제8장 표본분포

Jinseog Kim
Dongguk University
jinseog.kim@gmail.com

2017-06-07

모집단과 표본

- 모집단(population): 관심의 대상으로서 조사 가능한 정보나 개체의 모든 집합
- 표본(sample): 모집단의 정보를 알아내기 위해 모집단을 대표할 수 있는 모집단의 부분집합
- 랜덤추출(random sampling): 모집단에 속한 각각의 원소가 뽑힐 가능성이 동일하도록 표본을 추출하는 방법
  - 복원추출(ramdom sampling with replacement, RSWR)
  - 비복원 추출(ramdom sampling without replacement, RSWOR)
- 랜덤표본(random sample): 랜덤추출법에 의해 뽑힌 표본
- 모수(parameter): **관심있는** 모집단의 수치적 특성치
  - 예: 모평균, 모분산, 모중앙값, 상위 5%지점의 백분위수 등
- 통계량(statistic): 표본 자료에서 얻어지는 수치적 특성치
  - 예: 표본평균, 표본분산 등
- 추정량(estimator): 관심있는 모수를 추정하기 위한 통계량
  - 예: 모평균(모수) $\Rightarrow$ 표본평균(추정량)
- 모집단분포(population distribution): 모집단 자료값들의 분포
- 표본분포(sampling distribution): 통계량의 분포
- 표준오차(standard error, 또는 s.e): 통계량의 표준편차

예제

- 우리 대학에서 흡연하는 학생들의 비율을 추정하기 위해서 200명을 조사하였다.
  1. 모집단:우리대학 학생 개개인의 흡연여부
  2. 모수:우리대학 전체 학생들의 흡연비율
  3. 표본: 200명의 학생들의 흡연여부
  4. 통계량 (또는 흡연비율에 대한 추정량): 200명 학생들의 흡연비율

정규 모집단 분포, 통계량, 표본분포

- 정규모집단: $X \sim N(\mu, \sigma^2)$
- 랜덤표본 : $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$
- 모수: 모평균 (μ)
- 통계량: 표본평균

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots, +X_n)$$

- 표본분포(통계량의 분포)

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

- 표본평균의 표준오차 (standard error 또는 s.e)

$$se(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

예제 7

- 평균이 25, 분산이 36인 정규모집단으로부터 크기 9인 임의표본을 추출하였다. 표본평균 $\bar{X}$가 28보다 클 확률을 구하시오.

$$\bar{X} \sim N(25, \frac{36}{9} = 4) \Rightarrow \frac{\bar{X} - 25}{2} \sim N(0, 1)$$

$$\begin{aligned} P(\bar{X} \geq 28) &= P(\frac{\bar{X} - 25}{2} \geq \frac{28 - 25}{2}) \\ &= P(Z \geq 1.5) = 1 - P(Z < 1.5) \\ &= 1 - 0.9332 = 0.0668 \end{aligned}$$

예제 8

- 어떤 작업을 완료하는 데 걸리는 시간은 평균이 30분 표준편차가 9분인 정규분포를 따른다. 25명의 작업자를 임의로 추출했을 때 평균작업 시간이 28분에서 33분 사이일 확률을 구하시오

$$\bar{X} \sim N(30, \frac{9^2}{25} = 4) \Rightarrow \frac{\bar{X} - 30}{9/5} \sim N(0, 1)$$

$$\begin{aligned} P(28 < \bar{X} < 33) &= P(\frac{28 - 30}{9/5} < \frac{\bar{X} - 30}{9/5} < \frac{33 - 30}{9/5}) \\ &= P(-1.11 < Z < 1.67) \\ &= P(X < 1.67) - P(Z < -1.11) = 0.819 \end{aligned}$$

이항(베르누이) 모집단 분포, 통계량, 표본분포

- 베르누이 모집단: $X \sim Bernoulli(p)$
- 랜덤표본 : $X_1, X_2, \ldots, X_n \sim Bernoulli(p)$
  - $E(X) = E(X_1) = p$, $V(X) = V(X_1) = p(1-p)$
- 모수: 모비율 (p)
- 통계량: 표본비율

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- 표본분포(통계량의 분포)

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

- 표본비율의 표준오차

$$se(\bar{X}) = \sqrt{\frac{1}{n^2} V(\sum X_i)} = \sqrt{\frac{1}{n^2} np(1-p)} = \sqrt{\frac{p(1-p)}{n}}$$

요약

모집단 분포	통계량	표본분포
$N(\mu, \sigma^2)$	$\bar{X}$	$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
$Ber(p)$	$\bar{X}_i$	$\sum_{i=1}^{n} X_i \sim Bin(n, p)$

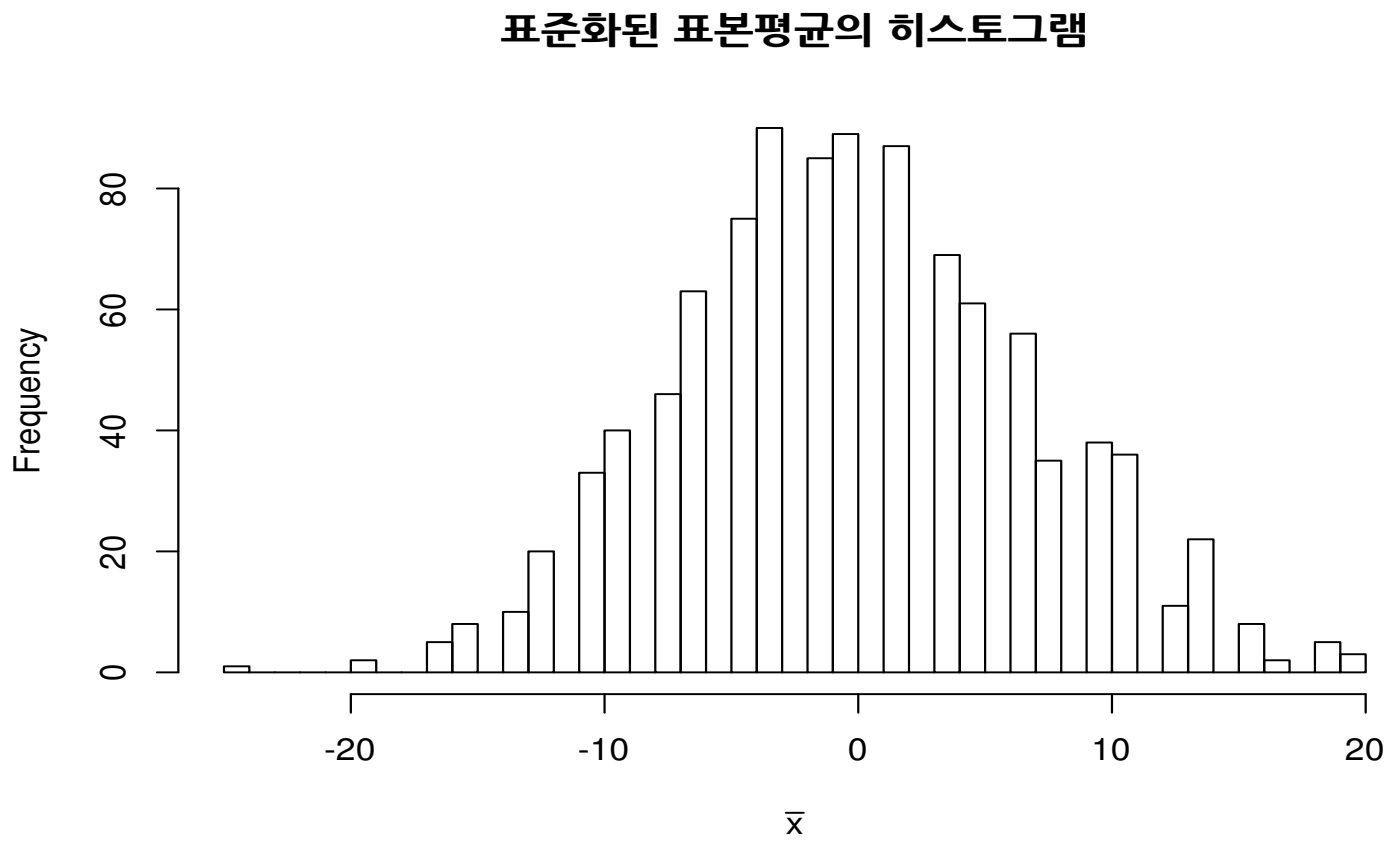
중심극한 정리 (Central Limit Theorem, CLT)

- 중심극한 정리 (Central Limit Theorem, CLT): 모평균이 μ이고, 모분산이 σ^2인 모집단(임의의 분포)으로부터 크기 n인 표본의 표본평균을 $\bar{X}$라고 할 때, 표본의 수가 충분히 커지면, 표준화된 표본평균의 확률분포는 표준정규분포에 수렴한다.
- 랜덤표본 : $X_1, X_2, \ldots, X_n$, $E(X_i) = \mu$, $V(X_i) = \sigma^2$

$$\frac{\bar{X} - (\bar{X}\text{의 평균})}{\bar{X}\text{의 표준편차}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1), n \to \infty.$$

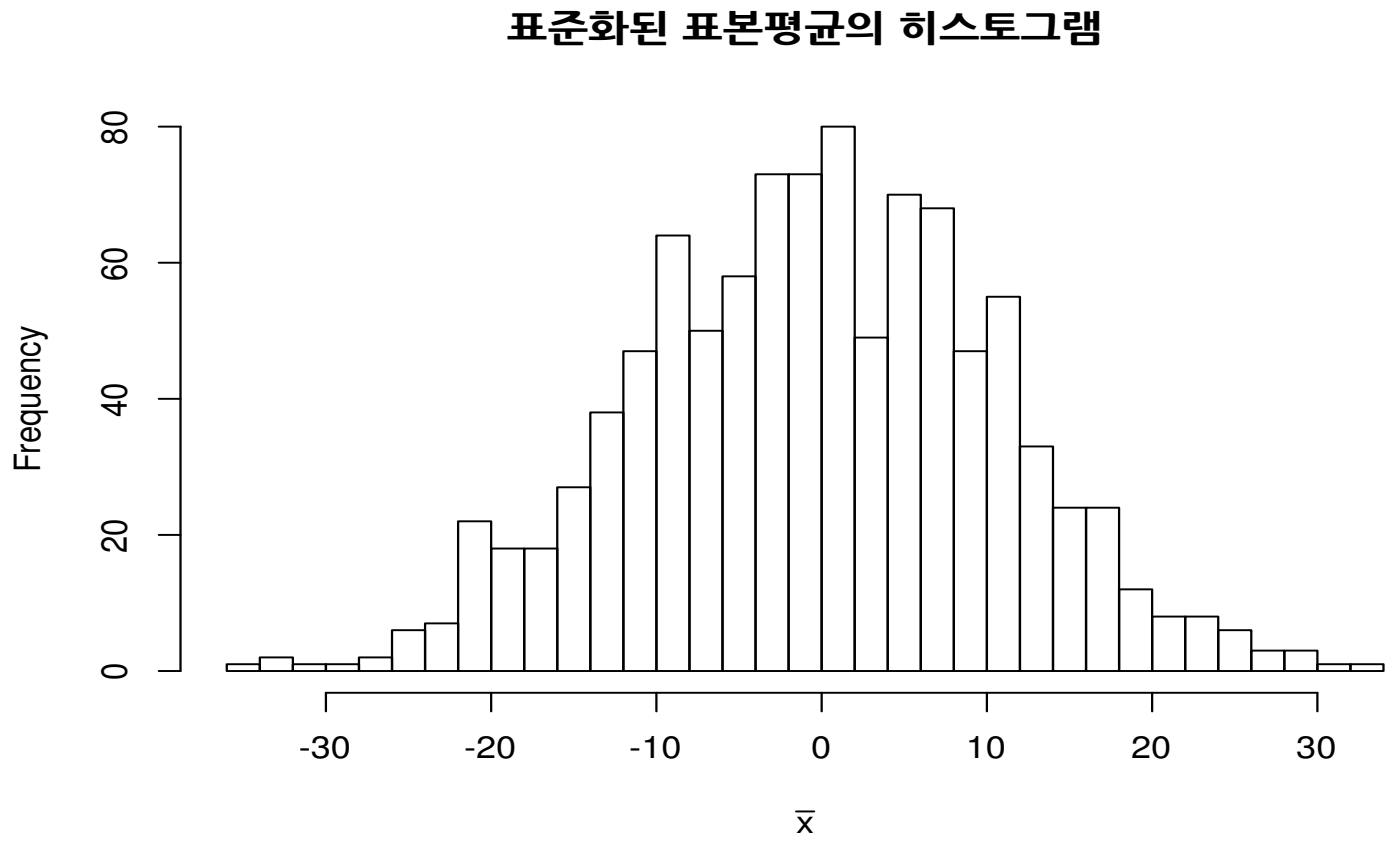
중심극한 정리 (Central Limit Theorem, CLT)

- 베르누이 분포의 랜덤표본: $X_i \sim Ber(p), i = 1, 2, \ldots, n.$
- $n = 100, p = 0.3$, 표준화된 표본평균을 1000번 계산



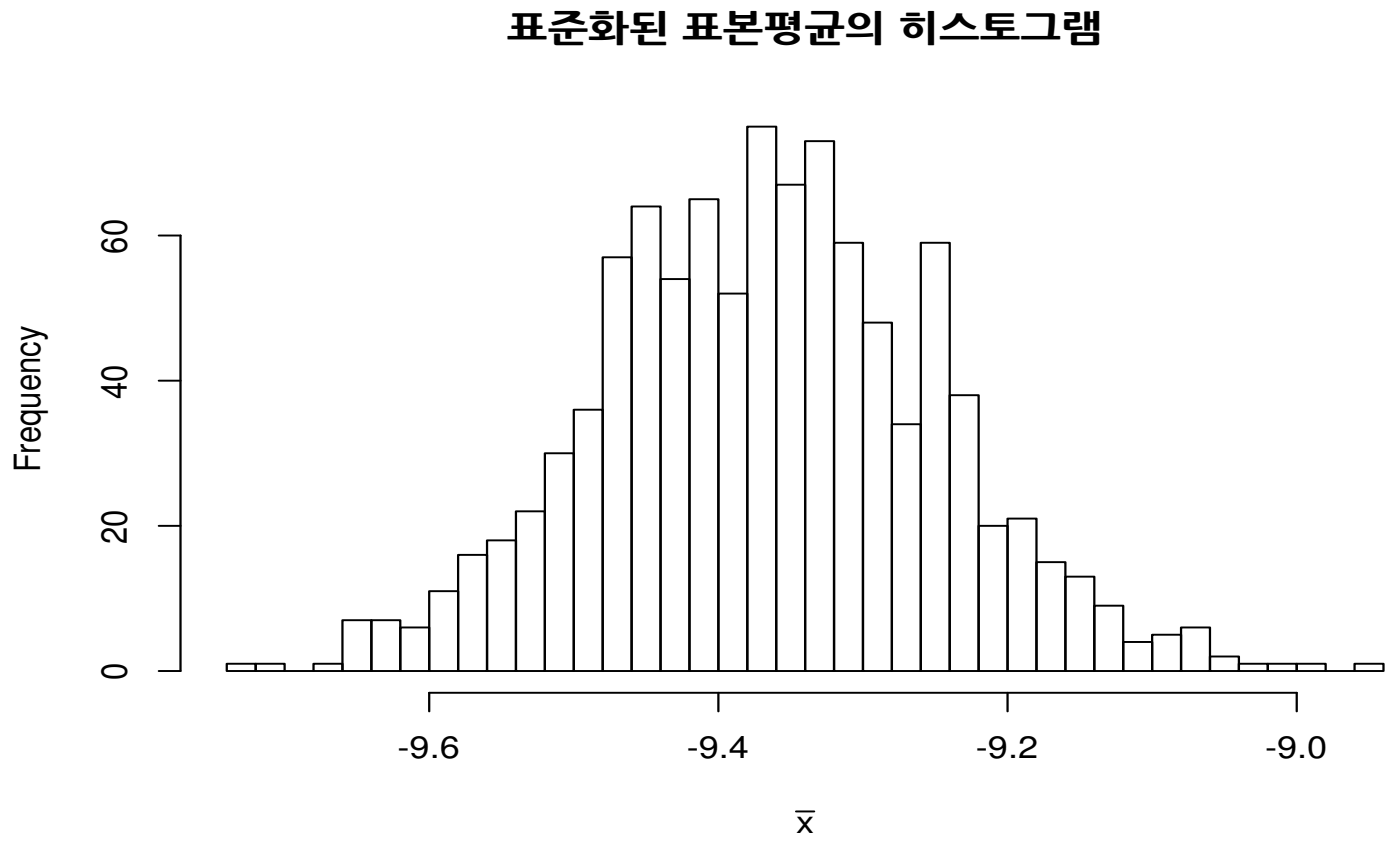
중심극한 정리 (Central Limit Theorem, CLT)

- 균등 분포의 랜덤표본: $X_i \sim U(0,1), i = 1, 2, \ldots, n.$
- $n = 100$, 표준화된 표본평균을 1000번 계산



중심극한 정리 (Central Limit Theorem, CLT)

- 지수 분포의 랜덤표본: $X_i \sim Exp(3), i = 1, 2, \ldots, n.$
- $n = 100$, 표준화된 표본평균을 1000번 계산



t분포

1. 정규모집단에서 랜덤표본의 표본평균($\bar{X}$)에 대한 분포는 정규분포
   - 표준화된 표본평균의 분포는 표준정규분포를 따름

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

2. 모표준편차 σ를 모르는 경우?? $\Rightarrow$ 모표준편차 σ를 랜덤표본으로 추정 즉,

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

이 때, 표준화된 통계량의 분포는?

$$T = \frac{\bar{X}-\mu}{s/\sqrt{n}} : \text{표준정규분포를 따르지 않음}$$

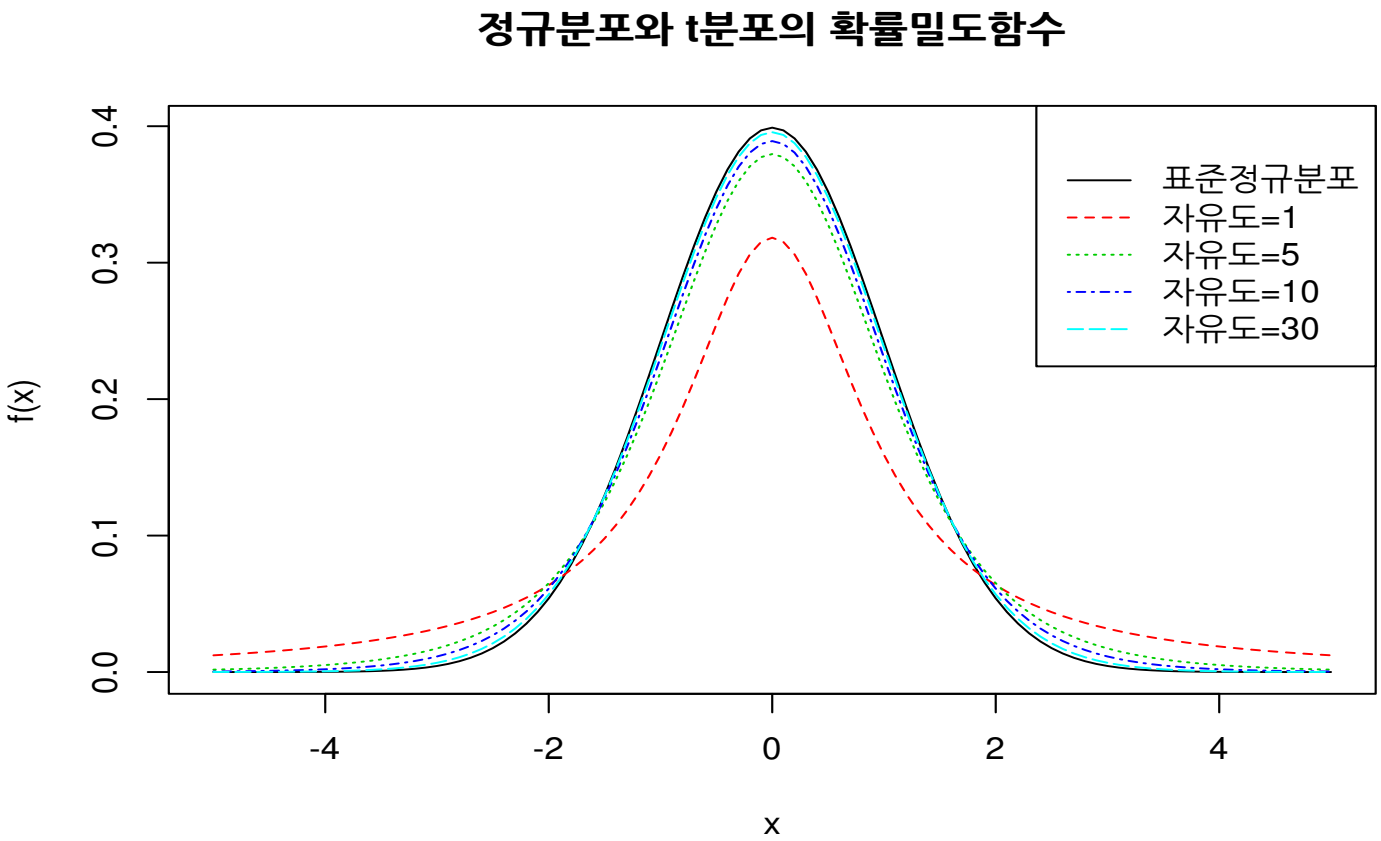
t분포

1. 1908년 아일랜드 양조회사 기사, W.S. Gosset(1876~1937) 고안 ⇒ Student t 분포 (필명)

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1), \text{ 여기서 } n-1 \text{ 을 자유도(degree of freedom)}$$

- t분포의 특징
  1. 평균은 0
  2. 평균 0에 대해 좌우 대칭
  3. 표준정규분포에 비해 변동이 큼(분산이 큼)
  4. 자유도에 의해 분포형태가 결정됨
  5. 자료의 수(표본의 크기) n이 커지면 (즉, 자유도가 커지면) 표준정규분포로 근사 ($n \geq 30$)

자유도에 따른 t분포의 모양



χ^2분포 (카이제곱 분포)

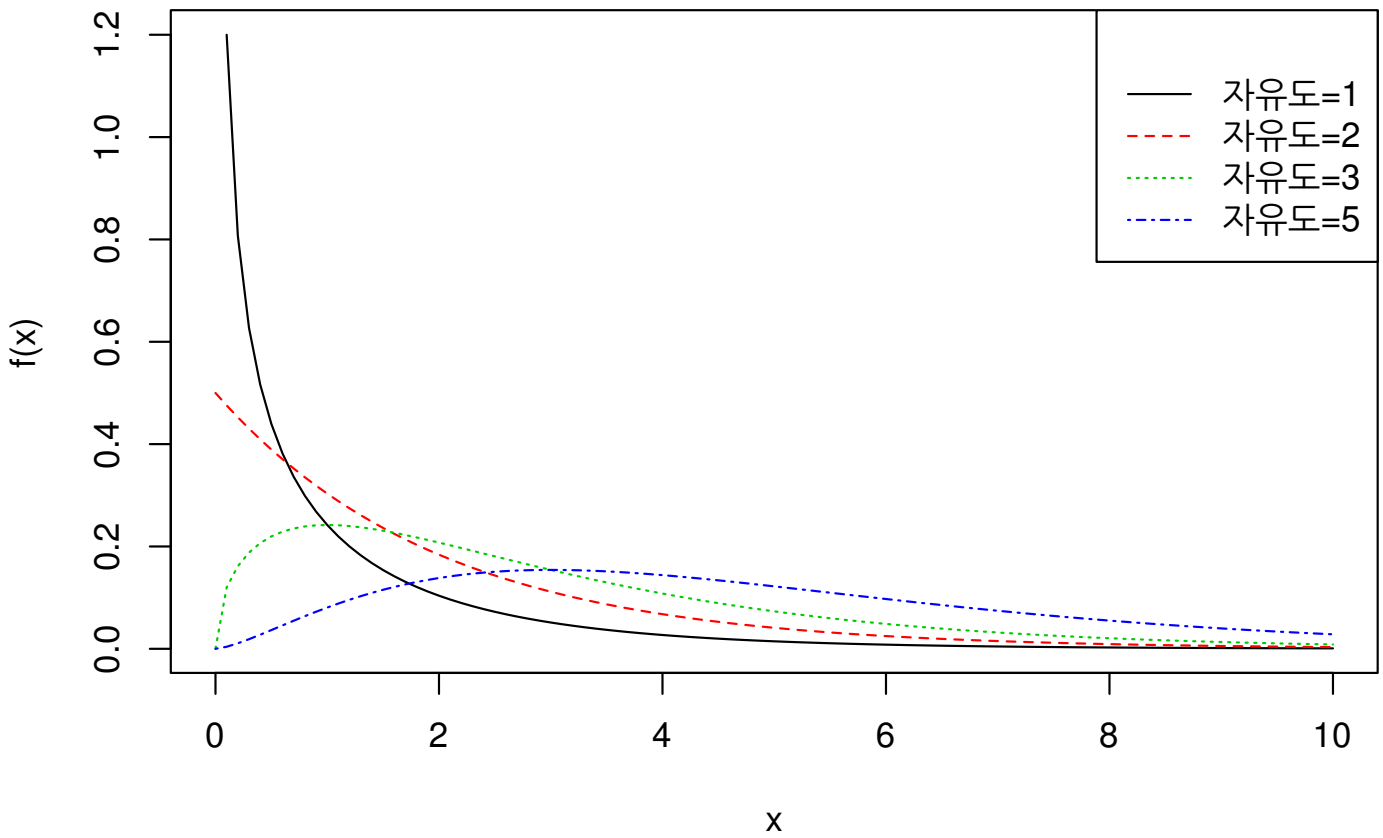
- k개의 독립적이고 표준정규분포($N(0,1)$)를 따르는 확률변수 $X_1, \ldots, X_k$가 있을 때,

$$X^2 = \sum_{i=1}^{k} X_i^2$$

  - X^2의 분포를 자유도가 k인 카이제곱 분포라고 하고
  - $\chi^2(k)$로 표현함
- 표본분산의 분포와 관련됨

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

자유도에 따른 χ^2 분포의 모양



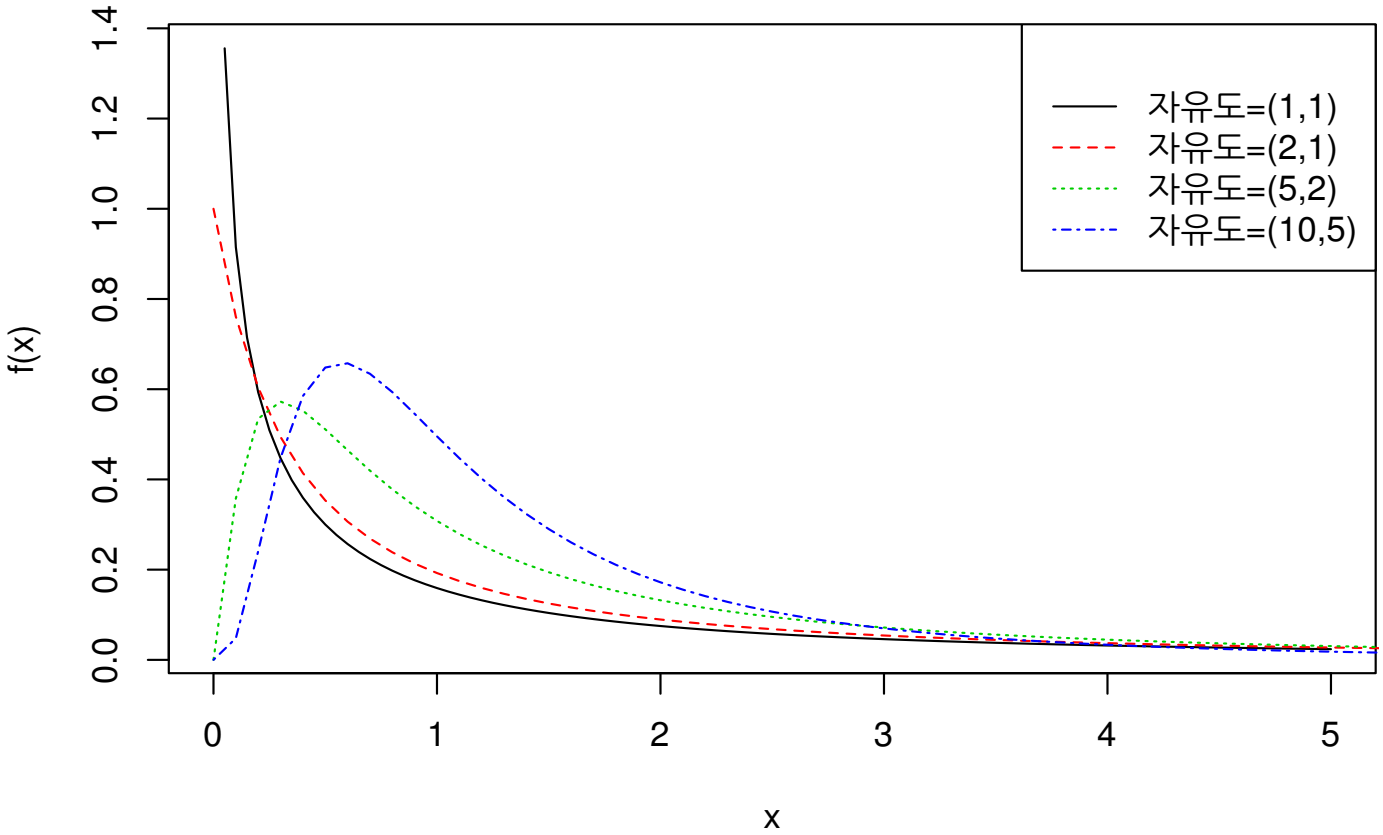
F 분포

- F 분포(F-distribution)
- 서로 독립인 확률변수 V_1, V_2 각각 자유도가 k_1, k_2이고 서로 독립인 카이제곱 분포를 따를 때,
- 아래의 확률변수 F를 자유도가 k_1, k_2인 F-분포

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

- 응용
  - 두 모집단간의 분산의 차이 검정
  - 분산분석 (여러 모집단의 분산차이 검정)

자유도에 따른 F 분포의 모양



연습문제 8

- D 대학 학생의 몸무게 분포가 $N(45, 4^2)$따른다고 할 때 49명의 표본을 추출하여 몸무게를 측정할 경우 표본의 평균 몸무게가 46kg 이상일 확률을 구하시오.

연습문제 16

- 16 어떤 도시에서 가구의 60%가 저녁 뉴스시간에 M방송을 시청한다고 하자. 만약 그 도시에서 150 가구를 랜덤하게, 즉 임의로 추출한다면, 그때 그 시간에 M 방송을 시청하는 표본비율이 0.52보다 클 확률은 얼마인가?