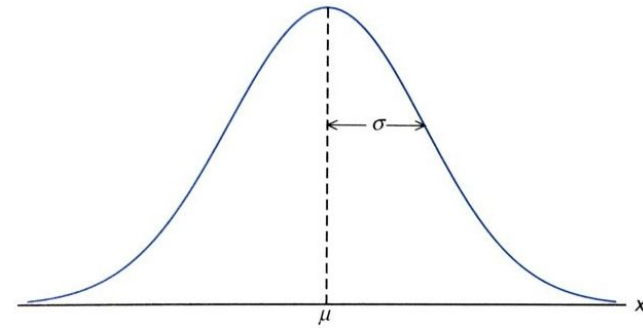
분산은  $Var(X) = \frac{(b-a)^2}{12}$



# 정규 분포(Normal distribution)

- 정규분포(Normal distribution, Gaussian distribution):  
가장 일반적으로 발견되는 양방향 대칭의 종 모양(Bell curve)으로 생긴 분포,  
수집된 자료의 분포를 근사할 때 대부분 정규분포를 사용합니다  
(중심 극한정리에 의해 독립적인 확률 변수들의 평균이 정규분포에 가까워지므로)

- 모수:  $\mu$ (평균)과  $\sigma^2$ (분산)  
평균: 분포가 모이는 중심  
분산: 평균을 중심으로 데이터들이 퍼진 정도  
보통  $N(\mu, \sigma^2)$ 로 표현합니다



- pdf는  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

- 정규분포에서의 확률은

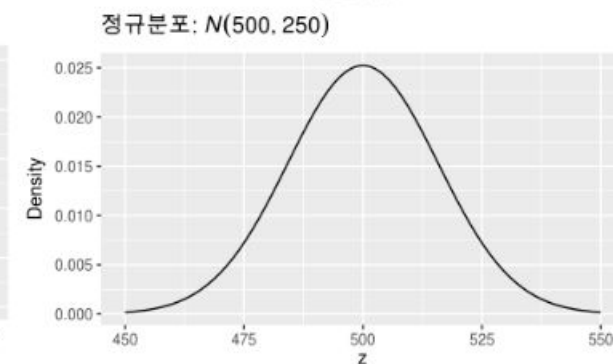
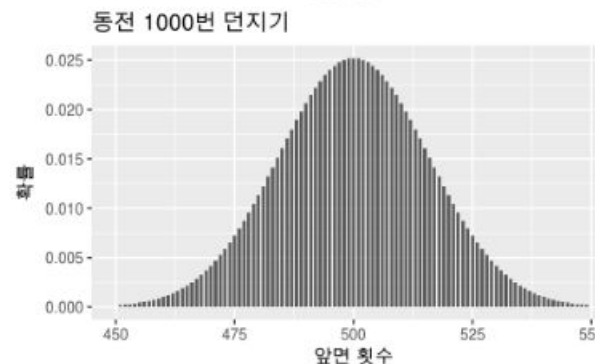
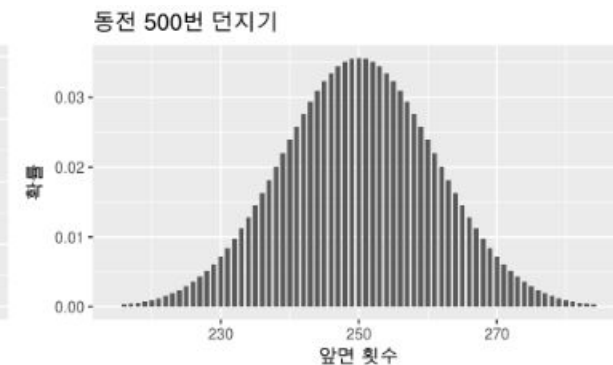
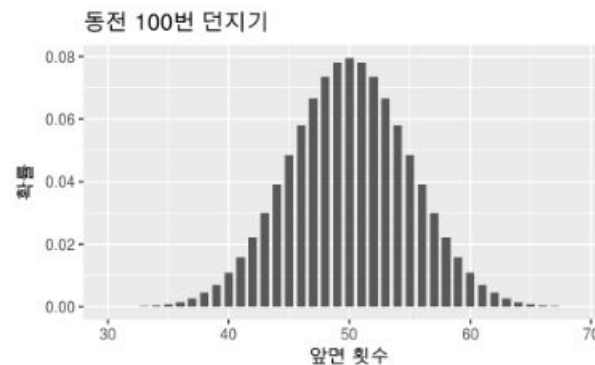
$$P(a \leq x \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



# 이항분포의 정규분포 근사

- 이항분포의 pmf  $f(x) = nCx p^x (1 - p)^{(n-x)}$   
 $n \rightarrow \infty$  극한으로 보낼 경우
$$nCx p^x (1 - p)^{(n-x)} \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(x-np)^2}{2np(1-p)}}$$
- e.g. 동전을 던졌을 때 앞이 나오는 경우  $X=1$ 인 확률변수  $X$ 의 경우  
 $X \sim \text{Bin}(n, p)$  ( $n$ =시행횟수)

$X$ 의 확률(히스토그램)을 그리면  
 $N(np, np(1-p))$ 의 분포에 가까워지는 것을  
알 수 있습니다  
(드무아브르-라플라스 정리)

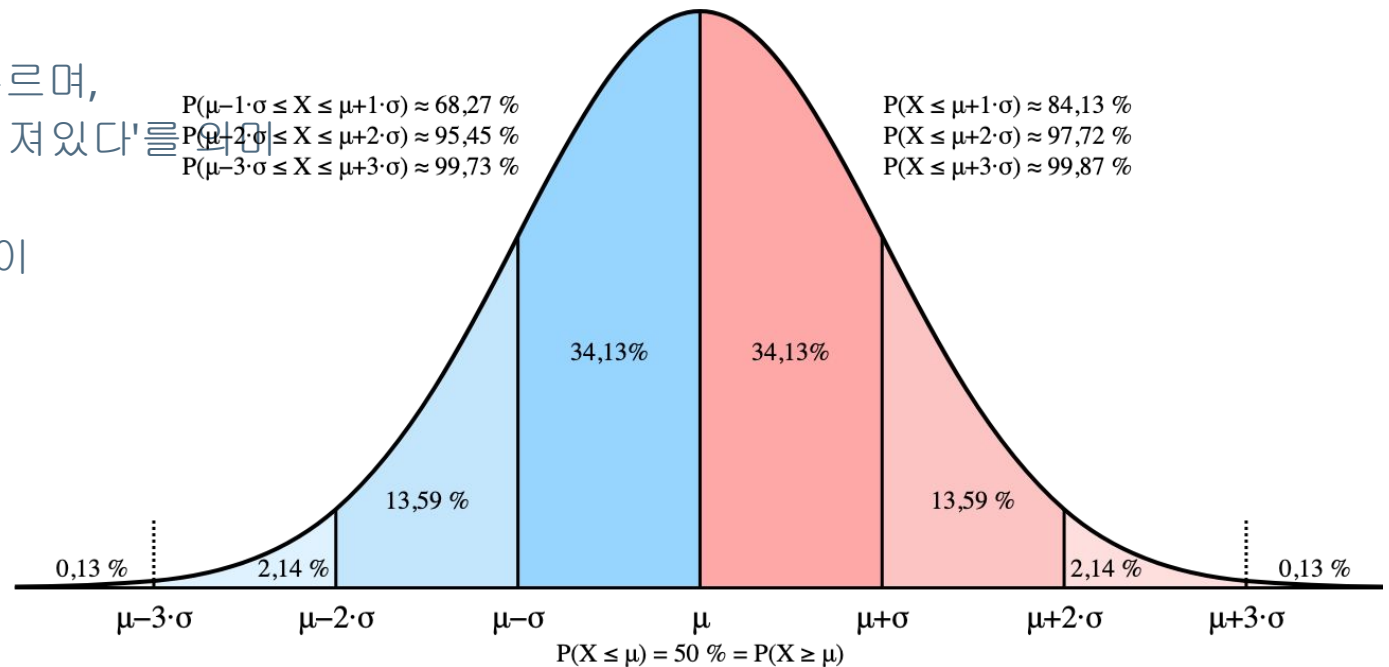


# 표준 정규 분포

- 표준 정규분포(Standard normal distribution): 서로 다른 parameter를 가진 집단들을 비교하기 위해 정규분포를 표준화한 분포 즉, 평균이 0이고 표준편차가 1인 분포로 표준화

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

- 표준화된 개별 데이터는 z-score라고 부르며, '평균으로부터 표준편차의 z배정도 떨어져있다'를 의미
- 표준 정규분포를 따르는 확률 변수 z 값이 (-2, 2) 사이에 위치할 확률은 약 95%



# 표준 정규 분포표

- 확률 변수 X가 정규분포를 따른다는 가정 하에  
표준화를 통해서 z값을 구한 다음  
표준 정규 분포표를 이용해  $P(Z < z)$  확률값을 구할 수 있습니다

- e.g.  $P(Z < 1.96) = 0.975$   
 $P(Z > 1.96) = 0.025$   
정규분포는 평균값 기준 양쪽 대칭이므로  
 $P(Z < -1.96) = 0.025$   
따라서  
 $P(-1.96 < Z < 1.96) = 1 - P(Z > 1.96) - P(Z < -1.96)$   
 $= 0.95$

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	0.5160	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	0.7852
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	0.8810	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	0.9772	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	0.9864	0.9868	0.9871	0.9875	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	0.9925	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	0.9979	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

## 표본 평균과 표본 분산

- 모집단으로부터 random sample을  $n$ 개 추출했을 때  
 $n$ 개의 random sample들의 평균과 분산을 각각 표본평균/표본분산  
단, 이 때 random sample들은 iid(independent and identically distributed)여야 함.

- 표본평균(sample mean)  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

$$\text{표본분산(sample variance)} S^2 = \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2$$

- 표본평균을 새로운 확률 변수라고 생각하고 표본 평균의 평균과 분산을 다시 구해보면

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}(X_1 + \dots + X_n)\right) \\ &= \frac{1}{n}[E(X_1) + \dots + E(X_n)] = \\ &= \frac{1}{n}n\mu = \mu \end{aligned}$$

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n}(X_1 + \dots + X_n)\right) \\ &= \frac{1}{n^2}[\text{Var}(X_1) + \dots + \text{Var}(X_n)] \\ &= \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

## Bessel's correction

- 추정량(Estimate): 모수를 추정하기 위한 표본 통계량(e.g. 표본평균, 표본분산)
- 불편향(Unbiased): 표본 추정량의 기댓값이 모수와 같다. ( $E(\bar{X}) = \mu$ )

- 왜 '표본'분산은  $n$ 이 아닌  $(n-1)$ 로 나눠주는걸까?  
→ 만약 표본분산이  $(n-1)$ 이 아닌  $n$ 으로 나눠주는 값이라면

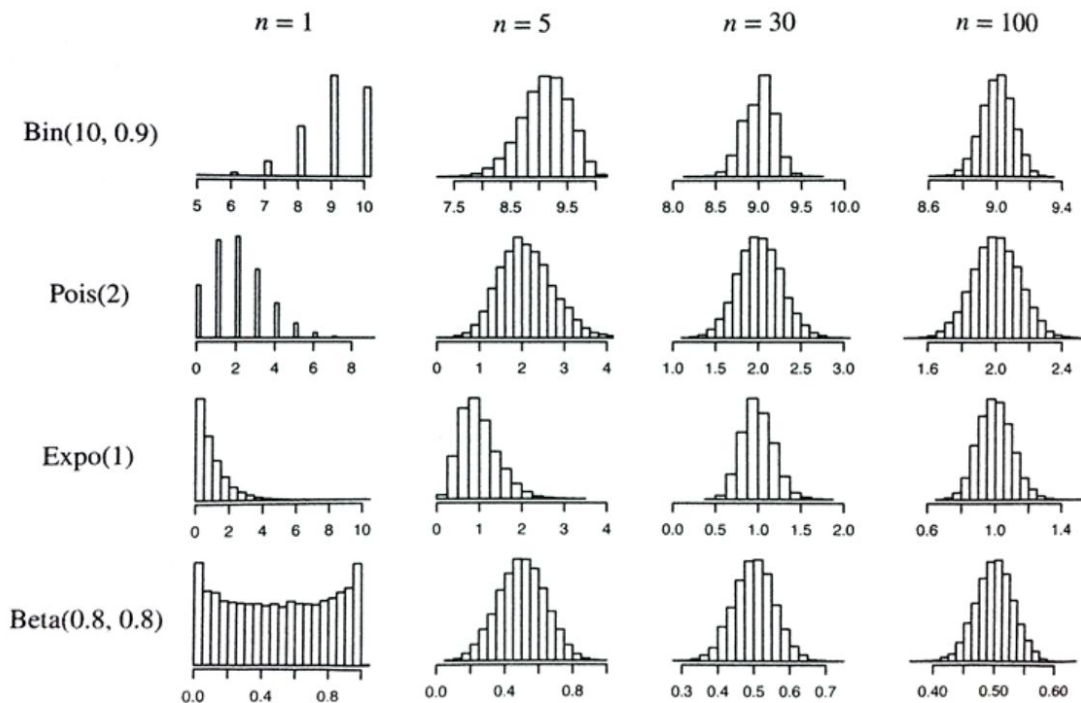
$$\begin{aligned} E(s^2) &= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + \sum_{i=1}^n \bar{X}^2\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i^2) - E(\bar{X}^2) \\ &= \frac{1}{n} \sum_{i=1}^n (\mu^2 + \sigma^2) - \left(\mu^2 + \frac{\sigma^2}{n}\right) \\ &= \frac{n-1}{n} \sigma^2 \end{aligned}$$

- 즉,  $n$ 으로 나눠지면 표본분산은 불편추정량(Unbiased estimate)이 아니게 되므로  $(n-1)$ 으로 나눠주게 됩니다(Bessel's correction)

# 중심극한정리

- 중심극한정리(Central Limit Theorem):  
평균  $\mu$ 와 분산  $\sigma^2$ 인 임의의 모집단에서 크기가  $n$ 인 표본  $(X_1, \dots, X_n)$ 에 대해  
표본 평균  $\bar{X}$ 의 분포는  $n \rightarrow \infty$ 일때(충분히 클  $N(\mu, \frac{\sigma^2}{n})$ 에 근사하고  
 $Z$ 의 분포는  $N(0, 1^2)$ 에 근사한다.

- 오른쪽에서 볼 수 있듯이, 서로 다른 여러 분포들의  
표본 평균은  $n$ 이 커질수록 정규분포에 근사하는  
것을 볼 수 있습니다.
- 표본들의 합에 대해서도 중심극한정리가 적용됩니다  
즉,  $W_n = X_1 + \dots + X_n = n\bar{X}$  일때  
 $n \rightarrow \infty$ 이면  $\bar{X}$ 는 평균  $\mu$ 와 분산  $\sigma^2$ 인 정규분포를 따르며  
 $W_n \sim N(n\mu, n\sigma^2)$ 인 정규분포를 따릅니다.
- $n$ 개의 표본이 특정 분포를 따르는 것이 아니고,  
 $n$ 개 표본의 평균값이  $n \rightarrow \infty$ 이면 (보통은  $n > 30$ )  
정규분포를 따르는 것입니다.



# 기술통계

## 정량적 데이터 분석이란?

---

- 정량적 데이터 분석은 숫자로 표현되는 수치 데이터를 이용하여 주어진 데이터를 분석하는 과정입니다
- 통계 수치를 구하여 이 값으로부터 여러 정보를 발견해내며, 다음과 같은 통계 수치를 주로 활용합니다
  - 평균, 중앙값(median), 최빈값(mode) 를 통해 데이터가 어느 값을 중심으로 뭉쳐있는지를 확인합니다
  - 분산, 표준편차, 분위수, Q1(25분위수), Q3(75분위수) 를 통해 데이터가 어떤 형태로 퍼져있는지를 확인합니다



# 정량적 데이터 분석 (통계수치)

## describe() - 요약통계

전반적인 주요 통계를 확인할 수 있습니다.  
기본 값으로 수치형(Numerical) 칼럼에 대한  
통계표를 보여줍니다

- **count**: 데이터 개수
- **mean**: 평균
- **std**: 표준편차
- **25%, 50%, 75%**:  
25분위(Q1), 50분위(median), 75분위(Q3)  
크기 순서로 나열했을때의  
'25%번째', ... 값

문자열 칼럼에 대한 통계표도 확인할 수 있습니다  
(.describe(include='object'))

- **unique**: 고유 데이터의 값 개수
- **top**: 가장 많이 출현한 데이터 개수
- **freq**: 가장 많이 출현한 데이터의 빈도수

```
df.describe()
```

	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
df.describe(include='object')
```

	sex	embarked	who	embark_town	alive
count	891	889	891	889	891
unique	2	3	3	3	2
top	male	S	man	Southampton	no
freq	577	644	537	644	549

## 정량적 데이터 분석 (통계수치)

- `df.count()` - 개수  
(column당) 데이터의 개수

```
df.count()
survived      891
pclass        891
sex           891
age           714
sibsp         891
parch         891
fare          891
embarked      889
class         891
who           891
adult_male    891
deck          203
embark_town   889
alive         891
alone         891
dtype: int64

# 단일 column의 데이터 개수를 구하는 경우
df['age'].count()
714
```

## 정량적 데이터 분석 (통계수치)

- `df.mean()` - 평균  
(column당) 데이터의 평균
- 기술통계 함수들에서  
`skipna=False`로 설정시 NaN값이 있는 column은  
NaN 값으로 출력됩니다.

```
df.mean()
survived      0.383838
pclass        2.308642
age           29.699118
sibsp         0.523008
parch         0.381594
fare          32.204208
adult_male    0.602694
alone         0.602694
dtype: float64

df['age'].mean()
29.69911764705882

# 조건별 평균
condition = (df['adult_male'] == True)
df.loc[condition, 'age'].mean()
33.17312348668281

df.mean(skipna=False)
survived      0.383838
pclass        2.308642
age           NaN
sibsp         0.523008
parch         0.381594
fare          32.204208
adult_male    0.602694
alone         0.602694
dtype: float64
```

## 정량적 데이터 분석 (통계수치)

- `df.median()` - 중앙값(50분위수)

데이터를 오름차순 정렬하여 중앙에 위치한 값입니다

이상치(outlier)가 존재하는 경우, `mean()`보다 `median()`을 대표값으로 더 선호합니다

- 짝수개의 데이터가 있는 경우에는 가운데 2개 중앙 데이터의 평균 값을 출력 합니다

```
df.median()
survived      0.0000
pclass        3.0000
age           28.0000
sibsp         0.0000
parch         0.0000
fare          14.4542
adult_male    1.0000
alone         1.0000
dtype: float64

pd.Series([1, 2, 3, 4, 999]).median()
3.0

# 짝수개의 데이터가 있는 경우
pd.Series([1, 2, 3, 4, 5, 999]).median()
3.5
```

## 정량적 데이터 분석 (통계수치)

- `df.sum()` - 합계
- 문자열 `column`은 모든 데이터가 붙어서 출력될 수 있습니다

```
df.loc[:, ['age', 'fare']].sum()
```

```
age      21205.1700
```

```
fare      28693.9493
```

```
dtype: float64
```

```
df['who'].sum()
```

```
'manwomanwomanwomanmanmanmanchildwomanchildchildwomanmanmanchild  
womanchildmanwomanwomanmanmanchildmanchildwomanmanmanwomanman  
anwomanwomanmanmanmanmanmanwomanchildwomanwomanmanchildwoma...'
```

## 정량적 데이터 분석 (통계수치)

- df.var() - 분산

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

where 평균 =  $\bar{X}$

데이터의 값들이 평균으로부터 얼마나 많이 흩뿌려져있는지를 나타냅니다

- df.std() - 표준편차

$$\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

```
fare_mean = df['fare'].values.mean()
my_var = ((df['fare'].values - fare_mean) ** 2).sum() /
(df['fare'].count() - 1)
my_var
2469.436845743116
```

```
df['fare'].var()
2469.436845743116
```

```
np.sqrt(df['fare'].var())
49.6934285971809
```

```
df['fare'].std()
49.6934285971809
```

## 정량적 데이터 분석 (통계수치)

- `df.agg([통계함수1, ...])`

복수의 통계 함수 적용할때 사용하는 함수입니다

```
df['age'].agg(['min', 'max', 'count', 'mean'])
```

```
min      0.420000
```

```
max      80.000000
```

```
count    714.000000
```

```
mean     29.699118
```

```
Name: age, dtype: float64
```

```
# 복수의 칼럼에 agg 적용
```

```
df[['age', 'fare']].agg(['min', 'max', 'count', 'mean'])
```

	age	fare
min	0.420000	0.000000
max	80.000000	512.329200
count	714.000000	891.000000
mean	29.699118	32.204208

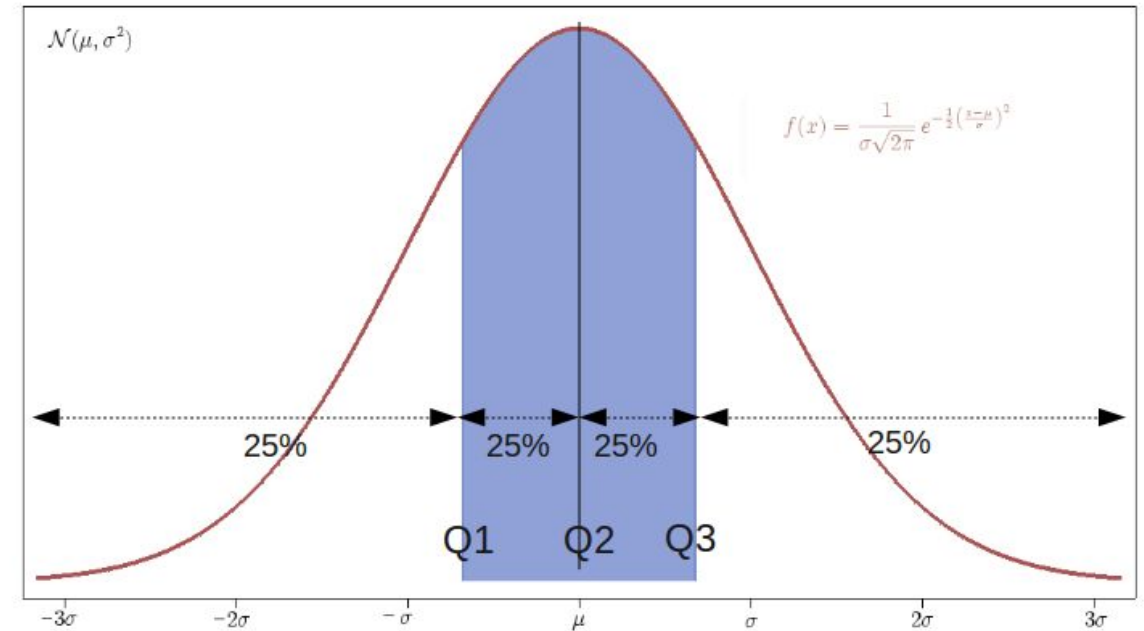
# 정량적 데이터 분석 (통계수치)

- `df.quantile()`

Quantile이란 주어진 데이터를 동등한 확률구간으로 분할하는 지점을 말합니다  
10%의 경우 0.1을, 80%의 경우 0.8을 대입하여 값을 구합니다

- `df['column_name'].unique()` - 칼럼 내 고유값

```
df['age'].quantile(0.1)
14.0
```



```
df['who'].unique()
array(['man', 'woman', 'child'], dtype=object)
```



## 정량적 데이터 분석 (통계수치)

- `df.mode` - 최빈값  
최빈값은 가장 많이 출현한 데이터를 의미합니다

- `df.corr()` - 상관관계

`corr()`로 칼럼별 상관관계를 확인할 수 있습니다

- -1~1 사이의 범위를 가집니다
- -1에 가까울 수록 반비례 관계, 1에 가까울수록 정비례 관계를 의미합니다

```
df['who'].mode()
```

```
0    man
dtype: object
```

# 카테고리형 데이터에도 적용 가능합니다.

```
df['deck'].mode()
```

```
0    C
```

```
Name: deck, dtype: category
```

```
Categories (7, object): ['A', 'B', 'C', 'D', 'E', 'F', 'G']
```

```
df.corr()
```

	survived	pclass	age	sibsp	parch	fare	adult_male	alone
survived	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307	-0.557080	-0.203367
pclass	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500	0.094035	0.135207
age	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067	0.280328	0.198270
sibsp	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651	-0.253586	-0.584471
parch	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225	-0.349943	-0.583398
fare	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000	-0.182024	-0.271832
adult_male	-0.557080	0.094035	0.280328	-0.253586	-0.349943	-0.182024	1.000000	0.404744
alone	-0.203367	0.135207	0.198270	-0.584471	-0.583398	-0.271832	0.404744	1.000000

# 가설검정

## 통계적 추정

---

- 통계적 추정: 모집단의 도 $\theta$ 수를 표본들의 통계값을 이용해서 추정하는 방법
- 점추정(Point estimation): 모집단의 특성을 단일한 값으로 추정  
이때 점추정값 $\hat{\theta}$ 는 표본값  $X_1, \dots, X_n$ 들의 함수입니다.
- 편향(Bias): 추정량의 기댓값과 모수의 차 $E(\hat{\theta}) - \theta$
- 평균제곱오차(Mean squared error  $E[(\hat{\theta} - \theta)^2]$ )

## 최대우도 추정량

---

- 우도함수(Likelihood function): 확률변수  $X_1, \dots, X_n$ 의 결합확률밀도함수  $f(x_1, \dots, x_n; \theta)$ 를 모수  $\theta$ 에 대한 함수로 볼 때, 이를 우도함수  $L(x_1, \dots, x_n; \theta)$ 라고 합니다  
즉, 결합확률밀도함수가 모수에 대한 함수일때 우도함수라고 합니다
- 확률 변수  $X_1, \dots, X_n$ 가 서로 독립이고 확률밀도함수  $f(x; \theta)$ 에서 얻은 표본이라면
$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$
- 최대우도추정량(Maximum likelihood estimate):  
 $X_1, \dots, X_n$ 를 표본으로 얻을 확률, 즉 우도함수  $L(x_1, \dots, x_n; \theta)$ 가 가장  $\hat{\theta}$  높은  
즉, MLE(최대우도추정량)은 주어진 관찰값을 가장 잘 설명하는 모수 추정량이 됩니다

## 정규분포의 MLE 유도

- 정규 분포의 likelihood function은

$$P(x|\theta) = \prod_{i=1}^n f_{\mu, \sigma^2}(x_i) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- Log 함수는 단조증가함수이므로 Log likelihood function의 최대값에서의  $\theta$ 는 Likelihood function의 최대값에서의  $\theta$ 와 같습니다.
- $\Pi$ (Product)에 Log 함수를 취해주면  $\Sigma$ 가 되어 미분을 하기 편해집니다
- 함수를  $\theta$ 에 대해 편미분한 값 = 0이 되는 지점에서 얻어지는  $\theta$ 는 함수의 최대점에서  $\theta$ 이 가지는 값입니다. 이 때 함수가 Likelihood function인 경우  $\theta$ 는 MLE가 됩니다.

$$\begin{aligned} \frac{\partial L(\theta|x)}{\partial \mu} &= -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{\partial}{\partial \mu} (x_i^2 - 2x_i\mu + \mu^2) & \frac{\partial L(\theta|x)}{\partial \sigma} &= -\frac{n}{\sigma} - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \frac{\partial}{\partial \sigma} \left( \frac{1}{\sigma^2} \right) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (-2x_i + 2\mu) & &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = \frac{1}{\sigma^2} \left( \sum_{i=1}^n x_i - n\mu \right) = 0 \end{aligned}$$

- 즉, 정규분포에서  $\mu$ 와  $\sigma^2$ 의 MLE는 다음과 같다  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$   $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$   
분산의 MLE는 편향추정량임을 알 수 있습니다.

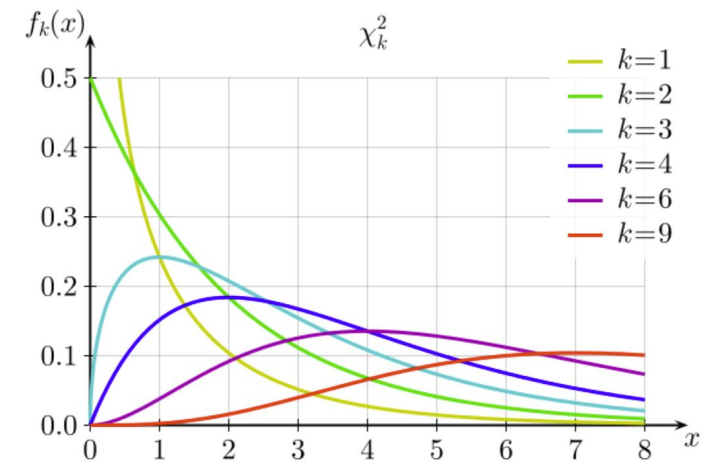
# 구간추정

---

- 점추정량은 추정된 값이 실제 모수와 얼마나 가까운지 알 수 없습니다.
- 구간추정(Interval estimation): 모수가 있을 것으로 예상되는 구간을 정해놓고, 해당 구간에 실제 모수가 있을 것으로 예상되는 확률을 구하는 것
- 신뢰도(Confidence level): 설정한 구간에 실제로 모수  $\theta$ 가 있을 확률  
e.g. 확률구간  $[a,b]$ 에 대해  $P(a < \theta < b) = 1 - \alpha$ 일때  $1 - \alpha$ 를 신뢰도,  $(1 - \alpha) * 100\%$  를 신뢰구간

# 카이제곱 분포

- 연속확률변수  $X$ 의 pdf  $f(x)$ 가 다음과 같을 [  $X$ 는 자유도  $\nu$ 인 카이제곱 분포  $\chi^2(\nu)$ 를 따른 
$$f(x;\nu) = \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, (x>0)$$
- 서로 독립인 확률변수  $X_1, X_2, \dots, X_n$ 가 각 자유도가  $\nu_1, \nu_2, \dots, \nu_n$ 인  $\chi^2$ 분포를 따를 때 확률 변수들의 합  $Y = X_1 + X_2 + \dots + X_n \sim \chi^2(\nu_1 + \nu_2 + \dots + \nu_n)$
- 연속확률변수  $X \sim N(\mu, \sigma^2)$ 일 때 확률변수  $Y = (X-\mu)^2/\sigma^2 \sim \chi^2(1)$
- $N(\mu, \sigma^2)$ 으로부터  $n$ 개의 표본을 추출해 구한 표본  $X_1, X_2, \dots, X_n$ 에 대해 
$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n)$$
- 반대로,  $\chi^2$ 분포를 따르는  $n$ 개의 표본의 합은  $n$ 이 커질수록 정규분포를 따른다 (by 중심극한정리)



# t-분포

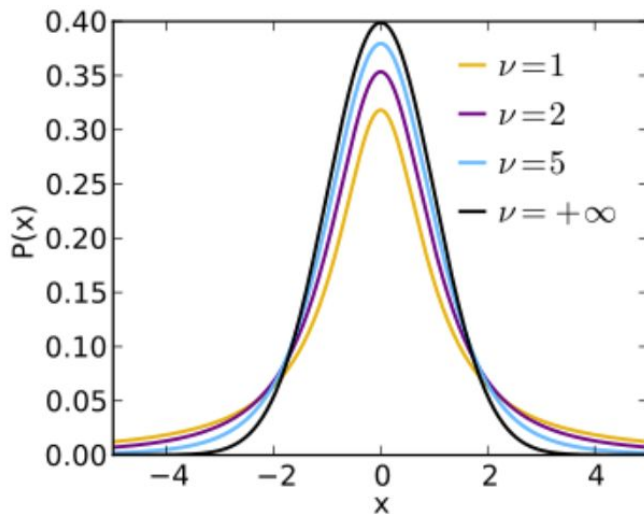
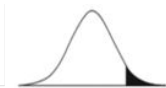
- 확률변수  $T$ 의 pdf가 다음과 같을 때  
 $T$ 는 자유도  $\nu$ 를 가진 t-분포를 따른다  
$$f(t)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\cdot\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},\quad-\infty<t<\infty$$

- 표준정규분포  $Z$ 를 찾기 위해서는 모분산  $\sigma^2$ 를 알아야 합니다  
→ 현실적으로 모분산을 알 수 없으므로 표본분산  $S^2$ 를 사용하고,  
이 때 정규분포 대신 t-분포를 이용합니다(보통은  $n<30$ 인 경우)

- $N(\mu, \sigma^2)$ 으로부터  $n$ 개의 표본( $X_i$ )을 추출해 구한 표본분산  $S^2$ 에 대해

$$T=\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}\sim\chi(n-1)$$

- t-분포에서의 확률은 z-분포와 비슷하게 표로 정리되어 있습니다

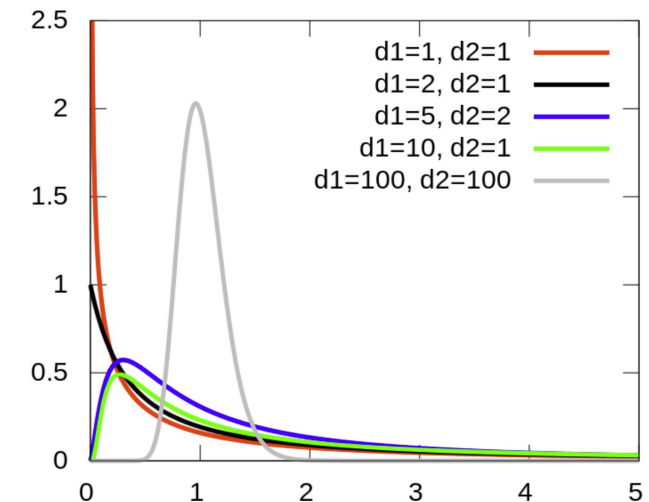



dof	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
70	0.254	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
130	0.254	0.676	1.288	1.657	1.978	2.355	2.614	2.856	3.154	3.367
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291



# F-분포

- 정규분포를 이루는 모집단에서 독립적으로 추출한 표본들의 분산 비율을 나타낼 때 F-분포를 사용할 수 있습니다
- 2개 이상의 표본 평균들이 동일한 모평균을 가진 집단에서 추출되었는지 OR 서로 다른 모집단에서 추출된 것인지를 판단할 때 사용합니다
- F-분포: 서로 독립인 두 확률변수 U와 V가 각각 자유도가  $v_1, v_2$ 인 카이제곱 분포를 따를 때, 새로운 확률 변수  $F = (U/v_1)/(V/v_2)$  는 자유도가  $(v_1, v_2)$ 인 F-분포를 따른다
- 확률변수 F가 자유도  $(v_1, v_2)$ 인 F-분포를 따를 때,  $1/F$ 는 자유도  $(v_2, v_1)$ 인 F-분포를 따릅니다
$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_{\alpha}(v_2, v_1)}$$
- 모분산이 각각  $\sigma_1^2$ 인,  $\sigma_2^2$ 인 정규 모집단에서 서로 독립적으로 추출된 크기  $(n_1, n_2)$ 인 표본의 분산을 각각  $S_1^2, S_2^2$ 라 할 때  $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2) = S_1^2\sigma_2^2/S_2^2\sigma_1^2$ 는 자유도가  $(n_1-1, n_2-1)$ 인 F-분포를 따른다



## 모평균 구간추정

- 모분산  $\sigma^2$ 이 알려진 경우 모평균  $\mu$ 의  $100(1-\alpha)\%$  신뢰구간은

$$\bar{X} - z_{0.5\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{0.5\alpha} \frac{\sigma}{\sqrt{n}}$$

이 때  $z_{0.5\alpha}$ 는  $P(Z > z_{0.5\alpha}) = 0.5\alpha$ 를 만족하는 z값입니다

- 가장 많이 사용하는 95% 신뢰구간의 경우  $0.5\alpha = 0.025$ ,  $z_{0.5\alpha} = 1.96$  입니다

- 모분산을 모르는 경우(표본분산을 이용하는) 모평균  $\mu$ 의 신뢰구간은

$$\bar{X} - t_{0.5\alpha, (n-1)} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{0.5\alpha, (n-1)} \frac{S}{\sqrt{n}}$$

이 때  $t_{0.5\alpha, (n-1)}$ 는, 자유도가 (n-1)인 t-분포를 따르는 T에 대해  $P(T > t_{0.5\alpha, (n-1)}) = 0.5\alpha$ 를 만족하는 t값입니다

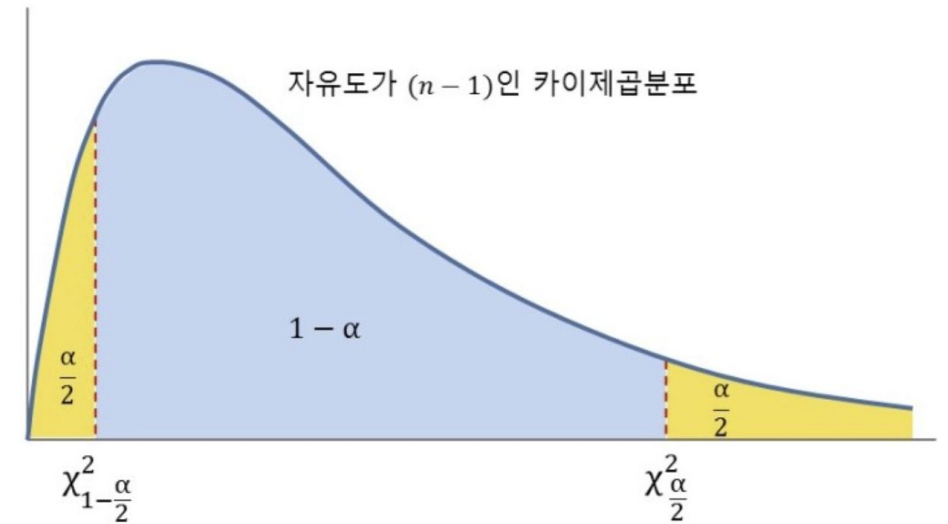
- t-분포는 표본수 < 30인 경우 사용합니다
- 표본수가 30개 이상인데 모분산을 모르는 경우에도, 표본 수가 충분히 많으면 모집단이 정규분포라는 조건 없이도 표본분산은 모분산에 매우 가까워지므로 z-분포를 이용해 신뢰구간을 구할 수 있습니다

## 모분산 구간추정

- 모분산  $\sigma^2$ 의  $100(1-\alpha)\%$  신뢰구간은

$$\frac{(n-1)S^2}{\chi_{0.5\alpha}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-0.5\alpha}^2}$$

여기서  $\chi_{0.5\alpha}^2, \chi_{1-0.5\alpha}^2$  는 자유도가  $(n-1)$ 인 카이제곱분포에서  $P(X > \chi_{0.5\alpha}^2) = 0.5\alpha, P(X > \chi_{1-0.5\alpha}^2) = 1-0.5\alpha$ 를 만족하는  $\chi^2$ 값입니다



## 모비율 구간추정

- 표본비율: 이항분포를 따르는 모비율이  $p$ 인 사건이  $n$ 개의 표본 중  $X$ 개가 나타났을 때 표본비율  $\hat{p} = X/n$ 의 분포는  $n \rightarrow \infty$ 에 가까워질 때 정규분포  $N(p, p(1-p)/n)$ 을 따르고  $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$  는 근사적으로 표준정규분포  $N(0,1^2)$ 를 따른다

- 표본비율의 기댓값  $E(\hat{p}) = p$ , 분산  $V(\hat{p}) = p(1-p)/n$
- 모비율이  $p$ 인 이항분포에서 충분히 많은  $n$ 개의 표본으로부터 나온 표본비율  $\hat{p}$ 에 대해 모비율  $p$ 의  $100(1-\alpha)\%$  신뢰구간은

$$\hat{p} - z_{0.5\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{0.5\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

# 가설검정

---

- 통계적 가설 검정: 표본에서 얻은 사실을 근거로 하여 모집단에 대한 가설이 맞는지 통계적으로 검정하는 분석 방법
- 가설을 먼저 세워야 합니다
- 귀무가설(Null hypothesis,  $H_0$ ): 직접 검정대상이 되는 가설  
먼저 증명된 적 없는 이 귀무가설이 일단 옳다는 가정 하에 검정을 시작하며,  
보통은 귀무가설이 진실일 가능성이 적기에 기각(reject)를 목표로 가설 세움
- 대립가설(Alternative hypothesis,  $H_1$ ): 귀무가설의 반대가 되는 가설  
보통은 새로운 주장 혹은 실제로 입증시키고 싶은 가설이며  
귀무가설이 기각되면 자동적으로 채택되도록 설계합니다

# 가설검정

- e.g. 대한민국 남성 100명을 대상으로 키를 sampling했을 때 표본평균 = 173이 나왔다고 하자.  
대한민국 남성 평균 키가 모표준편차가 12인 경우 170cm 이상이라고 할 수 있는지 알아볼 수 있는 가설은  
 $H_0: \mu = 170, H_1: \mu > 170$

- 유의수준(Level of significance,  $\alpha$ ): 귀무가설이 실제로 옳음에도 기각하는 오류  
즉, 귀무가설이 실제로는 맞지만 틀리다고 할 수 있는 확률, 위험부담  
보통 0.05로 값을 설정

- 임계값(Critical value): 유의수준이 주어졌을 때  
귀무가설의 채택과 기각 의사를 결정하는 기준이 되는 값

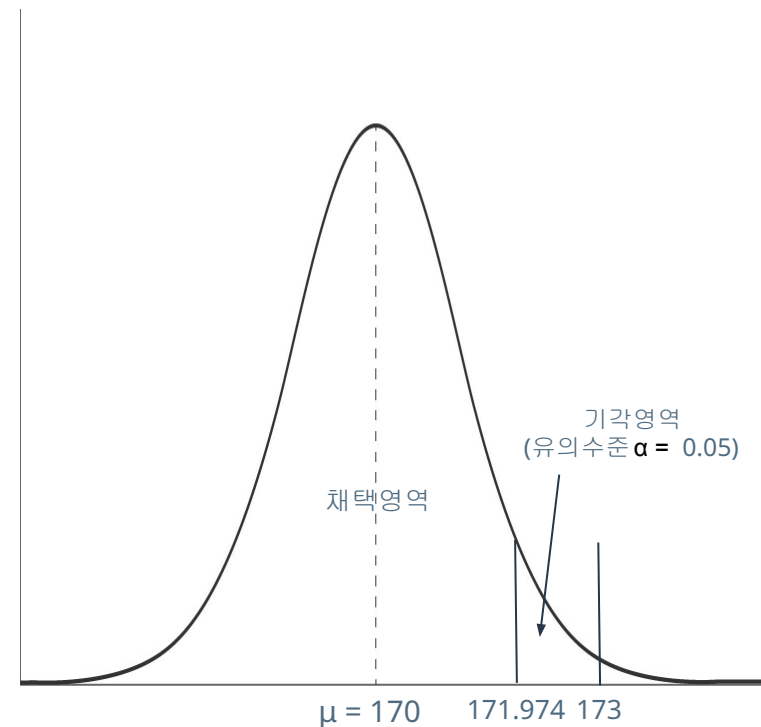
- e.g. 위 상황에서 유의수준  $\alpha=0.05$ 를 기준으로 임계값을 구해보면

표준 정규분포표를 이용해  $(1-\alpha)=0.95$ 에 해당하는 값을 찾으면  
 $z = 1.645$

( $H_1: \mu \neq 170$ 이 아닌  $\mu > 170$  이므로 단측검정입니다)

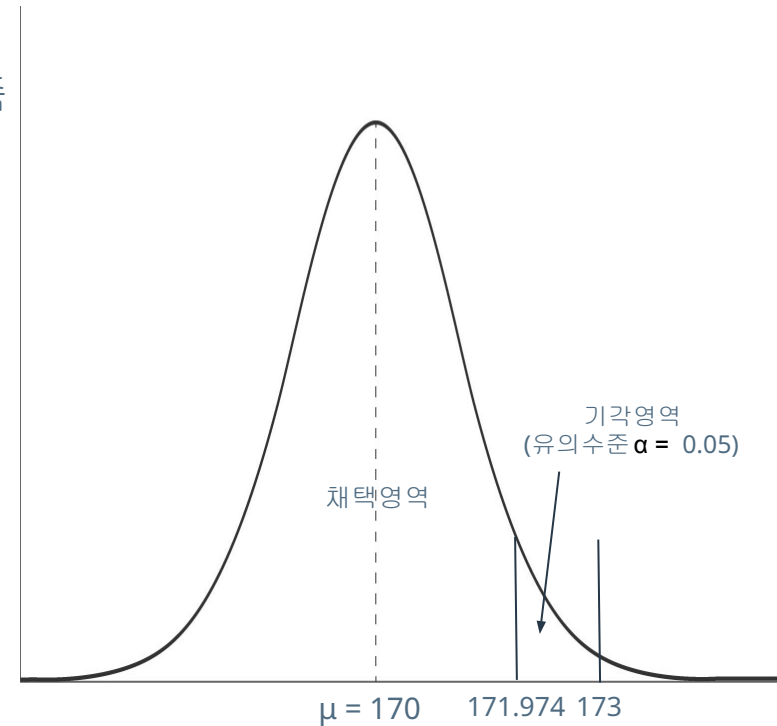
$$z = 1.645 = \frac{\text{임계값 } C.V - 170}{\frac{12}{\sqrt{100}}}$$

임계값 C.V 는 약 171.974가 됩니다



# 가설검정

- 임계값 171.974를 기준으로 표본평균은 173, 임계값보다 큼니다 (채택영역에 들어오지 못함)  
따라서 귀무가설은 기각되고, 대립가설이 채택됩니다  
즉, 대한민국 남자의 평균 키는 170cm 이상이라고 할 수 있습니다
- 단, 대립가설은 '귀무가설이 기각되었기 때문에 채택되었다' 라고 이야기해야됩니다  
즉, 대립가설이 참이라서가 아닌 다른 이유들로 귀무가설이 기각 되었다는 것을  
항상 기억해야됩니다  
e.g. 실험 설계의 오류, 귀무가설을 지지할 만한 충분한 증거가 부족



# 가설검정의 오류

- 가설검정에서 발생하는 두 가지 오류는 다음과 같습니다
- 제 1종 오류(type 1 error):  
귀무가설이 참임에도 이를 기각하는 오류
- 제 2종 오류(type 2 error):  
귀무가설이 거짓임에도 이를 채택하는 오류

	귀무가설 $H_0$ 참	귀무가설 $H_0$ 거짓
귀무가설 $H_0$ 채택	옳은 결정( $1 - \alpha$ )	제 2종 오류( $\beta$ -오류)
귀무가설 $H_0$ 기각	제 1종 오류( $\alpha$ -오류)	옳은 결정( $1 - \beta$ )

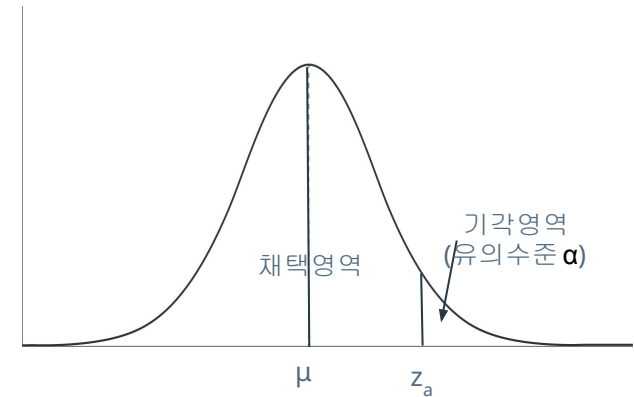
- $\alpha$ (유의수준) =  $P(\text{type 1 error}) = P(H_0\text{를 기각} \mid H_0\text{는 참})$   
 $\beta = P(\text{type 2 error}) = P(H_0\text{를 채택} \mid H_0\text{는 거짓})$
- $\alpha$ 와  $\beta$ 의 크기는 서로 상반되기 때문에 동시에 줄일 수 없습니다  
 다만, 표본의 크기를 증가시키면 분산이 작아지기 때문에 오류의 확률이 줄어듭니다



## 단측검정과 양측검정

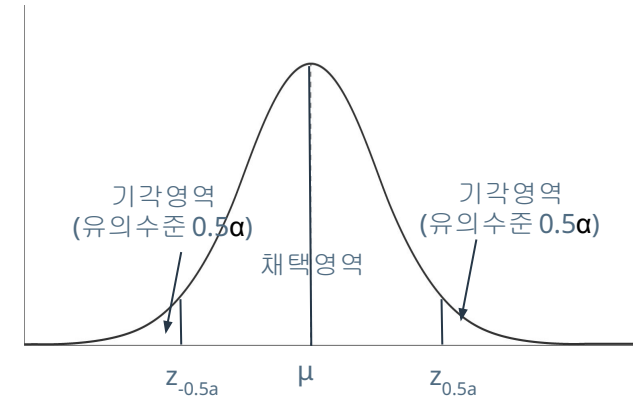
- 단측검정(One-sided test): 대립가설  $H_1$ 이 어떤 특정값 이상 / 이하 라고 설정되는 경우의 검정  
e.g.  $H_0: \mu = \mu_0, H_1: \mu > \mu_0$  /  $H_0: \mu \leq \mu_0, H_1: \mu > \mu_0$   
 $H_0: \mu = \mu_0, H_1: \mu < \mu_0$  /  $H_0: \mu \geq \mu_0, H_1: \mu < \mu_0$
- 단측검정의 경우 오른쪽 그림과 같이 유의수준을 잡고, 임계점을 계산해 기각영역을 설정합니다

e.g.  $H_0: \mu = \mu_0, H_1: \mu > \mu_0$  인 경우 기각역은  $z_\alpha$ ,  
 $H_0: \mu = \mu_0, H_1: \mu < \mu_0$  인 경우는  $-z_\alpha$   
→  $H_1: \mu > \mu_0$  인 경우  $z > z_\alpha$ 면 귀무가설 기각,  
 $H_1: \mu < \mu_0$  인 경우  $z < -z_\alpha$ 면 귀무가설 기각,



## 단측검정과 양측검정

- 양측검정(Two-sided test):  $H_0: \mu = \mu_0$ ,  $H_1: \mu \neq \mu_0$   
이 때  $\mu \neq \mu_0$ 는  $\mu > \mu_0$ 이거나  $\mu < \mu_0$ 임을 의미합니다
- 이 경우 유의 수준을 반으로 나누고 기각역을 양쪽에 설정합니다  
즉,  $|z| > z_{0.5\alpha}$ 이면 귀무가설 기각



## 모평균 가설검정

- 모평균의 구간 추정와 같이 모분산을 아는 경우와 모르는 경우로 나뉘서 접근합니다

- 모분산을 아는 경우 검정통계량  $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

단측검정인 경우  $Z > z_{\alpha}$ 이거나  $Z < -z_{\alpha}$ 면 귀무가설 기각

양측검정인 경우  $|Z| > z_{0.5\alpha}$ 이면 귀무가설 기각

- 모분산을 모르는 경우에는 t-분포를 이용해서 기각역을 구해야됩니다

검정통계량 T는 표본분산을 이용  $T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$

단측검정인 경우  $T > t_{\alpha, (n-1)}$ 이거나  $T < -t_{\alpha, (n-1)}$ 면 귀무가설 기각

양측검정인 경우  $|T| > T_{0.5\alpha, (n-1)}$ 이면 귀무가설 기각

## 모평균 차에 대한 가설검정

- 두 개의 모집단에 대한 가설검정도 진행할 수 있습니다.(e.g. 대한민국 남자들의 키 vs 여자들의 키)
- $X \sim N(\mu_1, \sigma_1^2)$ ,  $Y \sim N(\mu_2, \sigma_2^2)$  에서 각각  $n_1$ ,  $n_2$ 개 표본 추출하면 두 표본평균을 얻을 수 있습니다:

$$\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{n_1}), \bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$$

이 때,

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$$

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

(-1값이 Var() 밖으로 나오면서  $(-1)^2 = +1$ 이 되어 Var + Var)

따라서, 두 표본평균의 차이값은 다음 정규 분포를 따릅니다:

$$(\bar{X} - \bar{Y}) \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

- 두 집단의 모분산을 아는 경우에  $Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

두 집단의 모분산을 모르는 경우  $T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

를 이용하여 검정통계량을 계산 후

각각 표준정규분포, t-분포를 이용한 기각역과 비교하여 검정하면 됩니다

## 모분산 비에 대한 가설검정

---

- 두 모집단에서 나온 모분산이 같은지를 알아보는 가설 검정도 가능합니다
- 귀무가설  $H_0$ : 두 모집단의 분산이 같다 ( $\sigma_1^2 = \sigma_2^2$ )

이 때 두 분산의 비율  $S_1^2/S_2^2 = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$  ( $\sigma_1^2 = \sigma_2^2$ 이므로)는 자유도가  $(n_1-1, n_2-1)$ 인 F-분포를 따릅니다

- 검정통계량  $F = S_1^2/S_2^2$
- 양측검정:  $F \geq F_{0.5\alpha}(n_1-1, n_2-1), F \leq F_{1-0.5\alpha}(n_1-1, n_2-1)$
- 우측 단측검정:  $F \geq F_{\alpha}(n_1-1, n_2-1)$   
좌측 단측검정:  $F \leq F_{1-\alpha}(n_1-1, n_2-1)$

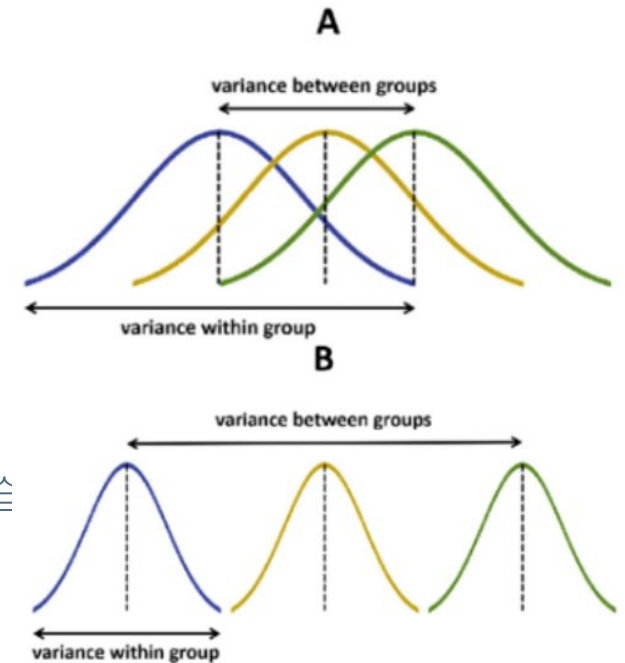
# ANOVA(analysis of variance, 분산분석)

- ANOVA: n개의 집단을 비교하는 통계적 분석 ( $n > 2$ )
- $n > 2$ 인 경우 n개의 집단에서 t 검정을 하는 경우 문제 발생  
e.g. n개의 집단에서 한 번이라도 type 1 error가 발생할 확률 =  $1 - 0.95^n > 0.05$   
즉, type 1 error의 누적을 해결하기 위해 ANOVA 사용
- 분산분석은 사용하기 전 3가지 조건을 만족해야 됩니다: 정규성, 독립성, 등분산성

정규성(normality): 모든 데이터가 정규분포를 따르는 모집단으로부터 추출됨  
정규분포를 따르지 않는 것으로 보이는 경우 Log 변환 등의 전처리가 필요할 수 있음

독립성(independency): 모든 데이터가 모집단으로부터 독립적으로 추출됨

등분산성(homoscedasticity): 모든 데이터는 분산이 동일한 모집단들로부터 추출됨  
(보통은 가장 큰 분산과 작은 분산의 비가 4:1 정도를 넘지 않으면 ANOVA를 적용해도 괜찮은 것으로 봄)



# ANOVA(analysis of variance, 분산분석)

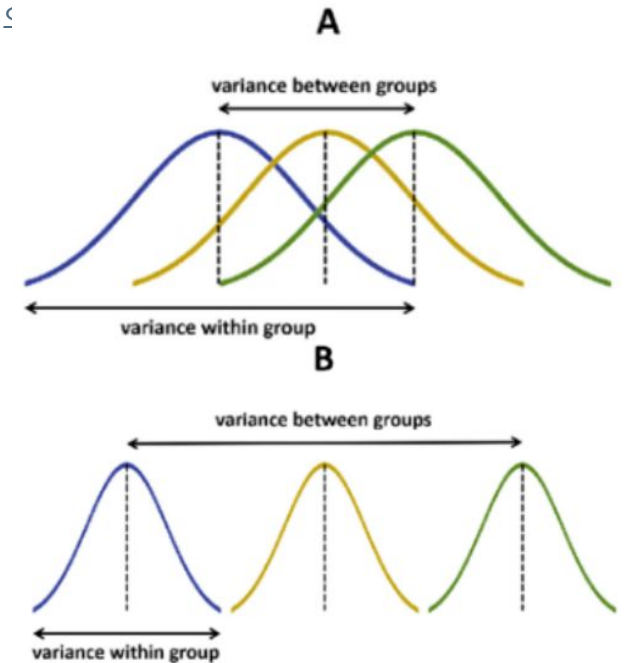
- 일원 분산분석(one-way ANOVA):  
집단의 종류(독립변수)가 하나이고 집단들이 가지는 평균값(종속 변수)이 하나인 경우  
그 집단간 모평균의 차이의 여부를 검증하는 방법
- 귀무가설  $H_0: \mu_{\text{group1}} = \dots = \mu_{\text{group m}} = \mu$  (m= 그룹 수)  
대립가설  $H_1$ : 적어도 한 쌍의 모평균은 같지 않음
- m개의 그룹에서, 각 그룹에서 추출한 n개의 표본에 대해서  
 $X_{ik}$ 를 i번째 그룹의 k번째 표본,  $\bar{X}_i$ 를 i번째 그룹 내 평균,  
전체 평균을  $\bar{X}$  이라고 할 때

Sum of square(제곱합)값들을 다음과 같이 구할 수 있습니다

SSB(Sum of squares between, 그룹간 제곱합)  $n \sum_{i=1}^m (\bar{X}_i - \bar{X})^2$

SSE(Sum of squares error, 그룹 내 제곱합)  $\sum_{i=1}^m \sum_{k=1}^n (X_{ik} - \bar{X}_i)^2$

SST(Sum of squares total, 전체 제곱합)  $= \sum_{i=1}^m \sum_{k=1}^n (X_{ik} - \bar{X})^2$



# ANOVA(analysis of variance, 분산분석)

- Sum of squares 값들을 이용해서 그룹 내 분산, 그룹 간 분산, 그리고 분산비를 구할 수 있습니다

- $MSB(\text{Mean squares between, 그룹간 분산}) = SSB / dfB = SSB / (m-1)$

$MSE(\text{Mean squares error, 그룹내 분산}) = SSE / dfE = SSE / m(n-1)$

분산비  $F = MSB / MSE$

(df: degree of freedom, 자유도)

- 여기서 구한 분산비  $F$ 는 자유도가  $(k-1, n-k)$ 인  $F$ 분포를 따릅니다

- $H_0$ 는 '각 집단간 평균이 같다' 이므로 맞다면  $MSB$ 가  $MSE$ 보다 작아야 됩니다  
즉, 이  $F$ 값과 유의수준  $\alpha$ , 자유도  $(k-1, n-k)$ 를 이용해서  
우측 단측검정( $F = MSB / MSE > 1$ )을 진행하면 됩니다.

우측 단측검정:  $F \geq F_{\alpha}(k-1, n-k)$

