강의교안 이용 안내

- 본 강의교안의 저작권은 **이윤환**과 **한빛아카데미(주)**에 있습니다.
- 이 자료를 무단으로 전제하거나 배포할 경우 저작권법 136조에 의거하여 벌금에 처할 수 있고 이를 병과(倂科)할 수도 있습니다.

제대로 알고 쓰는

R 통계분석

CHAPTER 04

표본분포

한빛아카데미 Hanbit Academy, Inc.

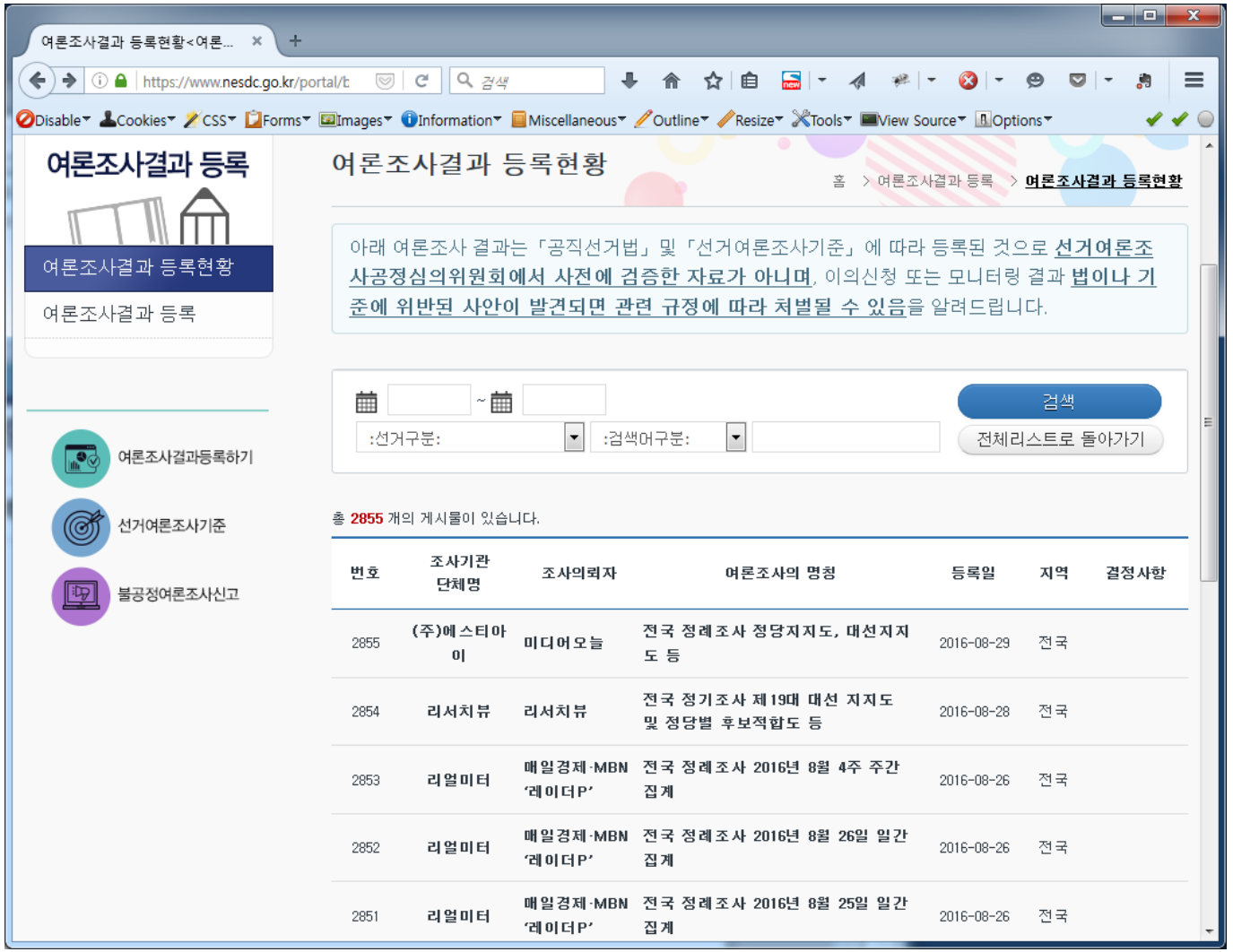
Contents

01. 표본분포

: 표본들로부터 모집단의 특성을 유추하는 배경

1. 표본들로부터 추출되는 특성이 표본추출에 따라 분포함을 학습한다.
2. 표본평균 $\bar{x}$의 분포에서 기댓값과 분산에 대해 학습한다.

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
여론조사의명칭	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

- 조사지역, 조사일시, 조사대상 및 표본의 크기를 명시하였습니다.
- 표본의 크기와 조사된 표본의 성별 및 연령대별 응답수를 명시하였습니다.

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

- 전체 중 27%는 무선 ARS 방법을 이용하여 조사하였음을 밝히고 있습니다.
- 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.

가중값 산출 및 적용 방법 ※ 추가가중은 기본가중 외에 과거선거 투표율 보정 등 추가적으로 수행했을 경우 등록	기본가중	산출방법	성별, 연령별, 지역별, 가중값 부여(2016년6월말 행정 자치부 주민등록 인구 기준)
		적용방법	림가중
	추가가중	산출방법	
		적용방법	

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 : $E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$
  - 복원추출의 경우 : $E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{\sigma^2}{n}$
  - 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

예제 4-1 표본평균 $\bar{x}$의 분포

준비파일 | 01.sampling.distribution.R

- 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
```

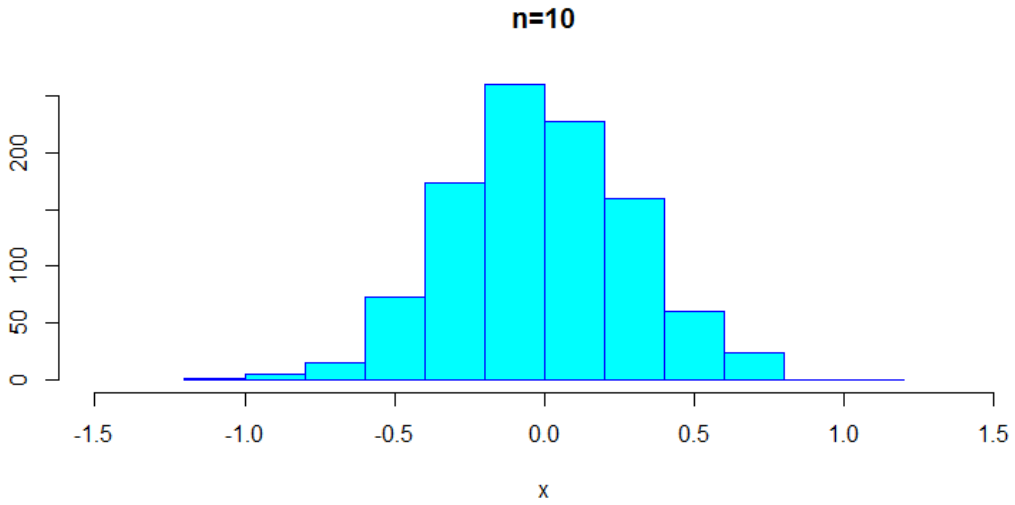
표본분포

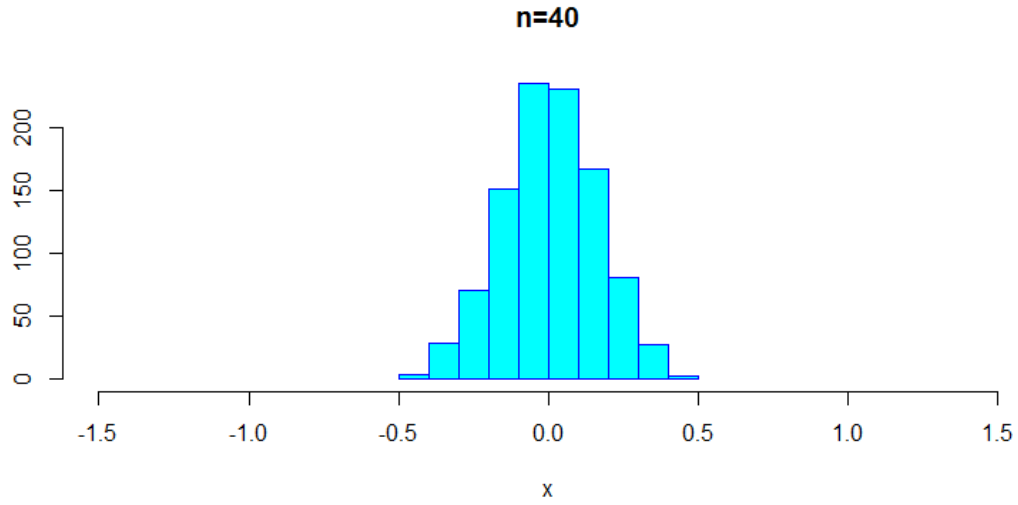
- **Step #4)** 표본 크기에 따른 표본평균 분포의 변화를 살펴봅니다.

```
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.

02. 중심극한 정리

: 표본평균의 분포가 궁금해요

1. 표본평균의 분포가 따르는 분포에 대해 학습한다.
2. 중심극한정리에 대해 학습한다.

중심극한정리

- **표본평균 $\bar{x}$ 분포는 어떤 분포를 따를까요?**
  - 앞서 표본평균 $\bar{x}$ 분포의 중요한 특성인 기댓값과 분산(표준편차)에 대해 알아봤습니다.
  - 그렇다면 표본평균 $\bar{x}$ 분포는 어떤 모양이 될지 알아봅시다.
  - 이 과정은 상급과정에서 수리적으로 복잡한 계산을 통해 증명하지만, 우리는 R을 통해 그래프를 그려가면서 어떤 분포와 닮아가는지 확인해 봅시다.
    - 먼저 아주 특수한 경우로 모집단이 정규분포를 따를 때 이로부터 추출한 표본들의 표본평균 $\bar{x}$ 분포가 어떤 분포를 따를지 살펴봅니다.
    - 그 다음으로 좀 더 일반적인 상황으로 모집단의 분포가 임의의 분포일 때 어떤 분포를 따를지 살펴보겠습니다.

중심극한정리

예제 4-2 정규분포로부터 추출된 표본평균 $\bar{x}$의 분포

준비파일 | 02.sampling.ND.R

- **모집단이 정규분포일 때**
  - 모집단이 정규분포일 때 이로부터 추출된 표본들의 표본평균의 분포는 어떤 모양을 따를지 살펴보겠습니다.
  - 서로 다른 두 정규분포에서 4개의 표본으로부터 평균을 구하는 것을 1,000번 실시하여 표본평균의 분포가 어떤 형태를 따르는지 확인해봅니다
- **Step#1)** 준비과정

```
1. set.seed(9)
2. n <- 1000
3. r.1.mean <- rep(NA, n)
4. r.2.mean <- rep(NA, n)
```

중심극한정리

- 1줄 : 난수생성의 초깃값을 9로 고정합니다.
- 2줄 : 표본추출 횟수 1,000을 변수 n에 저장합니다.
- 3, 4줄 : 모집단별로 표본평균이 저장될 두 변수 r.1.mean과 r.2.mean을 결측값(NA)으로 초기화합니다.

- **Step #2)** 두 정규분포 $N(3, 1^2)$과 $N(170, 6^2)$ 으로부터 표본 크기가 4인 표본을 1,000번 추출하고, 각 추출마다 평균을 저장합니다.

```
5. for (i in 1:n ) {
6.   r.1.mean[i] <- mean( rnorm(4, mean=3, sd=1) )
7.   r.2.mean[i] <- mean( rnorm(4, mean=170, sd=6) )
8. }
```

중심극한정리

- 5, 8줄 : 1:1000으로 생성되는 벡터의 원소 수만큼 반복문을 만듭니다. 이 반복문으로 인해 6, 7번째 줄을 1,000번 반복합니다.
- 6줄 : $N(3, 1^2)$ 으로부터 4개의 표본을 추출하고, 그 평균을 r.1.mean의 i번째 원소에 저장합니다.
- 7줄 : $N(170, 6^2)$ 으로부터 4개의 표본을 추출하고, 그 평균을 r.2.mean의 i번째 원소에 저장합니다.

- **Step #3)** 표본평균들의 분포에서 평균과 표준편차를 구합니다.

```
10.options(digits=4)
11.c(mean(r.1.mean), sd(r.1.mean))
12.c(mean(r.2.mean), sd(r.2.mean))
```

- 10줄 : 출력물의 자릿수를 4로 합니다.
- 11, 12줄 : 각 정규분포로부터 추출된 표본 크기가 4인 표본평균 분포의 평균과 표준편차를 출력합니다.

중심극한정리

```
> c(mean(r.1.mean), sd(r.1.mean))
[1] 3.0214 0.5096
> c(mean(r.2.mean), sd(r.2.mean))
[1] 170.032   2.835
```

- 표준정규분포로부터 추출한 표본평균의 분포는 그 평균이 모집단 평균에 가깝고, 표준편차는 모집단 정규분포의 표준편차를 표본 크기의 제곱근으로 나눈 값($\frac{\sigma}{\sqrt{4}}$, 모집단 표준편차의 반) 과 비슷합니다.

- **Step #4)** 표본평균의 분포에 대한 히스토그램을 그리고, 그 위에 각 표본평균의 분포가 따를 것으로 생각되는 분포의 확률도표를 그려봅니다.
  - 모집단이 정규분포일 때 이로부터 추출한 표본평균의 분포는 또 다른 정규분포를 따르는 것으로 알려져 있습니다.
    - 정규분포로 부터 추출된 경우 알려진 표본평균의 분포 : $\bar{X} \sim N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$

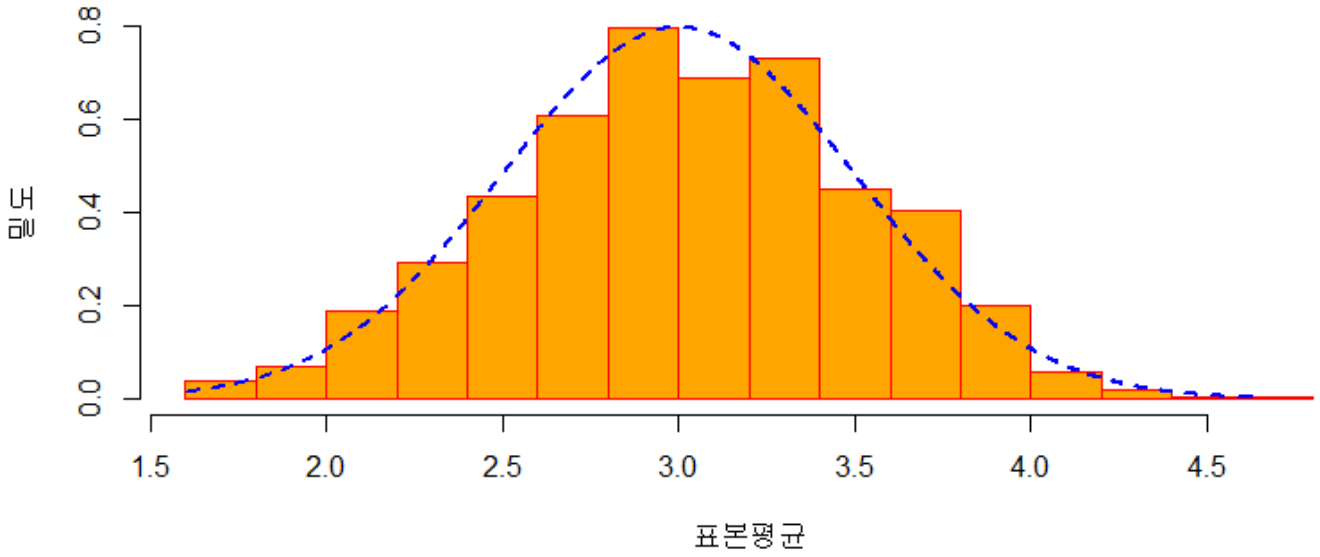
중심극한정리

```
14.hist(r.1.mean, prob=TRUE, xlab="표본평균", ylab="밀도", main="",
        col="orange", border="red")
15.x1 <- seq(min(r.1.mean), max(r.1.mean), length=1000)
16.y1 <- dnorm(x=x1, mean=3, sd=(1/sqrt(4)))
17.lines(x1, y1, lty=2, lwd=2, col="blue")


18.hist(r.2.mean, prob=TRUE, xlab="표본평균", ylab="밀도", main="",
        col="orange", border="red")
19.x2 <- seq(min(r.2.mean), max(r.2.mean), length=1000)
20.y2 <- dnorm( x=x2, mean=170, sd=(6/sqrt(4)) )
21.lines(x2, y2, lty=2, lwd=2, col="blue")
```

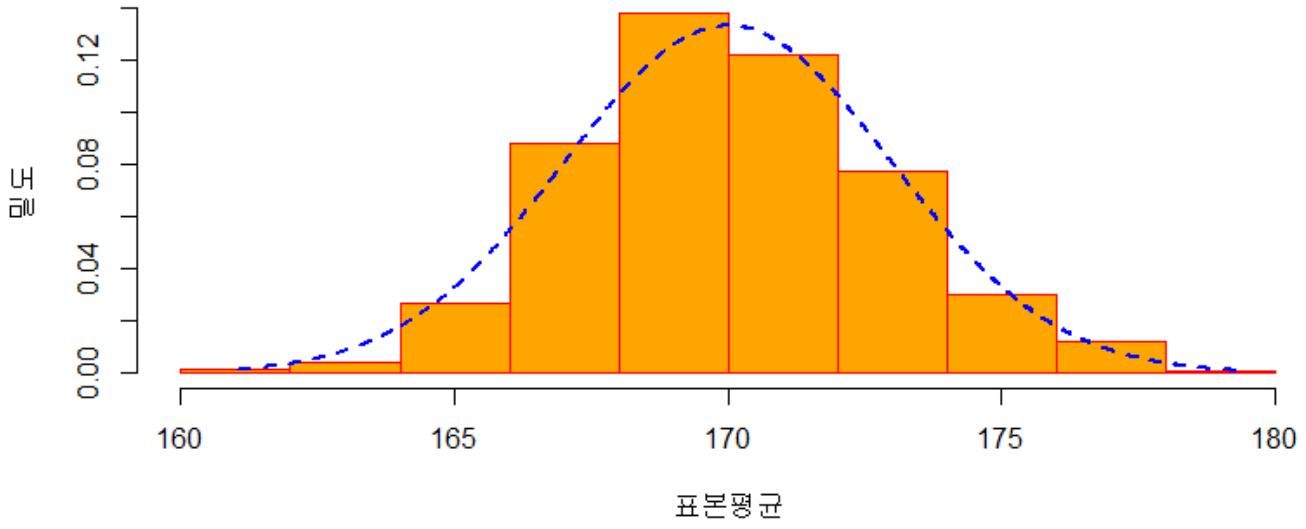
중심극한정리

- 14~17줄
  - $N(3, 1^2)$ 으로부터 표본 크기를 4로 하는 표본평균의 분포에서 평균은 모집단의 평균인 3이고, 표준편차는 $\frac{1}{\sqrt{4}}$입니다.
  - 표본평균의 히스토그램과 평균이 3이고 표준편차가 $\frac{1}{\sqrt{4}}$ 인 정규분포와 비교해 봅시다.
  - hist() 함수에서 prob으로 TRUE을 전달하면, 빈도에 대한 히스토그램이 아닌 상대빈도에 대한 히스토그램을 작성합니다.



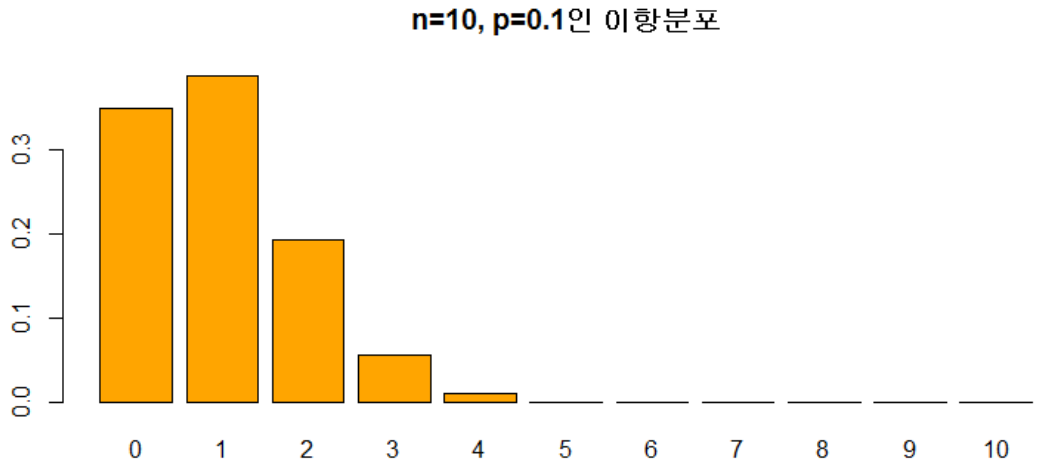
중심극한정리

- 19~22줄
  - $N(170, 6^2)$ 으로부터 표본 크기를 4로 하는 표본평균의 분포에서 평균은 모집단의 평균인 170이고, 표준편차는 $\frac{6}{\sqrt{4}}$입니다.
  - 표본평균의 히스토그램과 평균이 170이고 표준편차가 $\frac{6}{\sqrt{4}}$ 인 정규분포와 비교해 봅시다.



중심극한정리

- **모집단이 정규분포가 아닌 임의의 분포일 때**
  - 조건 : 모집단의 평균과 표준편차가 (그 값을 알지 못하나) 존재합니다.
  - 예) 모집단이 시행의 횟수가 10이고 성공의 확률이 0.1인 이항분포($B(10, 0.1)$)
    - $B(10, 0.1)$ 은 꼬리가 오른쪽으로 늘어진 모양을 갖습니다.
    - 기댓값 : $np = 10 \times 0.1 = 1$
    - 표준편차 : $\sqrt{np(1-p)} = \sqrt{10 \times 0.1 \times 0.9} \approx 0.9487$
    - 표본의 크기가 2, 4, 32로 늘려가면서 표본평균의 분포를 관찰합니다.



중심극한정리

예제 4-3 임의의 분포에서 추출된 표본평균 $\bar{x}$의 분포 준비파일 | 03.sampling.BD.R

- **Step #1) 자료 준비**

```
7. set.seed(9)
8. t <- 10
9. p <- 0.1
10.x <- 0:10
11.n <- 1000
12.b.2.mean <- rep(NA, n)
13.b.4.mean <- rep(NA, n)
14.b.32.mean <- rep(NA, n)
```

- 7~10줄 : 난수생성의 초기값을 9로, 시행 횟수 10을 변수 t에, 성공 확률 0.1을 변수 p에 저장하고 시행횟수가 10인 이항분포로부터 관찰 가능한 값을 변수 x에 저장합니다.
- 11줄 : 표본을 추출할 횟수 1,000을 변수 n에 저장합니다.
- 12~14줄 : 표본 크기에 따라 1,000번의 표본추출에서 관찰된 표본평균이 저장될 변수 b.2.mean, b.4.mean과 b.32.mean에 대해 각각 1,000개의 NA 값을 갖는 벡터로 준비합니다.

중심극한정리

- **Step #2) 표본 크기별로 1000번의 표본추출로 표본평균을 구합니다.**

```
16.for(i in 1:n) {
17.  b.2.mean[i] <- mean( rbinom(2, size=t, prob=p) )
18.  b.4.mean[i] <- mean( rbinom(4, size=t, prob=p) )
19.  b.32.mean[i] <- mean( rbinom(32, size=t, prob=p) )
20.}
```

  - 16, 20줄 : 17~19줄을 1000번 반복(1000번의 표본추출)하는 반복문입니다.
  - 17줄 : $B(10, 0.1)$로부터 2개의 표본을 추출하고,
    그 평균을 b.2.mean의 i번째 원소에 저장합니다.
  - 18줄 : $B(10, 0.1)$로부터 4개의 표본을 추출하고,
    그 평균을 b.4.mean의 i번째 원소에 저장합니다.
  - 19줄 : $B(10, 0.1)$로부터 32개의 표본을 추출하고,
    그 평균을 b.32.mean의 i번째 원소에 저장합니다.

중심극한정리

- **Step #3)** 표본평균들의 분포에서 평균과 표준편차를 구합니다

```
22.options(digits=4)
23.c(mean(b.2.mean), sd(b.2.mean))
24.c(mean(b.4.mean), sd(b.4.mean))
25.c(mean(b.32.mean), sd(b.32.mean))
```

  - 22줄 : 출력물의 자릿수를 4로 합니다.
  - 23~25줄 : $B(10, 0.1)$ 로부터 1,000번 추출된 표본 크기가 2, 4, 32인 표본평균 분포의 평균과 표준편차를 출력합니다.
    - 출력물에서 이항분포로부터 추출한 표본평균의 분포는 그 평균이 이항분포의 평균과 비교해 봅시다.
    - 표준편차는 모집단 이항분포의 표준편차를 표본 크기의 제곱근으로 나눈 값들과 비교해 봅시다.
      - $\frac{0.9487}{\sqrt{2}} \approx 0.6708, \frac{0.9487}{\sqrt{4}} \approx 0.4743, \frac{0.9487}{\sqrt{32}} \approx 0.1677$

중심극한정리

```
> c(mean(b.2.mean), sd(b.2.mean))
[1] 1.0090 0.6763
> c(mean(b.4.mean), sd(b.4.mean))
[1] 1.006 0.481
> c(mean(b.32.mean), sd(b.32.mean))
[1] 0.9989 0.1624
```

- **Step #4)** 각 표본평균 분포의 히스토그램을 그리고, 그 위에 각 표본평균의 분포가 따를 것으로 알려진 정규분포의 확률도표를 작성합니다.
  - 앞서 모집단이 정규분포일 경우와 마찬가지로 정규분포를 따를 것으로 생각해봅시다. $\bar{X} \sim N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$
    - n=2 일 때는 $N(1, \left(\frac{0.9473}{\sqrt{2}} \approx 0.6708\right)^2)$, n=4 일 때는 $N(1, \left(\frac{0.9473}{\sqrt{4}} \approx 0.4743\right)^2)$, n=32 일 때는 $N(1, \left(\frac{0.9473}{\sqrt{32}} \approx 0.1677\right)^2)$

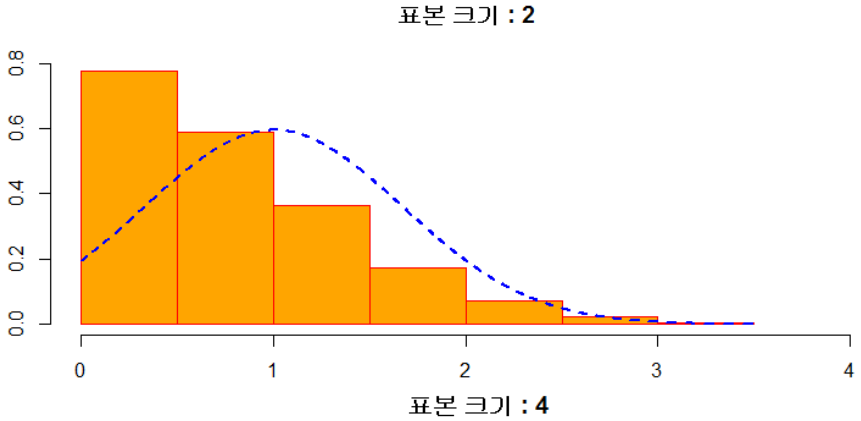
중심극한정리

```
27.hist(b.2.mean, prob=T, xlim=c(0, 4), main="표본 크기 : 2",
                ylab="", xlab="", col="orange", border="red")
28.x1 <- seq(min(b.2.mean), max(b.2.mean), length=1000)
29.y1 <- dnorm( x=x1, mean=1, sd=0.9/sqrt(2) )
30.lines(x1, y1, lty=2, lwd=2, col="blue")
31.
32.hist(b.4.mean, prob=T, xlim=c(0, 4), ylim=c(0, 1.2),
   main="표본 크기 : 4", ylab="", xlab="", col="orange", border="red")
33.x2 <- seq(min(b.4.mean), max(b.4.mean), length=1000)
34.y2 <- dnorm( x=x2, mean=1, sd=0.9/sqrt(4) )
35.lines(x2, y2, lty=2, lwd=2, col="blue")
36.
37.hist(b.32.mean, prob=T, xlim=c(0, 4), main="표본 크기 : 32",
   ylab="", xlab="", col="orange", border="red")
38.x3 <- seq(min(b.32.mean), max(b.32.mean), length=1000)
39.y3 <- dnorm( x=x3, mean=1, sd=0.9/sqrt(32) )
40.lines(x3, y3, lty=2, lwd=2, col="blue")
```

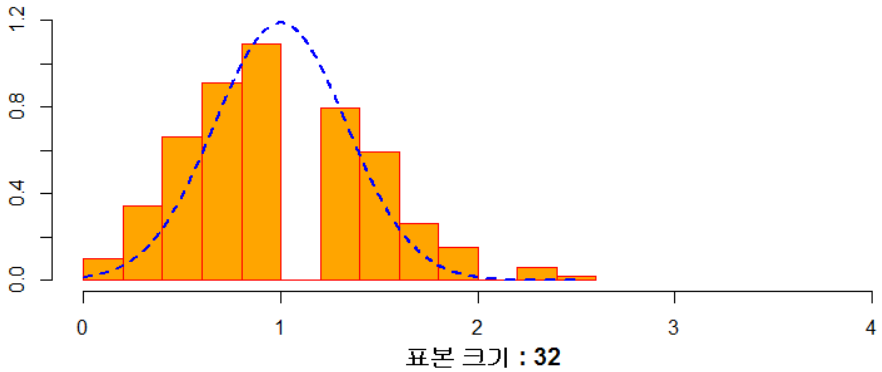
중심극한정리

- 27~30줄 : $B(10, 0.1)$ 로부터 표본 크기를 2로 하는 표본평균의 분포의 평균은 모집단의 평균인 1이고, 표준편차는 $\frac{0.9473}{\sqrt{2}} \approx 0.6708$을 가집니다.
  - 평균이 1이고 표준편차가 약 0.6708인 정규분포와 비교해봅시다
- 32~35줄 : $B(10, 0.1)$ 로부터 표본 크기를 4로 하는 표본평균의 분포의 평균은 모집단의 평균인 1이고, 표준편차는 $\frac{0.9473}{\sqrt{4}} \approx 0.4743$ 을 가집니다.
  - 평균이 1이고 표준편차가 약 0.4743인 정규분포와 비교해봅시다
- 37~40줄 : $B(10, 0.1)$ 로부터 표본 크기를 32로 하는 표본평균의 분포의 평균은 모집단의 평균인 1이고, 표준편차는 $\frac{0.9473}{\sqrt{32}} \approx 0.1677$을 가집니다.
  - 평균이 1이고 표준편차가 약 0.1677인 정규분포와 비교해봅시다.

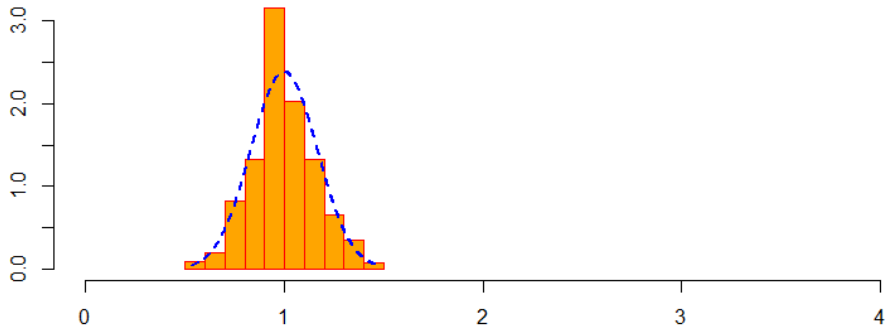
중심극한정리



여전히 표본평균의 분포는
오른쪽으로 늘어져 있습니다.



n=2 일 때 보다는 비교적 좌우대칭으로
보이지만, 중간중간 빈 구간이 보입니다.



n 이 증가할 수록 좌우대칭을 보이고,
점점 정규분포와 닮아갑니다.

중심극한정리

- **중심극한정리**
  - 표본의 개수가 증가할수록 표본평균의 분포가 정규분포와 닮아감을 확인해 보았습니다.
  - 이와 같은 성질을 수리적으로 밝혀낸 것이 중심극한정리입니다.
    - 모집단의 분포와 상관없이 평균과 표준편차가 μ와 σ로 존재하는 모집단에서 추출할 때 표본의 크기 n이 충분히 크면, 표본평균의 분포가 근사적으로 정규분포를 따릅니다.
  - 중심극한정리는 모집단의 분포에 대한 사전 지식 없이도 표본평균의 분포를 알 수 있게 하여 통계학에서 유용하게 사용합니다.

중심극한정리

참고 **중심극한정리(CLT, Central Limit Theorem)**

모집단의 분포와 상관없이 모집단의 평균 μ와 표준편차 σ가 존재할 때 표본 크기 n이 충분히 크다면, 표본평균의 분포는 다음과 같이 근사적으로 정규분포를 따릅니다.

$$\overline{X} \simeq N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) \tag{4.4}$$

또한, 표본평균의 분포가 정규분포를 따르므로 다음과 같이 표준화하여 사용할 수 있습니다.

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1^2) \tag{4.5}$$

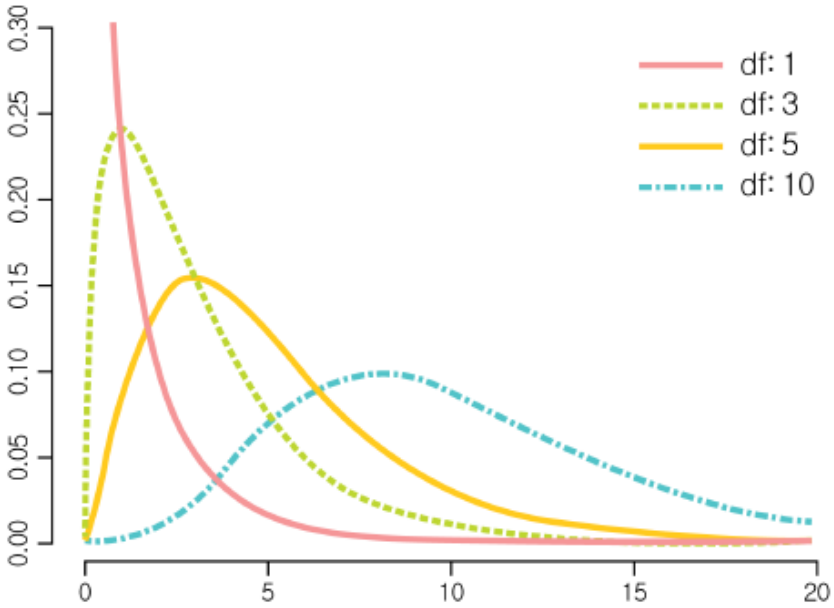
03. 다양한 표본분포

: 상황에 맞는 도구를 사용합시다!

1. 다양한 표본분포에 대해 학습한다.
2. 다양한 표본분포에 대한 R 함수를 확인해본다.

χ^2-분포

- **χ^2-분포**
  - 표본분산과 관련이 있는 분포
  - χ^2-분포의 모수 : 자유도 k
    - 자유도가 작을 때 꼬리가 오른쪽으로 길게 늘어지는 형태이고,
      자유도가 증가할 수록 정규분포와 유사하게 평균을 중심으로 좌우대칭 형태를 갖습니다.



χ^2-분포

- **χ^2-분포**
  - 표준정규분포로부터 독립적으로 추출한 k개의 확률표본 $Z_1, Z_2, \ldots, Z_k$에 대해,
    - 각각 확률표본의 제곱은 각각 자유도가 1인 χ^2-분포를 따릅니다. $(Z_i{}^2 \sim \chi^2(1))$
    - 확률표본들의 제곱의 합 $Z_1{}^2 + Z_2{}^2 + \ldots + Z_k{}^2 = \sum_{i=1}^{k} Z_i{}^2$은 자유도가 k인 χ^2-분포를 따릅니다.
    - 자유도가 k인 χ^2-분포의 기댓값과 분산은 다음과 같습니다. $(X \sim \chi^2(k))$
      - $E(X) = k, \quad Var(X) = 2k$

χ^2-분포

- **표본분산과 χ^2-분포**

① $X_1, X_2, \ldots, X_n$ 은 정규분포 $N(\mu, \sigma^2)$ 으로부터 추출한 n개의 확률표본입니다.

② 표본평균을 $\bar{X}$, 표본분산을 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 이라 하고, 확률변수 $V = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}$ 이라 할 때,

③ ②에 의해, $V = \frac{(n-1)S^2}{\sigma^2}$ 이 됩니다. $((n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2)$

- V는 자유도가 (n-1) 인 χ^2-분포를 따릅니다. $(V \sim \chi^2(n-1))$

④ 표본분산 S^2의 기댓값

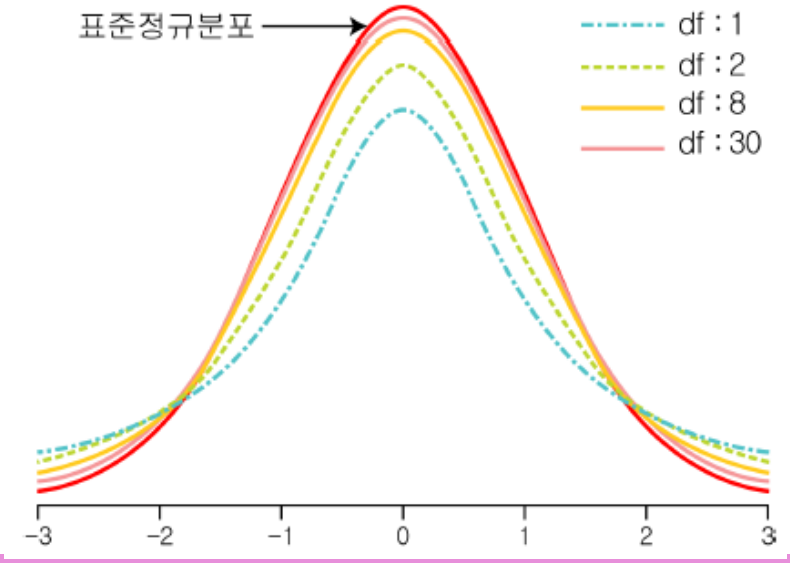
- $V = \frac{(n-1)S^2}{\sigma^2}$ 으로부터 $S^2 = \frac{\sigma^2}{n-1}V$ 이고 기댓값은 다음과 같습니다.

$$E(S^2) = E\left(\frac{\sigma^2}{n-1}V\right) = \frac{\sigma^2}{n-1}E(V) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2, \quad V \sim \chi^2(n-1), E(V) = n-1$$

- 표본분산 S^2의 기댓값은 모분산 σ^2 입니다.

t-분포

- ***t*-분포**
  - 정규분포처럼 평균을 중심으로 좌우 대칭입니다.
  - ***t*-분포의 모수는 자유도이며, 자유도에 따라 분포의 모양이 달라집니다.**
    - ***t*-분포는 정규분포와 비슷한 형태지만, 평균 주변에서 상대적으로 밀도가 낮고 양 끝으로 갈수록 꼬리 부분이 두툼한 형태를 갖습니다.**
    - **또한 자유도가 증가할수록 표준정규분포를 닮아갑니다.**



t-분포

- **t-분포**
  - 표본평균의 분포와 관련이 있습니다.
  - 두 개의 확률변수 Z와 V가 각각 표준정규분포와 자유도가 k인 χ^2-분포를 따르고($Z \sim N(0,1^2), V \sim \chi^2(k)$) 서로 독립인 경우 통계량 T를 다음과 같이 정의할 때,

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

  - 통계량 T는 자유도가 k인 t-분포를 따릅니다($T \sim t(k)$).
  - **t-분포의 기댓값과 분산 : $X \sim t(k)$**
    - $E(X) = 0\,, \quad Var(X) = \frac{k}{k-2}, \quad (k > 2)$

t-분포

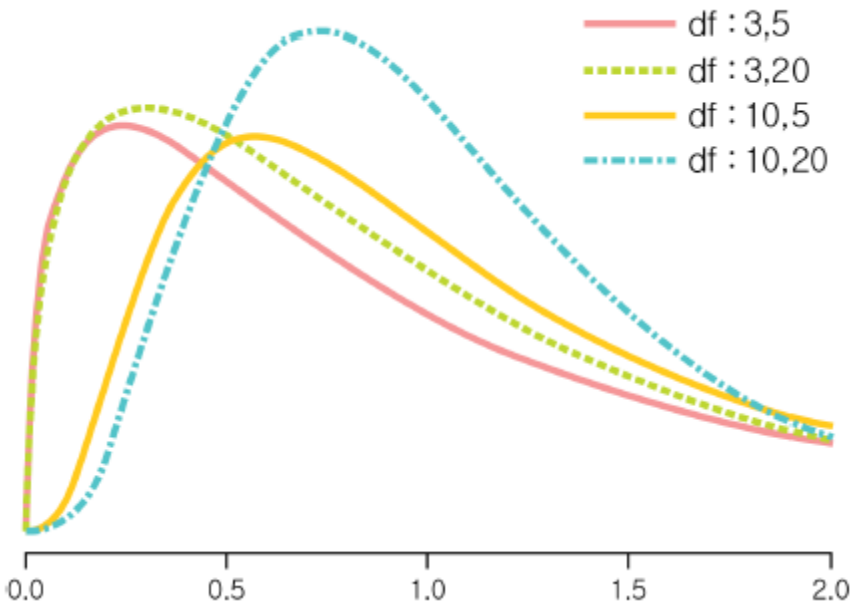
- **표본평균의 분포로써의 t-분포**
  - 정규분포로부터 추출된 n개의 확률표본 $X_1, X_2, \dots, X_n$ 의 평균들의 분포는 정규분포를 따름($N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$)을 앞에서 살펴봤습니다.
    - 표본평균들의 분포는 정규분포를 따르므로 $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$
  - 정규분포로부터 추출된 n개의 확률표본 $X_1, X_2, \dots, X_n$의 표본평균을 $\bar{X}$, 표본분산을 S^2(표준편차 S) 이라 하면 다음의 통계량 T는 자유도가 (n-1)인 t분포를 따릅니다.

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

    - 정규분포의 분산(표준편차)를 알지 못하는 경우 정규분포를 이용한 계산을 할 수 없어 표본들의 분산(표준편차)을 이용한 t-분포를 사용합니다.
    - 표본의 개수가 클수록 t-분포의 자유도가 증가하여 표준정규분포로 근사한 계산을 할 수 있습니다.

F -분포

- **F-분포**
  - 두 집단의 분산을 비교할 경우에 유용하게 사용하는 분포입니다.
  - 두 χ^2-분포를 이용하는 분포로 두 χ^2-분포의 모수들을 모수로 사용합니다.



F -분포

- **F-분포**
  - 서로 독립인 두 개의 확률변수 V_1, V_2가 각각 자유도가 k_1, k_2 인 χ^2-분포를 따르고($V_1 \sim \chi^2(k_1), V_2 \sim \chi^2(k_2)$), 각각의 확률변수를 각각의 자유도로 나눈 통계량 F를 다음과 같이 정의할 때

$$F = \frac{V_1/k_1}{V_2/k_2}$$

  - 통계량 F는 자유도가 (k_1, k_2)인 F-분포를 따릅니다. ($F \sim F(k_1, k_2)$)

F -분포

- **두 표본분산 분산비로써로의 F-분포**
  - F-분포는 독립인 두 χ^2-분포의 비율을 이용하는 것으로 두 모집단의 분산 비율을 알고자 할 때 사용할 수 있습니다.
    - 두 개의 정규분포$(N(\mu_1, {\sigma_1}^2), N(\mu_2, {\sigma_2}^2)$ 에서 확률표본 $X_1, X_2, \dots, X_n$ 과 $Y_1, Y_2, \dots, Y_m$을 서로 독립으로 추출했을 때, 각 확률표본의 통계량 $V_1 = \frac{(n-1){S_1}^2}{{\sigma_1}^2}$ 은 자유도가 (n-1)인 χ^2-분포를, $V_2 = \frac{(m-1){S_2}^2}{{\sigma_2}^2}$은 자유도가 (m-1)인 χ^2-분포를 따릅니다(S_1^2, S_2^2은 각 확률표본의 표본분산).
    - 이 때 V_1, V_2가 서로 독립이고 각각을 자유도 (n-1)과 (m-1)로 나눈 값의 비율인 다음의 통계량 F는 자유도가 (n-1, m-1) 인 F분포를 따릅니다. $(F \sim F(n-1, m-1))$

$$F = \frac{V_1/(n-1)}{V_2/(m-1)} = \frac{\left.\frac{(n-1){S_1}^2}{{\sigma_1}^2}\right/ n-1}{\left.\frac{(m-1){S_2}^2}{{\sigma_2}^2}\right/ m-1} = \frac{\left.{S_1}^2\right/{\sigma_1}^2}{\left.{S_2}^2\right/{\sigma_1}^2}$$

F -분포

- **F-분포**
  - 통계량 $F = \frac{{s_1}^2/{\sigma_1}^2}{{s_2}^2/{\sigma_1}^2}$에서, $P(a < F < b) = P\left(a < \frac{{s_1}^2/{\sigma_1}^2}{{s_2}^2/{\sigma_1}^2} < b\right) = P\left(a\frac{{S_2}^2}{{S_1}^2} < \frac{{\sigma_2}^2}{{\sigma_1}^2} < b\frac{{S_2}^2}{{S_1}^2}\right)$
  - 자유도가 k인 t-분포를 따르는 통계량 $T = \frac{Z}{\sqrt{V/k}}$에 대해 다음이 성립하여 $T^2 \sim F(1,k)$ 입니다.

$$T^2 = \frac{Z^2}{V/k} = \frac{Z^2/1}{V/k} = \frac{V_1/1}{V/k}, \qquad Z^2 = V_1 \sim \chi^2(1)$$

  - 확률변수 X가 자유도가 (n, m)인 F-분포를 따를 때 기댓값과 분산
    - $E(X) = \frac{m}{m-2}, \qquad m \geq 3$
    - $Var(X) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}, \qquad m \geq 5$

R에서 χ^2-분포, t-분포, F-분포

함수	시작문자	함수명	함수 형태
확률함수 $P(X=x)$	d	chisq t f	dchisq(x, df) dt(x, df) df(x, df1, df2)
분포함수 $P(X \le x)$	p	chisq t f	pchisq(x, df) pt(x, df) pf(x, df1, df2)
분위수함수 $P(X \le x) = q$	q	chisq t f	qchisq(q, df) qt(q, df) qf(q, df1, df2)
난수생성함수	r	chisq t f	rchisq(n, df) **rt(n, df)** rf(n, df1, df2)

R에서 χ^2-분포, t-분포, F-분포

- **자유도가 3인 χ^2-분포 ($X \sim \chi^2(3)$)**
  - $P(X \leq 3)$: 기댓값 이하일 확률

```
> pchisq(3, df=3)
[1] 0.6084
```

  - $P(X \leq x) = 0.95$: 어떤 값 이하의 확률이 0.95인지
    - $1 - P(X > x) = 0.95$

```
> qchisq(0.95, df=3)
[1] 7.815
```

  - 자유도가 3인 χ^2-분포로 부터 10개의 난수 추출(표본 추출)

```
> rchisq(10, df=3)
 [1] 1.5351 0.6144 0.4465 2.0851 7.4897
 [6] 1.3151 1.8537 2.4792 6.6432 1.0466
```

R에서 χ^2-분포, t-분포, F-분포

- **자유도가 5인 t-분포 ($X \sim t(5)$)**
  - $P(X \leq 0)$: 기댓값 이하일 확률 (t-분포는 좌우대칭입니다.)

```
> pt(0, df=5)
[1] 0.5
```

  - $P(-2.571 \leq X \leq 0)$

```
> pt(0, df=5) - pt(-2.571, df=5)
[1] 0.475
```

  - $P(X \leq x) = 0.975$: 어떤 값 이하의 확률이 0.975인지

```
> qt(0.975, df=5)
[1] 2.571
```

  - 자유도가 5인 t-분포로 부터 10개의 난수 추출(표본 추출)

```
> rt(10, df=5)
 [1] -0.4034 -1.7081  0.7053 -0.7799 -0.1970
 [6]  0.2080  1.3272 -0.5517  0.9401  0.4124
```

R에서 χ^2-분포, t-분포, F-분포

- **자유도가 $(3, 5)$인 F-분포 ($F \sim F(3, 5)$)**
  - $P(X \leq 5.409)$

```
> pf(5.409, df1=3, df2=5)
[1] 0.95
```

  - $P(X \leq x) = 0.95$: 어떤 값 이하의 확률이 0.95인지
    - $1 - P(X > x) = 0.95$

```
> qf(0.95, df1=3, df2=5)
[1] 5.409
```

  - 자유도가 $(3, 5)$인 F-분포로 부터 10개의 난수 추출(표본 추출)

```
> rf(10, df1=3, df2=5)
 [1] 1.4447 0.4273 2.7416 0.6569 0.4024
 [6] 1.0139 0.7866 0.7381 0.2702 0.7502
```

5장을 위한 준비

: R에서의 함수 (사용자 정의 함수)

모집단의 분산

- R에서 제공하는 함수 중 분산 및 표준편차 함수는 표본의 분산과 표준편차만을 구하는 함수로 제공됩니다.
- 모집단의 분산 및 표준편차를 구해 봅시다.
  - 물론 2장에서 그 차이를 보았으며 만들어 보았습니다.

예제 4-4 모집단의 분산

준비파일 | 08.var.pop.R

- 야구공을 만드는 회사에서는 KBO가 정한 반발계수에 맞춰 새롭게 공 10개를 시제품으로 만들고 다음과 같이 반발계수를 관찰했습니다.

0.4196	0.4172	0.4237	0.4182	0.4324
0.4365	0.4354	0.4156	0.4172	0.4414

모집단의 분산

- **실습 내용**
  - 위의 10개의 값들을 모집단으로 가정한 모집단의 분산을 구합니다.
    - 실제 자료는 KBO의 2015년 리그 공인구 수시검사 결과에서 나타난 4개 업체 10개의 샘플로부터 구한 반발계수로 실습을 위해 모집단으로 가정합시다.
- **Step #1) 자료준비**

```
1. options(digits=5)
2. cor <- c(0.4196, 0.4172, 0.4237, 0.4182, 0.4324,
            0.4365, 0.4354, 0.4156, 0.4172, 0.4414)
```

  - 1줄 : 출력물의 자릿수를 5자리로 맞춰줍니다.
  - 2줄 : 예제의 10개 자료를 벡터 변수 cor에 저장합니다.

모집단의 분산

- **Step #2)** 분산을 구하기 위한 각종 값을 생성합니다.

```
4. m <- mean(cor)
5. dev <- cor - m
6. num <- sum( dev^2 )
7. denom <- length(cor)
8. denom2 <- length(cor) - 1
```

  - 4줄 : 평균을 구해 변수 m에 저장합니다.
  - 5줄 : 개별관찰 값과 평균의 차이, 즉 편차를 구해 변수 dev에 저장합니다.
  - 6줄 : 편차들의 제곱을 모두 더해 변수 num에 저장합니다.
  - 7줄 : '자료의 개수'를 변수 denom에 저장합니다(모분산).
  - 8줄 : '자료의 개수 1'을 변수 denom2에 저장합니다(표본분산).

모집단의 분산

- **Step #3)** 모분산과 표본분산을 비교합니다

```
10.(var.p <- num / denom)
11.(var.s <- num / denom2)
12.var(cor)
```

  - 10줄 : 편차 제곱합을 '자료의 개수'로 나눈 모분산을 출력합니다.
  - 11줄 : 편차 제곱합을 '자료의 개수 - 1'로 나눈 표본분산을 출력합니다.
  - 12줄 : R이 제공하는 분산함수(var)와 비교해 봅시다.

```
> (var.p <- num / denom)
[1] 8.4608e-05
> (var.s <- num / denom2)
[1] 9.4008e-05
> var(cor)
[1] 9.4008e-05
```

  - 만일 다른 모집단의 분산을 구해야 하면, 이와 같이 또 여러줄의 코드를 해당 자료에 맞춰 작성해야 할까요?

사용자 정의 함수

예제 4-5 사용자 정의 함수(모분산)

준비파일 | 09.user.func.R

- **실습 내용**
  - 야구공을 만드는 회사에서 반발계수를 구하기 전에 다음과 같이 시제품 10개의 공의 크기와 공의 무게를 측정했습니다.
  - 공의 크기와 공의 무게의 모분산을 구하는 경우를 생각해 봅시다.
    - 앞에서 처럼 공의 크기에 대해 계산하고 공의 무게에 대해 계산해야 한다면 참 번거로울 것입니다.
    - R이 제공하지 않는 함수를 직접 만드는 방법에 대해 알아봅시다.

크기 (둘레,mm)	234	234	234	233	233
	233	233	231	232	231
무게 (g)	146.3	146.4	144.1	146.7	145.2
	144.1	143.3	147.3	146.7	147.3

사용자 정의 함수

- **Step #1)** 사용자 정의 함수 var.p를 만듭니다.

```
1. options(digits=4)
2. var.p <- function(x) {
3.   n <- length(x)
4.   m <- mean(x)
5.   num <- sum( (x - m)^2 )
6.   denom <- n
7.   var <- num / denom
8.   return( var )
9. }
```

사용자 정의 함수

- 1줄 : 출력물의 자릿수를 4자리로 합니다.
- 2~9줄 : R은 이름을 기반으로 모든 자원(변수, 함수 등)을 만듭니다.
  - var.p는 R에서 사용할 자원의 이름입니다.
  - 할당 연산자(<-)를 통해 var.p로 부를 자원을 할당합니다.
  - R에서 function은 지시어로, 사용자 정의 함수를 의미합니다.
  - function 뒤의 소괄호 ( ) 사이에 함수가 작동하는 필요로 하는 정보를 받습니다(전달인자).
  - var.p가 작동하기 위해 하나의 변수를 필요로 하고 있으며, 전달되어 오는 변수에 대해 함수 내에서 x라는 이름으로 사용할 것입니다.
    - 변수 안에 무엇이 들어있을지 확인하는 코드는 없지만, 일단 벡터를 받는 것으로 할 것입니다.
  - 중괄호 { } 사이에 함수가 수행하는 코드가 들어갑니다
    - 만일 함수 수행 코드가 한 줄일 경우 중괄호 없이 사용할 수 있습니다.

사용자 정의 함수

- 3줄 : 전달된 자료의 개수를 구하고, 변수 n에 저장합니다.
- 4줄 : 전달된 자료의 평균을 구하고, 변수 m에 저장합니다.
- 5줄 : 편차의 제곱들을 합하고, 변수 num에 저장합니다.
- 6줄 : 자료의 개수를 변수 denom으로 저장합니다.
- 7줄 : num / denom으로 모분산을 구하고, 변수 var에 저장합니다.
- 8줄 : return() 함수를 이용하여 사용자 정의 함수를 호출(사용)한 곳으로 함수 내부에서 계산한 var 값을 반환합니다.

사용자 정의 함수

- **Step #3) 분산을 구할 자료들을 생성합니다.**

```
11.radius <- c(234, 234, 234, 233, 233, 233, 233, 231, 232, 231)
12.weight <- c(146.3, 146.4, 144.1, 146.7, 145.2, 144.1, 143.3,
               147.3, 146.7, 147.3)
```

  - 11, 12줄 : 공의 크기는 radius에, 공의 무게는 weight에 저장합니다.

- **Step #4) 사용자 정의함수를 사용하고 표본분산과 비교해 봅시다.**

```
14.var.p(radius)
15.var(radius)
16.var.p(weight)
17.var(weight)
```

  - 위에서 만든 함수를 먼저 R에 실행하고 이 부분을 실행해야 합니다.

사용자 정의 함수

- R이 내장하고 있는 함수와 마찬가지로 함수 이름과 함수가 필요로 하는 정보(전달인자)를 작성하는 것으로 함수를 사용합니다.
  - 함수의 사용은 함수의 호출이라고도 부르며, 함수가 완료될 때까지 기다렸다가 함수가 반환하는 값이 있으면 그 값을 가져오고 다음으로 진행합니다.
- 15, 17줄 : 사용자 정의 함수 var.p()를 사용(호출)합니다.
  - 각각 radius, weight를 전달인자로 var.p() 함수의 결과를 확인해 봅시다.
- 16, 18줄 : R이 내장하고 있는 표본분산 함수(var())와 결과를 비교합니다.

```
> var.p(radius)
[1] 1.16
> var(radius)
[1] 1.289
> var.p(weight)
[1] 1.908
> var(weight)
[1] 2.12
```

사용자 정의 함수

- **사용자 정의 함수의 정의와 사용 과정**

① 함수의 이름을 부릅니다.

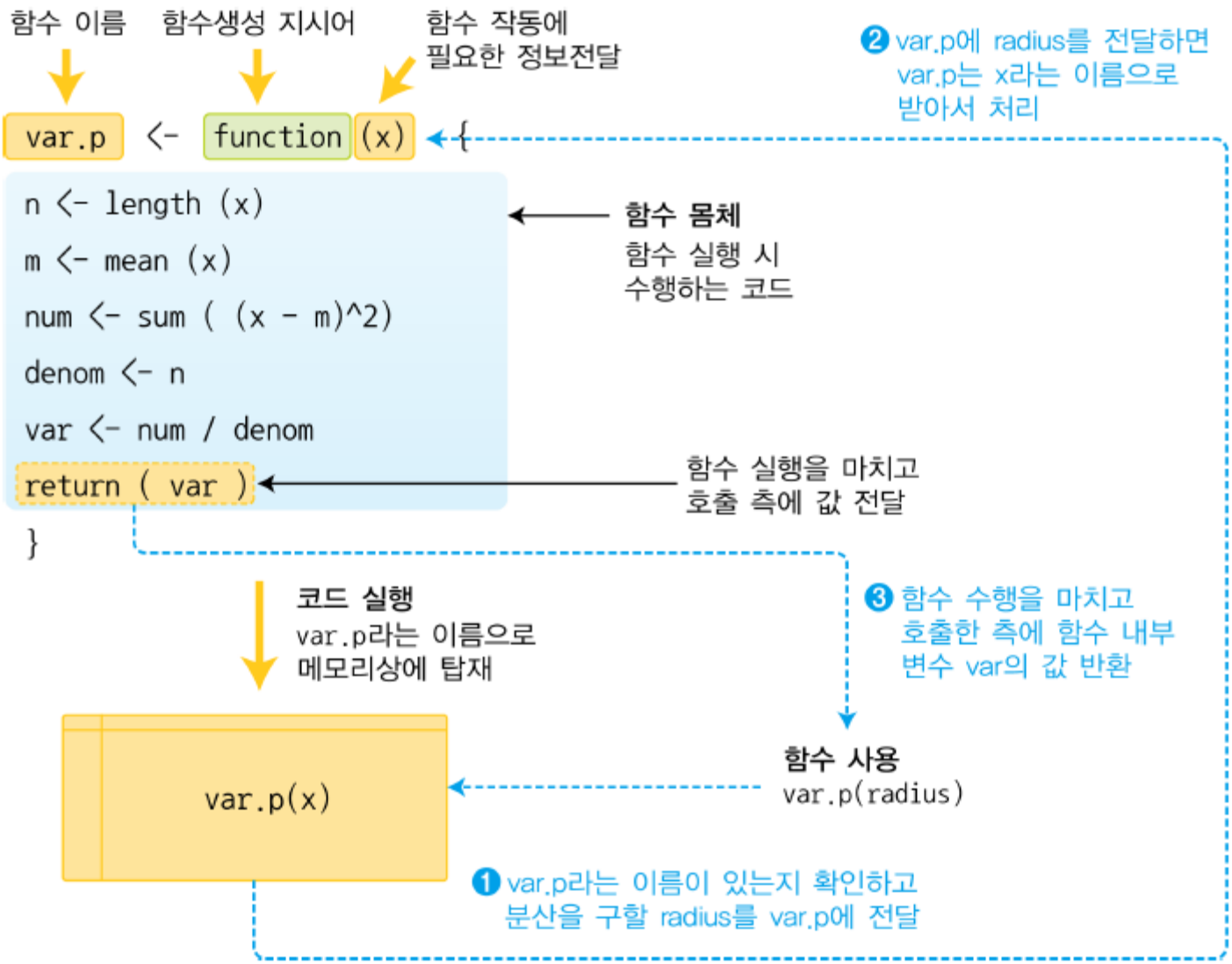
- R은 함수 이름을 듣고 자기의 작업영역에서 그 이름을 가진 함수를 찾아 함수의 이름을 부를 때 같이 전달된 정보(전달인자)를 함수에게 넘겨줍니다.
  - 함수 수행 시 전달인자가 필요 없는 함수도 있는데, 이는 주로 정해진 기능들을 수행하는 함수에 사용합니다(R 내장함수 getwd()는 전달인자로 전달하는 값 없이 현재 작업하고 있는 작업경로를 표시합니다).

② 전달된 정보를 함수는 받아들여 처리합니다.

③ 함수를 종료하고 호출한 곳으로 돌아갑니다.

- 함수를 호출하고 함수가 작동할 때까지 함수를 호출한 측은 잠시 대기하고 있습니다.
- 이 상태에서 함수가 끝나면 반환값이 있을 경우 그 값을 현재 위치로 가져오고, 없을 경우에는 원래 코드의 다음으로 계속 진행합니다.
  - 함수 호출 시 잠시 흐름이 변경됩니다.

사용자 정의 함수


함수 이름
함수생성 지시어
함수 작동에 필요한 정보전달
❷ var.p에 radius를 전달하면 var.p는 x라는 이름으로 받아서 처리
var.p <- function (x) {
n <- length (x)
m <- mean (x)
num <- sum ((x - m)^2)
denom <- n
var <- num / denom
return (var)
}
함수 몸체
함수 실행 시 수행하는 코드
함수 실행을 마치고 호출 측에 값 전달
코드 실행
var.p라는 이름으로 메모리상에 탑재
❸ 함수 수행을 마치고 호출한 측에 함수 내부 변수 var의 값 반환
var.p(x)
함수 사용
var.p(radius)
❶ var.p라는 이름이 있는지 확인하고 분산을 구할 radius를 var.p에 전달

사용자 정의 함수

예제 4-6 여러 개의 전달인자와 기본 전달인자

준비파일 | 10.user_func2.R

- **실습 내용**
  - 2장에서 mean(x, na.rm=TRUE) 을 사용하였습니다.
    - 여기서 전달인자들은 콤마(,)로 구분하여 여러 개를 전달한 경우입니다.
    - 또한 na.rm=TRUE의 경우 이를 명시적으로 알려주지 않으면 mean() 함수는 na.rm 의 값으로 FALSE를 사용하도록 되어 있습니다.
  - 위와 같이 여러 개의 전달일자는 사용하는 방법과 기본 전달인자로 불리우는 함수가 작동하는 데 필요한 정보의 기본값이 있는 함수를 만들어 보겠습니다.
  - 실습에서는 모분산을 구하는 함수에 적용할 것이며 자료는 다음과 같습니다.
    - 앞서 사용한 야구공의 크기 자료 중 7번째 값을 NA로 바꾼 자료입니다.

234	234	234	233	233
233	NA	231	232	231

사용자 정의 함수

- **Step #1) 사용자 정의 함수 var.p2를 만듭니다.**

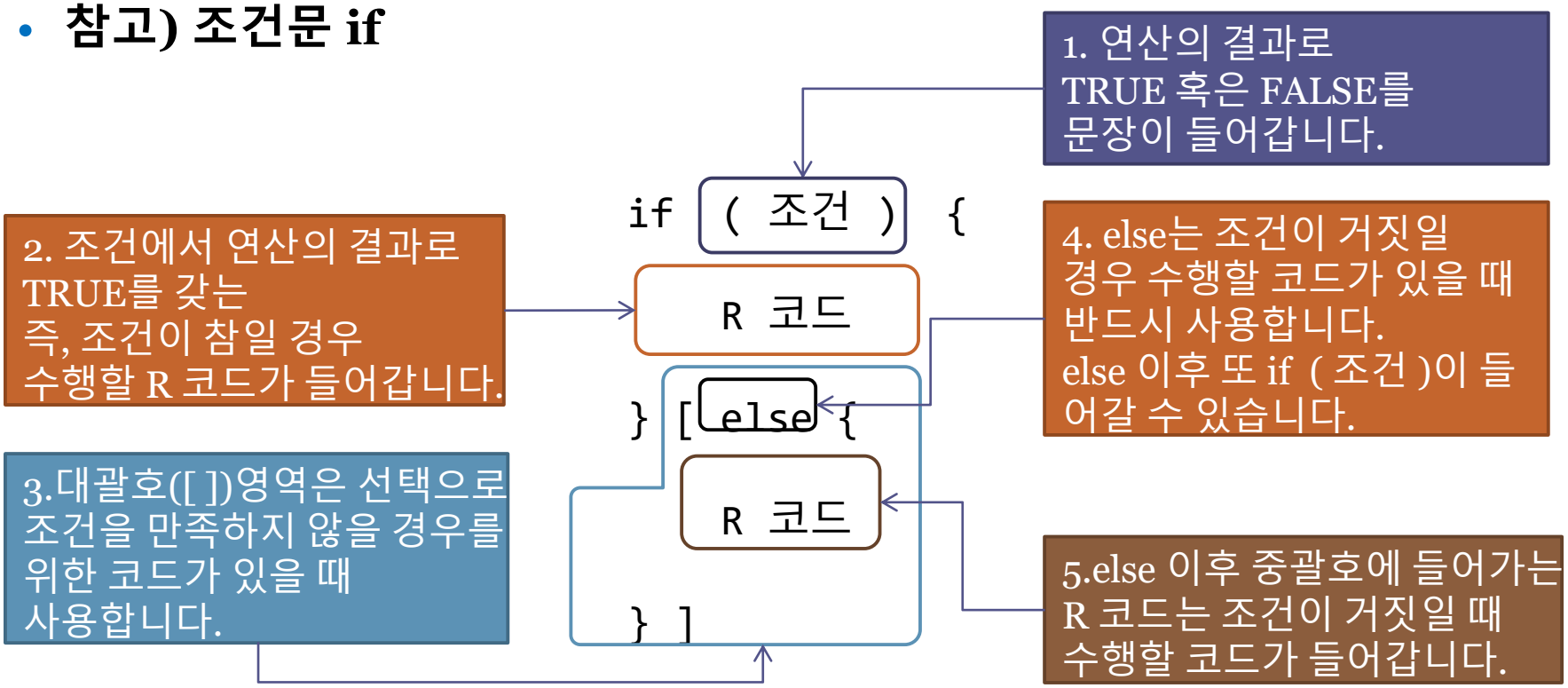
```
1. options(digits=4)
2. var.p2 <- function(x, na.rm=FALSE) {
3.   if(na.rm == TRUE){
4.     x <- x[!is.na(x)]
5.   }
6.   n <- length(x)
7.   m <- mean(x, na.rm=na.rm)
8.   num <- sum( (x - m)^2, na.rm=na.rm )
9.   denom <- n
10.  var <- num / denom
11.  return( var )
12.}
```

사용자 정의 함수

- 1줄 : 출력물의 자릿수를 4자리로 합니다.
- 2~12줄 : var.p2는 두 개의 전달인자를 갖고 있습니다.
  - 여러 개의 전달인자를 필요로 할 때 각각의 전달인자를 콤마(,)로 구별합니다.
  - 두 번째 전달인자를 받는 na.rm은 함수 정의 시 FALSE로 되어 있습니다.
    - 이는 사용자가 해당 전달인자를 전달하지 않더라도 FALSE 값을 갖습니다.
    - 만일 사용자가 해당 값을 바꾸면 그에 맞춰 바뀝니다.
- 3~5줄 : 전달인자 na.rm 값이 TRUE라면, 전달된 값에서 결측값을 제거합니다. (조건문 if 다음 슬라이드 참고)
  - is.na() 함수는 R 내장함수로 전달하는 값이 결측이면 TRUE를, 그렇지 않으면 FALSE를 반환하는 함수로 벡터가 전달되면 TRUE와 FALSE로 구성된 벡터를 반환합니다.
  - 4줄의 x[!is.na(x)]는 벡터 x에서 결측이 아닌 값들만 갖고 오도록 합니다.
    - 사용자가 na.rm에 TRUE를 전달한 경우 결측인 자료들은 제외하게 합니다.

사용자 정의 함수

- **참고) 조건문 if**



- 3~5줄에 사용된 if 문은 변수 na.rm의 값이 TRUE인 경우 4번째 줄을 수행하고 그렇지 않은 경우에는 아무것도 하지 않고 다음으로 넘어갑니다.
  - 즉, na.rm이 TRUE일 경우에는 x에서 결측을 뺀 자료로 다시 저장하고 na.rm이 TRUE가 아닐 경우에는 아무것도 하지 않고 x를 전달받은 대로 사용합니다.

사용자 정의 함수

- **if 사용 예) 기존 벡터의 값이 짝수이면 1, 홀수이면 0이 되도록 새로운 벡터를 생성해 봅시다.**

```
1. x <- c(2, 3, 6, 7, 8)
2. oe <- rep(NA, length(x))
3.
4. for( i in 1:length(x) ) {
5.   if ( x[i] %% 2 == 0 ) {
6.     oe[i] <- 1
7.   } else {
8.     oe[i] <- 0
9.   }
10.}
11.
12.oe
```

5줄부터 9줄까지 사용한 if의 조건 "x[i] %% 2 == 0" 은 벡터 x의 i번째 원소가 짝수 (2로 나눈 나머지가 0이면 짝수) 인지를 판별합니다.

짝수이면 6줄을 수행하여 oe의 i번째 원소를 1로 합니다.

홀수이면 7줄의 else 이후를 실행하여 oe의 i번째 원소를 0으로 합니다.

```
> oe
[1] 1 0 1 0 1
```

출력을(oe) 살펴보면, 벡터 x의 짝수인 2, 6, 8의 위치에 해당하는 oe의 값이 1이고, 나머지는 0임을 확인할 수 있습니다.

사용자 정의 함수

- 6줄 : 전달된 자료의 개수를 구하고, 변수 n에 저장합니다.
- 7줄 : 전달된 자료의 평균을 구하고, 변수 m에 저장합니다.
  - R의 내장함수 mean()도 na.rm을 전달인자로 가지며, 사용자가 전달 한 na.rm을 mean() 함수에 전달합니다.
- 8줄 : 편차의 제곱들을 합하고, 변수 num에 저장합니다.
  - R의 내장함수 sum()도 na.rm을 전달인자로 가지며, 사용자가 전달 한 na.rm을 sum() 함수에 전달합니다.
  - 7줄도 마찬가지로 사용자가가 na.rm 값을 전달하지 않으면 함수 정의시 사용한 ra.rm=FALSE 를 기본 값으로 가집니다. (기본전달인자)
- 9줄 : 자료의 개수를 변수 denom으로 저장합니다.
- 10줄 : num / denom으로 모분산을 구하고, 변수 var에 저장합니다.
- 11줄 : return() 함수를 이용하여 사용자 정의 함수를 호출(사용)한 곳으로 var 값을 반환합니다.

사용자 정의 함수

- **Step #2) 작성한 var.p2를 사용합니다.**
  - 전달인자를 다르게 하여 사용해 봅시다.
  - 14줄 : 공의 크기를 radius에 저장합니다.
  - 15줄 : na.rm을 전달하지 않으면 함수 정의 시 기본값으로 na.rm=FALSE가 됩니다.
    - mean()과 sum()은 na.rm=FALSE로 했을 때 결측도 하나의 값으로 합과 평균을 구하나, 결측이 포함된 자료는 R에서는 NA로 처리됩니다.
    - 이에 결과가 결측을 나타내는 NA로 나옵니다(이를 이용하여 NA가 있는지 확인할 수 있습니다).
  - 16줄 : na.rm=TRUE, 즉 결측 자료는 제외하고 모집단의 분산을 구합니다(결측을 제외한 만큼 표본의 수도 줄어듭니다).

사용자 정의 함수

```
> var.p2(radius)
[1] NA
> var.p2(radius, na.rm=TRUE)
[1] 1.284
```

- **정리**
  - 함수는 자주 사용하는 코드를 하나의 단위로 만들어 필요로 할 때 마다 사용할 수 있습니다.
  - 사용자 정의 함수는 함수를 사용하기 전에 R에 실행을 해야 합니다.
  - 전달인자를 여러 개를 사용할 경우 함수 정의시 콤마를 이용하여 구분합니다.
  - 기본전달인자는 함수 정의시 함수 수행에 필요한 정보를 저장하는 것으로 사용자가 다른 값을 전달할 경우 사용자가 전달한 값을 사용하며, 사용자가 해당 전달인자를 넘기지 않은 경우에는 기본값이 사용됩니다.
  - 함수 사용에는 더 많은 내용들이 있습니다만, 우리는 아주 기초적인 부분만 살펴보았습니다.

Q & A

수고하셨습니다.