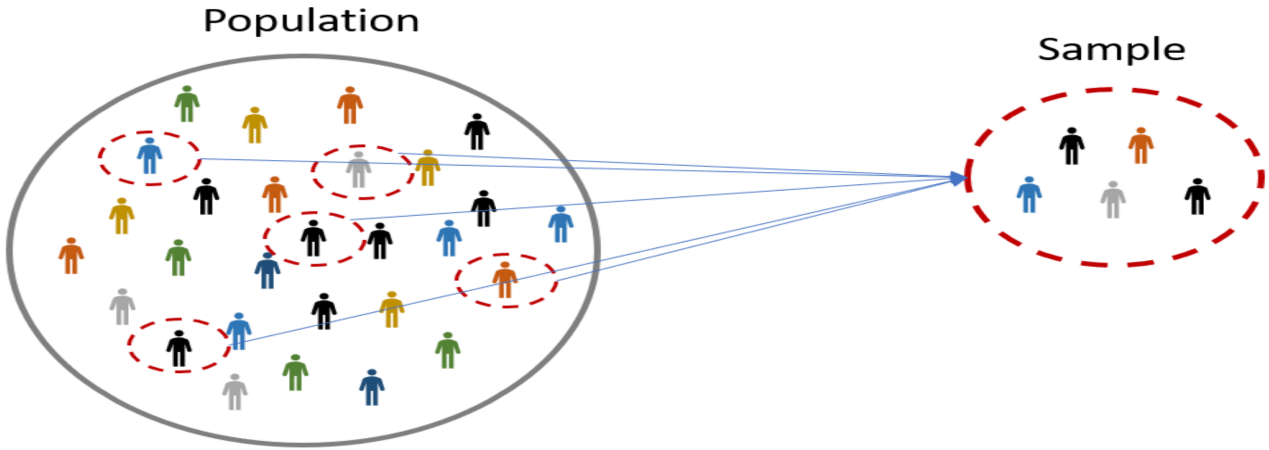
4장. 표본분포

박명현

서울대학교 학부대학

표본분포(Sampling distribution)



- 표본의 추출과정을 알아야 모집단 추론 가능

  - 시간과 공간의 제약, 사용 가능한 비용, 조사원 등 경제적 제약

- 모집단을 추론하기 위해서 표본평균이나 표본분산 등의 값과 오차를 계산하려면 표본통계량의 분포가 필요(표본분포)

모집단과 표본

모집단(population)

- 모수(parameter): 모집단의 특징을 나타내는 대표값

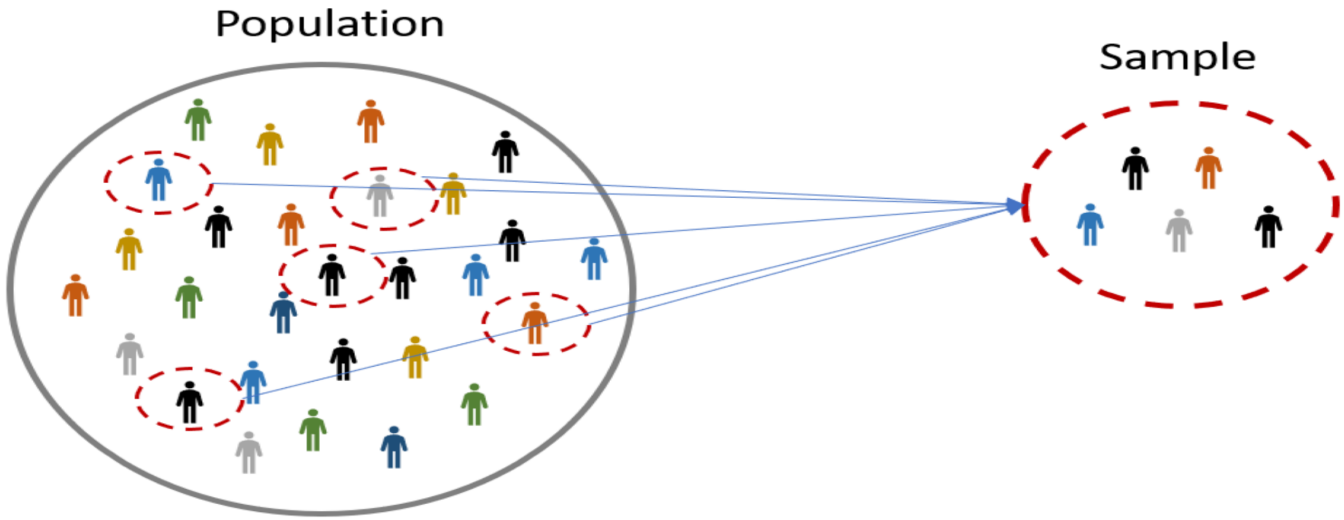
  예) 모평균 μ, 모비율 p, 모분산 σ^2

표본(sample)

- 통계량(statistic): 표본의 관측값들로부터 계산되는 수치함수
- 추정량(estimator): 모수의 추정을 위해 사용되는 통계량
  - 주로 대문자로 표시. 예) 표본평균 $\bar{X}$, 표본분산 S^2
  - 'hat'을 붙여 표기. 예) $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = S^2$, $\hat{p}$
- 추정치(estimate): 관측치로부터 계산된 모수의 예측값(관찰값).
  - 소문자로 표시. 예) $\bar{x}$, s^2, $\hat{p}$

예) 모평균 μ 추정시 표본 $X_1, \cdots, X_n$의 표본평균 $\bar{X}$는 추정량이고, 실제 자료 $x_1, \cdots, x_n$의 표본평균 $\bar{x}$는 추정값이 된다.

모집단과 표본



- 추정량을 하나의 [ ]로 취급
  - 어떤 표본이 추출되느냐에 따라서 하나의 통계량이 여러 가지 다른 값으로 실현될 수 있다.
  - 하지만 하나의 표본이 추출되고 나면 이제 통계량은 하나의 값으로 실현된다.(추정치)

단순랜덤추출법(simple random sampling)

- 단순랜덤복원추출(sampling with replacement):
  - 한 번 선택한 추출단위는 다시 모집단에 되돌려놓고 다음 단위를 추출
  - 크기가 N인 유한모집단에서 크기 n인 표본을 복원추출하는 경우에는 매번 동일하게 $1/N$의 확률로 1개씩 선택
  - 무한모집단의 랜덤표본을 정의하기 위한 개념적 도구로 이용됨.
- 단순랜덤비복원추출(sampling without replacement):
  - 선택한 추출단위를 다시 모집단에 되돌려 놓지 않고 다음 단위를 뽑는 방식
  - 각 항목은 $1/N, 1/(N-1), \cdots, 1/(N-n+1)$의 확률로 1개씩 선택
  - 일반적으로, 모집단의 크기가 표본의 크기에 비하여 충분히 클 때 유한모집단에서 표본을 단순랜덤비복원추출하는 것은 단순랜덤복원추출하는 것과 별 차이가 없다.

단순랜덤복원추출

Example

N개의 제비 중 M개의 당첨제비(Success)가 있을 때 2개의 제비를 단순랜덤복원추출하는 경우를 생각해 보자.

- X_i: $i$th번째 시행의 결과를 나타내는 베르누이확률변수, $i = 1, 2$

$$X_i = \begin{cases} 1, & if\ Success \\ 0, & if\ Fail \end{cases}$$

- '독립적':

$$P(X_2 = 1) = P(X_2 = 1|X_1 = 1) = P(X_2 = 1|X_1 = 0) = M/N$$

$$P(X_2 = 0) = P(X_2 = 0|X_1 = 1) = P(X_2 = 0|X_1 = 0) = (N - M)/N$$

- '동일한 분포': $P(X_1 = 1) = P(X_2 = 1) = M/N$

=> 두 확률변수 X_1, X_2는 서로 독립이며 동일한 분포를 가진다.

단순랜덤비복원추출

Example

N개의 제비 중 M개의 당첨제비(Success)가 있을 때 2개의 제비를 단순랜덤비복원추출하는 경우를 생각해 보자.

- '독립적':

  $P(X_1 = 1) = M/N$, $P(X_2 = 1|X_1 = 1) = (M-1)/(N-1)$,
  $P(X_2 = 1|X_1 = 0) = M/(N-1)$ 이므로

$$\begin{aligned} P(X_2 = 1) &= P(X_2 = 1 \cap X_1 = 1) + P(X_2 = 1 \cap X_1 = 0) \\ &= P(X_2 = 1|X_1 = 1)P(X_1 = 1) + \\ &\quad P(X_2 = 1|X_1 = 0)P(X_1 = 0) \\ &= \frac{M-1}{N-1}\frac{M}{N} + \frac{M}{N-1}\frac{N-M}{N} = \frac{M}{N}. \end{aligned}$$

  $P(X_2 = 1) \neq P(X_2 = 1|X_1 = 1)$이므로 두 시행은 독립적이지 않음

- '동일한 분포':

  $P(X_1 = 1) = P(X_2 = 1) = M/N$

  => 두 확률변수 X_1, X_2는 동일한 분포를 가지지만 서로 독립은 아니다.

- if $N \gg 2$ 이면,

  (모집단의 크기가 표본의 크기에 비하여 충분히 클 때)

  $P(X_2 = 1) \simeq P(X_2 = 1|X_1 = 1)$이 되어, X_1, X_2는 독립에 가깝게 됨.

- 즉, 유한모집단에서의 랜덤표본 $\{X_1, X_2, \cdots, X_n\}$의 성질은 단순랜덤복원추출에 의한 표본의 성질과 유사함을 알 수 있다.

랜덤표본(random sample)

- 랜덤표본

  - X_i: i번째로 뽑힌 추출단위의 특성값을 나타내는 확률변수

    $$i = 1, \cdots, n$$

  (1) 유한모집단인 경우 단순랜덤비복원추출로 뽑은 표본을 랜덤표본이라고 한다. 이 때, 크기 n인 랜덤표본을 $\{X_1, X_2, \cdots, X_n\}$로 나타낸다.

  (2) 무한모집단인 경우 서로 독립이며, 모집단 분포와 동일한 분포를 갖는 확률변수들의 집합 $\{X_1, X_2, \cdots, X_n\}$를 랜덤표본이라 한다. 즉, 모집단의 분포를 $f(x)$라 할 때 다음과 같이 표기한다.

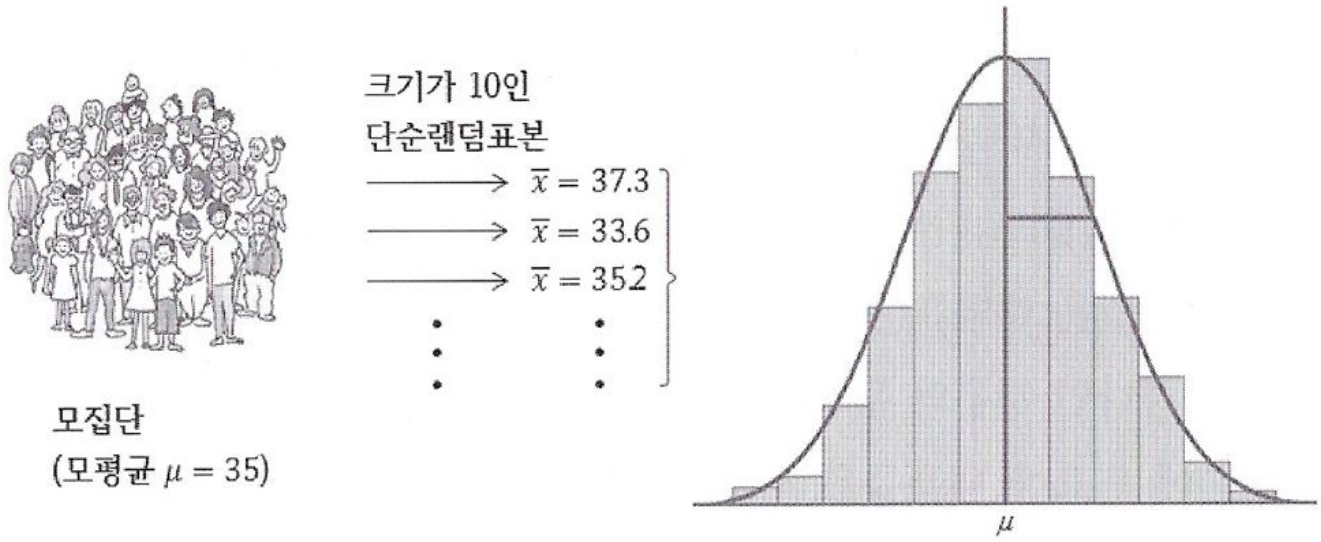
  $$X_i \sim_{iid} f(x), \ (i.i.d : \textit{independent and identically distributed})$$

- 5장 이후에서는 무한모집단에서의 랜덤표본을 가정

표본분포(sampling distribution)

- 표본분포: 통계량의 확률분포

Figure: 모집단으로 구한 표본평균 $\bar{X}$의 표본분포



- 평균이 $\mu = 35$인 모집단에서 $n = 10$인 단순랜덤표본을 추출
- 각 표본에서 사람들 나이의 평균 $\bar{X}$을 계산
- 표본에 따라 그 평균값이 다르게 나타남
- 계산된 표본평균들로 히스토그램을 그려보면 $\bar{X}$은 μ를 중심으로 분포하고 있음을 알 수 있음.

표본분포

Example

모집단 $\{2, 2, 4, 5\}$에서 크기 2인 표본을 단순랜덤비복원 추출하였을 때, 표본평균의 표본분포를 구하여라.

Table: 가능한 표본과 표본평균의 값

가능한 표본	$\{2,2\}$	$\{2,4\}$	$\{2,5\}$	$\{2,4\}$	$\{2,5\}$	$\{4,5\}$
표본평균	2	3	3.5	3	3.5	4.5
표본이 뽑힐 확률	1/6	1/6	1/6	1/6	1/6	1/6

Table: 표본평균 $\bar{X}$의 표본분포

표본 평균의 값	2	3	3.5	4.5	합
확률	1/6	1/3	1/3	1/6	1

- 모수: $\mu = 3.25$, 추정량: $\hat{\mu} = \bar{X}$, 추정치: $\bar{x} = 2, 3, 3.5, 4.5$

표본분포

Example

{찬성,반대,찬성,찬성,반대}와 같은 모집단에서 크기 3인 표본을 단순랜덤비복원추출로 뽑아 찬성의 모비율 p(이 경우, $p = 0.6$)를 추측하는 경우 표본비율의 표본분포를 구하여라.

- 가능한 표본: {찬성,찬성,찬성},{찬성,찬성,반대},{찬성,반대,반대}

Table: 표본비율 $\hat{p}$의 분포표

표본비율($\hat{p}$)	1	2/3	1/3
확률($p(\hat{p})$)	1/10	6/10	3/10

- 표본비율 $\hat{p}$은 표본추출의 결과에 따라 하나의 수값을 대응시키는 확률변수임과 동시에 통계량이며
- 위의 표에서 보이는 것처럼 표본비율 $\hat{p}$의 확률분포를 표본분포라 한다.

표본평균의 분포(무한모집단)

- 일반적인 모집단(주로 무한모집단)에서 표본평균의 표본분포에 대한 특징에 대해서 알아보자.

Theorem

모평균이 μ, 모분산이 σ^2인 무한모집단에서 크기 n인 랜덤표본 $X_1, \cdots, X_n$을 추출하였을 때, 랜덤표본의 표본평균 $\bar{X}$에 대하여 다음이 성립한다.

$$(1)\ E(\bar{X}) = \mu$$

$$(2)\ Var(\bar{X}) = \frac{\sigma^2}{n}, \quad sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

표본평균의 분포(무한모집단)

pf) (1) 랜덤표본의 정의로부터 $E(X_i) \equiv \mu$이므로,

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n} E(X_1 + \cdots + X_n) \\ &= \frac{1}{n} \{E(X_1) + \cdots + E(X_n)\} \\ &= \mu \end{aligned}$$

pf) (2) $X_1, \cdots, X_n$이 서로 독립이고 $Var(X_i) \equiv \sigma^2$이므로,

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{n^2} Var(X_1 + \cdots + X_n) \\ &= \frac{1}{n^2} \{Var(X_1) + \cdots + Var(X_n)\} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

표본평균의 분포(무한모집단)

- 추출한 표본에 따라 표본평균의 값이 달라지므로 표본평균의 분포가 생긴다.
- 표본평균 $\bar{X}$의 평균은 모집단의 평균과 같으며($E(\bar{X}) = \mu$) 표본의 크기 n이 클수록 그 분산이 0에 가까워져서

$$Var(\bar{X}) = \sigma^2/n$$

결국 $\bar{X}$는 모집단의 평균인 μ근처에 밀집되어 분포한다.

표본평균의 분포(정규분포 모집단)

- 모집단에 대한 통계적 추론시 모분포에 대한 정규성 가정이 많음

Theorem

모집단의 분포가 정규분포 $N(\mu, \sigma^2)$일 때 표본평균 $\bar{X}$는 정규분포 $N(\mu, \frac{\sigma^2}{n})$을 따른다.

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

- 정규모집단에서 단순임의추출한 표본평균의 분포는 표본의 개수와 상관없이 항상 정규분포를 따른다.
- 표준화:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \iff \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

표본평균의 분포(임의의 모집단)

Theorem (중심극한정리(central limit theorem))

평균이 μ, 분산이 σ^2인 임의의 모집단에서 표본의 크기 n이 충분히 크면, 랜덤표본의 표본평균 $\bar{X}$는 근사적으로 정규분포 $N(\mu, \frac{\sigma^2}{n})$을 따른다. 즉, n이 충분히 클 때 다음이 성립한다.

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

- 표본의 크기 n이 충분히 클 때에는 모집단의 분포가 어떤 형태이든 간에, 표본평균 $\bar{X}$의 분포는 근사적으로 정규분포를 따른다.
- 경험적으로 $n > 30$이면 근사정도가 만족할 만하다고 알려져 있다.

중심극한정리 예시

Example

모집단 분포가 $0, 1, \cdots, 9$의 각 정수를 택할 확률이 동일하게 0.1인 이산확률분포라 하자. 이 모집단에서 크기 5인 랜덤표본을 100개 추출한 결과와 각 표본평균 $\bar{X}$의 값이 아래표에 주어져 있다.

표 4–5 [예 4–4]의 모집단 분포에서 추출한 크기 5의 랜덤표본

표본 번호	관측값	합계	평균	표본 번호	관측값	합계	평균
1	4, 7, 9, 0, 6	26	5.2	51	4, 7, 3, 8, 8	30	6.0
2	7, 3, 7, 7, 4	28	5.6	52	2, 0, 3, 3, 2	10	2.0
3	0, 4, 6, 9, 2	21	4.2	53	4, 4, 2, 6, 3	19	3.8
4	7, 6, 1, 9, 1	24	4.8	54	1, 6, 4, 0, 6	17	3.4
5	9, 0, 2, 9, 4	24	4.8	55	2, 4, 5, 8, 9	28	5.6
6	9, 4, 9, 4, 2	28	5.6	56	1, 5, 5, 4, 0	15	3.0
7	7, 4, 2, 1, 6	20	4.0	57	3, 7, 5, 4, 3	22	4.4
8	4, 4, 7, 7, 9	31	6.2	58	3, 7, 0, 7, 6	23	4.6
9	8, 7, 6, 0, 5	26	5.2	59	4, 8, 9, 5, 9	35	7.0

- 이러한 100개의 $\bar{X}$값들에 대한 히스토그램을 그려보면 다음과 같다.

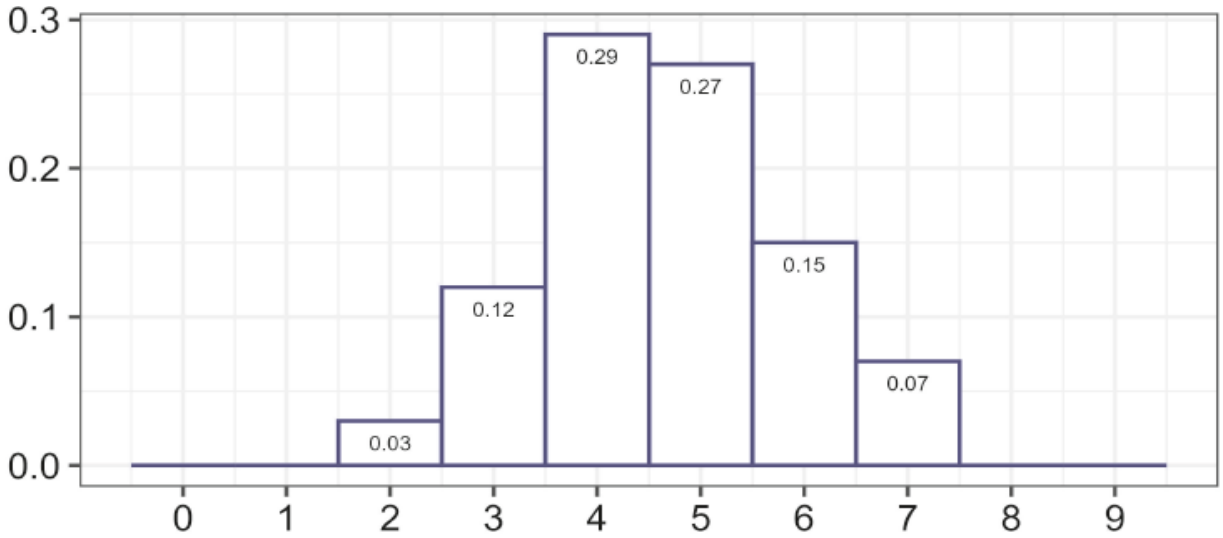
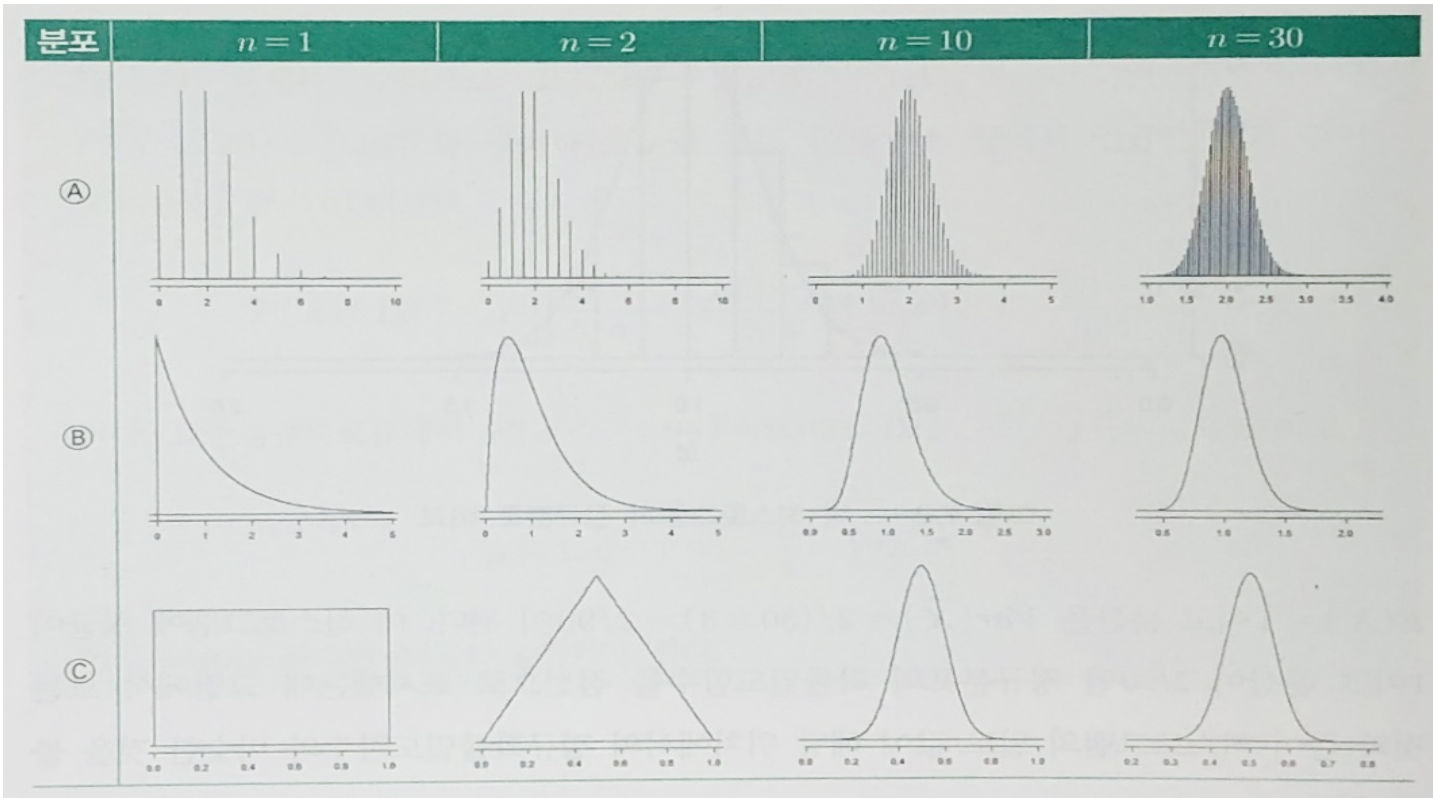


그림 4–5 [표 4–5]에 주어진 $\overline{X}$ 값들의 히스토그램

- 모집단의 형태가 정규분포가 아니더라도 표본의 크기가 5인 경우에 조차 히스토그램의 형태가 정규분포의 형태에 가까움을 알 수 있다.

중심극한정리(central limit theorem)

- 표본의 크기 n에 따른 표본평균 $\bar{X}$의 분포
  - 다양한 형태의 모집단에서 확률표본을 뽑았을 때
  - 표본크기가 작은 경우에는 모집단의 형태에 영향을 받으나
  - n이 커지면서 표본평균의 분포가 정규분포와 비슷해짐



표본평균의 분포 - 정리

- 표본평균 $\bar{X}$의 분포 (단순임의추출에 의한)

1. 모집단이 $N(\mu, \sigma^2)$ 일 때:

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

=> 정확한 분포. 정규분포의 성질임(수리적으로 정확한 결과)

2. 모평균이 μ, 모분산이 σ^2, 표본크기 $n \geq 30$일 때:

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

=> 중심극한정리에 의한 근사분포.

표본평균의 분포(유한모집단)

- 모평균이 μ, 모분산이 σ^2, 크기가 N인 유한모집단에서 비복원으로 크기가 n인 랜덤표본 $X_1, \cdots, X_n$을 추출한 경우에는 다음이 성립한다.

$$E(\bar{X}) = \mu, \ Var(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

- n 고정시키고 N을 무한대로 늘리면

$$Var(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \xrightarrow{N->\infty} \frac{\sigma^2}{n}$$

표본평균의 분포(유한모집단)

- N을 고정시키고 표본크기 n을 늘리면, 결국 전수조사가 된다. 이 때 분산은 점점 줄어들어 0으로 수렴하고 표준오차 또한 감소한다. 즉, 표본추출의 통계적 의미가 없어진다.

$$Var(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \xrightarrow{n->N} 0$$

- 유한모집단인 경우에도 N과 n이 충분히 크면 다음이 성립한다. (중심극한정리 적용)

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

중심극한정리 예시

Example

한 식품 회사에서 새로 개발한 음료의 평균 칼로리가 168kcal이고 표준편차가 6kcal로 알려져 있다. 음료 100개를 단순랜덤추출했을 때, 표본평균이 167kcal 이상 169kcal 이하일 확률을 구하라.

sol) $n = 100(> 30)$이므로 $\bar{X}$는 근사적으로 정규분포를 따르고 평균은 $\mu = 168$, 표준편차는 $\sigma/\sqrt{n} = 6/\sqrt{100} = 0.6$이다.

$$\bar{X} \dot{\sim} N(168, 0.6^2)$$

- $Z \sim N(0,1)$로 표준화하면

$$\begin{aligned} P(167 \leq \bar{X} \leq 169) &= P\left(\frac{167-168}{0.6} \leq \frac{\bar{X}-168}{0.6} \leq \frac{169-168}{0.6}\right) \\ &\doteq P(-1.67 \leq Z \leq 1.67) \\ &= 0.9525 - 0.0475 = 0.9050 \end{aligned}$$

t분포

Definition (t분포:t distribution)

표준정규분포 $N(0,1)$을 따르는 확률변수를 Z라 하고 이와는 독립이며 자유도 k인 카이제곱분포 χ^2를 따르는 확률변수를 V라 할 때,

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

의 분포를 자유도 k인 t분포라고 한다. 이 때 기호로는 다음과 같이 표기한다.

$$T \sim t(k)$$

- 자유도는 분포가 퍼져있는 정도를 나타내는 것으로 자유도가 클수록 가운데로 모이고 표준정규분포로 수렴한다.

t분포

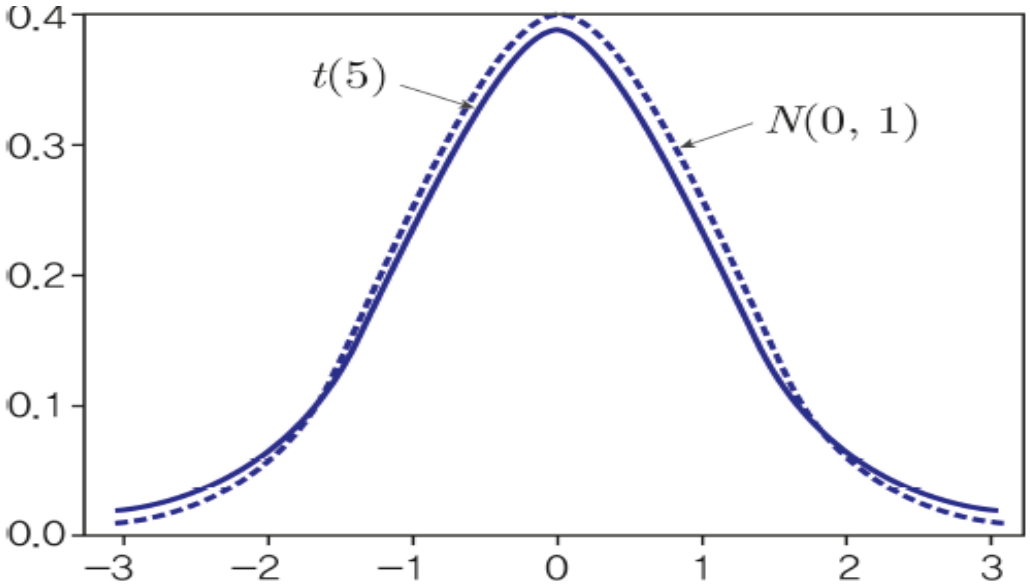


그림 4–8 표준정규분포 $N(0,\ 1)$과 자유도 5 인 t 분포의 비교

- t분포는 표준정규분포와 유사한 모양.
- 0을 중심으로 좌우 대칭형.
- 표준정규분포에 비하여 두터운 꼬리를 갖고 있음.

t분포의 분위수

- $T \sim t(k)$일 때 다음의 식을 만족하는 $t_\alpha(k)$를 자유도가 k인 t분포의 $(1-\alpha)$분위수라고 한다.

$$P\{T \geq t_\alpha(k)\} = \alpha$$

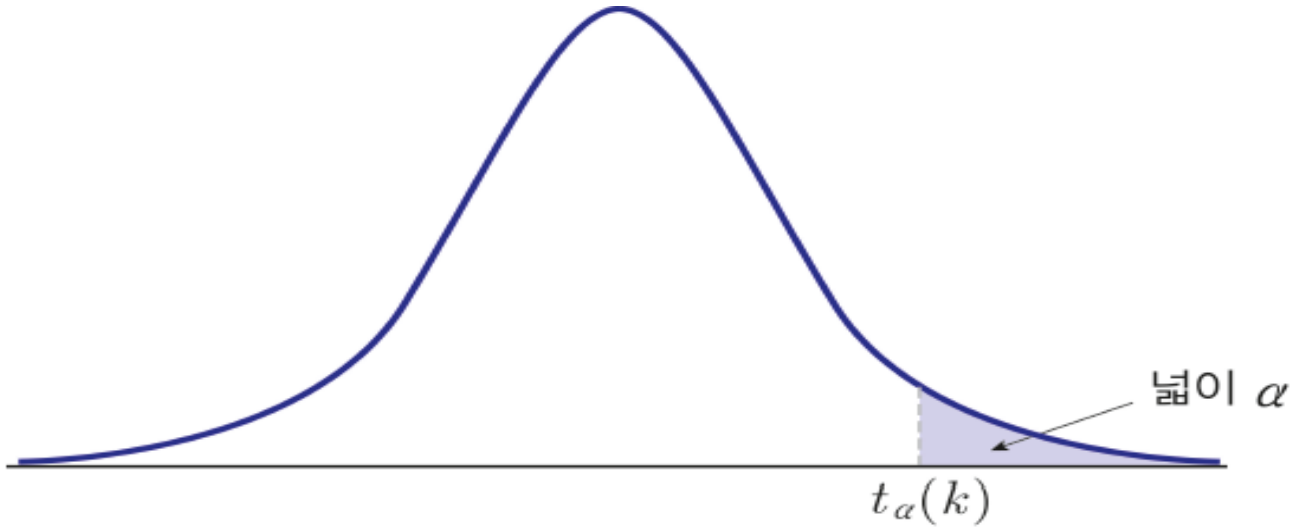
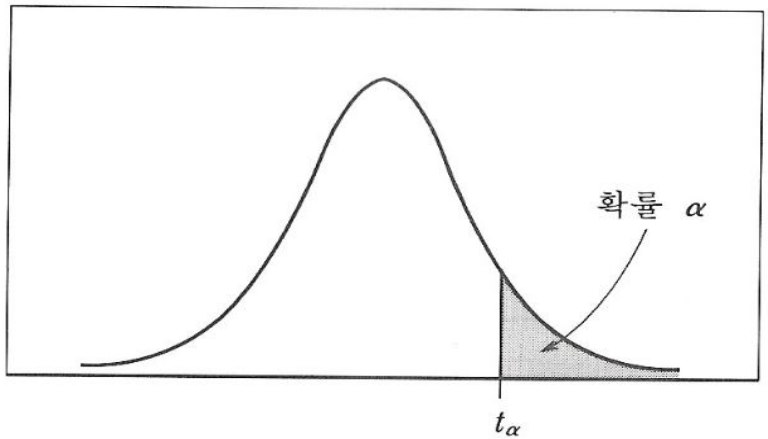


그림 4–9 $t_\alpha(k)$의 위치

t분포표

- t분포표: 자유도별로 각 수준에 대한 여러가지 분위수가 나와있다.
  예) $t_{0.1}(4) = 1.5333$, $t_{0.05}(4) = 2.132$, $t_{0.01}(4) = 3.747$

Figure: t분포표



표의 숫자는 주어진 α 값에 대한 t_α의 값을 나타낸다.

	α											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041

스튜던트화된 표본평균의 분포

- $X_1, \cdots, X_n \sim_{i.i.d} N(\mu, \sigma^2)$일 때, 다음이 성립함을 다룬 바 있다.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad or \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

그런데 σ가 미지인 경우에는, σ대신에 표본표준편차 S를 대입하여 스튜던트화(studentization)된 표본평균의 분포를 이용할 수 있다.

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

Theorem

$X_1, \cdots, X_n$이 정규모집단 $N(\mu, \sigma^2)$으로부터의 랜덤표본일 때, 표본평균 $\bar{X}$에 대하여 다음이 성립한다.

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

pf) $\frac{\bar{X}-\mu}{S/\sqrt{n}}$의 분포를 유도해 보자.

- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$이며 $\bar{X}$와 S^2는 서로 독립임이 알려져 있으므로
- t분포의 정의에 따라 $\frac{\bar{X}-\mu}{S/\sqrt{n}}$는 다음과 같이 자유도 $(n-1)$인 t분포를 따르게 된다.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/\left(\frac{\sigma}{\sqrt{n}}\right)}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim t(n-1)$$

표본비율의 분포

- 비율의 표본분포
  - 정부 정책에 대한 찬성률
  - 벤처기업의 파산 비율
  - 생산제품의 불량률

- 비율은 평균의 특수한 경우임

  예) 다섯 번 시도 중 두 번 성공, 세 번 실패면 성공 비율은 0.4

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

표본비율의 분포

- $X_1, \cdots, X_n$이 성공확률이 p인 베르누이 분포 $B(1,p)$를 따르는 무한모집단으로부터의 랜덤표본일때 다음과 같이 정의된다.

$$X_i = \begin{cases} 1, & (success) \\ 0, & (fail) \end{cases} \quad \sim_{iid} B(1,p), for\ i = 1, \cdots, n$$

- n개 표본에서 "성공"의 갯수 X는 베르누이 시행 X_i들의 합으로 표현되며 성공확률을 p로하는 이항분포를 따르게 된다.

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

표본비율의 분포

- 이때 표본비율(sample proportion)을 다음과 같이 정의하자.

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

- 표본비율 $\hat{p}$의 분포는 베르누이시행의 평균분포 또는 이항분포 확률변수를 n으로 나눈 확률변수의 분포가 되며 평균과 분산은 다음과 같다.

$$E(\hat{p}) = \frac{E(X)}{n} = \frac{np}{n} = p$$

$$Var(\hat{p}) = \frac{Var(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

표본비율의 분포 예시

Example

남자 6명, 여자 4명인 모집단에서 $n = 2$인 표본을 뽑아 모집단에서의 남자의 비율을 추정하려 한다.

sol) 모집단 남성비율은 0.6, 모분산은 0.6(1-0.6)=0.24

- 추출 가능한 모든 표본의 경우의 수: $\binom{10}{2} = 45$

표본구성	(남,남)	(남,여)	(여,여)
경우의 수	15회	24회	6회
확률	0.3333	0.5333	0.1333
남성의 비율($\hat{p}$)	1	0.5	0

- $E(\hat{p}) = 1 \times 0.3333 + 0.5 \times 0.5333 + 0 \times 0.1333 = 0.6$
- 따라서 $E(\hat{p}) = p$이다. 즉, 표본비율 $\hat{p}$의 기대값은 모집단의 모비율 즉, 모수 p와 같다.

표본비율의 분포

- 표본비율 $\hat{p}$의 근사분포:

  중심극한정리로부터 n이 클 때 다음과 같이 근사적으로 정규분포를 따르게 된다. (이항분포의 정규근사)

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

$$or$$

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1) \tag{1}$$

이항분포의 정규근사

- 이항분포에서 n이 크면 확률의 계산이 힘듬
- 이산형(discrete)인 이항분포의 여러 확률을 중심극한정리를 적용시켜 연속형(continuous)의 정규분포로 근사적으로 구할 수 있다.
- $E(\hat{p}) = p$, $Var(\hat{p}) = \frac{p(1-p)}{n}$이므로 중심극한정리에 의하여 다음이 성립하고,(by,(1))

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{X - np}{\sqrt{np(1-p)}}$$

이므로 다음의 정리가 성립한다.

이항분포의 정규근사

Theorem (이항분포의 정규근사)

이항분포 $B(n,p)$를 따르는 확률변수 X는 n이 클 때 근사적으로 정규분포 $N(np, np(1-p))$를 따른다. 즉, 다음이 성립한다.

$$\frac{X-np}{\sqrt{np(1-p)}} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

- 근사의 정밀도는 $np > 5$이고 $n(1-p) > 5$일 때 상당히 좋은 것으로 알려져 있다.

이항분포의 정규근사

- $B(15, 0.4)$의 히스토그램과 이를 근사하는 $N(9, 3.6)$의 곡선

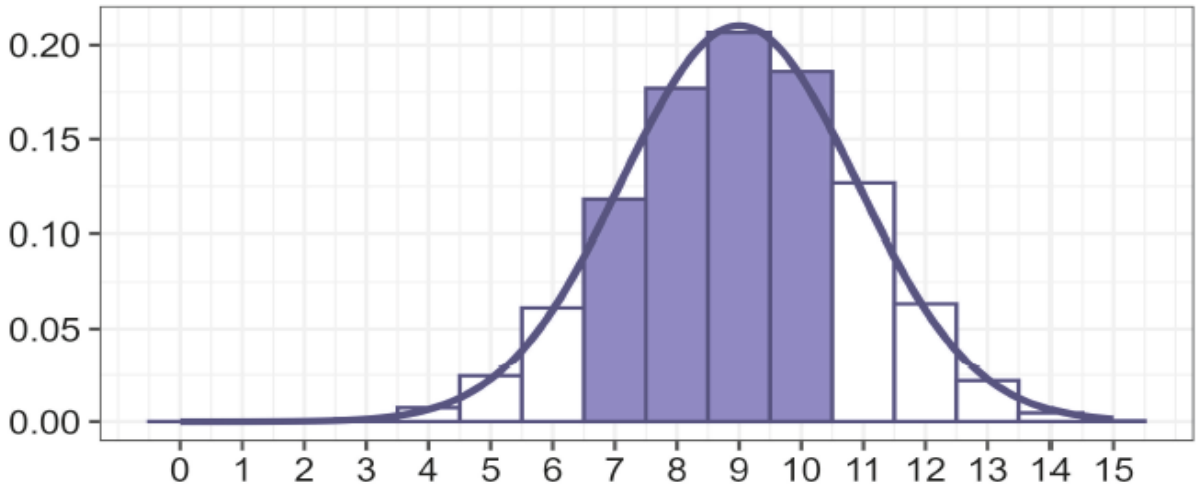


그림 4–6 이항분포 $B(15,\ 0.6)$의 히스토그램과 이를 근사하는 정규분포곡선

- n은 15로서 그리 크지 않더라도 근사정도가 좋음.
- 일반적으로 이항분포가 대칭형일 때는 근접 속도가 매우 빠르다.
- $n = 30$ 정도면 비대칭형일 때도 상당히 근접한다.
- 극단적인 비대칭($p \approx 0$ *or* $p \approx 1$)인 경우에는 n이 아주 커야한다.

이항분포의 정규근사 - 연속성 수정

- 연속성수정(continuity correction):

  [그림 4-6]에서 이항분포의 확률 $P(7 \leq X \leq 10)$은 히스토그램의 빗금친 부분의 넓이가 되므로, 이를 $P(6.5 \leq X \leq 10.5)$의 정규분포 $N(9, 3.6)$의 넓이로써 근사시켜 근사의 정밀도를 높일 수 있다.

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

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

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

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

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

이항분포의 정규근사 예시

Example

X가 이항분포 $B(100, \frac{1}{5})$일 때, 다음 확률의 근사값을 구하여라.
(a) $P(X \leq 21)$
(b) $P(15 \leq X \leq 19)$

sol) $np = 100 \times \frac{1}{5} = 20$, $\sqrt{np(1-p)} = \sqrt{100 \times \frac{1}{5} \times \frac{4}{5}} = 4$이므로,

$$(a)\ P(X \leq 21.5) = P(\frac{X-20}{4} \leq \frac{21.5-20}{4})$$
$$= P(Z \leq 0.375)$$
$$= 0.6462$$

$$(b)\ P(14.5 \leq X \leq 19.5) = P(\frac{14.5-20}{4} \leq \frac{X-20}{4} \leq \frac{19.5-20}{4})$$
$$= P(-1.375 \leq Z \leq -0.125)$$
$$= 0.4503 - 0.0846$$

정규분포 분위수대조도

- z_α : 표준정규분포의 제 $100(1-\alpha)$백분위수로, $Z \sim N(0,1)$일 때, 주어진 α값에 대하여 $P(Z > z) = \alpha$를 만족하는 분위수의 값.
- 정규분포 분위수대조도(or quantile-quantile plot; Q-Q plot):

  정규모집단의 가정을 검토하는 방법으로 표준정규분포의 분위수와 이에 대응하는 자료 분포의 분위수를 좌표로 하는 점들을 그림으로 나타낸 것.
- 점들이 직선 주위에 밀접하여 나타나면 모집단의 분포가 정규분포라는 가정이 타당하다고 할 수 있다.

정규분포 분위수대조도

Example

$x_1, x_2, \cdots, x_{100}$이 정규분포 $N(47, 10^2)$를 따르는 모집단으로부터의 표본인지를 검토하여 보자.

- 표준정규분포의 분위수를 $\{z_{0.99}, z_{0.98}, \cdots\}$라고 하고

  자료의 값들을 크기순으로 정열한 $x_{(1)} < x_{(2)} < \cdots < x_{(100)}$에서 $x_{(i)}$를 대략 이 자료 분포의 $(i/100)$ 분위수라고 하자.

정규분포 분위수대조도

- 이 때, 다음과 같은 관계식에 의해서

$$\frac{X-\mu}{\sigma} = Z \iff X = \sigma Z + \mu$$

$x_{(1)}, x_{(2)}, \cdots$에 대응하는 정규분포 $N(47, 10^2)$의 분위수는 각각 다음과 같이 주어진다.

$$10 \times z_{0.99} + 47, \ 10 \times z_{0.98} + 47, \cdots$$

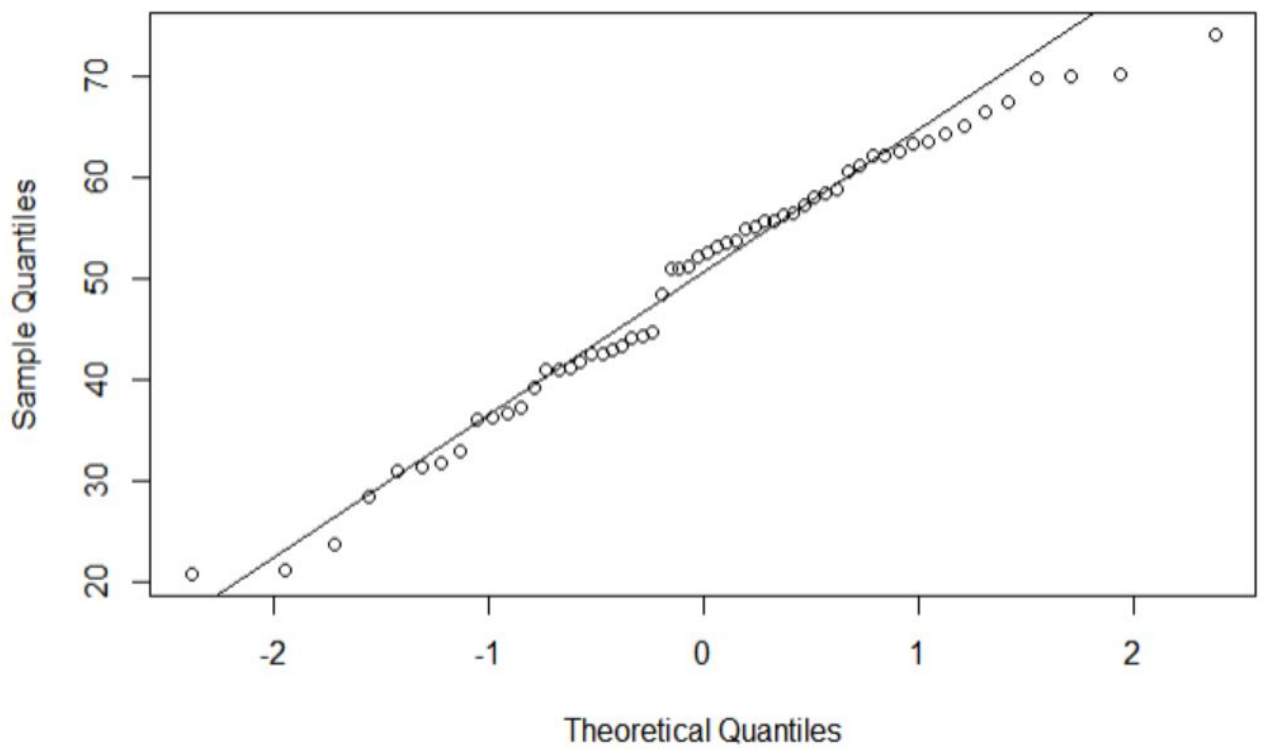
- 실제 모집단의 분포가 $N(47, 10^2)$인 경우에는 위의 서로 대응하는 분위수들은 서로 가깝게 나타날 것이다.

정규분포 분위수대조도

- 그러면 자료분포의 분위수와 이에 대응하는 표준정규분포의 분위수를 좌표로 하는 점들을 그림으로 그려보자.

$$(z_{0.99}, x_{(1)}),\ (z_{0.98}, x_{(2)}), \cdots$$

Figure: 정규분포 분위수대조도 출력 결과

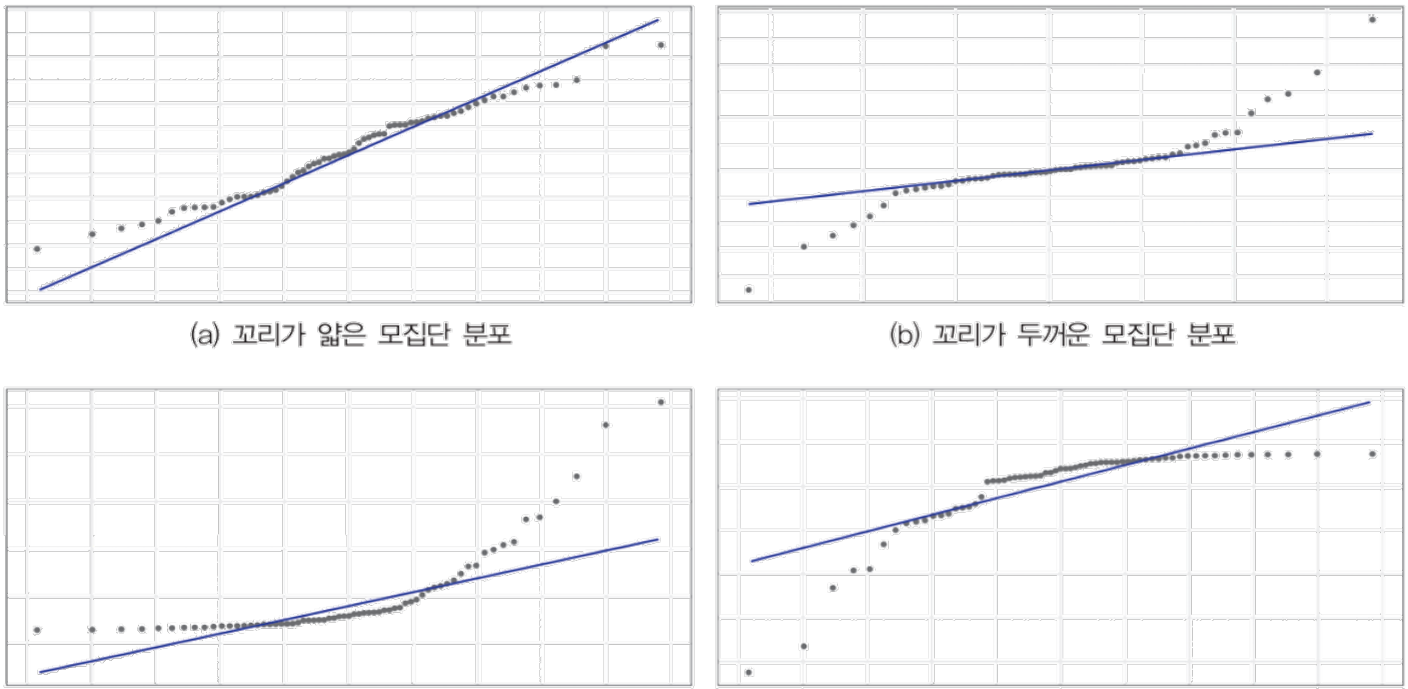


정규분포 분위수대조도

- 자료분포의 분위수와 이에 대응하는 표준정규분포의 분위수를 좌표로 하는 점들은 직선 주위에 밀집하여 나타날 것이다.
- 위 그림의 경우에는 정규모집단의 가정이 타당하다고 할 수 있다.
- 통계패키지에서는 표본의 크기가 n일 때 자료 분포의 분위수 $x_{(i)}$에 대하여 표준정규분포의 (i/n) 분위수 대신 흔히 $\left(\frac{i-3/8}{n+1/4}\right)$ 분위수를 대응시켜 그린다.

정규분포 분위수대조도

- 정규분포 분위수대조도에서 점들이 직선 주위에 가까이 나타나지 않는 경우, 모집단이 정규분포가 아닐 가능성이 큼
- 아래 그림은 이러한 경우에 추측할 수 있는 전형적인 모집단 분포의 형태



(a) 꼬리가 얇은 모집단 분포

(b) 꼬리가 두꺼운 모집단 분포

(c) 오른쪽 꼬리가 긴 모집단 분포

(d) 왼쪽 꼬리가 긴 모집단 분포

그림 4-12 정규분포 분위수대조도와 모집단 분포의 형태