지역사회건강조사에서 소지역추정법으로 동읍면 흡연율 계산

1. 서언

지역사회건강조사는 보건소단위로 주민의 주요건강지표를 생산하기 위하여 보건소별로 19세 이상 성인 900여명을 층화집락추출법으로 선정하여 CAPI(computer assissted personal interviewing)를 이용한 조사이다. 보건소내의 동읍면과 주택유형을 층화변수로 사용하여 보건소별 표본크기 900명을 동읍면에 대해서 비례배분한 후에 동읍면내에서 주택유형(아파트와 단독주택)에 대해서 비례배분하고 표본지점인 통반리를 가구수를 기준으로 확률비례추출법으로 선정한다. 그 후에 선정된 표본지점에서 평균적으로 5가구를 계통추출법으로 선정하고 추출된 조사대상가구에서는 19세 이상 모든 성인을 대상으로 개별방문면접조사를 통해서 주민의 건강실태를 조사한다(이계오 외4인, 2013). 보건소단위에서는 일정수준의 정도(precision)를 갖는 건강지표를 생산하고 있으나 보건소내 동읍면단위로 주요건강지표 생산의 요구가 많아지고 있다. 동읍면단위의 조사된 표본규모는 적게는 10여명에서 많게는 수백명에 이르고 있으므로 표본규모가 30명이하인 동읍면의 흡연율과 같은 건강지표를 전통적인 통계추정법으로 생산할 경우에는 추정치의 표본분산이 너무 커서 이용할 수 없으므로 Ghosh and Rao(1994)가 설명한 소지역추정법(small area estimation)과 같은 특별한 추정법으로 동읍면단위 통계치의 계산이 필요하여 SAS를 이용한 계산방법을 설명하고자 한다.

2. 소지역추정법

표본설계 당시에 통계 생산 단위로 고려되지 않았으나 조사 후에 소영역에 대한 통계를 생산하고자 할 때 소영역에 할당된 표본규모가 작기 때문에 추정값의 분산이 커지게 되므로 이를 보완하기 위해서 주변의 조사정보를 이용하거나 다른 source의 보조정보를 이용 또는 모집단의 통계적 모형 구조를 이용하는 추정기법을 소지역추정법(small area estimation)이라 한다(이계오 외3인, 2001; Gonzalez, Placek and Scott, 1993).

지역사회건강조사에서도 보건소단위로 건강지표를 산출하기 위한 표본설계를 하

였으나 조사완료 후에 소영역인 동읍면단위의 건강지표를 안정적으로 생산하기 위해서는 소지역추정법을 적용할 필요가 있으며 보건소내의 동읍면별 건강지표를 산출하는데 적용할 수 있는 소지역추정법을 설명하겠다(이계오, 2000).

(1) 직접추정량(Direct estimator)

지역사회건강조사에서 수집한 자료중에서 동읍면별로 해당되는 자료만을 이용하여 건강지표를 추정하는 것이며 조사된 데이터세트에 조사항목별로 조사된 사례에 가중치가 부여되었으므로 가중값(w_j^i)와 관찰값(y_j^i)를 이용한 표본설계 기반의 직접추정량과 분산의 추정식을 아래와 같이 나타낼 수 있다.

$$\hat{\bar{Y}}_{\mathrm{D}}^{\mathrm{i}} = \frac{\sum_{j=1}^{n_i} w_j^i y_j^i}{\sum_{j=1}^{n_i} w_j^i} \quad (1)$$

여기서, n_i는 i 동읍면 표본 수이고 w_j^i는 추출률과 응답률을 고려한 승수이며 y_j^i는 관찰값이다. 식(1)에 주어진 추정량의 분산 추정식은 식(2)로 나타낼 수 있다.

$$\widehat{Var}\left(\hat{\bar{Y}}_D^i\right) = \frac{1}{n_i(n_i-1)\bar{w}_i^2}\sum_{j=1}^{n_i}\left(w_j^i\right)^2\left(y_j^i - \hat{R}_\iota\right)^2 \quad (2)$$

여기서$\hat{R}_\iota = \frac{\sum_{j=1}^{n_i} w_j^i y_j^i}{\sum_{j=1}^{n_i} w_j^i}$이고, $\bar{w}_{\mathrm{i}} = \frac{1}{n_i}\sum_{j=1}^{n_i} w_j^i$ 이다.

(2) 합성추정량(Synthetic estimator)

보건소내 동읍면별 성별과 연령대별 주민등록인구를 보조정보로 사용하여 동읍면별 흡연율의 추정치를 좀더 정확하게 산출하는 추정량이 합성추정량(synthetic estimator)이다(이계오, 2000 ; Ghosh and Rao, 1994). 보건소내의 동읍면의 사회생활환경과 인구구성비를 집락변수(clustering variable)로 사용하여 동읍면들을 2-3개의 동질적인 집락으로 구분한 후에 각 집락별로 집락내의 동읍면들은 성별과 연령대별로 흡연율과 같은 건강지표가 유사할 것이라는 가정을 할 수 있으므로 성별*연령대별 흡연율과 동읍면의 주민등록인구수를 결합하여 동읍면의 흡연율을 아래와 같이 계산할 수 있다.

$$\hat{\bar{Y}}_S^i = \frac{\sum_{k=1}^{n_c} r_{jk} N_{jk}^i}{\sum_{j=1}^{n_c} N_{jk}^i} \quad (3)$$

여기서 $r_{jk} = \frac{\sum_{l1}^{n_g} w_{jkl} y_{jkl}}{\sum_{l=1}^{n_g} w_{jkl}}$ 는 j그룹 내 k범주의 평균 추정값을, N_{jk}^i는 j그룹 내 i동읍면, k범주의 주민등록인구를, n_g는 그룹 내의 k범주 표본 수를, n_c는 j그룹 내의 범주의 수를 의미한다.

식(3)에 주어진 추정량의 분산의 추정식은 식(4)와 같다.

$$\widehat{Var}\left(\hat{\bar{Y}}_S^i\right) = \sum_{k=1}^{n_c} \left(Z_{jk}^i\right)^2 \widehat{Var}\left(r_{jk}^i\right) \quad (4)$$

$$= \sum_{k=1}^{n_c} \frac{\left(Z_{jk}^i\right)^2}{n_g\left(n_g - 1\right)\bar{w}_{jk}^2 \sum_{l=1}^{n_g} w_{jkl}\left(y_{jkl} - r_{jk}\right)^2}$$

여기서 $z_{jk}^i = \frac{N_{jk}^i}{\sum_{k=1}^{n_c} N_{jk}^i}$이다.

합성추정량은 일종의 편향 추정량이지만 보건소내의 동읍면별로 집락화를 잘 했을 경우에는 각 집락(그룹)내에서 동읍면별로 성별-연령대별 범주의 특성이 유사하게 될 것이므로 이 경우에는 편향을 무시해도 될 것이다. 만일에 편향이 무시할 정도가 아니라면 식(4)에 주어진 분산의 추정식은 식(3)의 추정량의 추정오차를 과소 추정하게 될 가능성이 있기 때문에 집락화를 통한 그룹의 구분에 유의해야 한다.

(3) 복합추정량(composite estimator)

식(1)에 주어진 직접추정량은 불편 추정량이지만 표본 크기가 크지 않기 때문에 표준오차가 클 뿐만 아니라 추정값이 불안정하고, 식(3)에 주어진 합성 추정량은 편향을 갖기 때문에 두 추정량의 문제점을 보완하여 보다 안정된 추정량을 얻는 방법은 두 추정량의 가중 평균 형식의 추정법을 고려할 수 있다. 식(1)과 식(3)에 주어진 추정량들의 가중평균형식을 복합추정량(composite estimator)이라 하고 아래와 같이 계산할 수 있다.

$$\hat{\bar{Y}}_C^i = \alpha \hat{\bar{Y}}_D^i + (1-\alpha)\hat{\bar{Y}}_S^i \quad (5)$$

여기서 α는 $\mathrm{MSE}(\hat{\bar{Y}}_C^i)$를 최소화하는 가중값이 되어야하므로 다음과 같이 계산된다.

$$\alpha = \frac{\widehat{Var}(\hat{\bar{Y}}_S^i)}{\widehat{Var}(\hat{\bar{Y}}_D^i)+\widehat{Var}(\hat{\bar{Y}}_S^i)} \quad (6)$$

α의 최적값은 $\hat{\bar{Y}}_C^i$의 평균제곱오차를 최소화하는 값이 되어야 하지만 합성추정량 $\hat{\bar{Y}}_S^i$의 편향의 크기가 무시될 정도로 동읍면의 집락화가 잘 되고 직접추정량과 합성추정량이 서로 독립이라는 가정에서 식(6)에 의해서 α를 계산한다.

3. SAS이용 흡연율 계산

2013년 지역사회건강조사 자료를 이용하여 보건소내 동읍면단위의 흡연율을 추정하는 방법을 설명하기 위해서 앞에서 설명한 소지역추정법을 서울시 강남구 22개 동별 흡연율 계산에 적용하여 계산과정을 설명하겠다.

(1) 직접추정량

2013년에 조사된 강남구 보건소의 데이터에서 22개 동별 표본수의 분포를 보면 표본수가 가장 작은 동은 개포4동으로 24명이고 가장 많은 동은 역삼1동으로 58명이다. 22개별로 흡연율을 식(1)에 주어진 직접추정량과 식(2)에 주어진 분산추정식으로 계산하기 위해서 아래와 같은 SAS코드를 사용하였다(R코드 프로그램은 이계오(2014) 연구보고서 참조).

```
/*서울시 강남구 보건소 데이터 구성-연령그룹 및 흡연여부 변수 생성*/
data abc.seoul_gangnam_data;
      set abc.chs13;
      length age_group $8.0
      keep josa_year dong sm_a0100 sma_01z2 sma_03z2 age age_group
sex            wt;
      rename dong=읍면동;;\
      if 19<=age<=39 then age_group="19-39세";
      if 40<=age<=59 then age_group="40-59세";
      if 60<=age then age_group="60세이상";
  **7.현재흡연율(다른 조사항목 변수 설명은 질병관리본부(2012) 참조)
      ==========================================
      변수명 :  sm_a0100 (현재 흡연율 산출 변수)
     분석 데이터 변수명 : sma_01z2(평생 흡연여부) sma_03z2(현재 흡연여부)
      ==========================================;
      if sma_01z2 = 1 then do ;
            if sma_03z2 in (1,2) then sm_a0100 = 1 ;
            else if sma_03z2 = 3 then sm_a0100 = 0 ;
      end ;
      else if sma_01z2 = 2 then do ;
            sm_a0100 = 0 ;
```

```
    end ;
    if bogun_cd=001; (강남구 보건소 코드)
run;

/*서울시 강남구 직접추정량과 분산추정*/
proc surveymeans data=abc.seoul_gangnam_data;
    var sm_a0100; (현재 흡연율 산출 변수)
    domain 읍면동;
    weight wt; ( 표본설계가중치)
    ods output Domain=abc.direct_estimator;
run;
```

(2) 합성추정량

동별 흡연율의 합성추정량을 계산하기 위해서는 강남구의 22개 동을 3개의 집락으로 구분하고 각 집락별로 성별*연령대별(19-39세, 40-59세, 60세이상) 의 흡연율을 계산한 다음에 동별 흡연율과 분산을 각각 식(3)과 식(4)를 이용하여 계산하는데 세부적인 계산과정은 아래와 같다.

① 동별 2013년7월말기준 성별*연령대별 주민등록인구소아 구성비를 계산한다.

<표 1> 동별 성별 연령대의 주민등록 인구구성비

동	남19_39세	남40_59세	남60세이상	여19_39세	여40_59세	여60세이상
신사동	0.188	0.174	0.094	0.243	0.192	0.109
논현1동	0.241	0.149	0.066	0.309	0.155	0.080
논현2동	0.220	0.158	0.083	0.268	0.175	0.095
압구정동	0.172	0.178	0.106	0.209	0.206	0.128
청담동	0.198	0.182	0.085	0.228	0.209	0.098
삼성1동	0.203	0.191	0.091	0.200	0.217	0.099
삼성2동	0.203	0.189	0.069	0.257	0.198	0.084
대치1동	0.140	0.262	0.073	0.159	0.287	0.078
대치2동	0.175	0.238	0.074	0.173	0.259	0.080
대치4동	0.211	0.188	0.058	0.257	0.216	0.069
역삼1동	0.256	0.155	0.064	0.310	0.145	0.070
역삼2동	0.194	0.199	0.067	0.247	0.210	0.083

도곡1동	0.198	0.195	0.081	0.223	0.214	0.090
도곡2동	0.167	0.211	0.090	0.203	0.239	0.091
개포1동	0.193	0.193	0.087	0.189	0.234	0.105
개포2동	0.209	0.204	0.061	0.211	0.233	0.083
개포4동	0.207	0.204	0.070	0.208	0.226	0.085
세곡동	0.188	0.193	0.105	0.188	0.194	0.132
일원본동	0.182	0.226	0.065	0.195	0.253	0.079
일원1동	0.209	0.179	0.091	0.196	0.205	0.120
일원2동	0.205	0.189	0.077	0.193	0.221	0.114
수서동	0.205	0.144	0.101	0.185	0.191	0.173

② 22개 동을 성별*연령대별 인구구성비와 흡연율 등을 집락화변수로 사용하여 k-means방법으로 집락화 한다. 3개 집락으로 구분된 결과를 주민등록인구수와 사회생활환경여건의 유사성 등을 고려하여 집락구분을 검토한 후에 확정한다.

```
/*서울시 강남구2013년 성별_연령대별 인구구성비 흡연율을 통한 군집 분석*/
/*데이터 불러오기*/

proc import out=ABC.seoul_gangnam_cluster_data datafile="D:\2014연구활
                동\sas강의\서울강남구_군집화데이터" datafiel=dbms=excel
                replace;  rage="서울$“  getnames=yes;  mixed=no;
                scantext=yes; usedate=ye; scantime=yes;
run;
/*k-평균 집락화*/
proc fastclus data=abc.seoul_gangnam_cluster_data
     maxc=3 out=abc.seoul_gangnam_kcluster;
          var  남2013_19_39세  남2013_40_59세  남2013_60세이상  여
          2013_19_39세 여2013_40_59세 여2013_60세이상;
     id 읍면동;
run;

/*군집 엑셀로 보내기*/
proc exprot data=abc.seoul_gangnam_kcluster OUTFILE= "D:\2014연구활
                동\sas강의\서울강남구_군집분석" label dbms=excel replace;
```

```
run;

/*군집 수정하기*/
proc import out=abc.seoul_gangnam_rcluster datafile="D:\2014연구활동\sas
                강의\서울강남구_수정군집화" dmbs=excel replace; rage="서
                울$" getnames=yes;
mixed=no; scantext=yes; usedate=yes; scantime=yes;
run;
```

③ 3개 집락별로 집락내의 성별*연령대별 흡연율을 계산한다.

```
/*서울시 강남구 데이터에 수정군집 통합*/
proc sort data=abc.seoul_gangnam_data; by 읍면동; run
proc sort data=abc.seoul_gangnam_rcluster; by 읍면동; run
data abc.seoul_gangnam_data2;
    merge abc.seoul_gangnam_data abc.seoul_gangnam_rcluster;
    by 읍면동;
    group1=수정군집||"_"||sex||"_"||age_group;
    group1=compress(group1);
run;
proc print data=abc.seoul_gangnam_data2; run;
```

<표 2> 수정군집, 성별과 연령대와 흡연율 산출변수 포함 데이터 구조 예시

OBS	읍면동	sex	sma_01z2	sma_03z2	age	wt	age_group	sm_a0100	Cluster	수정군집	group1
1	논현1동	1	1	1	26	592.15	19-39세	1	2	1	1_1_19-39세
2	논현1동	1	1	1	36	647.7	19-39세	1	2	1	1_1_19-39세
3	논현1동	1	1	1	35	647.46	19-39세	1	2	1	1_1_19-39세
4	논현1동	1	1	1	25	592.15	19-39세	1	2	1	1_1_19-39세
5	논현1동	1	2	.	19	592.15	19-39세	0	2	1	1_1_19-39세
6	논현1동	1	2	.	32	647.46	19-39세	0	2	1	1_1_19-39세
7	논현1동	1	1	1	32	647.46	19-39세	1	2	1	1_1_19-39세
8	논현1동	1	1	1	39	647.46	19-39세	1	2	1	1_1_19-39세

```
/* 집락내 성별*연령대별 흡연율 추정치 계산 */
proc surveymeans data=abc.seoul_gangnam_data2 mean;
        var sm_a0100;
        domain group1;
        weight wt;
        ods output Domain=abc.com_estimator_r;
run;
```

<표 3> 그룹별(군집, 성별과 연령대) 현재 흡연율 추정치와 표준오차

group	Variable	Mean	Std Error of Mean
1_1_19-39세	sm_a0100	0.465403	0.074305
1_1_40-59세	sm_a0100	0.474124	0.075564
1_1_60세이상	sm_a0100	0.241961	0.084827
1_2_19-39세	sm_a0100	0.167799	0.056109
1_2_40-59세	sm_a0100	0.074688	0.032533
1_2_60세이상	sm_a0100	0.031734	0.031296
2_1_19-39세	sm_a0100	0.364547	0.074057
2_1_40-59세	sm_a0100	0.419048	0.062085
2_1_60세이상	sm_a0100	0.160079	0.062115
2_2_19-39세	sm_a0100	0.137315	0.047784
2_2_40-59세	sm_a0100	0.05688	0.029818
2_2_60세이상	sm_a0100	0	0
3_1_19-39세	sm_a0100	0.412643	0.068541
3_1_40-59세	sm_a0100	0.317444	0.065528
3_1_60세이상	sm_a0100	0.207356	0.08519
3_2_19-39세	sm_a0100	0.064959	0.03821
3_2_40-59세	sm_a0100	0	0
3_2_60세이상	sm_a0100	0.134244	0.056051

④ 22개동별 흡연율의 합성추정치와 분산을 계산한다.

```
/*집락내 성별*연령대별 흡연율 추정치 계산 데이터 준비 */
/*서울시 강남구 데이터에 수정군집 통합*/
proc sort data=abc.seoul_gangnam_data; by 읍면동; run
proc sort data=abc.seoul_gangnam_rcluster; by 읍면동; run
data abc.seoul_gangnam_data2;
      merge abc.seoul_gangnam_data abc.seoul_gangnam_rcluster;
      by 읍면동;
      group1=수정군집||"_"||sex||"_"||age_group;
      group1=compress(group1);
run;


/*서울 강남구 22개 동별 합성추정량*/
proc surveymeans data=abc.seoul_gangnam_com mean;
      domain 읍면동;
      var mean;
      weight 인구;
      ods output Domain=abc.seoul_gangnam_comestimator_mean;
run;
```

<표 4> 동별 현재 흡연율의 합성추정치와 표준오차

동	Mean	Std Error of Mean
개포1동	0.185167	0.070934
개포2동	0.188473	0.074533
개포4동	0.18973	0.073423
논현1동	0.264952	0.073975
논현2동	0.258582	0.073317
대치1동	0.210825	0.076593
대치2동	0.213915	0.072523
대치4동	0.212771	0.06561
도곡1동	0.209736	0.065891
도곡2동	0.205124	0.067728
삼성1동	0.259836	0.076925

삼성2동	0.261522	0.075489
세곡동	0.190447	0.066024
수서동	0.186681	0.066364
신사동	0.251396	0.072852
압구정동	0.245006	0.073481
역삼1동	0.219177	0.061773
역삼2동	0.210728	0.065803
일원1동	0.190545	0.069906
일원2동	0.188713	0.071434
일원본동	0.183653	0.073479
청담동	0.256171	0.075264

```
/* 22개 동별 합성추정량의 분산추정*/
proc sort data=abc.seoul_gangnam_data2; by group1; run
data abc.seoul_gangnam_comvar;
      merge abc.seoul_gangnam_data2 abc.com_estimator_r;
      by group1;
      sum_wj_yj_rj=((wt*wt)*((sm_a0100-mean)*(sm_a0100-mean)));
run;
/*그룹내 k범주의 가중치의 합*/
proc surveymeans data=abc.seoul_gangnam_comvar;
      domain group1;
      var wt;
      ods output domain=abc.seoul_gangnam_comvar_1;
run;

/*그룹내 k 범주의 분산을 위한 합*/
proc tabulate data=abc.seoul_gangnam_comvar;
      class group1;
      var sum_wj_yj_rj;
      table group1*sum_wj_yj_rj;
      ods output table=abc.seoul_gangnam_comvar_2;
run;
```

```
data abc.seoul_gangnam_comvar_data1;
                    merge                                   abc.seoul_gangnam_pop
                    abc.seoul_gangnam_comvar_1
                    abc.seoul_gangnam_comvar_2;
      by group1;
           keep   읍면동   수정군집   성별   연령그룹   인구   N   Mean
           sum_wj_yj_rj_sum group1;
run;
```

<표 5> 그룹별(군집, 성별과 연령대) 합성추정치의 분산

OBS	읍면동	수정 군집	성별	연령그룹	인구	group1	N	Mean (wt)	sum_wj_yj_ rj_Sum
1	신사동	1	1	19-39세	3071	1_1_19-39세	47	629.3595	4825692
2	논현1동	1	1	19-39세	5629	1_1_19-39세	47	629.3595	4825692
3	논현2동	1	1	19-39세	4267	1_1_19-39세	47	629.3595	4825692
4	압구정동	1	1	19-39세	4014	1_1_19-39세	47	629.3595	4825692
5	청담동	1	1	19-39세	5060	1_1_19-39세	47	629.3595	4825692
6	삼성1동	1	1	19-39세	2577	1_1_19-39세	47	629.3595	4825692
7	삼성2동	1	1	19-39세	5224	1_1_19-39세	47	629.3595	4825692
8	신사동	1	1	40-59세	2841	1_1_40-59세	48	470.9592	2914779
9	논현1동	1	1	40-59세	3476	1_1_40-59세	48	470.9592	2914779
10	논현2동	1	1	40-59세	3061	1_1_40-59세	48	470.9592	2914779
11	압구정동	1	1	40-59세	4153	1_1_40-59세	48	470.9592	2914779
12	청담동	1	1	40-59세	4659	1_1_40-59세	48	470.9592	2914779
13	삼성1동	1	1	40-59세	2427	1_1_40-59세	48	470.9592	2914779
14	삼성2동	1	1	40-59세	4852	1_1_40-59세	48	470.9592	2914779
15	신사동	1	1	60세이상	1527	1_1_60세이상	27	414.0746	898420.2
16	논현1동	1	1	60세이상	1540	1_1_60세이상	27	414.0746	898420.2
17	논현2동	1	1	60세이상	1613	1_1_60세이상	27	414.0746	898420.2
18	압구정동	1	1	60세이상	2462	1_1_60세이상	27	414.0746	898420.2
19	청담동	1	1	60세이상	2179	1_1_60세이상	27	414.0746	898420.2

```
proc surveymeans data=abc.seoul_gangnam_comvar_data1 sum;
     domain 읍면동;
     var 인구;
```

```
    ods output domain=abc.seoul_gangnam_comvar_data2;
run;

proc sort data=abc.seoul_gangnam_comvar_data1; by 읍면동; run;
proc sort data=abc.seoul_gangnam_comvar_data2; by 읍면동; run;

/*합성추정량 분산추정*/
data abc.seoul_gangnam_comvar_data;
        merge abc.seoul_gangnam_comvar_data1
        abc.seoul_gangnam_comvar_data2;
    by 읍면동;
    drop varname varlabel stddev DomainLabel;
    Zjk=인구/Sum;
    Var=((Zjk*Zjk)/((N*(N-1))*(mean*mean)))*sum_wj_yj_rj_sum;
run;
```

<표 6> 동별 그룹별 합성추정치의 분산계산

OBS	읍면동	수정군집	성별	연령그룹	인구	group1	N	Mean	sum_wj_yj_rj_Sum	Sum	Zjk	Var
1	개포1동	3	1	19-39세	3677	3_1_19-39세	54	615.2674	5180109	19069	0.19283	0.000178
2	개포1동	3	1	40-59세	3677	3_1_40-59세	58	461.3913	3071663	19069	0.19283	0.000162
3	개포1동	3	1	60세이상	1659	3_1_60세이상	29	406.8678	1009267	19069	0.087	5.68E-05
4	개포1동	3	2	19-39세	3601	3_2_19-39세	64	583.2234	2031942	19069	0.18884	5.28E-05
5	개포1동	3	2	40-59세	4455	3_2_40-59세	78	419.5206	0	19069	0.23363	0
6	개포1동	3	2	60세이상	2000	3_2_60세이상	37	411.813	728625.3	19069	0.10488	3.55E-05
7	개포2동	3	1	19-39세	5720	3_1_19-39세	54	615.2674	5180109	27359	0.20907	0.000209
8	개포2동	3	1	40-59세	5590	3_1_40-59세	58	461.3913	3071663	27359	0.20432	0.000182
9	개포2동	3	1	60세이상	1656	3_1_60세이상	29	406.8678	1009267	27359	0.06053	2.75E-05

10	개포2동	3	2	19-39세	5760	3_2_19-39세	64	583.2234	2031942	27359	0.21053	6.57E-05
11	개포2동	3	2	40-59세	6368	3_2_40-59세	78	419.5206	0	27359	0.23276	0
12	개포2동	3	2	60세이상	2265	3_2_60세이상	37	411.813	728625.3	27359	0.08279	2.21E-05

```
proc surveymeans data=abc.seoul_gangnam_comvar_data sum;
      domain 읍면동;
      var var;
      ods output domain=abc.seoul_gangnam_comvariance;
run;

/* ⑤ 동별 합성치와 분산추정치의 통합 데이터세트 */
/*합성추정량과 분산*/
data abc.seoul_gangnam_estimator_com;
            merge abc.seoul_gangnam_comestimator_mean
            abc.seoul_gangnam_comvariance;
      by 읍면동;
      drop DomainLabel VarName stderr StdDev;
      rename mean=Y_s sum=var_Y_s;
run;
```

(3) 복합추정량

앞에서 계산한 동별 흡연율의 직접추정치와 합성추정치를 가중평균으로 결합하여 복합추정치를 계산하는데 가중치를 계산하는 방법으로 다음 3가지를 적용한다.

먼저 식(5)로 주어진 복합추정량의 평균제곱오차 $\mathrm{MSE}(\hat{Y}_c^i)$를 최소화하는 α_i는 아래와 같다.

$$\alpha_{i(opt)} = \frac{MSE(\hat{Y}_S^i)}{MSE(\hat{Y}_S^i)+V(\hat{Y}_D^i)} \quad (7)$$

최적 가중값 $\alpha_{i(opt)}$의 추정값은 다음 식으로 계산된다.

$$\hat{\alpha}_{i(opt)} = \frac{MSE(\hat{Y}_S^i)}{\left(\hat{Y}_S^i - \widehat{\overline{Y}^i}\right)^2} \quad (8)$$

모든 소영역에 공통 가중값을 부여하는 방법으로써 초기 공통 가중값 α을 이용하여 $\mathrm{MSE}(\hat{Y}_S^i)$들의 평균을 최소화하는 가중값은 아래와 같다.

$$\hat{\alpha}(opt) = 1 - \frac{\sum_i \hat{V}(\hat{Y}_D^i)}{\sum_i (\hat{Y}_S^i - \hat{Y}_i)^2} \quad (9)$$

각 소영역에 배정된 표본 크기에 의존하는 가중값은 다음과 같이 계산된다.

$$\alpha_i(\delta) = \begin{pmatrix} 1, & \hat{N}_i \geqq \delta N_i \\ \frac{\tilde{N}_i}{\delta N_i} & (그외) \end{pmatrix} \quad (10)$$

단, N_i는 i 소영역의 크기이며 $\hat{N}_i = N(\frac{n_i}{n})$이다. $\tilde{N}_i$ 는 직접추정량이며 δ는 합성추정량의 기여도를 조정하는 값이므로 주관적으로 결정한 값이다. 예를 들어 캐나다 노동력 통계조사에서는 δ = 2/3을 사용하므로 본 계산에서도 δ = 2/3을 사용한다 (Singh, Gambino and Mantel, 1994).

위에서 주어진 3종의 가중치별로 동별 흡연율의 복합추정치를 계산한 후에 적합한 추정방법을 선택할 것이며 복합추정치의 세부계산 절차는 아래와 같다.

① 22개 동별로 계산된 직접추정치(Y_d)와 합성추정치(Y_s)를 통합하여 데이터 세트를 구성한다.

```
/*직접추정량 합성추정량 (추정치와 분산)*/
data abc.seoul_gangnam_estimators;
    merge                          abc.seoul_gangnam_estimator_direct
abc.seoul_gangnam_estimator_com;
    by 읍면동;
run;
```

<표 7> 동별 현재 흡연율의 직접추정치와 합성추정치 계산결과

OBS	읍면동	N	Y_d	Var_Y_d	수정군집	Y_s	var_Y_s
1	개포1동	51	0.149427	0.003041	3	0.185167	0.000485
2	개포2동	52	0.163193	0.004069	3	0.188473	0.000506
3	개포4동	24	0.155198	0.00694	3	0.18973	0.000511
4	논현1동	41	0.383208	0.006272	1	0.264952	0.000828
5	논현2동	28	0.443943	0.009638	1	0.258582	0.000742
6	대치1동	30	0.186235	0.005946	2	0.210825	0.000528
7	대치2동	53	0.070319	0.00102	2	0.213915	0.000541
8	대치4동	31	0.357597	0.007662	2	0.212771	0.000595
9	도곡1동	44	0.185543	0.003435	2	0.209736	0.000548
10	도곡2동	54	0.075201	0.001131	2	0.205124	0.000506
11	삼성1동	35	0.211848	0.005182	1	0.259836	0.000694
12	삼성2동	51	0.240701	0.004155	1	0.261522	0.000737
13	세곡동	25	0.108247	0.003576	3	0.190447	0.000522
14	수서동	40	0.259566	0.00515	3	0.186681	0.000516
15	신사동	36	0.217951	0.005245	1	0.251396	0.000682
16	압구정동	36	0.156356	0.005147	1	0.245006	0.000639
17	역삼1동	58	0.365498	0.004569	2	0.219177	0.000717
18	역삼2동	55	0.267025	0.005251	2	0.210728	0.000562
19	일원1동	38	0.218381	0.005347	3	0.190545	0.000513
20	일원2동	44	0.166962	0.003452	3	0.188713	0.0005
21	일원본동	46	0.241629	0.004493	3	0.183653	0.00049
22	청담동	49	0.13509	0.002371	1	0.256171	0.000692

② 첫 번째 가중치를 사용한 복합추정치1(Y_c1)을 계산한다.

```
/*복합추정량1*/
data abc.seoul_gangnam_estimator_c1;
      set abc.seoul_gangnam_estimators;
      alpha1=Var_Y_s/(Var_Y_d+Var_Y_s);
      Y_c1=(alpha1*Y_d)+((1-alpha1)*Y_s);
```

```
Var_Y_c1=((alpha1*alpha1)*Var_y_d)+(((1-alpha1)*(1-
alpha1))*var_Y_s);
        sumvar_Ys_Yd=(var_Y_s+var_Y_d);
run;

/*수정군집별 직접추정량의 분산과 직접추정량분산+합성추정량*/
proc surveymeans data=abc.seoul_gangnam_estimator_c1 sum;
        domain 수정군집;
        var Var_Y_d;
        ods output domain=abc1;
run;
data abc1;
        set abc1;
        rename sum=sum1;
run;
proc surveymeans data=abc.seoul_gangnam_estimator_c1 sum;
        domain 수정군집;
        var sumvar_Ys_Yd;
        ods output domain=abc2;
run;
data abc3;
        merge abc1 abc2;
        by 수정군집;
        alpha2=1-(sum1/sum);
        keep 수정군집 alpha2;
run;
proc sort data=abc.seoul_gangnam_estimator_c1; by 수정군집; run;
```

③ 두 번째 가중치를 사용한 복합추정치2(Y_c2)을 계산한다.

```
/*복합추정량2*/
data abc.seoul_gangnam_estimator_c2;
      merge abc.seoul_gangnam_estimator_c1 abc3;
      by 수정군집;
      Y_c2=(alpha2*Y_d)+((1-alpha2)*Y_s);
      Var_Y_c2=((alpha2*alpha2)*Var_y_d)+(((1-alpha2)*(1-
alpha2))*var_Y_s);
run;

proc surveymeans data=abc.seoul_gangnam_pop sum; domain 읍면동; var
                              인구; ods output domain=abc4; run;
proc surveymeans data=abc.seoul_gangnam_pop sum; domain 수정군집; var
                              인구; ods output domain=abc5; run;
proc surveymeans data=abc.seoul_gangnam_estimator_c2 sum; domain 수정
                              군집; var N; ods output domain=abc6; run;
proc sort data=abc4; by 읍면동; run;
data abc5; set abc5; rename Sum=집락인구수; run;
proc sort data=abc5; by 수정군집; run;
data abc6; set abc6; rename Sum=집락표본수; run;
proc sort data=abc6; by 수정군집; run;
proc sort data=abc.seoul_gangnam_estimator_c2; by 읍면동; run;
data abc.seoul_gangnam_estimator_c3_1;
      merge abc.seoul_gangnam_estimator_c2 abc4;
      by 읍면동;
      drop DomainLabel VarName VarLabel StdDev;
      rename sum=주민등록인구수;
run;
proc sort data=abc.seoul_gangnam_estimator_c3_1; by 수정군집; run;
data abc.seoul_gangnam_estimator_c3_2;
      merge abc.seoul_gangnam_estimator_c3_1 abc5 abc6;
      by 수정군집;
      drop DomainLabel VarName VarLabel StdDev;
run;
```

④ 세 번째 가중치를 사용한 복합추정치3(Y_c3)을 계산한다.

```
/*복합추정량3*/
data abc.seoul_gangnam_estimator_c3;
      set abc.seoul_gangnam_estimator_c3_2;
      hat_N_i=집락인구수*(N/집락표본수);
      if hat_N_i>=((2/3)*주민등록인구수) then
            alpha3=1
      else alpha3=hat_N_i/((2/3)*주민등록인구수);
      Y_c3=(alpha3*Y_d)+((1-alpha3)*Y_s);
      Var_Y_c3=((alpha3*alpha3)*Var_y_d)+(((1-alpha3)*(1-
alpha3))*var_Y_s);
run;
```

⑤ 첫 번째 가중치와 세 번째 가중치의 평균가중치를 사용한 복합추정치4(Y_c4)를 계산한다.

```
/*복합추정량4*/
data abc.seoul_gangnam_estimator_c4;
      set abc.seoul_gangnam_estimator_c3;
      alpha4=(alpha1+alpha3)/2
      Y_c4=(alpha4*Y_d)+((1-alpha4)*Y_s);
      Var_Y_c4=((alpha4*alpha4)*Var_y_d)+(((1-alpha4)*(1-
alpha4))*var_Y_s);
run;

/*직접추정량_합성추정량_복합추정량*/
data abc.estimator_total;
      set abc.seoul_gangnam_estimator_c4;
            keep 읍면동 Y_d Var_Y_d Y_s var_y_s alpha1 Y_c1 var_Y_c1
            alpha3 Y_c3 var_y_c3 Y_c4 var_y_c4;
run;
proc print data=abc.estimator_total; run;
```

⑥ 4종의 복합추정법에 따른 추정결과 요약

<표 8> 동별 현재 흡연율의 4종 복합추정치의 계산결과

읍면동	표본수	Y_d	V_d	Y_c1	V_c1	Y_c2	V_c2	Y_c3	V_c3	Y_c4	V_c4
논현1동	41	0.3832	0.0063	0.2787	0.0007	0.2787	0.0007	0.3832	0.0063	0.3310	0.0021
논현2동	28	0.4439	0.0096	0.2718	0.0007	0.2802	0.0007	0.4439	0.0096	0.3579	0.0029
삼성1동	35	0.2118	0.0052	0.2542	0.0006	0.2542	0.0006	0.2119	0.0052	0.2330	0.0018
삼성2동	51	0.2407	0.0042	0.2584	0.0006	0.2591	0.0006	0.2407	0.0042	0.2495	0.0015
신사동	36	0.2180	0.0052	0.2476	0.0006	0.2475	0.0006	0.2180	0.0052	0.2328	0.0018
압구정동	36	0.1564	0.0051	0.2352	0.0006	0.2347	0.0006	0.1564	0.0051	0.1958	0.0017
청담동	49	0.1351	0.0024	0.2288	0.0005	0.2421	0.0006	0.1351	0.0024	0.1820	0.0010
대치1동	30	0.1862	0.0059	0.2088	0.0005	0.2079	0.0005	0.1862	0.0059	0.1975	0.0018
대치2동	53	0.0703	0.0010	0.1642	0.0004	0.1965	0.0004	0.0703	0.0010	0.1173	0.0005
대치4동	31	0.3576	0.0077	0.2232	0.0006	0.2303	0.0006	0.3576	0.0077	0.2904	0.0023
도곡1동	44	0.1855	0.0034	0.2064	0.0005	0.2068	0.0005	0.1855	0.0034	0.1960	0.0012
도곡2동	54	0.0752	0.0011	0.1650	0.0003	0.1894	0.0004	0.0752	0.0011	0.1201	0.0005
역삼1동	58	0.3655	0.0046	0.2390	0.0006	0.2369	0.0006	0.3655	0.0046	0.3023	0.0016
역삼2동	55	0.2670	0.0053	0.2162	0.0005	0.2175	0.0005	0.2670	0.0053	0.2416	0.0017
개포1동	51	0.1494	0.0030	0.1803	0.0004	0.1816	0.0004	0.1494	0.0030	0.1648	0.0011
개포2동	52	0.1632	0.0041	0.1857	0.0005	0.1859	0.0005	0.1632	0.0041	0.1744	0.0014
개포4동	24	0.1552	0.0069	0.1874	0.0005	0.1863	0.0005	0.1610	0.0048	0.1742	0.0016
세곡동	25	0.1082	0.0036	0.1800	0.0005	0.1822	0.0005	0.1083	0.0036	0.1441	0.0012
수서동	40	0.2596	0.0051	0.1933	0.0005	0.1940	0.0005	0.2596	0.0051	0.2264	0.0016
일원1동	38	0.2184	0.0053	0.1930	0.0005	0.1934	0.0005	0.2184	0.0053	0.2057	0.0017
일원2동	44	0.1670	0.0035	0.1860	0.0004	0.1865	0.0004	0.1670	0.0035	0.1765	0.0012
일원본동	46	0.2416	0.0045	0.1894	0.0004	0.1895	0.0004	0.2416	0.0045	0.2155	0.0015

<표 8>에 동별 현재 흡연율의 4종 복합추정치 계산결과를 정리하였는데 직접추

정치에 비해서 모두 분산을 작아졌으나 현재 흡연율 추정치가 동별 표본크기에 따라서 변화가 크게 나타났는데 이는 복합추정방법의 특징이 추정치를 안정화시키기 때문이다. 따라서 직접추정치의 분산은 줄이고 합성추정치의 편향을 보정하는 관점에서 볼 때 4번째 복합추정량이 동별 추정치의 안정화에서 유용한 것으로 생각된다.

보건소단위로 건강지표를 산출할 목적으로 조사를 설계하여 자료를 수집한 후에 소영역인 동읍면별 건강지표를 소지역추정법으로 산출하는 수치적인 사례를 설명하였는데 다양한 분야에서 광역단위의 통계를 산출할 목적으로 조사한 후에 소영역이나 세부영역의 통계를 생산할 필요가 있을 경우에 응용할 수 있는 계산과정을 SAS코드로 예시하였으므로 앞으로 소영역의 통계생산에 도움이 될 수 있기를 바란다.

4. 참고 문헌

[1] 이계오외 4인(2013), “2013년지역사회건강조사 전국표본설계 및 표본관리”, 최종결과보고서

[2] 이계오(2000), “시군구 실업자 추정을 위한 소지역 추정법”, 응용통계연구, 제13권 2호, 275-285

[3] 이계오외 3인(2001), “소지역 통계 추정법”, 통계청 연구결과 보고서

[4] 이계오외 1인(2014), “지역사회건강조사 동읍면단위 통계생산 프로그램(알고리즘) 개발연구”, 질병관리본부 연구결과 보고서

[5] 질병관리본부(2012), “지역사회건강조사 원시자료 이용지침서”, 질병관리본부

[6] J.F. Gonzalez, P.J.Placek, and C.Scott(1993), "Synthetic estimation in followback surveys at the national center for health statistics", Statistical Policy Working Paper 21, Chapter2.

[7] Ghosh, M. and Rao, J.N.K (1994) Small area estimation: an appraisal. Statistical Science, 9, 55-93

[8] Singh, M.P., Gambino, J. and Mantel, H.J. (1994) Issues and strategies for small area data. Survey Methodology, 20, 3-22