김은영

제6장 반복표본에서의 변동: 표본분포

6.1 서론

(1) 통계적 추론(statistical inference)

① 모수(parameter): 모집단(population)의 수치적 특성치

② 통계량(statistic): 표본(sample) 관측값들의 실수값 함수

(예) 표본평균, 표본중앙값, 표본표준편차

(2) 표본과 통계량에 관한 세 가지 중요점

① 표본은 모집단의 일부분이므로 통계량의 값은 모수의 참값과 정확히 일치할 것이라고 기대할 수 없다.

② 통계량의 관측값은 많은 가능한 표본들 중에서 (실제 추출된) 특정한 하나의 표본에 의해서만 결정된다.

③ 표본을 반복적으로 추출한다면 그로부터 계산된 통계량의 값들은 달라지므로 그 값들은 변동을 가지게 된다.

6.2 통계량의 표본분포

(1) 표본분포(sampling distribution): 통계량의 확률분포

(2) 통계량의 표본분포는 모집단의 분포[$f(x)$]와 표본의 크기[n]에 의해 결정된다.

[예제 6.1]

세 명의 대학원생이 수강하는 어떤 과목의 담당교수는 매주 쪽지 시험을 치른다. 어떤 주 쪽지 시험 점수가 2점, 3점, 4점이었다고 하자. 두 명의 학생을 복원추출방법으로 선택하여 그 점수의 평균을 $\overline{X}$라고 할 때, $\overline{X}$의 표본분포를 구하라.

【풀이】

X : 임의로 추출한 한 학생의 점수

< 모집단의 분포 >

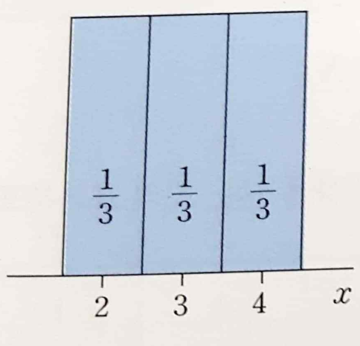
x	$f(x)$
2	1/3
3	1/3
4	1/3

< 크기가 2인 모든 가능한 표본과 그에 해당하는 $\overline{X}$의 값 >

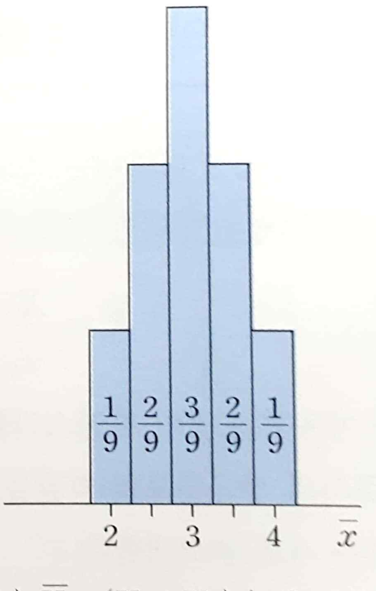
(x_1, x_2)	$(2,2)$	$(2,3)$	$(2,4)$	$(3,2)$	$(3,3)$	$(3,4)$	$(4,2)$	$(4,3)$	$(4,4)$
$\overline{x}$	2	2.5	3	2.5	3	3.5	3	3.5	4

< $\overline{X}$의 분포 >

$\overline{X}$	확률
2	1/9
2.5	2/9
3	3/9
3.5	2/9
4	1/9



(a) 모분포



(b) $\overline{X}=(X_1+X_2)/2$의 표본분포

6.3 표본평균의 분포와 중심극한정리

(1) 모평균(μ)에 대한 추론 절차는 표본평균($\overline{X}$)과 그 표본분포를 이용

(2) $\overline{X}$의 표본분포의 평균과 표준편차는 μ와 σ를 이용하여 다음과 같이 나타낸 수 있다.

크기 n인 확률표본의 표본평균 $\overline{X}$의 분포에 대해 다음이 성립한다.

$E(\overline{X}) = \mu$	$Var(\overline{X}) = \frac{\sigma^2}{n}$	$sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

[예제 6.2]

① [표 6.1]의 모집단의 분포에 대한 모평균과 모표준편차를 계산하라.

【풀이】

x	$f(x)$	$xf(x)$	$x^2f(x)$
2	$\frac{1}{3}$	$2\times\frac{1}{3}=\frac{2}{3}$	$2^2\times\frac{1}{3}=\frac{4}{3}$
3	$\frac{1}{3}$	$3\times\frac{1}{3}=\frac{3}{3}$	$3^2\times\frac{1}{3}=\frac{9}{3}$
4	$\frac{1}{3}$	$4\times\frac{1}{3}=\frac{4}{3}$	$4^2\times\frac{1}{3}=\frac{16}{3}$
total	1	3	$\frac{29}{3}$

$$\mu = \sum_{i=1}^{3} x_i f(x_i) = x_1 f(x_1) + x_2 f(x_2) + x_3 f(x_3) = \frac{2}{3} + \frac{3}{3} + \frac{4}{3} = \frac{9}{3} = 3$$

$$\sigma^2 = \sum_{i=1}^{3} x_i^2 f(x_i) - \mu^2 = \left[x_1^2 f(x_1) + x_2^2 f(x_2) + x_3^2 f(x_3)\right] - \mu^2$$

$$= \left[\frac{4}{3} + \frac{9}{3} + \frac{16}{3}\right] - 3^2 = \frac{29}{3} - 3^2 = \frac{2}{3}$$

$$\sigma = \sqrt{\frac{2}{3}}$$

② [표 6.2]에 주어진 $\overline{X}$의 분포에 대하여 평균과 표준편차를 계산하라.

【풀이】

$\overline{x}$	$f(\overline{x})$	$\overline{x}f(\overline{x})$	$\overline{x^2}f(\overline{x})$
2	$\frac{1}{9}$	$2\times\frac{1}{9}=\frac{2}{9}$	$2^2\times\frac{1}{9}=\frac{4}{9}$
2.5	$\frac{2}{9}$	$2.5\times\frac{2}{9}=\frac{5}{9}$	$2.5^2\times\frac{2}{9}=\frac{12.5}{9}$
3	$\frac{3}{9}$	$3\times\frac{3}{9}=\frac{9}{9}$	$3^2\times\frac{3}{9}=\frac{27}{9}$
3.5	$\frac{2}{9}$	$3.5\times\frac{2}{9}=\frac{7}{9}$	$3.5^2\times\frac{2}{9}=\frac{24.5}{9}$
4	$\frac{1}{9}$	$4\times\frac{1}{9}=\frac{4}{9}$	$4^2\times\frac{1}{9}=\frac{16}{9}$
total	1	3	$\frac{84}{9}$

$$E(\overline{X})=\sum_{i=1}^{5}\overline{x}_i f(\overline{x}_i)=\frac{2}{9}+\frac{5}{9}+\frac{9}{9}+\frac{7}{9}+\frac{4}{9}=\frac{27}{9}=3$$

$$Var(\overline{X})=\sum_{i=1}^{5}\overline{x}_i^2 f(\overline{x})-\left[E(\overline{X})\right]^2$$

$$=\left(\frac{4}{9}+\frac{12.5}{9}+\frac{27}{9}+\frac{24.5}{9}+\frac{16}{9}\right)-3^2=\frac{84}{9}-3^2=\frac{1}{3}$$

$$sd(\overline{X})=\frac{1}{\sqrt{3}}$$

③ 그 결과를 이용하여 식 $E(\overline{X})=\mu$와 $sd(\overline{X})=\frac{\sigma}{\sqrt{n}}$을 확인해 보라.

【풀이】

$$E(\overline{X})=3=\mu$$

$$sd(\overline{X})=\frac{1}{\sqrt{3}}=\frac{\sigma}{\sqrt{n}}=\sqrt{\frac{2}{3}}\times\frac{1}{\sqrt{2}}=\frac{1}{\sqrt{3}}$$

[$\overline{X}$의 표본분포의 모양에 대한 두 가지 중요한 점]

❑ 모집단의 분포가 정규분포일 때는 $\overline{X}$의 정확한 분포모양을 알 수 있다. 즉, 평균 μ, 표준편차 σ인 정규 모집단으로부터의 표본평균 $\overline{X}$는 평균 μ, 표준편차 $\sigma/\sqrt{n}$인 정규분포를 따른다.

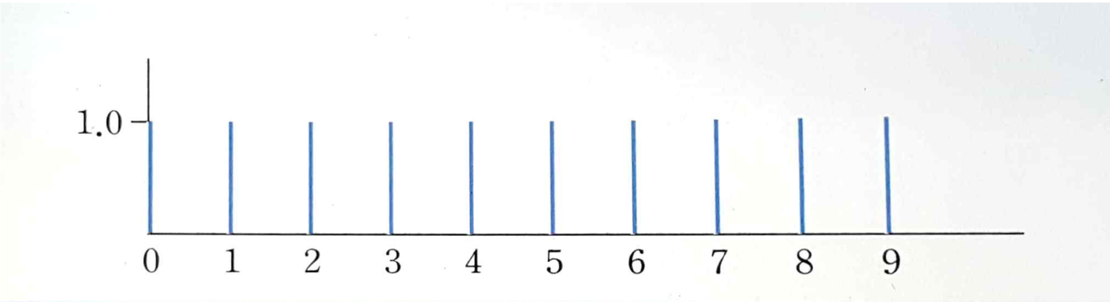
❑ 모집단이 정규분포를 따르지 않을 때, 표본평균 $\overline{X}$의 분포는 모집단의 분포가 무엇이냐에 따라 달라진다. 그러나 중심극한정리에 의하면 표본의 크기가 크면 모집단의 분포에 상관없이 $\overline{X}$의 분포는 근사적으로 정규분포를 따른다는 것이 알려져 있다.

(3) 중심극한정리(Central Limit Theorem)

모집단의 분포가 무엇이든 관계없이 n이 크면 표본평균 $\overline{X}$의 분포는 근사적으로 정규분포가 된다. 즉, 평균 μ, 표준편차 σ인 임의의 모집단으로부터 표본평균 $\overline{X}$는 n이 크면 근사적으로 평균 μ, 표준편차 $\sigma/\sqrt{n}$인 정규분포를 따른다. 따라서, $Z=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$는 근사적으로 표준정규분포 $N(0,1)$를 따른다.

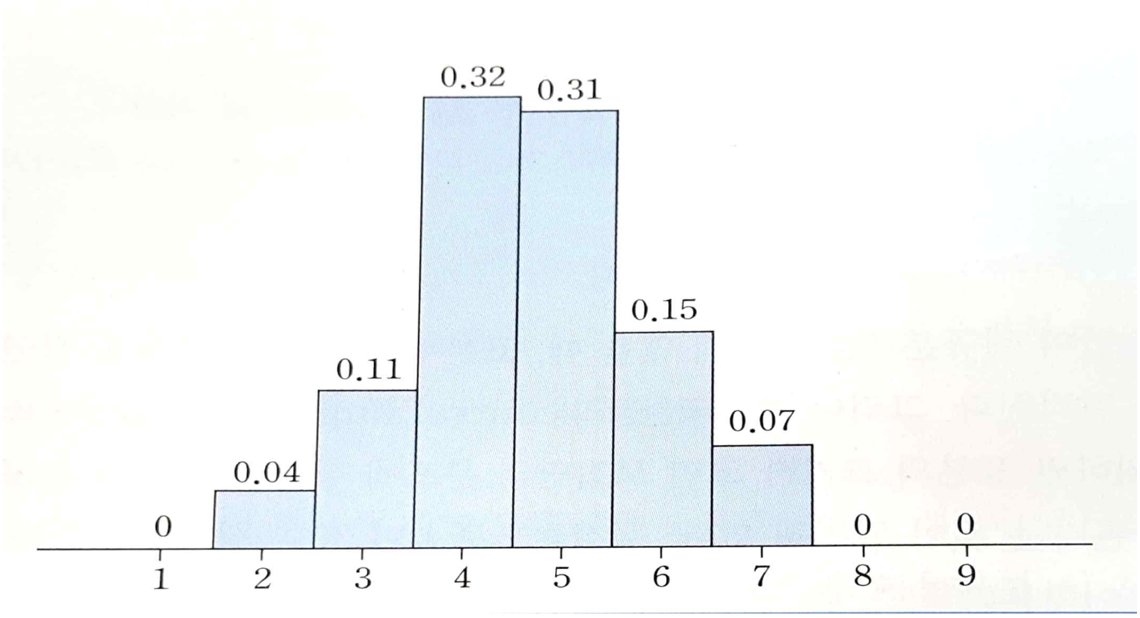
[예제 6.3] (중심극한정리의 우수성)

모집단이 정수 $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$에 각각 0.1씩의 확률을 갖는 이산균일분포를 따른다고 하자. 이 확률모형은 전화번호의 끝자리 숫자의 분포에 적절한 모형일 것이다. 이 분포의 선 그래프는 아래 그림과 같다. 이때 모평균 $\mu=4.5$, 모표준편차 $\sigma=2.872$이다.



$n=5$인 100개의 표본

표본번호	관측값	합	평균	표본번호	관측값	합	평균
1	8,3,7,3,5	26	5.2	51	8,0,4,5,4	21	4.2
2	6,1,3,7,5	22	4.4	52	8,7,7,7,8	37	7.4
3	4,8,5,3,7	27	5.4	53	6,3,4,7,2	22	4.4
4	5,8,1,9,7	30	6.0	54	9,4,5,5,3	26	5.2
5	3,7,5,1,7	23	4.6	55	7,1,3,1,9	21	4.2
6	6,6,4,3,6	25	5.0	56	2,2,7,0,3	14	2.8
7	7,2,6,1,6	22	4.4	57	3,2,4,3,6	18	3.6
8	8,6,6,4,2	26	5.2	58	0,2,7,2,3	14	2.8
9	2,6,7,0,7	22	4.4	59	4,1,7,4,7	23	4.6
10	3,7,2,1,8	21	4.2	60	7,6,4,8,9	34	6.8
11	1,8,5,3,8	25	5.0	61	5,6,3,0,6	20	4.0
12	9,3,5,4,1	22	4.4	62	3,7,0,3,3	16	3.2
13	3,3,2,3,7	18	3.6	63	9,5,7,6,1	28	5.6
14	1,9,2,4,0	16	3.2	64	1,7,6,7,1	22	4.4
15	2,3,1,6,7	19	3.8	65	2,7,5,8,8	30	6.0
16	0,2,3,3,2	10	2.0	66	4,2,7,7,8	28	5.6
17	8,8,7,8,4	35	7.0	67	5,6,0,6,7	24	4.8
18	2,3,5,7,8	25	5.0	68	3,8,4,6,6	27	5.4
19	4,2,3,6,7	22	4.4	69	1,6,7,4,8	26	5.2
20	5,3,7,8,2	25	5.0	70	6,4,6,3,6	25	5.0
21	3,8,7,6,4	28	5.6	71	5,6,3,5,2	21	4.2
22	4,2,5,8,9	28	5.6	72	5,6,8,9,5	33	6.6
23	0,5,4,4,8	21	4.2	73	1,5,4,5,4	19	3.8
24	2,5,8,1,7	23	4.6	74	4,6,7,3,7	27	5.4
25	8,1,4,5,1	19	3.8	75	6,6,8,5,3	28	5.6
26	1,7,3,8,6	25	5.0	76	8,7,6,6,3	30	6.0
27	3,1,6,1,1	12	2.4	77	7,9,3,4,1	24	4.8
28	5,3,6,2,3	19	3.8	78	8,5,5,1,4	23	4.6
29	8,5,5,2,4	24	4.8	79	3,3,4,2,4	16	3.2
30	9,6,2,2,4	23	4.6	80	3,7,4,3,0	17	3.4
31	3,1,6,5,4	19	3.8	81	2,8,3,8,6	27	5.4
32	3,7,6,8,2	26	5.2	82	5,4,0,4,9	22	4.4
33	2,9,1,1,0	13	2.6	83	3,0,4,6,8	21	4.2
34	6,6,3,6,8	29	5.8	84	1,5,9,3,7	25	5.0
35	8,7,6,8,7	36	7.2	85	3,6,9,7,3	28	5.6
36	8,6,3,3,2	22	4.4	86	7,3,5,6,2	23	4.6
37	6,0,1,1,1	9	1.8	87	2,1,7,6,8	24	4.8
38	1,1,6,0,8	16	3.2	88	8,2,2,1,3	16	3.2
39	3,2,6,8,6	25	5.0	89	9,4,3,3,4	23	4.6
40	9,5,4,9,8	35	7.0	90	1,6,8,0,5	20	4.0
41	4,0,2,6,4	16	3.2	91	6,4,0,7,4	21	4.2
42	6,6,7,1,0	20	4.0	92	6,8,6,5,5	30	6.0
43	1,9,8,7,9	34	6.8	93	4,8,0,2,1	15	3.0
44	9,1,8,3,1	22	4.4	94	7,1,3,7,3	21	4.2
45	6,9,1,8,5	29	5.8	95	9,7,4,3,8	31	6.2
46	5,5,5,5,6	26	5.2	96	8,7,4,8,5	32	6.4
47	4,1,8,8,1	22	4.4	97	5,2,6,5,3	21	4.2
48	4,0,0,2,4	10	2.0	98	0,9,4,5,7	25	5.0
49	6,8,3,4,3	24	4.8	99	7,8,1,8,4	28	5.6
50	2,7,7,3,3	22	4.4	100	3,5,6,1,4	19	3.8



[예제 6.4]

주어진 정보: $\mu = 82$, $\sigma = 12$

(1) 표본의 크기가 $n=64$인 확률표본을 추출할 때, 표본평균이 80.8과 83.2 사이에 있을 확률은?

【풀이】 $n=64(>30)$ ⇒ 중심극한정리 적용가능

$$P(80.8 < \overline{X} < 83.2) = P\left(\frac{80.8-82}{12/\sqrt{64}} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{83.2-82}{12/\sqrt{64}}\right)$$

$= P(-0.8 < Z < 0.8) = 0.7881 - 0.2119 = 0.5762$

(2) 표본의 크기가 $n=100$인 확률표본을 추출할 때, 표본평균이 80.8과 83.2 사이에 있을 확률은?

【풀이】 $n=100(>30)$ ⇒ 중심극한정리 적용가능

$$P(80.8 < \overline{X} < 83.2) = P\left(\frac{80.8-82}{1.2/\sqrt{100}} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{83.2-82}{1.2/\sqrt{100}}\right)$$

$= P(-1.0 < Z < 1.0) = 0.8413 - 0.1587 = 0.6826$

구간 $(80.8, 83.2)$의 중심은 모평균 $\mu = 82$

이 구간에 표본평균 $\overline{X}$가 속할 확률은 $n=64$일 때보다 $n=100$일 때 더 크다.