Statistics I

Homework 2

학 과	응용수학과
교수님	김경수 교수님
학 번	2014110374
이 름	정상만
제출일	2018. 6. 18.

13.5 Problem)

a) : 잡은 오타의 수 라 하면, $C \quad Bin(10,0.7)$

M : 놓친 오타의 수 라 하면, $M \quad Bin(10,0.3)$

b) $M=3) = \left.\begin{matrix}10\\3\end{matrix}\right) 0.3)\ (0.7)^7 = (120)(0.027)(0.08235) = 0.2668.$

3 이상인 경우에 이항분포표를 이용하면 $P(M \quad 3) = 0.6172.$

13.11 Problem)

a) 이항분포의 기댓값(평균) 은 $\mu = np = (1520)(0.31) = 471.2$

표준편차는 $\sigma = \quad np(1-p) = \sqrt{}1520(0.31)(1-0.31) = 18.0313$

b) $np = 471.2 \quad 10$ 이고 $n(1-p) = (1520)(0.69) = 1048.8 \quad 10$ 이므로,

n은 정규분포로 근사하기에 충분한 크기를 갖는다. 대학에선 475명을 모집하므로

모집 수가 이보다 클 때의 이항확률은 다음과 같다.

$P(X \quad 476) = P\ Z \quad \left.\begin{matrix}476-471.2\\18.0313\end{matrix}\right) = P(Z \quad 0.27) = 0.3936.$

c) 정확한 확률은 0.4045이므로, 정규근사는 0.0109 정도 낮다.

14.1 Problem)

a) $\begin{matrix}\sigma\\\sqrt{n}\end{matrix} = \begin{matrix}34\\\sqrt{51000}\end{matrix} = 0.1506$

b) 표본평균 x에 대하여 95%는 μ의 $2\sigma = 2(0.1506) = 0.3012$ 내에 위치한다.

c) 95% 신뢰구간은 153 ± 0.3012 혹은 $(152.7, 153.3)$ 이다.

14.8 Problem)

a)

신뢰수준 z^*값

90%	95%	99%
1.645	1.960	2.576

위의 신뢰수준을 참조하여 세 개 신뢰구간을 구한다.

$\bar{x} = 26.8,\ \sigma = 7.5\ \ n = 654$ 인 정규분포하는 모집단을 가지는 표본분포이므로

90% 신뢰구간은 $26.8 \pm (1.645)\ \begin{matrix}7.5\\\sqrt{654}\end{matrix} = 26.8 \pm 0.4824$

95% 신뢰구간은 $26.8 \pm (1.960)\frac{7.5}{\sqrt{654}} = 26.8 \pm 0.5748$

99% 신뢰구간은 $26.8 \pm (2.576)\frac{7.5}{\sqrt{654}} = 26.8 \pm 0.7555$

b) a) 에서 구한 신뢰구간의 오차한계는 아래와 같다.

신뢰수준 z^*에 따른 오차한계

90%	95%	99%
0.4824	0.5748	0.7555

표본의 크기와 모표준편차가 동일하다면 오직 * 값에 영향을 받고, z^*값이 커지면 그만큼 오차한계가 점점 증가하게 된다.

14.9 Problem)

a) 14.8 Problem에서 100명의 단순 무작위 표본을 가진다고 하면 95% 신뢰구간에 대한 오차한계는 $1.960)\ \frac{7.5}{100} = 1.4700$ 이다.

b) a) 와 마찬가지로, $n=400, n=1600$인 경우 95% 신뢰구간에 대한 오차한계를 구하면

$n=400 \quad 1.96\frac{7.5}{\sqrt{400}} = 0.7350$, $n=1600 \quad 1.96\ \frac{7.5}{\sqrt{1600}} = 0.3675$ 이다.

c)

95% 신뢰수준에서 n에 따른 오차한계

100	400	1600
1.4700	0.7350	0.3675

동일한 신뢰수준과 모표준편차를 가질 때 표본크기가 점점 커지면 오차한계는 감소한다.

15.9 Problem)

a) 뽑은 표본의 크기가 18이므로 모표준편차가 60일 때 모집단이 정규분포를 따르고, $\mu=0$이라 가정했으므로 표본평균의 분포는 $0,\ \frac{\sigma}{\sqrt{18}}\Big) = N(0, 14.1421)$ 이다.

b) 양측 검정이므로 P값은 $P = 2P(x \quad 17) = 2P\ Z \quad \frac{17-0}{14.1421}\Big) = 0.2302$이다.

15.15 Problem)

: 나쁜 소식을 접한 고객들이 주는 팁의 평균 백분율이라 가정하자.
가설은 : $= 20$ 대 H_1 : $\mu < 20$ 라 하자.

x의 표준편차는 $\frac{2}{20} = 0.4472$ 이므로, 따라서 검정 통계량은 $z = \frac{18.19 - 20}{0.4472} = -4.05$.
가설 H_1에 의하면 단측검정이므로, P값은 $P(Z \quad -4.05)$ 0 이다.
P값이 거의 0에 가까우므로, H_0을 기각, 따라서 나쁜 소식을 접했을 때 팁의 평균백분율이
전체 팁의 백분율보다 낮다는 강력한 증거가 된다.

15.17 Problem)

검정통계량 $z = 1.876$이므로 표 C에서 $z^* = (1.645, 1.960)$ 사이에 위치한다.
이 때 P-값은 $H_a : \mu \quad 0$이므로 양측검정이고, 따라서 $P-value$: $(0.05, 0.10)$ 내에 위치한다. 그러므로
이 검정은 유의수준 5%, 1% 에서 모두 유의하지 않다.

16.5 Problem)

아니다. 신뢰구간은 $\overline{x}$ 장래값의 범위를 설명하지 못한다.

16.9 Problem)

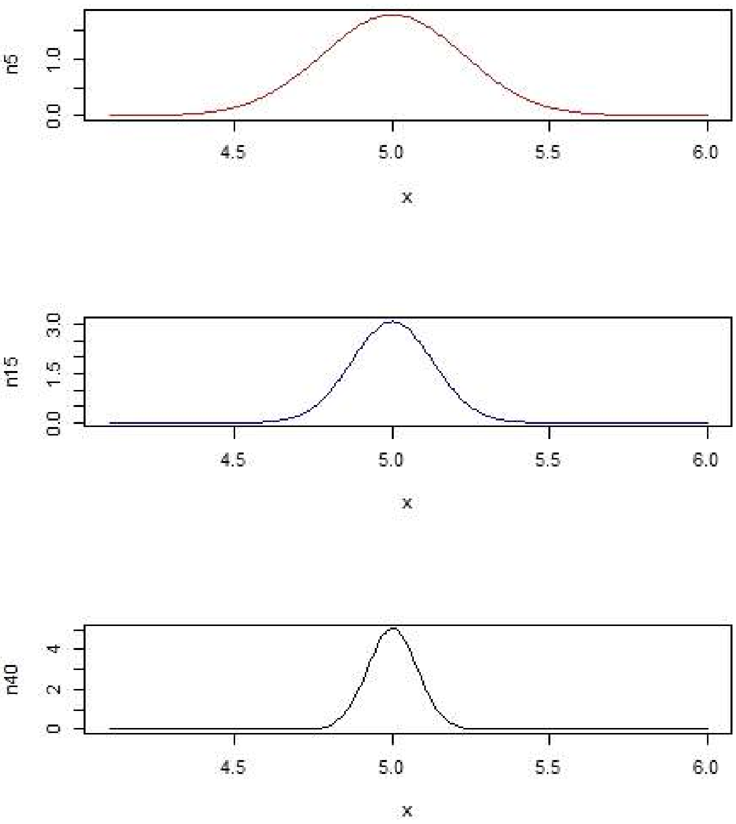
문제에 의하면 H_0 : $\mu = 5.0$ 대 H_1 : $\mu < 5.0$, $\sigma = 0.5$, $\overline{x} = 4.8$.

a)

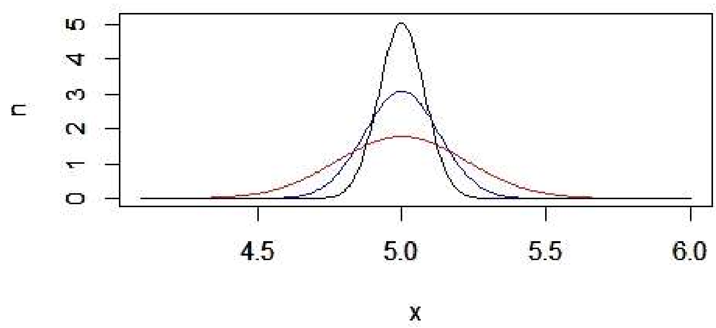
각 n 에 대한 검정통계량 z 에서의 P-value $(Z \quad z)$

n	z	P
5	-0.89	0.1867
15	-1.55	0.0606
40	-2.53	0.0057

b) R 프로그램(R studio)을 이용하여 정규분포를 그려보면 다음과 같은 분포를 갖는다.



[n=5, n=15, n=40 에 대한 정규분포]



[n의 모든 경우에 따른 정규분포]

16.14 Problem)

a) 검사력(power)=1-제 2종 오류의 확률

b) 검사력, 즉 귀무가설을 기각할 확률이 0.24이므로, 대립가설이 참이라 주장하기 어렵다. 즉, 귀무가설을 기각하는데 실패할 확률이 높으므로 제 2종 오류를 범할 가능성이 높다. 따라서 전도성이 10.15인 강철봉을 판매하는 것인 대립가설을 적절하게 보호하지 못한다.

16.15 Problem)

a) 더 많아야 한다.

b) 유의 수준이 작아질수록 귀무가설을 기각할 확률이 낮아진다. (P-값에 의하면)
반대로 유의수준이 높아질수록 귀무가설을 기각할 확률이 높아진다.
따라서 유의수준이 높아질수록 검정력이 증가한다.

c) 10.2는 이전에 언급한 10.15 보다 더 큰 차이가 나므로 검사력이 증가한다.

18.3 Problem)

a) 유의수준 $\quad = 0.05$에서 $t(4)$ 분포에 기초한 단측검정의 임계값 $t^* = 2.132$

b) $t(26)$분포에 기초한 98% 신뢰구간에 대한 임계값 $t^* = 2.479$

18.7 Problem)

4	9.1
5	1.0 4.8
6	0.8 3.3 3.4 4.4 4.5 5.0

스템 플롯을 보면 왼쪽으로 기울어짐을 알 수 있고, 이탈값은 존재하지 않는다.
그런데 표본 수가 적으므로 정규성에 벗어난다고 볼 수 없다.
$t^* = 1.860$, (자유도는 8) 90% 신뢰구간은 $59.5889 \pm 1.860 \frac{6.2553}{9} = (55.71\%, 63.47\%)$.
따라서 고대 공기 중에 질소의 평균 백분율은 55.71%와 63.47% 사이에 위치한다고 90% 신뢰한다.

18.10 Problem)

$n = 9$, $x = 59.5889$, $\frac{s}{\sqrt{n}} = \frac{6.2553}{\sqrt{9}} = 2.0851$

1. 가설 : $\quad : \mu = \mu_0 = 78.1$ 대 $H_a : \mu \quad \mu_0$
2. 검정통계치 : $t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{59.5889 - 78.1}{2.0851} = -8.8778$
3. 기각역 : $R :\ |T| > t_{0.0005}(8) = 5.041\}$ $\quad R : \{T > 5.041,\ T < -5.041\}$
4. 결론 : $t = -8.8778 < -t_{0.0005}(8) = -5.041$ 이므로 t는 기각역에 속한다. 따라서 H_0을 기각할 수 있다.

즉, 유의수준 0.001에서 고대 공기에 함유된 질소의 백분율은 현재의 78.1% 과 다르다고 주장할 수 있다.

19.7 Problem)

신뢰구간을 구하자. 우선 $=17.50$, $\overline{x_2}=13.67$, 표준오차 $E=1.813$이다. 자유도 df=9-1=8 로 두면, 이 경우에 임계값 $t^*=3.355$가 된다. 따라서 벌목이 이루어지지 않은 구획과 벌목이 이루어진 구획에서 나무종류의 평균 숫자 차이에 대한 99% 신뢰구간은 $(\overline{x_1}-\overline{x_2})\pm t^* SE=(-2.253, 9.913)$.

19.9 Problem)

향기가 없는 경우		향기가 있는 경우
9 8	6	
3 2 2	7	
9 6 5	7	6
4 4	8	
7 7 6 5	8	8 9
3 2 2 2 1	9	2 3 4
8 6	9	5 7 8
3 1	10	1 2 3 4
9 7 7 6	10	5 5 6 6 7 8 8 9 9 9
	11	4
8 5	11	6
1	12	1 4
	12	6 9
	13	
	13	7

a)
다음 스템 플롯에서 보면 정규성에 벗어나지 않고 적당하게 정규분포하는 것 처럼 보인다. 앞선 논의에 의하면, 자료를 독립적인 단순 무작위 표본으로 취급하는 것을 정당화할 수 있다. 따라서 t 절차를 사용할 수 있다.
μ_1, μ_2를 각각 향기가 없는 경우 음식점에서 보낸 모집단 평균시간, 향기가 있는 경우 음식점에서 보낸 모집단 평균 시간이라고 하자. 가설은 $H_0 : \mu_1 = \mu_2$ 대 $H_a : \mu_1 < \mu_2$ 이다. $\overline{x_1}=91.27$, $\overline{x_2}=105.700$, $s_1=14.930$, $s_2=13.105$, $n_1=30$, $n_2=30$ 이고, 표준오차 $SE=\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}=3.627$ 및 $t=\frac{\overline{x_1}-\overline{x_2}}{SE}=-3.98$ 이 된다.
이제 자유도 df=29 로 두면, (30-1 과 30-1 중 작은 수) 표 C에서 P값은 $P<0.0005$ 가 된다.
따라서 라벤다 향기가 있는 경우 고객들은 음식점에서 평균적으로 더 많은 시간을 보낸다는 매우 강한 증거가 있다.

b)

향기가 없는 경우		향기가 있는 경우
9	12	
	13	
	14	
99999999999999	15	
	16	
	17	
555555555555	18	55555555555
	19	
5	20	7
9	21	5599999999
	22	3558
	23	
	24	99
5	25	59

지출 자료에 대한 스템 플롯을 만들면, 분포가 한쪽으로 기울어졌으며 많은 틈새가 존재한다.
가설은 $\mu_1 = \mu_2$ 대 $H_a : \mu_1 < \mu_2$ 이다. 여기서 μ_1, μ_2를 각각 향기가 없는 경우 음식점에서 지출한 모집단 평균 총액, 향기가 있는 경우 음식점에서 지출한 모집단 평균 총액이라 하자.

$x_1 = 17.5133$, $\overline{x_2} = 21.1233$, $s_1 = 2.3588$, $s_2 = 2.3450$, $n_1 = 30$, $n_2 = 30$ 이고, 표준오차 $SE = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = 0.6073$

및 $t = \frac{\overline{x_1} - \overline{x_2}}{SE} = -5.95$ 이 된다. 자유도를 마찬가지로 df=29로 두면, 표 C에서 P값은 $P < 0.0001$ 이 된다.
따라서 라벤다 향기가 있는 경우 고객들은 음식점에서 평균적으로 더 많은 금액을 지출한다는 매우 강한 증거가 있다.

20.3 Problem)

a) 평균이 $p = 0.70$ 및 표준편차 $\frac{0.70(1-0.70)}{1500} = 0.0118$ 을 갖는 근사한 정규분포이다.

b) 평균이 $p = 0.70$ 및 표준편차 $\sqrt{\frac{0.70(1-0.70)}{6000}} = 0.0059$ 을 갖는 근사한 정규분포이다.
이 때, 표본크기가 네 배가 되면 p의 표준편차는 절반으로 줄어든다.

20.5 Problem)

a) 이 표본조사는 거주지 전화번호가 없거나 무선 전화만을 사용하는 사람들뿐만 아니라 북쪽 지역 주민들을 배제하였다.

b) $\hat{p} = \frac{1288}{1505} = 0.8558$, $SE_p = \frac{\hat{p}(1-\hat{p})}{n} = 0.009055$ 이므로, 95% 신뢰구간은
$0.8558 \pm (1.96)(0.009055) = (0.8381, 0.8736)$ 이 된다.

20.9 Problem)

a) $= \frac{20}{20} = 1$ 이므로, 따라서 $E = 0$이다. 그러므로 오차한계는 신뢰수준에 관계없이 0이 된다.
이와 같은 크기가 큰 표본 방법에 의하면 유용하지 못한 신뢰구간 (1,1)을 얻게 된다.

b) 플러스 4 추정값은 $\tilde{p} = \frac{22}{24} = 0.9167$, $SE_{\tilde{p}} = 0.0564$ 이다. 이제 p에 대한 95% 신뢰구간을 구하면
$0.9167 \pm 1.96(0.0564) = (0.8062, 1.0272)$ 이 된다. 비율은 1을 초과할 수 없기 때문에 p에 대한 95% 신뢰구간은 $(0.8061, 1)$ 이 된다.

20.11 Problem)

$n = \frac{z^*}{m}\Big)^2 p^*(1-p^*) = \left(\frac{1.645}{0.04}\right)^2 (0.75)(1-0.75) = 317.1$ 이다. 따라서 $n = 318$로 택하면 된다.

20.13 Problem)

p : 얼굴 생김새가 가장 좋은 입후보자가 당선되는 횟수의 비율이라 하자.
가설은 $H_0 : p = 0.50$ 대 $H_a : p > 0.50$. 표본은 32번의 시도로 구성되기 때문에 16번의 성공(얼굴 생김새가 가장 좋은 입후보자가 당선되는경우)과 16번의 실패(당선되지 않는 경우)를 기대하게 된다. 표본은 $\hat{p}$의 표본분포를 설명하기 위해서 정규분포로의 근사를 사용할 만큼 충분히 크다. 또한 표본이 단순 무작위 표본이라고 가정하자.
그러면 $\hat{p} = \frac{22}{32} = 0.6875$, $SE_{\hat{p}} = \frac{0.50(1-0.50)}{32} = 0.0884$, $z = \frac{\hat{p} - p_0}{SE} = \frac{0.6875 - 0.50}{0.0884} = 2.12$, $P = 0.0170$ 이다.
따라서 얼굴 생김새가 가장 좋은 입후보자가 당선되는 횟수의 비율이 0.50을 초과한다는 강한 증거가 존재한다.

21.3 Problem)

p_1 은 권장 수준에 부합되는 남학생의 비율을 나타내고 p_2 는 여학생의 비율을 나타낸다. 두 개 표본 모두에서 많은 성공과 실패를 갖고 있어서 크기가 큰 표본 방법을 사용하는 것이 합리적이다. $p_1 = \frac{3594}{7881} = 0.4560$, $\widehat{p_2} = \frac{2261}{8164} = 0.2769$, $SE = 0.0075$이고 오차한계는 $(2.576)(SE) = 0.0193$이다. 신체 화동의 권장 수준에 부합되는 남학생과 여학생 사이 비율의 차이에 대한 99% 신뢰구간은 0.1598 부터 0.1984 까지, 또는 16.0% 부터 19.8% 까지이다.

21.7 Problem)

가설은 $H_0 : p_1 = p_2$ 대 $H_a : p_1 < p_2$. 가장 작은 횟수가 96이므로 유의성 검정 절차가 타당하다. $\widehat{p_1} = \frac{96}{578} = 0.1661$, $\widehat{p_2} = \frac{656}{2992} = 0.2193$, $\hat{p} = \frac{96+656}{578+2992} = 0.2106$, $SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{578} + \frac{1}{2992}\right)} = 0.01853$, $z = \frac{0.1661 - 0.2193}{SE} = -2.87$,
따라서 $P = 0.0021$. 머리 부상을 입은 알펜 스키어와 스노보더는 머리 부상을 입지 않은 알펜 스키어와 스노보더 보다 헬멧을 착용한 가능성이 더 적다는 강한 증거가 있다. ($\alpha = 0.01$ 수준에서 유의하다.)