제 4장. 표본분포

베르누이 분포

- **베르누이 시행 (Bernoulli trial)**

  : 시행의 결과가 성공(s) 또는 실패(f) 두 개 뿐인 시행.

  : 성공의 확률= $p = \Pr(s)$, 실패의 확률= $q = \Pr(f) = 1 - p$

- 베르누이 확률변수

  : 베르누이 시행의 표본공간 $S = \{s, f\}$ 에서 $X(s) = 1,\ \ X(f) = 0$ 인 확률변수

- 베르누이 분포

  : 베르누이 확률변수의 확률분포, $X \sim Bernoulli(p)$

  : 특성값이 이원적인 모집단의 분포를 나타낼 때 사용될 수 있음

x	0	1
$p(x)$	$1-p$	p

정규분포

- **정규분포 (normal distribution)**

  : 가우스(Gauss, 1777-1855)에 의해 제시, 가우스 분포 (Gauss distribution)

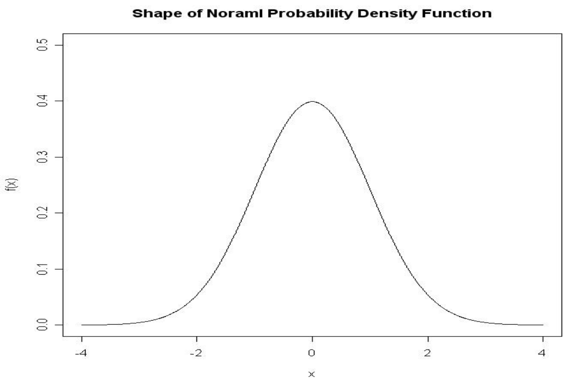
  : 특성값이 연속적이며 셀 수 없이 많은 무한모집단의 대표적인 분포

- **정규분포의 확률 밀도함수**

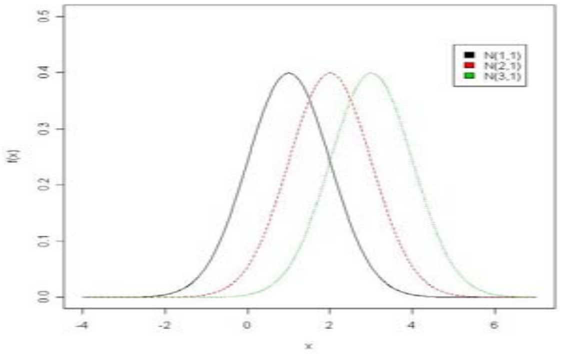
$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \qquad -\infty < x < \infty$$

- 정규분포의 특성
  - 평균 μ 를 중심으로 대칭이며, 대칭점에서의 높이가 가장 높음
  - 표준편차 σ 는 변곡점과 대칭점 사이의 거리를 나타냄

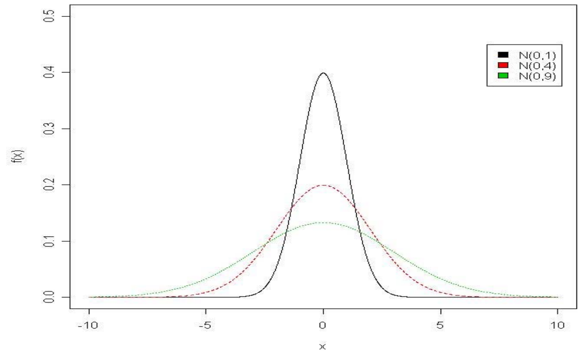
< 정규분포의 밀도함수>



<분산이 같고 평균이 다른 정규분포>



<평균이 같고 분산이 다른 정규분포>



정규분포의 성질

- X : 정규 모집단에서 관측된 관측값 (확률변수) $\Rightarrow$ $X \sim N(\mu, \sigma^2)$
  - $aX + b \sim N(a\mu + b, a^2\sigma^2)$
  - $X_1 \sim N(\mu_1, \sigma_1^2),\ X_2 \sim N(\mu_2, \sigma_2^2),\quad X_1, X_2 : indep.$

    $\Rightarrow a_1X_1 + a_2X_2 \sim N(a_1\mu_1 + a_2\mu_2,\ a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$

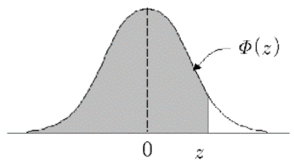
- 표준정규분포 (standard normal distribution)

  : 평균이 0이고 분산이 1인 정규분포.

  : 확률변수 $X \sim N(\mu, \sigma^2)$ 를 표준화 한 확률변수

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$



- 표준정규분포표

  : 표준정규분포 누적확률을 계산해 놓은 표

  : $P(Z < z)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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	0.5754
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	0.7258	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	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	0.7996	0.8023	0.8051	0.8079	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

정규분포의 확률 계산

- 정규분포의 확률 계산

: 일반적인 공식이 존재하지 않아, 정규분포와 표준정규분포 사이의 관계를 정의하고 컴퓨터로 계산된 표준 정규분포의 누적 확률분포 함수 값을 사용

- $$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$

- $$P(a \le Z \le b) = P(Z \le b) - P(Z \le a)$$
- $$P(Z \ge z) = 1 - P(Z \le z)$$

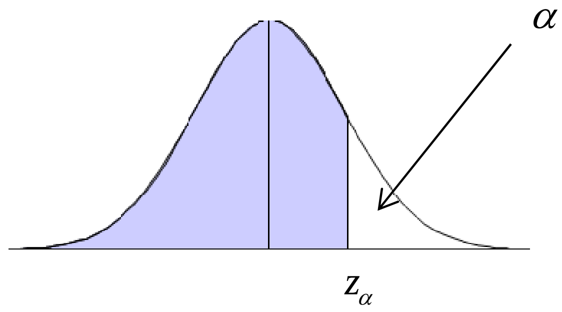
- 어느 학과 학생들의 통계학 성적 분포가 근사적으로 $N(60,10^2)$ 을 따를 때, 45점 이하인 학생에게 F 학점을 준다면 F 학점을 받게 될 학생들의 비율을 근사적으로 구하여라

표준정규분포의 백분위수

- 표준정규분포의 $100(1-\alpha)$ 백분위수 (z_α)

  : $Z \sim N(0,1)$ 일 때, 주어진 α 값에 대하여

  $P(Z > z) = \alpha$ 를 만족하는 z 값

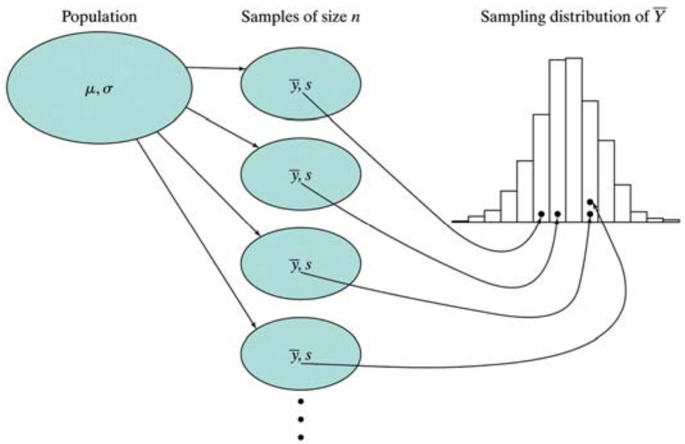


- $z_{0.005} = 2.58, \quad z_{0.025} = 1.96, \quad z_{0.05} = 1.645$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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	0.5754
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	0.7258	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	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	0.7996	0.8023	0.8051	0.8079	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	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	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	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767

표본 분포

- 예 : 유한 모집단 {2,2,4,5}에서 크기 2인 표본을 SRSWOR로 추출하여 모평균을 추정.



- **표본 분포 (sampling distribution)**
  : 통계량의 확률분포
  : 통계량의 표본분포를 통해
  통계량의 정확성을 계산해 낼 수 있음

- 통계량의 표본분포는 모집단의 분포와 표본의 추출방법에 의해 결정됨

랜덤 표본

- 예 4.1 : 10개 중 3개의 당첨제비가 있을 때, 두 개의 제비를 단순랜덤 복원 추출하는 경우,

  $X_i = i$ 번째 시행의 결과

- 예 4.2 : 10개 중 3개의 당첨제비가 있을 때, 두 개의 제비를 단순랜덤 비복원 추출하는 경우,

랜덤 표본

- **랜덤 표본 (random sample)**
  - 유한 모집단의 경우, 단순랜덤 비복원 추출로 뽑은 표본을 의미함
  - 그런데 모집단의 크기가 충분히 큰 경우에는 복원 - 비복원 추출의 차이가 거의 없음
  - 따라서 모집단의 크기가 큰 유한 모집단, 또는 무한 모집단의 랜덤 표본은 다음과 같은 성질을 갖게 된다.

  (1) $X_1, X_2, \ldots, X_n$ 각각의 분포가 모집단 분포와 동일하고
  (2) $X_1, X_2, \ldots, X_n$ 은 서로 독립

- **I.I.D.** : Independently and Identically Distributed
  - $X_1, \ldots, X_n \sim i.i.d.\ f(x)$

  ⇔ 확률변수들이 서로 독립이고 동일한 분포를 따른다

초기하분포

- **초기하분포 (hypergeometric distribution)**
  특성값 "1"의 개수가 D, "0"의 개수가 $N-D$ 인 크기 N의 유한모집단에서, 크기 n인 랜덤 표본을 뽑을 때, 표본에서 "1"의 개수의 확률분포를 초기하분포라고 하며, 확률밀도함수는 다음과 같다

$$p(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, \ldots, n.$$

  단, $n \le D, \;\; n \le N - D.$

- 초기하분포의 평균과 분산
  : $X \sim H(N, D, n)$ 일 때,

$$E(X) = np, \qquad p = D/N$$

$$V(X) = np(1-p)\frac{N-n}{N-1}$$

초기하분포

- 예제 4.7 : 찬성자의 수가 6명, 반대자의 수가 4명인 크기가 10인 모집단에서 크기가 3인 랜덤 표본을 선택하는 경우, 표본에서 찬성자 수의 평균과 분산을 구하여라

이항분포

- **이항분포 (binomial distribution)**

: 특성값 “1”의 비가 p 이며 “1”과 “0”으로 이루어진 무한모집단에서, 크기 n 인 랜덤 표본을 뽑을 때, 표본에서 “1”의 개수의 확률분포.

: 성공률 p인 베르누이 시행을 독립적으로 n번 반복 시행할 때, 성공의 횟수 X의 분포는 이항분포를 따르게 되고, 확률밀도 함수는 다음과 같다

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

이항분포

- 이항분포의 평균과 분산

$X \sim B(n, p)$ 일때,

$$E(X) = np$$
$$V(X) = np(1-p)$$

- 초기하분포의 이항분포 근사

$X \sim H(N, D, n)$ 에서 $N \to \infty$, $\dfrac{D}{N} \to p$ 이면 $X \to B(n, p)$ 이다.

이항분포

- 예제 4.8 : 5개 중 하나를 택하는 선다형 문제가 20문항 있는 시험에서 랜덤하게 답을 써 넣는 경우,
  - 정답이 하나도 없을 확률은?
  - 4개부터 6개 사이의 정답을 맞힐 확률은?
- 예제 4.9 : 어떤 제품을 생산하는 공정의 불량률이 5%로 알려져 있다. 오늘 생산한 10,000개의 제품 중 20개를 단순랜덤추출하여 조사할 때, 불량률이 10% 이상일 확률을 구하여라.

표본평균의 분포

- $\{X_1, X_2, \ldots, X_n\}$ 가 (μ, σ^2) 인 무한모집단으로부터의 랜덤표본일때,

$$E\left(\bar{X}\right) = \mu, \qquad Var\left(\bar{X}\right) = \frac{\sigma^2}{n}$$

- 표본평균의 표준오차

$$S.E.\left(\bar{X}\right) = \frac{\sigma}{\sqrt{n}}$$

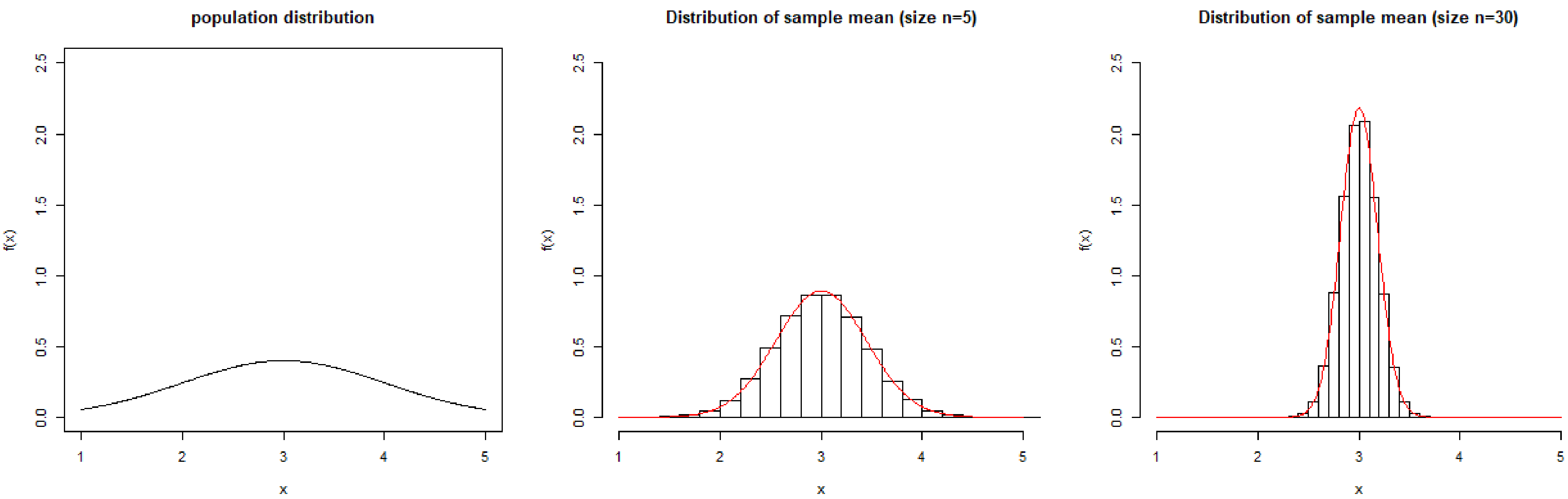
- 데이터가 많아질수록 표본평균의 표준편차는 작아진다.
- 즉, 표본평균이 모평균으로부터 벗어날 확률이 점점 작아진다는 것을 의미한다.
- 따라서 데이터가 많아질 수록 표본평균은 모평균을 더욱 정확하게 추측한다는 것을 알 수 있다.

- 표준편차(sd) : 자료가 가지고 있는 변동성 또는 흩어짐의 정도.
- 표준오차(se) : 표본에서 얻은 통계량이 가지고 있는 흩어짐의 정도.

표본평균의 분포 : 모집단이 정규분포인 경우

- 모집단이 $N(\mu,\sigma^2)$의 정규분포를 따르는 경우,

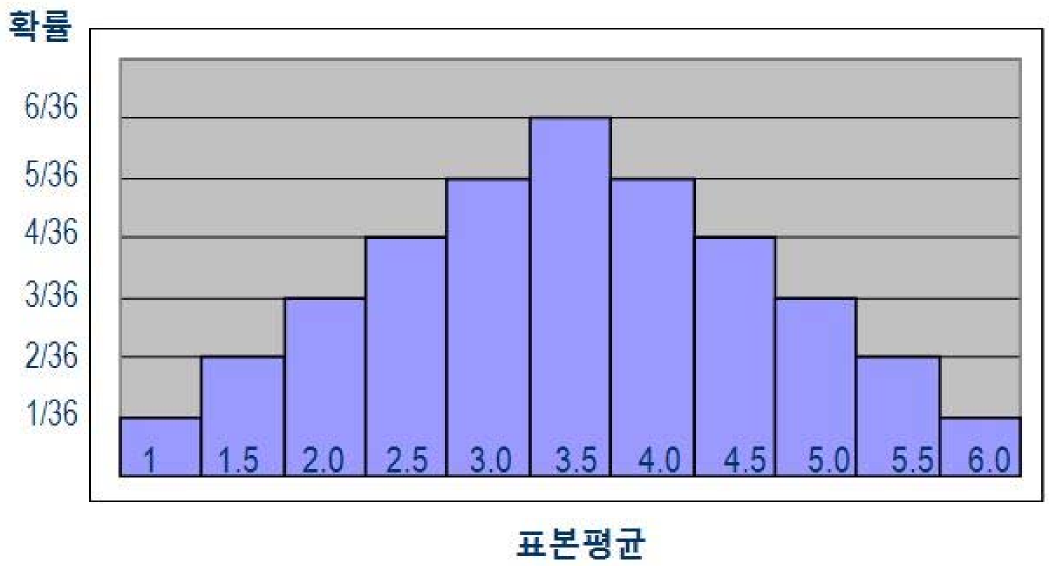
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



표본평균의 분포 : 모집단이 정규분포가 아닌경우

- Q : 모집단이 정규분포를 따르지 않는 경우의 표본평균의 분포는?
- 예 : 주사위 던지기 - 주사위 던지는 문제는 무한 모집단으로 볼 수 있다.

X	P(X=x)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6



- 확률변수 X = 주사위를 던졌을 때 나온 눈.

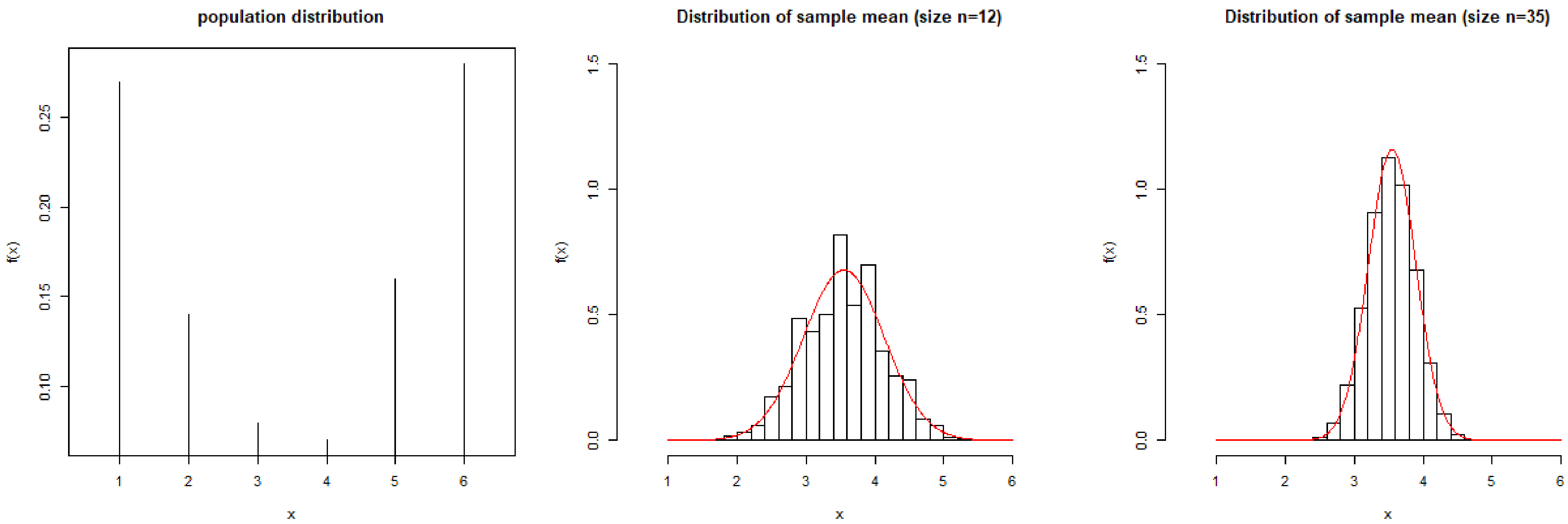
- 주사위 2개를 던졌을 때 나온 눈의 평균의 분포는?

중심극한정리

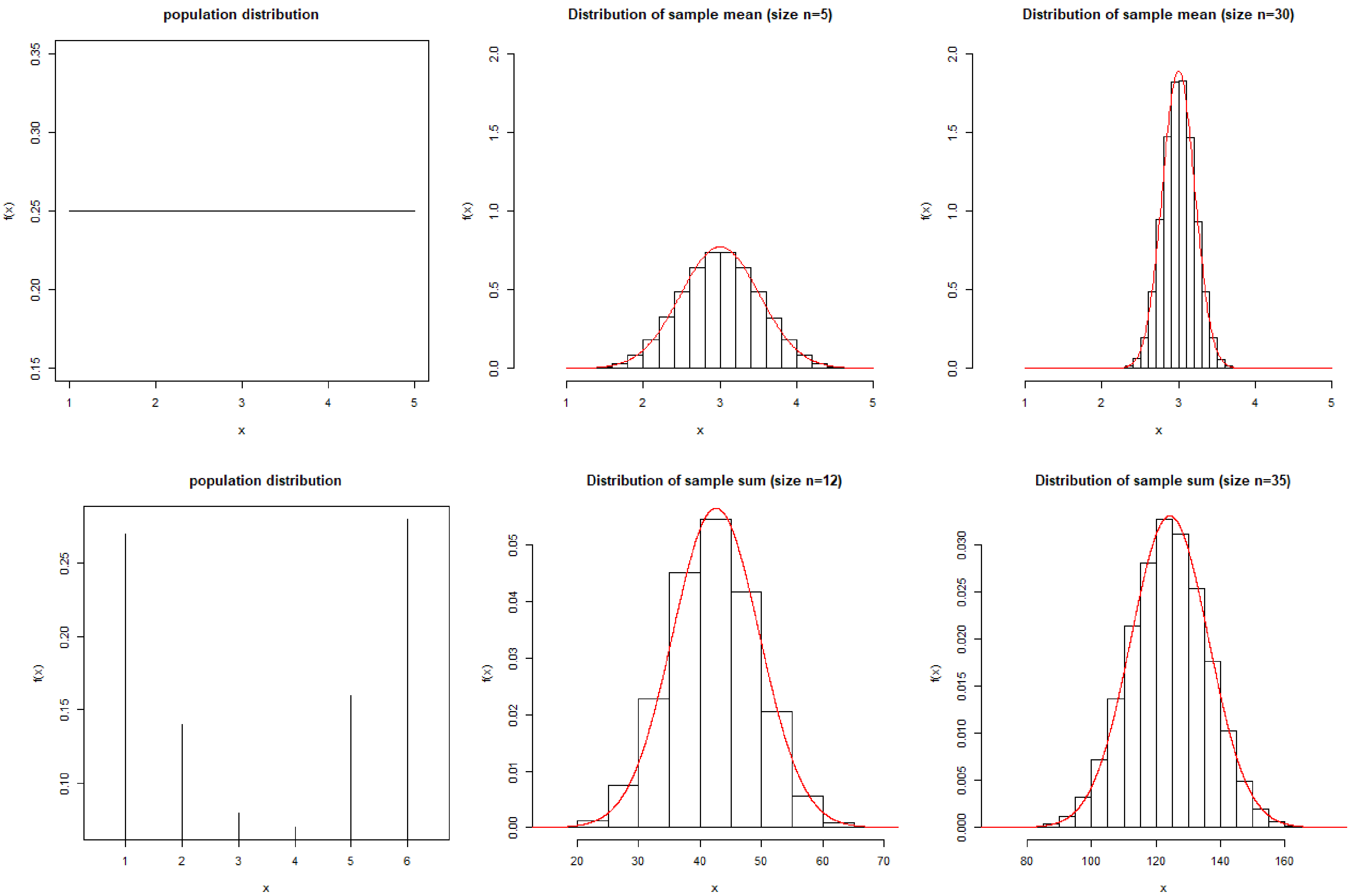
- **중심극한정리 (Central Limit Theorem)**

  : 모집단이 정규분포가 아니라도, 표본의 크기 n이 충분히 크면 표본평균의 분포는 근사적으로 정규분포를 따름

- 경험적으로 중심극한정리는 n 이 30이상이면 적용할 수 있는 것으로 알려져 있음
- 그림 예 : 비균등 주사위를 던졌을 때 표본평균의 분포



중심극한정리의 적용


population distribution
Distribution of sample mean (size n=5)
Distribution of sample mean (size n=30)
population distribution
Distribution of sample sum (size n=12)
Distribution of sample sum (size n=35)
f(x)
x

표본평균의 분포

- 표본평균의 분포
  - 모집단이 $N\left(\mu,\sigma^2\right)$ 일 때 :

$$\bar{X} \sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

  - 모집단이 $\left(\mu,\sigma^2\right)$ 이고, 표본의 크기가 30 이상일 때 :

$$\bar{X} \dot{\sim} N\left(\mu,\frac{\sigma^2}{n}\right)$$

- ABC 대학의 한 건물의 엘리베이터에는 다음과 같은 문구가 적혀 있다. “최대 적재 하중 2,700kg 혹은 40명.” 통계학을 수강하는 한 학생은 40명의 몸무게가 과연 2,700kg을 넘는 확률이 얼마나 될지 궁금 했다. ABC 대학 구성원들의 평균 몸무게는 64kg이고, 표준편차는 10kg인 것을 가정하라.

이항분포의 정규근사

- 이항분포 $X \sim B(n, p)$ 을 따르는 확률변수에서

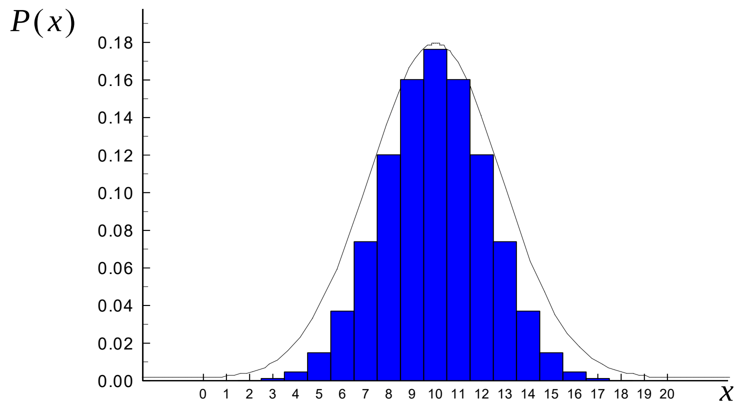
$$P(a \le X \le b) = \sum_{k=a}^{k=b} \binom{n}{k} p^k (1-p)^{n-k}$$ 을 계산하고자 할 때,

  - n 이 작은 경우
    : 이항분포표를 이용, 혹은 직접 계산

  - n 이 매우 크거나, 확률의 정확한 값을 알 필요가 없을 때
    : 정규분포를 이용한 근사 계산
    : **np>5와 n(1-p)>5** 의 조건이 필요

    예) p=0.5 이면, n≥10
    p=0.01 또는 p=0.99 이면, n≥500

<n = 20, p = 0.5 인 이항분포와 정규근사>



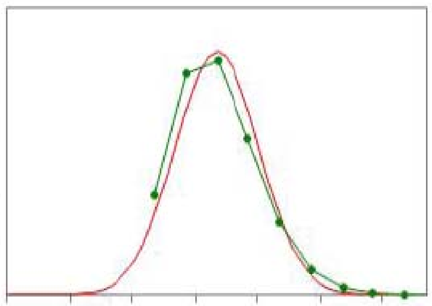
이항분포의 정규근사

- $X \sim B(n, p)$ 일 때, 확률 변수 X의 분포는 다음과 같다.

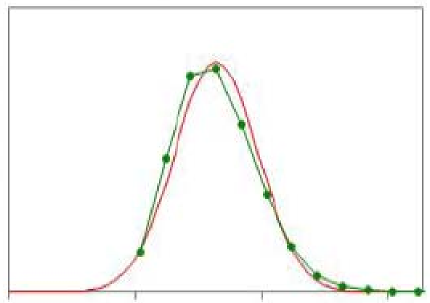
$$X \sim B(n, p) \quad \rightarrow \quad X \dot{\sim} N\left(np,\ np(1-p)\right)$$

단, $np > 5$, $n(1-p) > 5$.

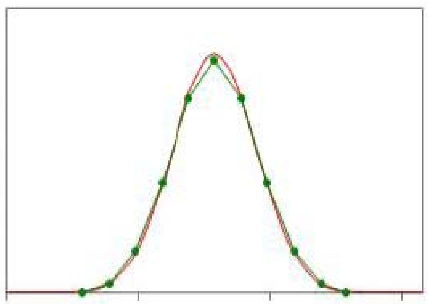
- 정규분포와 이항분포의 비교



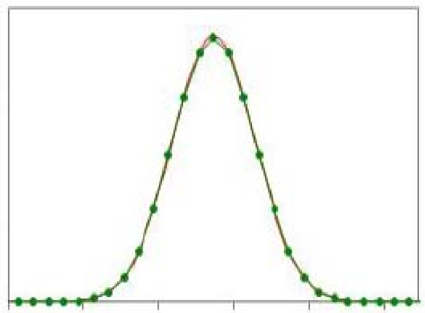
(a) $p = 0.1$, $n = 20$



(b) $p = 0.1$, $n = 30$



(c) $p = 0.5$, $n = 10$



(d) $p = 0.5$, $n = 30$

연속성 수정

- **연속성 수정 (continuity correction)**

: 이산형 분포를 연속형 분포로 이용하면서 생기는 오차를 보정함

: 연속성 수정계수 (continuity correction factor)

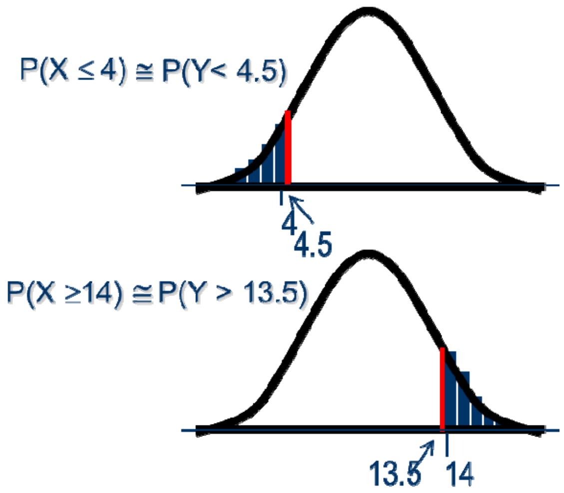
$$\mu = np,\ \sigma = \sqrt{np(1-p)}$$

$$P(a \le X \le b) \approx P\left(\frac{a-0.5-\mu}{\sigma} < Z < \frac{b+0.5-\mu}{\sigma}\right)$$

$$P(a \le X < b) \approx P\left(\frac{a-0.5-\mu}{\sigma} < Z < \frac{b-0.5-\mu}{\sigma}\right)$$

$$P(a < X \le b) \approx P\left(\frac{a+0.5-\mu}{\sigma} < Z < \frac{b+0.5-\mu}{\sigma}\right)$$

$$P(a < X < b) \approx P\left(\frac{a+0.5-\mu}{\sigma} < Z < \frac{b-0.5-\mu}{\sigma}\right)$$



이항분포의 정규근사 : 연속성 정정 예제

- 예제 : $X \sim B(20, 0.3)$ 일 때, $P(2 \leq X \leq 5)$ 의 값을 구하는 방법
  - 정확한 값
  - 이항분포의 정규근사
  - 연속성 정정을 이용한 이항분포의 정규근사

표본비율의 분포

- 표본비율의 예
  - 정부 정책에 대한 찬성률
  - 벤처기업의 파산 비율
  - 생산제품의 불량률
  - 같은 자동차를 다시 사겠다는 소비자의 비율

- 비율이란 1 또는 0으로 이루어져있는 자료의 평균 : 평균의 특수한 경우
  예) 다섯 번 시도 중 두 번은 성공이고 세 번은 실패일 때,

$$\hat{p} = 0.4 = \frac{1+0+1+0+0}{5} = \frac{\sum x_i}{n} = \overline{x}$$

- **표본비율은 표본평균의 특수한 경우**이므로 중심극한정리를 적용할 수 있다.

$$\hat{p} \dot{\sim} N\left(p,\ \frac{p(1-p)}{n}\right), \qquad np > 5 \ \& \ n(1-p) > 5$$

표본비율의 분포

- n개의 표본에서 '성공'의 수가 X 라면, X 는 베르누이 시행이 n개 모인 것.

$$X = \sum_{i=1}^{n} X_i \sim B(n, p)$$

  - 이항분포의 평균과 분산은

$$E(X) = np, \qquad V(X) = np(1-p)$$

  - **표본비율(sample proportion)**

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X}{n}$$

  - 따라서, 표본비율의 평균과 분산은

$$E(\hat{p}) = p, \qquad Var(\hat{p}) = \frac{p(1-p)}{n}$$

- 중심극한정리를 적용하면,

$$\hat{p} \stackrel{\cdot}{\sim} N\left(p, \ \frac{p(1-p)}{n}\right)$$

표본비율의 분포 : 예제

- 예 : 어느 대학의 신입생들을 대상으로 하는 통계학 강의는 70명씩 10개반으로 나누어 진행되고 있다. 신입생 전체 여학생의 비율이 0.54라고 할 때, 어느 특정 통계학 반의 여학생수가 과반이 될 확률은?

카이제곱분포

- **카이제곱분포 (chi-squared distribution)**

  : 확률변수 $Z_1, Z_2, \ldots, Z_k$ 이 표준정규분포 $N(0,1)$ 의 랜덤 표본일 때,

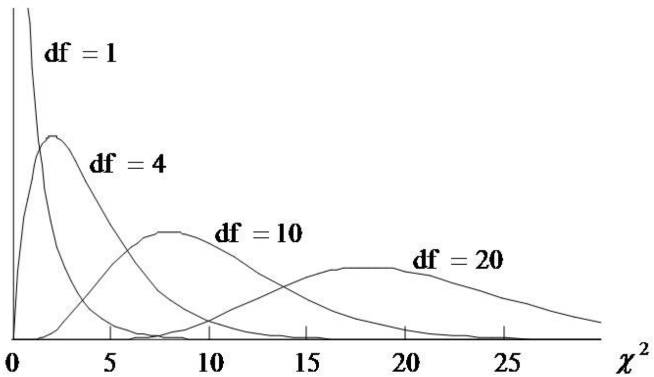
  - $Z_1^2 \sim \chi^2(1)$: 자유도 (degree of freedom) 1인 카이제곱분포
  - $V_1 \sim \chi^2(k_1)$, $V_2 \sim \chi^2(k_2)$ 이고, V_1, V_2 이 독립일 때,

$$V_1 + V_2 \sim \chi^2(k_1 + k_2)$$

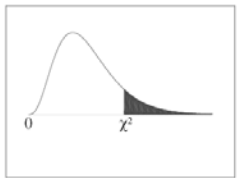
  - 따라서 $Z_1^2 + \ldots + Z_k^2 \sim \chi^2(k)$ 이고, 자유도 k 인 카이제곱분포라고 한다.

- $\chi_\alpha^2(k)$: 자유도가 k인 카이제곱분포의 $(1-\alpha)$ 분위수

$$V \sim \chi^2(k) \quad \to \quad P\{V \geq \chi_\alpha^2(k)\} = \alpha$$



Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589

표본분산의 분포

- $X_1, X_2, \ldots, X_n$ 이 $N(\mu, \sigma^2)$ 의 랜덤표본일 때, 표본분산 S^2 에 대해 다음이 성립한다

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \quad \rightarrow \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

두 정규모집단에서의 표본분산의 분포

- $X_1, X_2, ..., X_{n_1}$ 과 $Y_1, Y_2, ..., Y_{n_2}$ 이 서로 독립이고 각각 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ 에서의 랜덤표본일 때, 합동표본분산(pooled sample variance) S_p^2 에 대해 다음이 성립한다

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

단, $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

t-분포

- t - 분포 (t -distribution)

  서로 독립인 두 확률변수 $Z \sim N(0,1)$ 와 $V \sim \chi^2(k)$ 에 대하여 아래의 확률변수

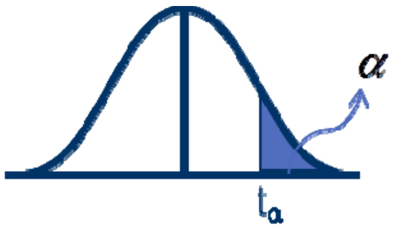
$$T = \frac{Z}{\sqrt{V/k}}$$

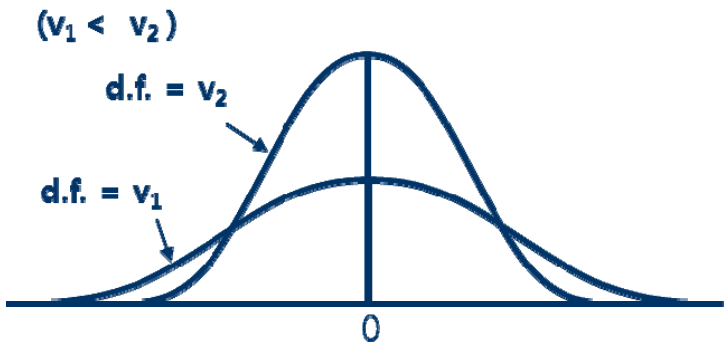
  의 분포를 자유도 k 인 t 분포라고 한다. ($T \sim t(k)$)

  - 표준정규분포처럼 0을 중심으로 좌우 대칭이지만 두꺼운 꼬리를 갖고 있음
  - 자유도(degree of freedom)라는 모수를 가지며, 자유도에 따라 분포의 모양 결정
  - 자유도가 커질수록 0을 중심으로 더욱 조밀하게 모이며, 자유도가 30 이상인 경우에는 표준정규분포와 거의 유사해짐 (자유도가 커질수록 표준정규분포로 수렴)

<자유도에 따른 분포의 모양>

  - $t_\alpha(k)$: 자유도가 k 인 t 분포의 $(1-\alpha)$ 분위수





스튜던트화된 표본평균의 분포

- 스튜던트화(studentization) 된 표본평균의 분포
  정규모집단 $N(\mu,\sigma^2)$ 으로부터의 랜덤표본 $X_1,\ldots,X_n$ 에 대하여 스튜던트화된 표본 평균의 분포는 다음과 같다.

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

자유도	$t_{.100}$	$t_{.05}$	$t_{.025}$	$t_{.01}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.92	4.303	6.965	9.925
.	.	.	.	.	.
.	.	.	.	.	.
20	1.325	1.725	2.086	2.528	2.845
.	.	.	.	.	.
.	.	.	.	.	.
200	1.286	1.653	1.972	2.345	2.601
	1.282	1.645	1.96	2.326	2.576

두 정규모집단에서의 t 분포

- $X_1, X_2, ..., X_{n_1}$ 과 $Y_1, Y_2, ..., Y_{n_2}$ 이 서로 독립이고 각각 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ 에서의 랜덤표본이라고 할 때, 다음이 성립한다.

$$\frac{\left(\bar{X} - \bar{Y}\right) - \left(\mu_1 - \mu_2\right)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

F 분포

- **F 분포 (F-distribution)**

  $V_1 \sim \chi^2(k_1),\ V_2 \sim \chi^2(k_2)$ 이고 V_1, V_2 가 서로 독립이면

$$F = \frac{V_1 / k_1}{V_2 / k_2} \sim F(k_1, k_2)$$

  이고, 자유도가 (k_1, k_2) 인 F 분포라고 한다.

  - $F_\alpha(k_1, k_2)$: 자유도가 (k_1, k_2) 인 F 분포의 $(1-\alpha)$ 분위수

- F 분포의 성질 : $F \sim F(k_1, k_2)$ 일 때, 다음이 성립한다.

$$\frac{1}{F} \sim F(k_2, k_1) \quad \rightarrow \quad F_{1-\alpha}(k_2, k_1) = \frac{1}{F_a(k_1, k_2)}$$

v_2	q	분자자유도 v_1					
		1	2	3	4	5	6
1	0.10	39.9	49.5	53.6	55.8	57.2	58.2
	0.05	161	200	216	225	230	234
	0.025	648	800	864	900	922	937
	0.01	4,052	5,000	5,403	5,625	5,764	5,859
2	0.10	8.53	9.00	9.16	9.24	9.29	9.33
	0.05	18.5	19.0	19.2	19.2	19.3	19.3
	0.025	38.5	39.0	39.2	39.3	39.3	39.3
	0.01	98.5	99.0	99.2	99.2	99.3	99.3
3	0.10	5.54	5.46	5.39	5.34	5.31	5.28
	0.05	10.1	9.55	9.28	9.12	9.01	8.94
	0.025	17.4	16.0	15.4	15.1	14.9	14.7
	0.01	34.1	30.8	29.5	28.7	28.2	27.9

두 정규모집단에서의 표본분산 비의 분포

- $X_1, X_2, ..., X_{n_1}$ 과 $Y_1, Y_2, ..., Y_{n_2}$ 이 서로 독립이고 각각 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ 에서의 랜덤표본이라고 할 때, 다음이 성립한다.

$$\frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

t 분포와 F 분포의 관계

- 확률변수 T 가 $T \sim t(k)$ 일 때 , 다음이 성립한다

$$T^2 \sim F(1,k)$$

정규분포 분위수 대조도

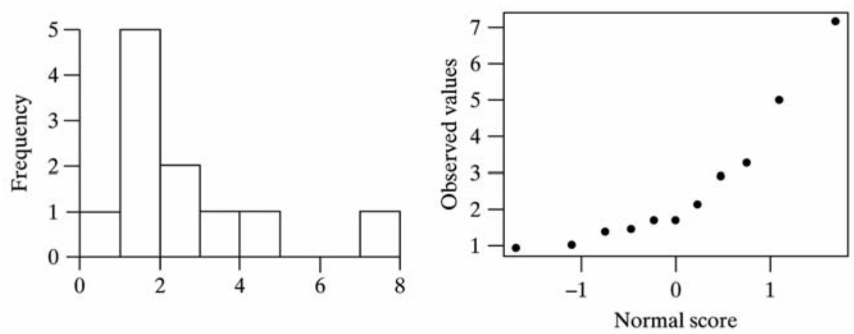
- 우리가 다루는 통계적 추론은 대부분 모집단의 분포가 정규분포라는 가정을 바탕으로 이루어지기 때문에, 이러한 정규모집단 가정의 검토가 매우 중요하다
- 정규모집단의 가정을 검토하는 방법으로 분위수 대조도를 이용하는 방법을 알아보자

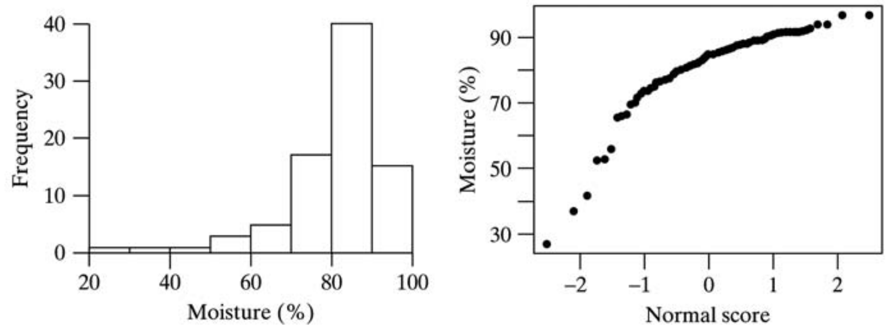
- **정규분포 분위수 대조도 (quantile-quantile plot : Q-Q plot)**
  : 정규분포의 이론적 분위수와 이에 대응하는 자료 분포의 실제 분위수를 좌표평면에 수평축과 수직축의 좌표로 대응하여 나타낸 것
  : 일반적으로 표준정규분포의 분위수와 자료분포의 분위수를 대조하여 나타냄
  : 점들이 직선 주위에 밀집하여 나타나면 모집단의 정규분포 가정을 만족하게 됨

  - 예 : $x_1, x_2, ..., x_{99}$ 이 $N(2, 3^2)$ 을 따르는 모집단에서 나온 표본인지 검토하는 경우

다양한 형태의 Q-Q plot

- skewed data의 Q-Q plot





- 꼬리가 긴 분포의 Q-Q plot