creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

**저작자표시**. 귀하는 원저작자를 표시하여야 합니다.

**비영리**. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

**변경금지**. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

**저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.**

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사학위논문

# 엑셀 VBA를 이용한
# 표본추출 프로그램

# Application Frame for Sampling with
# Excel VBA

2013년 2월

서울대학교 대학원
통 계 학 과
이 안 나

# 엑셀 VBA를 이용한
# 표본추출 프로그램

지도교수 조 신 섭

이 논문을 이학석사 학위논문으로 제출함
2012년 10월

서울대학교 대학원
통 계 학 과
이 안 나

이안나의 이학석사 학위논문을 인준함
2012년 12월

위 원 장 박 태성 (인)
부 위 원 장 조 신섭 (인)
위 원 이 재용 (인)

# Application Frame for Sampling with Excel VBA

by
An-na Lee

A thesis
submitted in partial fulfillment of
the requirement for the degree of Master of Science
in the Department of Statistics
Seoul National University
December, 2012

# 국 문 초 록

자료 분석을 위해서 주로 사용되는 통계 분석틀(package)로는 SAS, SPSS 등을 들 수 있다. 하지만 대부분 사용이 어렵거나, 영문으로 되어 있어 일반인들이 사용하기에는 어려움이 있으며, 고가의 비용을 들여서 구입해야 한다는 부담이 있다.

본 논문에서는 표본조사활동을 위한 분석틀 개발에 필요한 기본 개념들을 정리해 보고, 이러한 개념을 적용하여 그에 적합한 표본 추출틀을 개발하였다. 일반인들이 보다 편리하고 쉽게 통계 자료 분석을 할 수 있도록 입출력의 용이성 및 사용의 보편성을 갖춘 엑셀(EXCEL)을 개발의 기본도구로 삼았다.

주요어 : 통계 분석틀, 엑셀, 표본추출

학번 : 2010-23062

# 차 례

1. 서론

2. 표본추출

3. 기존 통계패키지에서 표본추출 실행방법

4. KESS에서 표본추출 실행방법

5. 맺음말 및 추후과제

참고문헌

Abstract

# 1. 서 론

사회가 발달함에 따라 다양하고 보다 빠른 정보가 요구된다. 이러한 요구를 충족시키기 위해서 연구대상 집단의 일부인 표본을 조사하여 전체집단에 대한 일반화된 추론결과를 신속하고 정확하게 도출하는 것이 필요하다. 이에 통계학자들은 컴퓨터 소프트웨어를 이용한 프로그램을 구현하기 위해 지속적인 노력을 하고 있다.

Scheaffer, Mendenhhall 그리고 Ott는 SAS 매크로를 활용하여 표본추출이 가능한 프로그램을 구현하였다.(1996) 표본추출 매크로는 단순임의추출, 계통추출, 층화추출 등이 가능하고 요약통계량이 제공된다.

서울대학교에서는 일반인들이 보다 편리하고 쉽게 통계 자료분석을 할 수 있도록 Excel에서 구동되는 교육용 한글 통계 소프트웨어인 KESS(2007)를 개발하여 제공하고 있다. 본 논문에서는 표본조사활동을 위한 프로그램을 Excel VBA를 통해 구현하였고, 이를 KESS에 부메뉴로 추가함으로써 KESS에서 표본추출이 가능하도록 하였다.

본 논문은 총 5장으로 구성되어 있다. 먼저 연구의 배경 및 목적을 제시하였고, 표본추출에 대한 설명, 기존 통계패키지에서 표본추출 실행방법, 마지막으로 KESS에서의 실행방법과 연구의 결론 및 향후 전망에 대해 기술하였다.

# 2. 표본추출

## 2-1. 단순임의추출(Simple Random Sampling)

단순임의추출이란 대상집단의 모든 구성요소에 동등한 확률을 부여하여 표본을 추출하는 방법을 말한다. 즉, 크기 $N$인 모집단에서 크기 $n$인 표본을 추출하고자 할 때 조사자의 주관적 판단을 배제하고 무작위로 표본을 추출하는 방법이다. 이는 확률추출의 가장 간단한 형태로서 다른 표본추출들의 이론적 기초를 제공한다. 모집단 평균 $\mu$의 추정치 $\hat{\mu}$에 대한 기댓값과 분산은 다음과 같다.

$$E(\hat{\mu}) = \mu$$
$$Var(\hat{\mu}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

분산의 추정치는 식 (2.1)과 같다. 모집단 분산 $\sigma^2$을 알고 있는 경우에는 이를 사용하여 분산을 구할 수 있으나, 대부분의 경우 모집단 분산은 모수로서 알 수 없으므로 표본 분산 $s^2$을 사용한다.

$$\widehat{Var}(\hat{\mu}) = \frac{N-n}{N}\frac{s^2}{n} \tag{2.1}$$

사용자가 목적에 맞는 표본의 크기를 미리 정하여 추출할 수도 있으나, 오차의 한계(bound of error)를 미리 정하고 모수

의 추정치와 모수와의 차이가 그 한계 내에 있도록 하는 경우도 있다. 모수와 그 추정치를 각각 $\theta$, $\hat{\theta}$라 하면, 오차의 한계 $B$와 표본의 크기 $n$을 다음과 같이 표현할 수 있다.

$$\left| \hat{\theta} - \theta \right| \leq B$$

$$n = \frac{N Z_{\alpha/2}^2 \sigma^2}{N B^2 + Z_{\alpha/2}^2 \sigma^2}$$

(2.2)

그런데 식 (2.2)에서 표본의 크기를 결정하려면 유의수준 $\alpha$와 모분산 $\sigma^2$의 값을 알고 있어야 한다. 일반적으로 많은 조사연구에서는 $\alpha = 0.05$로 놓으나 모분산의 값을 미리 아는 경우는 거의 없다. 이를 해결하기 위해 데이터들의 범위(range)를 기초로 모분산을 추정할 수 있다. 정규분포에서는 평균을 중심으로 $\pm 2\sigma$ 범위 내에 약 95%의 값들이 존재하므로 $\sigma$는 다음과 같이 결정되고,

$$\sigma \approx \frac{range}{4}$$

본 논문에서는 입력된 자료를 바탕으로 표본추출을 수행하므로 모분산 $\sigma^2$의 값을 미리 계산하여 표본의 크기를 정하도록 한다.

단순임의추출을 하려면 모집단 내 모든 원소들의 목록이 필요하며, 각 원소가 선택될 확률의 동일성과 독립성이 만족되어야 한다. 실제 추출과정에서는 조사대상 원소에 일련번호를 붙

인 후 난수표를 이용한 랜덤 메커니즘을 사용한다. 이렇게 추출된 표본은 모집단에 대한 대표성을 가질 수 있다.

단순표본추출의 아이디어는 단순하지만 실제로 시행하기는 어려운 경우가 많다. 한 학교의 재학생과 같이 모집단이 크지 않으면서 그 명단을 확보할 수 있는 경우는 한정되어 있기 때문이다. 예를 들어 전국 규모의 조사를 통해 임의표본을 추출하려 한다면, 조사원의 수 뿐만 아니라 많은 시간과 경비가 소요될 것이다.

## 2-2. 계통추출(Systematic Sampling)

계통추출은 조사자료 목록의 나열된 순서가 대체로 임의화되었다고 간주할 수 있는 상황에서 시행하는 추출기법이다. 계통표집의 장점은 시행의 편리성에 있으며, 다음과 같은 과정을 거친다.

① 표본크기 $n$을 결정한다.

② $\frac{n}{N}$에 가장 근접한 정수 $k$를 구한다.

③ 1 이상 $k$ 이하의 정수 $R$을 임의 선택한다.

④ $R, R+k, R+2k, ..., R+(n-1)k$번째 표본단위들을 추출한다.

즉, $R$번째 표본단위부터 시작하여 매 $k$번째 단위들을 순차적으로 추출하는 것이다. 이렇게 추출하면 약 $\frac{1}{k}$이 표본으로 추

출된다. 그런데 모집단의 크기 $N$이 항상 $k$로 나누어 떨어지는 정수는 아니므로, 시작점 $R$ 값에 따라 계통표본의 크기는 $\pm 1$ 만큼의 차이가 날 수 있다.

계통추출에서 모수 추정시 필요한 표본 크기의 결정은 단순임의추출에서 설명한 식들을 그대로 적용한다. 물론 이 경우에도 임의모집단에서의 추출을 가정한다.

계통추출은 표본추출이 쉽고 오류 없이 시행가능하며, 시간을 절약할 수 있는 장점이 있다. 하지만 계통추출은 모든 가능한 표본이 동시에 표본으로 선택될 수는 없다는 점에서 단순임의추출과 같지는 않다. 아울러 모집단의 나열이 어떤 순서로 이루어졌느냐에 따라 표본의 모집단 대표성이 좌우된다. 자료들이 특별한 순서나 주기성 없이 잘 섞여있다면 단순임의추출과 유사한 효과를 줄 수 있을 것이다.

## 2-3. 층화추출(Stratified Random Sampling)

특정 제품을 생산하는 기업들을 조사하는 상황을 가정해 보자. 대기업의 수는 상대적으로 적지만 총생산량에서 차지하는 비중은 크고, 중소기업의 수는 많지만 전체적으로 그 비중은 작다. 이 경우 단순임의추출을 통해 대기업이 많이 추출되지 않는다면 조사 결과는 왜곡될 가능성이 크다. 대기업과 중소기업을 처음부터 구분하여 독립적인 추출을 한다면 이러한 문제를 해결할 수 있을 것이다.

위의 예에서와 같이 관심 변수가 부모집단(Sub-population) 내에서 서로 다른 평균을 취한다는 정보를 갖고 있다면, 그 변수를 기준으로 모집단을 둘 이상의 부모집단으로 분할한 다음

각 부모집단마다 독립적인 단순임의추출을 시행하면 보다 정밀한 추정치를 얻을 수 있다. 이런 형태의 추출을 층화추출이라 하고, 나눠진 부모집단들을 층이라 한다.

층화추출을 시행할 때는 먼저, 전체 모집단 중 추출하고자 하는 표본의 크기가 $n$이라고 할 때 이 표본을 각 층에서 몇 개씩 추출할 것인가를 고려해야 한다. 각 층의 크기 $N_h$가 서로 다를 때는 여러 가지 배분법이 사용된다. 각 층의 크기에 관계없이 동일하게 배분하는 동일 배분법, 각 층의 크기에 비례하여 표본을 배분하는 비례 배분법, 주어진 표본추출비용 하에서 추정치의 분산을 최소화하도록 하는 최적 배분법, 그리고 각 층에서 한 단위 표본을 추출하는 비용이 동일하다는 가정 하에서 추정치의 분산을 최소화하도록 하는 네이만 배분법 등이 있다. 본 논문에서는 비례배분을 통한 알고리즘을 다루고 있으므로 이 배분법을 통한 분산의 추정치에 대해 설명하겠다.

비례배분은 모집단으로부터 추출하고자 하는 표본크기 $n$을 각 층의 크기 $N_h$에 비례하여 배분하는 방법으로, 각 층에서 추출하고자 하는 표본크기 $n_h$를 식 (2.3)과 같이 구하는 배분법이다.

$$n_h = n \times \frac{N_h}{N} \tag{2.3}$$

$$\widehat{Var}(\overline{X}) = \frac{N-n}{N} \sum_{h=1}^{l} \frac{N_h}{N} \frac{s_h^2}{n}$$

거의 모든 대규모 표본조사에서는 반드시 층화추출을 활용한

다. 적어도 층화의 기준으로 선택된 보조변수의 속성대로 균질한 그룹으로 나눠진 층에서, 각 층 내 분산은 전체 모집단 분산에 비하여 상당히 낮으리라 기대할 수 있기 때문이다. 이런 측면에서 층화추출은 단순임의추출에 비하여 관리가 용이하고 비용이 절감되는 효과가 있다.

층화의 기준이 될 수 있는 변수들은 많다. 조사 목적과 밀접한 관계가 있는 층화변수들을 잘 알고 있다면 이들 모두의 조합에 대해서 층화를 시도함이 당연하다. 그러나 변수가 많아지면 그만큼 층화가 복잡해진다. 아울러 층화변수들에 대한 정보가 없다면 층의 변수는 가급적 적게 잡아야 한다. 이에 본 논문에서 다루고자 하는 층화추출 알고리즘은 층의 변수를 일반적인 수준인 세 개까지로 한정하여 진행하였다.

## 3. 기존 통계패키지에서 표본추출 실행방법

### 3-1. SAS

SAS는 'Statistical Analysis System'의 약자로서, 1970년대 초에 통계분석용 패키지로 개발된 소프트웨어다. 예제 데이터를 통해 단순임의추출과 층화추출의 사용방법에 대해 설명하겠다. 2012년 한국프로야구(KBO) 타격 상위 25명의 포지션(좌타우타), 득점, 안타, 홈런, 도루, 타율을 조사하여 다음과 같이 자료를 입력하였다.

```
DATA best25;
   INPUT no form score hit homerun steal avg;
DATALINES;
1   1  61   151 16  3   0.363
2   1  77   137 25  21  0.314
3   1  61   158 5   10  0.314
4   1  79   138 23  2   0.312
5   1  50   113 12  3   0.31
6   1  84   150 21  6   0.307
7   1  39   112 4   4   0.307
8   1  82   152 11  30  0.305
9   1  61   118 1   4   0.304
10  1  60   117 3   8   0.303
11  1  55   128 18  3   0.3
12  1  52   126 5   5   0.3
13  1  85   142 26  20  0.3
14  1  62   128 5   32  0.294
15  1  48   114 15  4   0.292
16  1  47   127 7   6   0.291
17  1  76   136 31  20  0.29
18  2  60   141 3   20  0.288
19  2  86   139 2   44  0.283
20  2  63   124 5   30  0.281
21  2  36   105 4   5   0.279
22  2  39   100 5   1   0.279
23  2  64   115 2   25  0.274
24  2  40   109 11  7   0.274
25  2  41   109 19  0   0.273
;
RUN;
```

<그림 3-1> SAS 데이터 입력화면

단순임의추출과 계통추출의 경우 샘플 데이터(25개)에서 10개의 표본을 추출하는 명령문을 실행하였고, 층화추출(비례배분)

은 각 두 개의 층에서 50%식 추출하는 방법을 사용하였다. <그림 3-2>는 순서대로 단순임의추출, 계통추출, 층화추출을 시행하는 명령문이다.

```
proc surveyselect data = best25 method = SRS
                  sampsize = 10 out = bestofbest;
RUN;

proc surveyselect data = best25 method = SRS
                 rate=(0.5, 0.5) out = bestofbest;
     strata form;
RUN;

proc surveyselect data = best25 method = SYS
                  sampsize=5 out = bestofbest;
RUN;
```

<그림 3-2> 표본추출 명령문

명령문을 통해 세 가지 추출을 시행한 결과는 다음과 같다.

| no | form | score | hit | homerun | steal | avg |
|---|---|---|---|---|---|---|
| 1 | 1 | 61 | 151 | 16 | 3 | 0.363 |
| 3 | 1 | 61 | 158 | 5 | 10 | 0.314 |
| 6 | 1 | 84 | 150 | 21 | 6 | 0.307 |
| 7 | 1 | 39 | 112 | 4 | 4 | 0.307 |
| 9 | 1 | 61 | 118 | 1 | 4 | 0.304 |
| 13 | 1 | 85 | 142 | 26 | 20 | 0.300 |
| 15 | 1 | 48 | 114 | 15 | 4 | 0.292 |
| 16 | 1 | 47 | 127 | 7 | 6 | 0.291 |
| 19 | 2 | 86 | 139 | 2 | 44 | 0.283 |
| 22 | 2 | 39 | 100 | 5 | 1 | 0.279 |

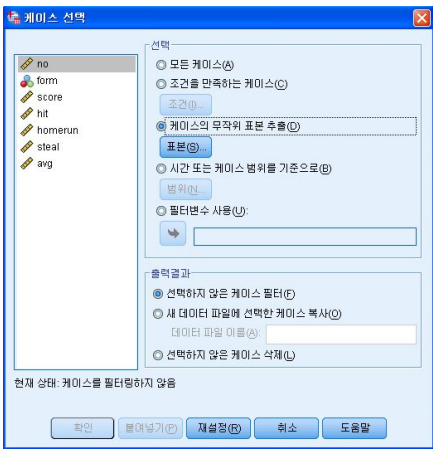
| no | form | score | hit | homerun | steal | avg |
|---|---|---|---|---|---|---|
| 3 | 1 | 61 | 158 | 5 | 10 | 0.314 |
| 8 | 1 | 82 | 152 | 11 | 30 | 0.305 |
| 13 | 1 | 85 | 142 | 26 | 20 | 0.300 |
| 18 | 2 | 60 | 141 | 3 | 20 | 0.288 |
| 23 | 2 | 64 | 115 | 2 | 25 | 0.274 |

| form | no | score | hit | homerun | steal | avg | Selection Prob | Sampling Weight |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 61 | 158 | 5 | 10 | 0.314 | 0.52941 | 1.88889 |
| 1 | 5 | 50 | 113 | 12 | 3 | 0.310 | 0.52941 | 1.88889 |
| 1 | 6 | 84 | 150 | 21 | 6 | 0.307 | 0.52941 | 1.88889 |
| 1 | 8 | 82 | 152 | 11 | 30 | 0.305 | 0.52941 | 1.88889 |
| 1 | 12 | 52 | 126 | 5 | 5 | 0.300 | 0.52941 | 1.88889 |
| 1 | 13 | 85 | 142 | 26 | 20 | 0.300 | 0.52941 | 1.88889 |
| 1 | 15 | 48 | 114 | 15 | 4 | 0.292 | 0.52941 | 1.88889 |
| 1 | 16 | 47 | 127 | 7 | 6 | 0.291 | 0.52941 | 1.88889 |
| 1 | 17 | 76 | 136 | 31 | 20 | 0.290 | 0.52941 | 1.88889 |
| 2 | 20 | 63 | 124 | 5 | 30 | 0.281 | 0.50000 | 2.00000 |
| 2 | 22 | 39 | 100 | 5 | 1 | 0.279 | 0.50000 | 2.00000 |
| 2 | 24 | 40 | 109 | 11 | 7 | 0.274 | 0.50000 | 2.00000 |
| 2 | 25 | 41 | 109 | 19 | 0 | 0.273 | 0.50000 | 2.00000 |

<그림 3-3> 표본추출 시행 결과창

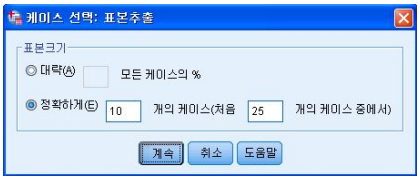
## ~~3-2. SPSS~~

SPSS(Statistical Package for the Social Science) 역시 널리 보급되어 있는 통계 소프트웨어로 특히 설문조사에 의해 수집된 데이터의 집계와 분석에 용이하다. SPSS를 통한 단순임의추출을 수행하려면 먼저 변수 지정을 통해 데이터를 입력해야 한다. 3-1장에서와 동일한 데이터를 입력한 후 상단 메뉴 중 '데이터' → '케이스 선택'을 차례로 선택하면 다음과 같은 대화상자가 열린다.




**<그림 3-4> SPSS 초기 입력 대화상자**

'케이스의 무작위 표본추출'을 통해 단순임의추출을 수행할 수 있다. 여기서는 사용자가 표본의 크기를 정할 수 있는 두 가지 형태를 제공하고 있다. 하나는 직접 표본의 크기를 입력하는 것이고, 나머지 하나는 전체에 대한 표본의 비율을 지정하여 추출하도록 하는 것이다.




<그림 3-5> SPSS 초기 입력 대화상자

표본의 크기를 10으로 지정하여 출력한 화면은 다음과 같다.

| | no | form | score | hit | homerun | steal | avg | filter_$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 61 | 151 | 16 | 3 | .363 | 0 |
| 2 | 2 | 1 | 77 | 137 | 25 | 21 | .314 | 1 |
| 3 | 3 | 1 | 61 | 158 | 5 | 10 | .314 | 1 |
| 4 | 4 | 1 | 79 | 138 | 23 | 2 | .312 | 1 |
| 5 | 5 | 1 | 50 | 113 | 12 | 3 | .310 | 0 |
| 6 | 6 | 1 | 84 | 150 | 21 | 6 | .307 | 1 |
| 7 | 7 | 1 | 39 | 112 | 4 | 4 | .307 | 1 |
| 8 | 8 | 1 | 82 | 152 | 11 | 30 | .305 | 1 |
| 9 | 9 | 1 | 61 | 118 | 1 | 4 | .304 | 1 |
| 10 | 10 | 1 | 60 | 117 | 3 | 8 | .303 | 0 |
| 11 | 11 | 1 | 55 | 128 | 18 | 3 | .300 | 0 |
| 12 | 12 | 1 | 52 | 126 | 5 | 5 | .300 | 0 |
| 13 | 13 | 1 | 85 | 142 | 26 | 20 | .300 | 0 |
| 14 | 14 | 1 | 62 | 128 | 5 | 32 | .294 | 0 |
| 15 | 15 | 1 | 48 | 114 | 15 | 4 | .292 | 0 |
| 16 | 16 | 1 | 47 | 127 | 7 | 6 | .291 | 1 |
| 17 | 17 | 1 | 76 | 136 | 31 | 20 | .290 | 0 |
| 18 | 18 | 2 | 60 | 141 | 3 | 20 | .288 | 1 |
| 19 | 19 | 2 | 86 | 139 | 2 | 44 | .283 | 0 |
| 20 | 20 | 2 | 63 | 124 | 5 | 30 | .281 | 0 |
| 21 | 21 | 2 | 36 | 105 | 4 | 5 | .279 | 0 |
| 22 | 22 | 2 | 39 | 100 | 5 | 1 | .279 | 0 |
| 23 | 23 | 2 | 64 | 115 | 2 | 25 | .274 | 0 |
| 24 | 24 | 2 | 40 | 109 | 11 | 7 | .274 | 0 |
| 25 | 25 | 2 | 41 | 109 | 19 | 0 | .273 | 1 |

<그림 3-6> SPSS 초기 입력 대화상자

~~특별한 결과창을 출력하여 주는 것이 아니라, 데이터 입력창에서 추출되지 않은 케이스들을 사선으로 제거표시하거나 새롭게 입력된 변수의 값을 0으로 지정해 주고 있다.~~

# 4. KESS에서 표본추출 실행방법

~~3장에서 살펴본 바와 같이 SAS의 표본추출 프로시저는 추출 방법이 매우 다양하며 대용량 데이터의 집합에서도 사용이 가능하다. 그러나 SAS를 이용하려면 먼저 SAS언어로 프로그램을 작성해야 하는데, 이는 고유 문법과 사용가능한 어휘가 정해져 있는 고급 언어에 속한다. SPSS의 경우 사용방법은 비교적 간단하나 표본추출 형태가 매우 제한적이다. 결국 이들 패키지는 기관 또는 기업에 소속되지 않은 개인이 접근하기 힘들며, 목적에 맞게 활용하기 위해서는 통계적인 이론과 사용방법을 따로 익혀야 하는 어려움이 있다.~~

통계 자료 분석 도구가 갖추어야할 중요한 요건 중의 하나는 사용자에게 편리한 인터페이스 환경을 제공해야 한다는 것이다. 또한 다양한 종류의 자료파일을 불러오거나 자료를 직접 입력할 수 있어야 하고, 분석 결과를 효과적으로 보고서화 할 수 있어야 한다. Microsoft Excel은 이러한 측면에서 매우 우수한 스프레드시트 프로그램이며, 자료의 관리 및 계산, 그래프의 작성 등을 손쉽게 할 수 있다는 점에서 일반인들에게 가장 널리 사용되고 있는 프로그램이다.

위와 같은 점에 착안하여 KESS (2007)는 Microsoft Excel을 개발의 도구로 삼았다. 따라서 KESS에서는 Excel에서 제공되는 모든 기능들을 그대로 이용할 수 있으며, 일반인들이 편리하게 통계분석을 할 수 있도록 한다는 점에서 통계 분석 도구로써 상당한 장점을 갖고 있다.

기초적인 통계 분석은 Excel의 데이터 분석 도구를 통해서도 가능하다. 그러나 KESS에서는 SAS, SPSS 등의 통계 패키지에서와 같이 통계분석 도구를 메뉴 형식으로 제공함으로써,

Excel의 데이터 분석 도구와는 차별화된 보다 통합적이고 체계화된 형태의 통계 분석 프로그램을 제공하고 있다. 그러나 현재까지 KESS에서는 표본조사 수행 메뉴를 제공하고 있지 않다.

이에 본 논문에서는 표본조사를 수행하는 프로그램을 구현하였고 Excel의 [데이터 분석] 도구와 동일한 형태의 데이터 입출력 방식을 이용하였다. 또한 본 논문에서는 표본조사 도구를 독립적으로 제공하는 것에 그치는 것이 아니라, KESS에 부메뉴로 표본조사를 포함시킴으로써 다른 분석들과의 연계가 가능하도록 하였으므로 여기에서 개발의 의의를 찾을 수 있다.

(1) 개발에 사용된 툴과 프로그램의 구성

개발에 사용된 툴로는 Microsoft Excel VBA(Visual Basic for Application)를 이용하였다. VBA는 Visua Basic 편집기라는 통합 프로그래밍 환경과 Basic 프로그래밍 언어를 조합한 것이다.

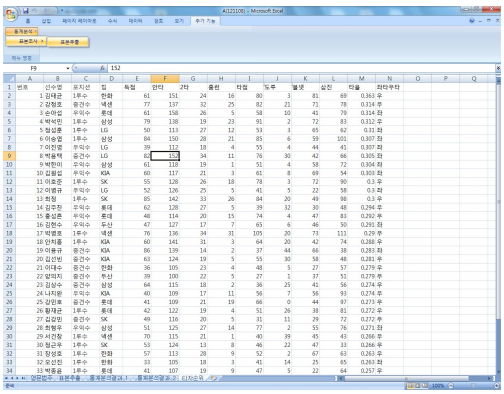
(2) 개발 알고리즘 및 출력물의 구성

표본추출 활동은 앞에서 언급되었듯이, 기본적으로 난수생성 알고리즘을 이용하여 수행한다.

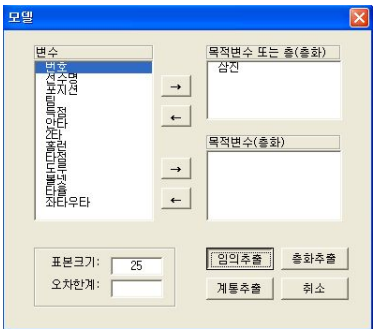
출력화면에서는 각 추출형태에 따른 자료들의 목록이 주어진다. 그리고 추출된 표본들의 기본통계량인 표본의 크기, 평균, 표준편차 등이 주어진다.

## 4-1. 단순임의추출

EXCEL을 실행하여 3장에서 예로 든 데이터를 입력한다. 이제 KESS를 실행하여 메뉴 중 '표본조사'를 선택하면 <그림 4-1>에서와 같이 팝업메뉴 '표본추출'이 나타난다. 이것을 클릭하면 <그림 4-2>에서와 같이 프로그램 수행을 위한 기본사항(추출형태, 목적변수 및 층의 선택, 표본의 크기, 오차의 한계 등)을 요구하는 팝업창이 뜬다. 여기에 사용자가 해당하는 변수를 선택하거나 값을 입력하면 된다.

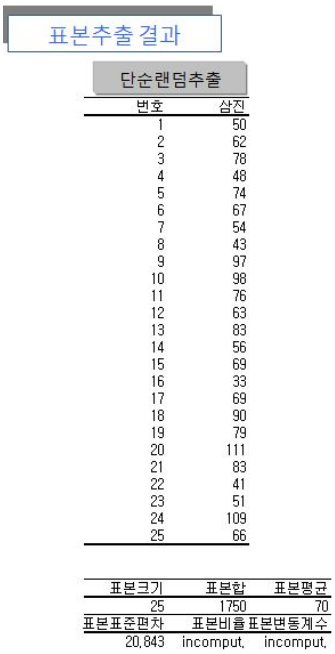


<그림 4-1> 표본추출 초기화면




<그림 4-2> 기본 값 입력을 위한 팝업창

표본의 크기를 지정하지 않고, 사용자가 지정하는 오차의 한계만 입력해도 2장에서 살펴보았던 수식을 통해 표본의 크기가 자동으로 계산되어 추출을 수행한다. '표본크기'와 '오차한계' 둘 중의 하나는 반드시 값을 입력해야 한다. 오른쪽 상단의 '목적변수 또는 층(층화)' 블록에는 조사하고자 하는 목적변수를 화살표를 통해 선택하도록 되어 있다. 마지막으로 '임의추출' 버튼을 클릭하면 <그림 4-3>과 같이 '_통계분석결과_'라는 새로운 시트에 추출된 자료와 기본 통계량이 출력된다.



표본추출 결과

단순랜덤추출

| 번호 | 삼진 |
|---|---|
| 1 | 50 |
| 2 | 62 |
| 3 | 78 |
| 4 | 48 |
| 5 | 74 |
| 6 | 67 |
| 7 | 54 |
| 8 | 43 |
| 9 | 97 |
| 10 | 98 |
| 11 | 76 |
| 12 | 63 |
| 13 | 83 |
| 14 | 56 |
| 15 | 69 |
| 16 | 33 |
| 17 | 69 |
| 18 | 90 |
| 19 | 79 |
| 20 | 111 |
| 21 | 83 |
| 22 | 41 |
| 23 | 51 |
| 24 | 109 |
| 25 | 66 |

| 표본크기 | 표본합 | 표본평균 |
|---|---|---|
| 25 | 1750 | 70 |
| 표본표준편차 | 표본비율 | 표본변동계수 |
| 20.843 | incomput. | incomput. |

**<그림 4-3> 단순임의추출 출력결과**

표본비율과 표본변동계수는 이분속성 데이터를 목적변수로 하는 경우에만 출력된다. 이분속성 데이터란 어떤 관심 있는 속성을 포함한 원소를 1, 아니면 0으로 표시한 {0, 1} 데이터를

말한다. 시트의 자료에서 '좌타우타' 데이터는 '좌'와 '우' 둘 중 하나의 값만을 가지는데, 자료 중 가장 위에 있는 값을 1로, 다른 하나의 값을 0으로 변환하여 임의추출을 수행하도록 설계하였다. 이 경우에는 <그림 4-4>와 같이 표본비율과 표본변동계수의 값도 출력창에 계산되어 나타난다.

표본추출결과(이분값 변환)

단순랜덤추출

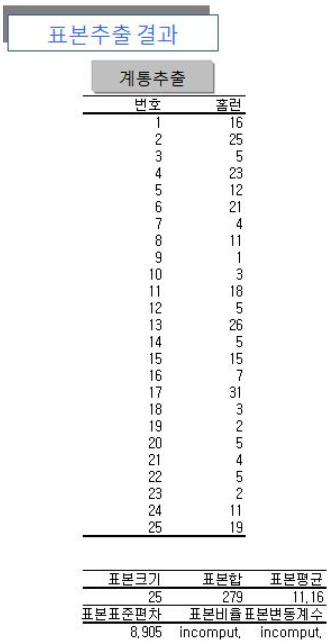
| 번호 | 좌타우타 | 이분변환값 |
|---|---|---|
| 1 | 좌 | 0 |
| 2 | 우 | 1 |
| 3 | 좌 | 0 |
| 4 | 우 | 1 |
| 5 | 우 | 1 |
| 6 | 좌 | 0 |
| 7 | 좌 | 0 |
| 8 | 우 | 1 |
| 9 | 우 | 1 |
| 10 | 우 | 1 |
| 11 | 우 | 1 |
| 12 | 좌 | 0 |
| 13 | 좌 | 0 |
| 14 | 좌 | 0 |
| 15 | 우 | 1 |
| 16 | 우 | 1 |
| 17 | 좌 | 0 |
| 18 | 우 | 1 |
| 19 | 좌 | 0 |
| 20 | 우 | 1 |
| 21 | 우 | 1 |
| 22 | 우 | 1 |
| 23 | 우 | 1 |
| 24 | 우 | 1 |
| 25 | 좌 | 0 |

| 표본크기 | 표본합 | 표본평균 |
|---|---|---|
| 25 | 15 | 0.6 |
| 표본표준편차 | 표본비율 | 표본변동계수 |
| 0.5 | 0.6 | 0.5 |

<그림 4-4> 단순임의추출 출력결과(이분속성 데이터)

## 4-2. 계통추출

KESS 실행 후 계통추출을 위한 선택 및 입력사항은 단순임의추출과 동일하다. '계통추출' 버튼을 클릭하면 작업을 수행하여 <그림 4-5>의 결과창을 출력한다.


표본추출 결과

계통추출

| 번호 | 홈런 |
|---|---|
| 1 | 16 |
| 2 | 25 |
| 3 | 5 |
| 4 | 23 |
| 5 | 12 |
| 6 | 21 |
| 7 | 4 |
| 8 | 11 |
| 9 | 1 |
| 10 | 3 |
| 11 | 18 |
| 12 | 5 |
| 13 | 26 |
| 14 | 5 |
| 15 | 15 |
| 16 | 7 |
| 17 | 31 |
| 18 | 3 |
| 19 | 2 |
| 20 | 5 |
| 21 | 4 |
| 22 | 5 |
| 23 | 2 |
| 24 | 11 |
| 25 | 19 |

| 표본크기 | 표본합 | 표본평균 |
|---|---|---|
| 25 | 279 | 11.16 |
| 표본표준편차 | 표본비율표본변동계수 | |
| 8.905 | incomput. | incomput. |

**<그림 4-5> 계통추출 출력결과**

목적변수를 '홈런'으로 하여 작업을 수행하였으므로 임의추출과 마찬가지로 표본크기, 표본합, 표본평균, 표본표준편차를 제외한 통계량이 'incomput.(incomputable-계산불가)'로 출력되었다. 숫자로 나타내어지지 않은 목적변수의 경우, <그림 4-6>과 같이 표본크기만이 출력된다.

표본추출 결과

계통추출

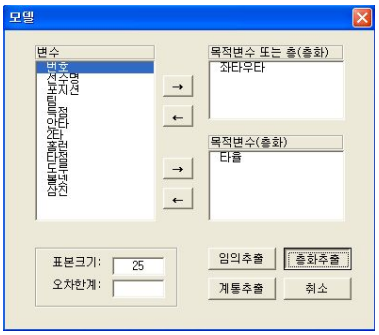
| 번호 | 선수명 |
|---|---|
| 1 | 김태균 |
| 2 | 강정호 |
| 3 | 손아섭 |
| 4 | 박석민 |
| 5 | 정성훈 |
| 6 | 이승엽 |
| 7 | 이진영 |
| 8 | 박용택 |
| 9 | 박한이 |
| 10 | 김원섭 |
| 11 | 이호준 |
| 12 | 이병규 |
| 13 | 최정 |
| 14 | 김주찬 |
| 15 | 홍성흔 |
| 16 | 김현수 |
| 17 | 박병호 |
| 18 | 안치홍 |
| 19 | 이용규 |
| 20 | 김선빈 |
| 21 | 이대수 |
| 22 | 양의지 |
| 23 | 김상수 |
| 24 | 나지완 |
| 25 | 강민호 |

| 표본크기 | 표본합 | 표본평균 |
|---|---|---|
| 25 | incomput. | incomput. |
| 표본표준편차 | 표본비율 | 표본변동계수 |
| incomput. | incomput. | incomput. |

<그림 4-6> 계통추출 출력결과(숫자가 아닌 목적변수)

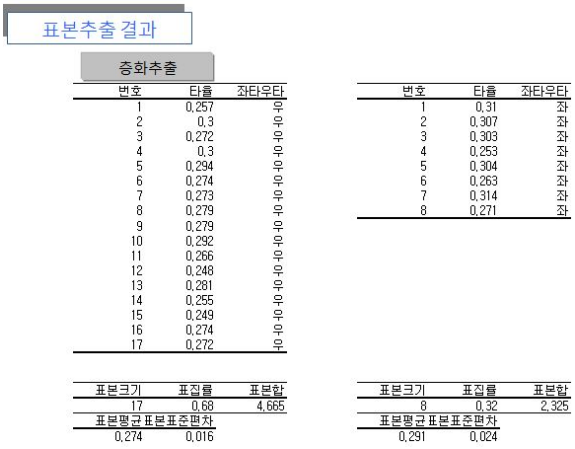
## 4-3. 층화추출

층화추출을 위한 초기입력화면은 <그림 4-7>와 같다.




**<그림 4-7> 층화추출 초기 입력화면**

앞선 두가지 추출과는 달리, 오른쪽 상단의 '목적변수 또는 층(층화)' 블록에 층으로 설정하고자 하는 변수를 선택하고, 그 아래 '목적변수(층화)' 블록에는 조사하고자 하는 목적변수를 입력한다. <그림 4-8>은 출력화면이다.




**<그림 4-8> 층화추출 출력화면**

## 5. 맺음말 및 추후과제

표본추출은 여러 통계 패키지에서 제공하고 있으나, 통계 비전공자들이 쉽게 활용할 수 있도록 Excel에서 구동되는 분석도구의 일환으로 제공하고자 하였다. 본 논문에서는 Excel에서 구동하는 표본추출틀을 개발하되, 사용자 입장에서 사용자의 요구를 만족시키고, 보다 편리하게 사용할 수 있도록 한다는 부분에 개발의 주안점을 두었다. 특히 작업 수행시에 초기 자료를 지정하는 데에 있어서 범위 지정이 다른 통계패키지에 비해 용이하다는 것이 큰 장점이다.

또한 본 논문에서는 표본추출을 KESS에 부메뉴로 추가시킴으로써 보다 넓은 차원에서 다양한 분야의 통계분석이 가능하도록 하여 KESS의 활용도를 높이는데 의의를 두었다.

하지만 본 논문에서는 층화추출의 경우 세 범주 이하로만 구분된 층을 변수로 지정할 수 있다는 데에 그 한계가 있으며, 추후에 다양한 범주를 가진 층을 기준으로 추출할 수 있도록 확장한다면 KESS가 표본추출의 도구로서 보다 높은 완성도를 가질 수 있을 것이다.

# 참 고 문 헌

W. Cochran (1977) Sampling Techniques, Wiley.

S. Thompson (2002) Sampling, Wiley.

S. Lohr (1999) Sampling: Design and Analysis, Dexbury Press.

R. Scheaffer (1990) Elementary Survey Sampling, Duxbury.

김현철 (2000) 표본의 추출과 분석, 교육과학사.

# Abstract

An-na Lee
Statistics, Time Series
The Graduate School
Seoul National University

Statistical packages such as SAS, SPSS have been wedely used for data analysis. Since most of them are developed in English, it is difficult to be used by general public. And they put some burden of cost on users.

In the thesis we summarize the fundamental concepts necessary for the development of statistical packages and develop an sampling package based on these concepts. Since most of the windows menu and results are provided in Korean, the package is more user-friendly. EXCEL is used as a basis tool for the development since it is the most widely used package and is easy to use for the input and output of the data.

keywords : Application Frame, Excel, Sampling

*Student Number* : 2010-23062