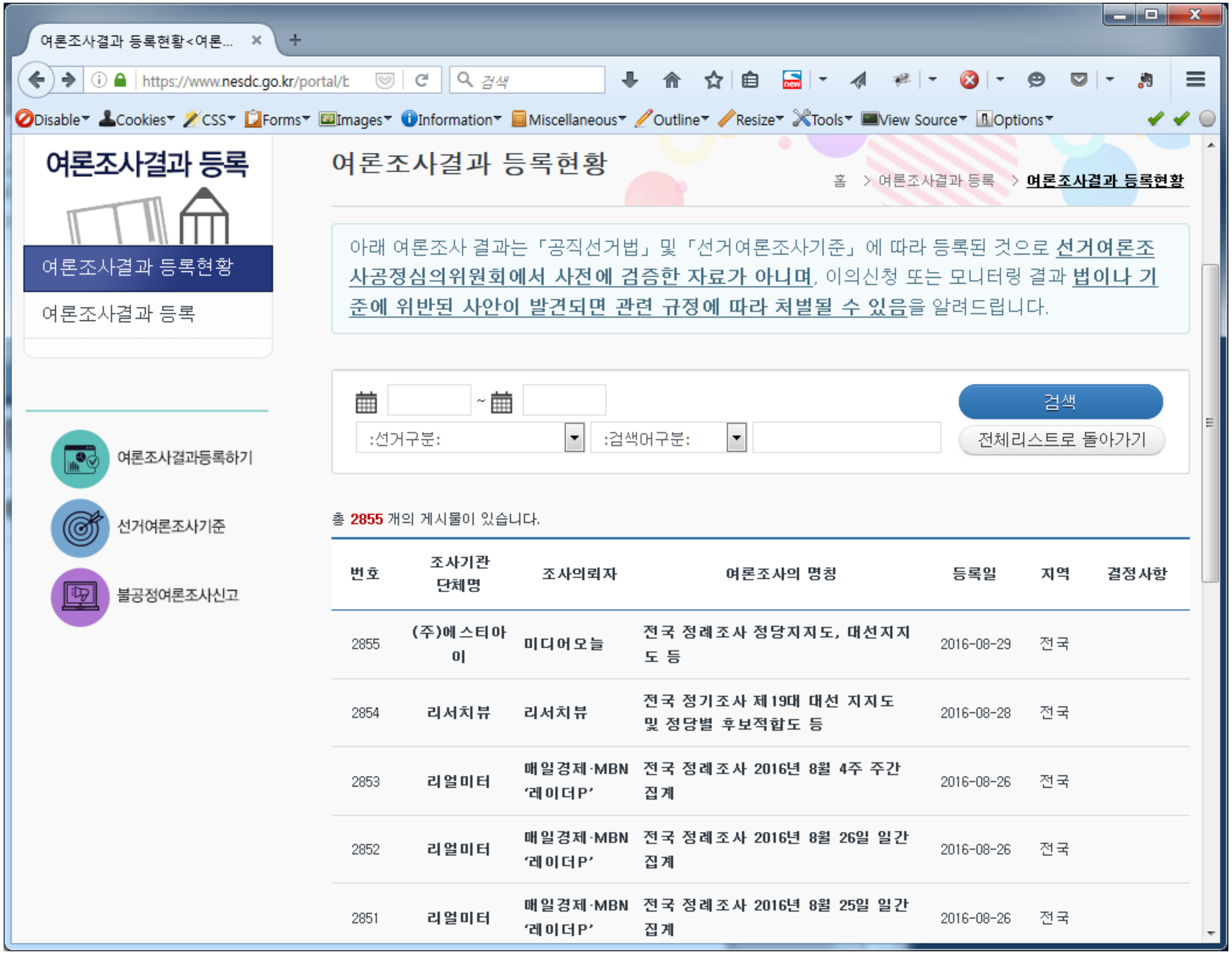
표본분포

개요



중앙선거여론조사공정심의위원회에 등록된 여론조사

개요

- 중앙선거여론조사 공정심의위원회
  - 중앙선거관리위원회의 선거여론조사 심의기구입니다.
  - 각 여론조사 기관들이 실시한 선거에 대한 여론조사 결과를 받아 심의하고, 그 내용을 홈페이지를 통해 알리고 있습니다.
  - 등록된 여론조사
    - 선거기간 : 특정 후보자와 정당을 얼마나 많은 유권자가 지지하는지
    - 국정지지도, 잠재적 대선후보군에 대한 지지도
    - 그 외 정치와 관련한 각종 주제들
- 선거여론조사
  - 모든 유권자를 조사하는 것이 가장 명확한 방법이지만,
  - 현실적으로 모든 유권자를 조사하는 것은 시간과 비용에 있어 지극히 힘듭니다.
  - 전체 유권자 집단을 잘 대표할만한 표본을 뽑아 조사합니다.
    - 조사자의 의도가 들어가지 않은 다양한 확률표본추출법을 사용합니다.

개요

- 여론조사 예
  - 위원회에 등록된 여론조사 개요를 함께 살펴봅시다.
    - 위원회가 공정성과 정확성을 위해 요구하는 기본 사항입니다.
    - 개요를 통해 여론조사 과정을 간략히 훑어봅시다.
  - 언론사에서 여론조사기관에 의뢰하여 전국을 대상으로 실시한 여론조사
    - 출처 : 중앙선거관리위원회 중앙선거여론조사공정심의위원회, http://goo.gl/LNGJbZ (단축주소), 매일경제, MBN(의뢰) 리얼미터(기관)
  - 조사의 명칭

등록 글번호		3043
여론조사의 명칭	선거구분	기타
	지역	전국
	선거명	정례조사 (2016년 8월 4주 주간집계)

개요

- 조사개요

조사지역	전국
조사일시	2016-08-22 13 시 - 19 시 2016-08-23 13 시 - 19 시 2016-08-24 13 시 - 19 시 2016-08-25 13 시 - 19 시 2016-08-26 13 시 - 19 시
조사대상 및 표본크기	전국에 거주하는 만 19세 이상 남녀 2,529명
성별·연령별 표본크기	**남성** 1722 명, **여성** 807 명 | 합계 : 2529명
	20대 이하 504 명, **30대** 440 명, **40대** 476 명, **50대** 485 명, **60대 이상** 624 명 | 합계 : 2529명

개요

- 조사방법 : 본 조사에서는 4가지 방법을 사용했으며, 그 중 한가지를 소개합니다.

조사방법 (2)		무선 ARS 27%
피조사자 선정 방법	표본추출틀	무선전화번호 기타 국번별, 0001~9999까지 랜덤 생성한 50만 전화 번호
	표본추출방법	RDD
	기타	151104개 번호 사용

- 표본추출틀은 표본 추출을 위한 모집단의 목록으로 이 조사에서는 무선전화번호를 사용하였음을 밝히고 있습니다.(1장의 미국대선여론조사의 표본추출틀과 비교)
- 표본추출방법인 RDD는 표본추출틀내에서 무작위(Random)로 전화번호 숫자(Digit)를 만들어 전화 연결 (Dialing) 하는 것을 말합니다.

개요

- 피조사자(표본) 접촉 현황
  - RDD를 통해 표본을 추출하고 각 표본들의 반응을 나타냅니다.
  - 연결이 안 된 사례, 거절 및 중간에 조사를 멈춘 사례수를 밝힙니다.
  - 응답완료된 사례수와 전체 연결 중 응답률을 밝힙니다.
    - 본 조사에서는 전화면접 방법이 18.2%로 가장 높았으며 전체 응답률은 9.8% 입니다.

피조사자 접촉 현황	비적격 사례수 (결번/사업체번호/팩스/대상지역 아님 /할당초과 등)	55810
	연결실패 사례수 (통화중/부재중 /접촉안됨)	83713
	연결 후 거절 및 중도 이탈 사례수(A)	10890
	연결 후 응답완료 사례수(B)	691
	합계	151104
	응답률(B/(A+B))	6%

개요

- 가중값 산출
  - 이와 같은 조사에서는 인구통계학적 특성 중 성별, 연령별, 지역별 응답이 실제 모집단 상황에 맞지 않아 가중치를 통해 보정을 합니다.
  - 가중치 산출 방법에 대한 많은 연구가 있으며, 조사기관의 연구자들이 본 조사와 가장 어울리는 방법을 사용했음을 밝힙니다.
    - 발표한 결과에 대한 과학적인 근거를 제시합니다.

<table>
<tr><td rowspan="4">가중값 산출 및 적용 방법
※ 추가가중은 기본가중 외에 과거선거 투표율 보정 등 추가적으로 수행했을 경우 등록</td><td rowspan="2">기본가중</td><td>산출방법</td><td>성별, 연령별, 지역별, 가중값 부여(2016년6월말 행정 자치부 주민등록 인구 기준)</td></tr>
<tr><td>적용방법</td><td>림가중</td></tr>
<tr><td rowspan="2">추가가중</td><td>산출방법</td><td></td></tr>
<tr><td>적용방법</td><td></td></tr>
</table>

개요

- 조사의 신뢰성과 여론조사 결과
  - 표본오차를 통해 조사의 신뢰도를 나타냅니다.
  - 이와 같이 조사된 결과를 '붙임자료'를 통해 공개합니다.

표본오차		95% 신뢰수준에 ±1.9%p
여론조사 결과	여론조사 결과 최초 공표·보도 예정일시	※ 붙임자료는 여론조사기관이 공개 지정한 **최초 공표·보도 예정일시(2016-08-29 07시 00분)**에서 24시간 후에 공개됩니다. 단, 「잡지 등 정기간행물의 진흥에 관한 법률」 제 2조에 따른 정기간행물에 여론조사 결과를 최초 공표·보도 하기로 한 경우는 최초 공표·보도 예정일시에서 48시간 후에 공개됩니다.
	붙임자료	[리얼미터] 주간집계 보도통계표 _ 2016년 8월 4주차 (22~26일)_최종.pdf
	결정 사항	

모수와 통계량

- 앞서 학습한 내용을 다시 한번 확인해 봅시다.
- 모수
  - 모집단의 특성을 나타내는 값입니다.
  - 예) 대한민국 유권자의 무당층 비율
    - 모집단 : 대한민국 유권자 전체
    - 모수 : 지지하는 정당이 없는 유권자의 비율
  - **모수는 알지 못하나 존재하는 값**으로 우리가 알고자 하는 대상이 됩니다.
- 통계량
  - 잘 알고 있다시피 관찰되는 표본의 특성입니다.
  - 통계량은 **수집된 표본에 따라 그 값이 달라집니다.**
  - 통계량에 표본으로부터 관찰된 값을 대입하여 구한 실측값을 "**통계**" 혹은 "**통계치**"라고 합니다.
  - 예) 대한민국 유권자의 A 정당에 대한 지지율 조사
    - 앞선 여론조사에서 표본 2,529명으로 부터 무당층은 19.5%로 조사되었습니다.

표본분포

- 표본조사를 실시하면 조사를 위해 표본을 모집단으로부터 한 번 추출하고 모집단에 대해 추출합니다.
  - 모집단의 크기가 N이고 표본의 크기가 n일 때 표본을 비복원으로 추출하는 경우의 수는 $\binom{N}{n}$ 가지로 모집단의 크기와 표본의 크기에 따라 다양합니다.
  - 표본분포는 표본의 크기가 n으로 정해졌을 때 추출될 수 있는 모든 표본으로부터 구한 통계량으로 구성된 확률분포입니다.
- 예) 다음과 같이 4장의 카드가 있을 때 2장의 카드를 뽑아 4장의 카드의 평균을 맞추는 게임이 있다고 해 봅시다.
  - 카드에는 10, 20, 30, 40 을 써 넣습니다. (평균은 25 입니다.)
  - 이제 게임에 참가하는 사람은 두 장의 카드를 뽑아 카드에 쓰여져 있는 숫자들로 평균을 맞추고자 합니다.

표본분포

- 참가자들이 4장 중 2장의 카드를 비복원추출로 뽑을 수 있는 경우의 수는 $\binom{4}{2} = \frac{4!}{2!\cdot 2!} = 6$으로 여섯 가지 경우가 있습니다.
  - 이 과정은 모집단이 모르는 숫자 4가지로 구성되어 있고, 이로부터 2개를 표본으로 뽑아 관찰하는 과정입니다.
- 여섯 가지 경우별로 평균(표본평균)을 구해보면 다음과 같습니다.

구분	경우 1		경우 2		경우 3		경우 4		경우 5		경우 6	
추출된 개별표본	10	20	10	30	10	40	20	30	20	40	30	40
표본평균 $(\bar{x})$	15		20		25		25		30		35	

표본분포

- 추출된 표본평균으로부터 모집단의 평균을 추측할 때
  - '경우 1'과 같이 표본평균($\bar{x} = 15$)이 모집단 평균($\mu = 25$)과 차이가 있을 때도 있고,
  - '경우 3'과 '경우 4'와 같이 표본평균이 모집단 평균과 일치할 때도 있습니다.
- 표본의 크기 n인 표본으로부터 구하는 표본평균 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$는 추출된 확률표본에 따라 값이 달라집니다.
  - 추출된 **확률표본에 따라 값이 결정되는 표본평균**은 표본평균 $\bar{x}$의 분포로부터 확률추출된 확률변수입니다.

표본분포

- 표본평균 $\bar{x}$의 분포
  - 4장의 카드에서 표본으로 2장의 카드를 뽑아서 구한 표본평균 $\bar{x}$ 의 분포를 구해봅시다.
    ① 표본으로 추출될 6가지의 경우 추출될 확률이 1/6 으로 동일합니다.
    ② 각 표본으로부터 구할 수 있는 표본평균 $\bar{x}$는 15, 20, 25, 30, 35의 5가지입니다.
    ③ 표본평균이 25가 될 확률은 '경우 3' 혹은 '경우 4'가 선택될 경우로 확률은 $\frac{1}{6}+\frac{1}{6}=\frac{2}{6}$ 입니다.
    ④ ③의 경우가 아닌 다른 표본평균이 나타날 확률은 $\frac{1}{6}$로 모두 동일합니다.
  - 이를 바탕으로 표본평균 분포의 확률분포와 그 특성을 다음의 표를 통해 확인해 봅시다.

표본분포

$\overline{X}=\overline{x}$	$p(\overline{X}=\overline{x})=p(\overline{x})$	❶ $\overline{x}\cdot p(\overline{x})$	❷ $\overline{x}^2\cdot p(\overline{x})$
15	$\frac{1}{6}$	$15\cdot\frac{1}{6}=\frac{15}{6}$	$15^2\cdot\frac{1}{6}=\frac{225}{6}$
20	$\frac{1}{6}$	$20\cdot\frac{1}{6}=\frac{20}{6}$	$20^2\cdot\frac{1}{6}=\frac{400}{6}$
25	$\frac{2}{6}$	$25\cdot\frac{2}{6}=\frac{50}{6}$	$25^2\cdot\frac{2}{6}=\frac{1250}{6}$
30	$\frac{1}{6}$	$30\cdot\frac{1}{6}=\frac{30}{6}$	$30^2\cdot\frac{1}{6}=\frac{900}{6}$
35	$\frac{1}{6}$	$35\cdot\frac{1}{6}=\frac{35}{6}$	$35^2\cdot\frac{1}{6}=\frac{1225}{6}$
합	1	$E(\overline{X})=\sum_{\overline{X}}\overline{x}\cdot p(\overline{x})=\frac{150}{6}=25$	$E(\overline{X}^2)=\sum_{\overline{X}}\overline{x}^2\cdot p(\overline{x})=\frac{4000}{6}$

표본분포

- 표본평균 $\bar{x}$ 분포의 기댓값과 분산을 구해봅시다.
  - ❶ 열의 합은 표본평균 $\bar{x}$ 분포의 기댓값이고, 그 값은 25로 모집단의 평균과 같습니다.
  - ❷ 열의 합은 $\bar{x}^2$ 의 기댓값으로, 이 값에서 기댓값의 제곱을 빼 표본평균 $\bar{x}$분포의 분산을 구합니다.
    - $Var(\bar{X}) = \frac{4000}{6} - 25^2 = \frac{4000}{6} - 625 = \frac{4000-3750}{6} = \frac{250}{6}$
    - 10, 20, 30, 40으로 구성된 모집단의 분산은 125 입니다.
    - N을 모집단의 수 n을 표본의 수, 모집단의 분산을 σ^2이라 할 때, $\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$을 계산해 봅시다. (모분산 $\sigma^2 = 125$)

$$\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} = \frac{4-2}{4-1} \cdot \frac{125}{2} = \frac{125}{3} = \frac{250}{6}$$

    - 이 값은 위에서 구한 표본평균 $\bar{x}$ 분포의 분산과 동일합니다.

표본분포

- 표본평균 $\bar{x}$ 분포의 기댓값과 분산
  - 비복원추출의 경우 : $E(\bar{X}) = \mu, \qquad Var(\bar{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$
  - 복원추출의 경우 : $E(\bar{X}) = \mu, \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$
  - 모집단의 크기 N이 표본크기 n에 비해 매우 크다면 $\frac{N-n}{N-1}$ 은 1에 가까워져 **근사적**으로 복원추출과 비복원추출의 표본평균 $\bar{x}$ 분포의 분산은 같아집니다
    - 일반적인 경우 모집단의 크기가 표본의 크기보다 매우 크므로 복원추출과 비복원추출로 인한 분산의 차이가 크지 않을 것으로 가정하며 이에 표본평균 $\bar{x}$ 가 이루는 분포의 특성을 다음과 같이 정리합시다.

    ① 기댓값은 모집단의 평균과 같습니다 : $E(\bar{X}) = \mu$

    ② 분산은 모분산을 표본의 수로 나눈 값과 같으며 : $Var(\bar{X}) = \frac{\sigma^2}{n}$

    ③ 표준편차는 분산의 제곱근입니다 : $sd(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

표본분포

- 표본평균 $\bar{x}$ 분포의 기댓값과 표준편차(분산)의 의미
  - 기댓값
    - 표본조사에서는 여러 번에 걸쳐 동일한 크기의 표본을 추출하는 것이 아닌 단 한 번 추출한 표본을 통해 모집단의 특성을 유추합니다.
    - 추출된 표본으로부터 구한 표본평균은 표본평균 $\bar{x}$의 분포에서 확률추출한 것으로 생각할 수 있습니다.
      - 4장의 카드에서 2장을 뽑는 예제에서 2장의 카드를 확률추출하여 (1, 2)가 나온 것은 표본평균 $\bar{x}$ 분포에서 1.5인 값을 확률추출한 것과 동일한 의미입니다.
    - 표본평균의 기댓값이 모집단의 평균과 같다는 성질은 표본을 추출하기에 앞서 추출된 표본으로부터 구한 표본평균이 모집단의 평균과 같을 것으로 기대할 수 있음을 나타냅니다.
  - 표준편차(분산)
    - 각 표본평균들이 기댓값(모집단 평균)에 대해 얼마나 흩어져 있는지를 나타냅니다.
      - 이 값이 작을 경우 표본을 통해 관찰한 표본평균이 모집단의 평균과 차이가 날 확률이 작을 것으로 봅니다
      - 표준편차($\frac{\sigma}{\sqrt{n}}$)를 반으로 줄이기 위해서는 표본의 수를 4배로 늘려야 합니다.

표본분포

- 표준정규분포로부터 표본 크기가 10과 40인 표본을 각각 1,000번씩 추출하고, 이로부 터 평균을 구해 특성을 살펴봅시다.
- Step #1) 표본의 크기별로 표본평균이 저장될 변수들을 초기화합니다.

```
1. m10 <- rep(NA, 1000)
2. m40 <- rep(NA, 1000)
```

  - 1, 2줄 : m10과 m40을 각각 결측값(NA) 1,000개로 구성된 벡터로 만듭니다.
    - 표본의 크기에 따라 1000번 씩 추출하는 과정에 각 표본평균이 저장될 공간을 미리 만들어 놓습니다.
    - 초기값으로 NA외에도 NULL 을 사용할 수 있습니다.

표본분포

- Step#2) 반복문을 이용하여 표본의 크기별로 1,000번씩 추출하고 각 표본의 평균을 저장합니다.

```
3. set.seed(9)
4. for( i in 1:1000) {
5.   m10[i] <- mean(rnorm(10))
6.   m40[i] <- mean(rnorm(40))
7. }
```

- 3줄 : 난수생성의 초깃값을 9로 지정합니다.
- 4, 7줄 : 1:1000으로 생성되는 벡터의 원소 수만큼 반복문을 만듭니다.
  - 1:1000으로 생성된 벡터의 크기만큼 5, 6번째 줄을 반복합니다. (1000번)
- 5줄 : 표준정규분포로부터 10개의 표본을 추출하고, 그 평균을 m10의 i번째 원소에 저장합니다.
  - 표준정규분포의 경우 rnorm() 함수에 평균과 표준편차를 지정하지 않아도 됩니다. (기본값)
- 6줄 : 표준정규분포로부터 40개의 표본을 추출하고, 그 평균을 m40의 i번째 원소에 저장합니다.

표본분포

- Step #3) 표본평균의 평균과 표준편차를 구합니다.

```
9. options(digits=4)
10.c(mean(m10), sd(m10))
11.c(mean(m40), sd(m40))
```

  - 9줄 : 출력물의 자릿수를 4로 합니다.
  - 10줄 : 표본 크기가 10인 표본평균 분포의 평균과 표준편차를 출력합니다.
  - 11줄 : 표본 크기가 40인 표본평균 분포의 평균과 표준편차를 출력합니다.
  - 표준정규분포로부터 추출한 표본평균의 분포는 그 평균이 0에 가깝고, 표준편차는 표본 크기가 커짐에 따라 줄어듭니다. 표본 크기가 10일 때보다 40일 때 절반가까이 줄어들었습니다(0.303과 0.161).

```
> c(mean(m10), sd(m10))
[1] -0.01214  0.30311
> c(mean(m40), sd(m40))
[1] 0.004212 0.160942
```

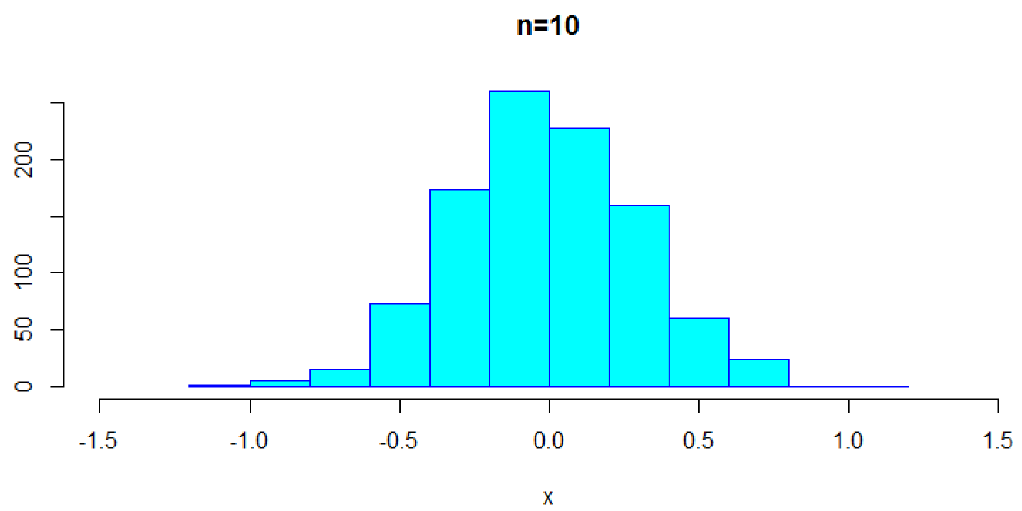
표본분포

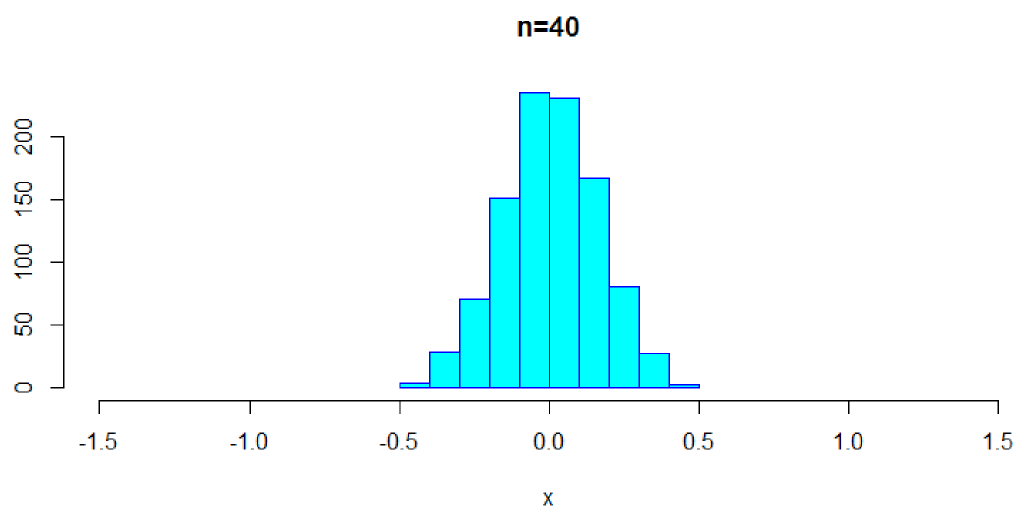
- Step #4) 표본 크기에 따른 표본평균 분포의 변화를 살펴봅니다.

```
13.hist(m10, xlim=c(-1.5, 1.5), main="n=10", xlab="x",
             ylab="", col="cyan", border="blue")
14.hist(m40, xlim=c(-1.5, 1.5), main="n=40", xlab="x",
             ylab="", col="cyan", border="blue")
```

- 13줄 : 표본 크기가 10인 표본평균들의 분포를 히스토그램으로 그립니다.
  - col로 전달되는 값으로 히스토그램의 막대 색을 지정합니다.
  - border 는 전달되는 값으로 히스토그램의 막대별로 경계선의 색을 지정합니다.
- 14줄 : 표본 크기가 40인 표본평균들의 분포를 히스토그램으로 그립니다.
- hist() 함수에 xlim을 통해 전달되는 전달인자는 그래프의 x축 범위를 (최솟값, 최댓값) 의 벡터로 전달합니다.
  - 이를 통해 두 히스토그램의 x축을 고정하고 표본의 크기별로 표본평균 $\bar{x}$ 분포의 퍼진 정도를 확인해 봅시다.

표본분포





표본 크기가 클수록
기댓값(모집단 평균) 주변에
많이 몰려 있으며
자료가 분포하는 전체 폭이
줄어듦을 알 수 있습니다.