한국영양학회지(*Korean J Nutr*) 2012; 45(6): 600 ~ 612 ISSN 0367-6463 / E-ISSN 2005-7121
http://dx.doi.org/10.4163/kjn.2012.45.6.600

국민건강영양조사 자료의 복합표본설계효과와 통계적 추론

정진은

안산대학교 식품영양학과

Complex sample design effects and inference for Korea National Health and Nutrition Examination Survey data

Chung, Chin-Eun

Department of Food and Nutrition, Ansan University, Ansan 426-701, Korea

ABSTRACT

Nutritional researchers world-wide are using large-scale sample survey methods to study nutritional health epidemiology and services utilization in general, non-clinical populations. This article provides a review of important statistical methods and software that apply to descriptive and multivariate analysis of data collected in sample surveys, such as national health and nutrition examination survey. A comparative data analysis of the Korea National Health and Nutrition Examination Survey (KNHANES) was used to illustrate analytical procedures and design effects for survey estimates of population statistics, model parameters, and test statistics. This article focused on the following points, method of approach to analyze of the sample survey data, right software tools available to perform these analyses, and correct survey analysis methods important to interpretation of survey data. It addresses the question of approaches to analysis of complex sample survey data. The latest developments in software tools for analysis of complex sample survey data are covered, and empirical examples are presented that illustrate the impact of survey sample design effects on the parameter estimates, test statistics, and significance probabilities (p values) for univariate and multivariate analyses. (*Korean J Nutr* 2012; 45(6): 600 ~ 612)

KEY WORDS: Korea National Health and Nutrition Examination Survey (KNHANES), sample design, design effect, stratification, clustering, weighting, sampling variance.

서 론

영양역학이나 지역사회 영양실태조사와 같은 연구를 위하여 수행되는 영양실태조사는 가족단위로 조사되는 층화다단확률 표본이 대부분이다. 우리나라의 국민건강영양조사는 국민의 건강 및 영양상태에 관한 통계 생산, 국민건강 증진 종합계획 목표지표 설정 및 평가, 국제기구에 제공하는 건강지표 산출을 위하여 수행되고 있다. 국민건강영양조사는 보건복지부, 질병관리본부 주관으로 전국규모의 검진조사, 영양조사, 건강설문조사 등을 통해 국민의 건강상태 및 건강과 관련된 위험요인과 식생활습관 등을 조사하고 있다.[1)] 국민건강영양조사 자료의 결과는 국민을 위한 건강 영양정책 수립, 국가 및 지역의 보건의료계획 수립 및 평가, 보건관련 프로그램 개발 및 평가, 기준치 설정, WHO나 OECD 등 국제기구가 요구하는 보건부문의 통계산출, 건강관련 삶의 질 향상을 위한 계획수립, 영양교육 프로그램 개발, 가공식품의 영양표시제도 확립, 한국인을 위한 영양섭취기준 제정, 한국인을 위한 영양 감시체계 구축, 그 외 모든 분야의 학술 연구 등에 활용되고 있다.

최근 국민건강영양조사 자료를 활용하여 여러 분야에서 보고서 및 논문의 수가 증가하는 추세임에 반하여, 그 자료의 특성과 표본추출법에 대한 이해 없이 올바른 통계패키지, 또는 통계프로그램을 사용하지 못하여 잘못된 결론을 도출하는 문제점이 제기되고 있다. 국민건강영양조사의 표본추출 방법은 단순임의추출 (simple random sampling)이 아니고, 층화집락계통추출법 (stratified, clustered, and systematic sampling)에 의해 2~3단계 표본 추출한 것이며 추출단위는 동읍면, 조사구, 가구이고 층화변수는 시도, 동읍면, 주택유형이다. 또한 표본이 모집단을 대표하여 국민 전체의 건강에 관한 의식 및 행태, 영양상태 등을 추정하기 위하여 추출율, 응답율, 모집단 인구

접수일 : 2012년 11월 12일 / 수정일 : 2012년 12월 1일 / 채택일 : 2012년 12월 14일
§To whom correspondence should be addressed.
E-mail: cechung@ansan.ac.kr

구조를 반영하는 가중치를 부여하고 있으므로 통계분석 시 이를 사용하여야 한다.[2-4] 그러므로 이에 적합한 올바른 통계패키지를 사용하여 올바른 결론을 도출해야 한다.

선행연구[5]에서는 2007~2009년 사이에 국내 학술지에 발표된 국민건강영양조사 자료를 이용하여 작성한 논문 21편에 대하여 통계적 활용기법과 사용한 통계패키지 등을 분석하였다. 대상논문은 한국영양학회지 18편, 대한지역사회영양학회지 2편, 한국식품영양학회지에 1편으로 총 21편이었고, 사용된 통계패키지로 SAS[6] 사용 논문은 11편, SPSS[7] 사용 논문은 4편, SAS와 SPSS를 함께 사용한 논문은 2편, SAS와 SUDAAN[8]을 함께 사용한 논문은 4편이었다. SAS를 사용한 논문에서는 연구방법에서 어떤 procedure를 사용했는지 밝히지를 않았고 대부분 단순임의추출법에 사용하는 일반 procedure를 사용한 것으로 추측되었다.

그러므로 본 연구에서는 국민건강영양조사와 같은 복합 층화 집락계통 추출에 의한 data를 분석할 경우 복합 표본설계효과가 통계적 추정 및 가설검정에 어떻게 영향을 미치는지 검토하고자 하였다. 평균, 백분율과 같은 점추정시 나타나는 복합표본의 효과를 알아보고, *t*-검정, chi-square 검정, 분산분석과 사후검정, 회귀분석, 로지스틱회귀분석과 같은 다변량 등의 통계적 가설검정 분석 시 나타나는 복합표본의 영향을 알아보고 저 하였다. 또한 통계패키지 또는 통계프로시져를 잘못 사용했을 경우의 결과 차이와 결론 도출의 오류 등을 파악하고 저 하였다.

연구방법

국민건강영양조사 data를 통계 분석할 경우 단순임의추출에 의한 통계패키지를 이용한 경우와 복합표본 설계효과를 고려한 통계패키지로 분석한 경우 2가지 분석법에 의한 결과를 비교하였다. 이를 위하여 국민건강영양조사 data를 실지로 통계 분석하였다. 당뇨병, 비만, 고혈압 등 질병유병률, 영양소섭취, 식품섭취의 평균값 등과 같은 점추정 결과에 대한 것과, 여러 가지 통계 분석 방법, 즉, *t*-검정, chi-square 검정, 분산분석과 사후검정, 회귀분석, 로지스틱회귀분석과 같은 다변량 등의 모수추정과 통계적 분석에 의한 가설검정 시 나타나는 복합표본의 설계효과를 비교해 보고 저 하였다.

연구대상자

본 연구는 2005년도 국민건강영양조사 자료 중에서 식이섭취조사에 참여한 20세 이상 성인의 data를 이용하였으며, 여러 자료 중 조사대상자의 성별, 나이, 교육정도, 결혼상태, 흡연 등 일반적인 data와, 영양소 섭취 data, 식품섭취 data, 당뇨병, 비만, 고혈압 등 질별data를 이용하였다.

분석 내용

2005년도 국민건강영양조사 자료를 이용하여 임의로 다음과 같은 내용을 분석하였다.

1) 우리나라 사람들의 당뇨병, 비만, 고혈압의 유병률.

2) 에너지, 단백질, 지질, 탄수화물의 영양소 섭취량과, 당류, 과일류, 지질류의 식품군 섭취량의 평균값과 표준오차.

3) 당뇨병유무와 연령군, 성별, 교육정도, 결혼상태, 흡연상태 등 일반 사항과의 관계를 알아보기 위해 chi-square test로 독립성 분석.

4) 당뇨군과 비당뇨군 간의 에너지, 단백질, 지질, 탄수화물, 당류, 과일류, 지질류의 섭취량의 차이를 *t*-test로 검정.

5) 결혼상태 (미혼, 결혼, 이혼)에 따라 비타민 C 섭취량이 차이가 있는지 알아보기 위하여 분산분석을 하였고 결혼상태에 따라 어느 군 간에 차이가 있는지를 알아보기 위한 Bonferroni의 multiple *t*-test로 사후검정.

6) 공복혈당수준을 종속변수로 하고, 에너지섭취량, 연령군, 성별, 교육정도, 결혼상태, 흡연상태 등 6가지 독립변수와의 관계식을 구해보기 위한 다중회귀 분석.

7) 당뇨병의 발병여부와 에너지, 단백질, 지질, 탄수화물, 연령군, 성별, 교육정도, 결혼상태, 흡연상태 등 9가지의 독립변수와의 로지스틱 회귀계수 분석.

8) 로지스틱 모형을 구하기 위하여 9가지 독립변수의 유의성을 알아보기 위해 Wald chi-square 검정.

통계분석

모든 data를 단순임의추출표본이라고 가정한 일반 통계패키지 중 SAS[6]를 선택하여 분석하였고, 복합층화집락계통 추출법에 의한 통계분석을 고려한 통계패키지 중 SUDAAN[8]을 이용하여 층화, 집락, 가중치를 고려하여 국민건강영양조사 data를 분석하였다. 또한 최근 SAS에 추가된 복합층화집락계통 추출법을 고려한 Proc Surveyfreq, Proc Survemean, Proc Surveyreg, Proc Surveylogistic 등의 복합표본설계 프로시져도 이용하여 분석하였다.

위와 같이 모든 내용을 분석하여 단순임의표본설계 SAS 프로시져, 복합표본설계 SAS 프로시져, SUDAAN을 이용한 결과를 비교하였고, 모든 분석 프로그램은 Appendix에 첨부하였다.

결 과

평균, 백분율 및 분산 추정에 대한 표본설계효과

단순 일변량 통계량의 복합표본 설계효과의 특성을 제시하기 위하여 국민건강영양조사 자료를 분석하여 평균, 백분율의

추정치, 표준오차, 설계효과를 계산하였다. 국민건강영양조사 자료를 이용하여 1차추출단위, 층화변수, 가중치를 지정하고 복합표본설계 SAS 프로시져를 사용한 결과, SUDAAN을 사용하여 분석한 결과, 단순임의표본설계 SAS 프로시져를 사용하여 자료를 분석한 결과의 평균, 백분율, 표준오차의 차이를 살펴본 결과는 다음과 같다.

Table 1에는 2005년도 한국 국민건강영양조사 자료를 분석하여 당뇨병, 비만, 고혈압의 유병률 (%)과 표준오차를 제시하였고, Table 2에는 2005년도 한국 국민건강영양조사 자료를 이용하여 한국 사람들의 에너지, 단백질, 지질, 탄수화물 섭취량과, 식품군 중 당류, 과일, 유지류의 섭취량을 분석하여 평균값과 표준오차를 제시한 결과이다. 단순임의표본설계 SAS 프로시져를 사용하여 계산한 표준오차와 층 (strata), 집락 (cluster), 가중치 (weight)를 고려한 SUDAAN과 복합표본설계 SAS 프로시져인 proc surveymeans를 사용하여 계산한 평균과 표준오차 값을 제시하였고, 설계효과 (또는 표본설계효과)의 값도 제시하였다. '설계효과 (design effect)'란 복합표본설계에 의한 통계량의 표준오차 [$SE(p)_{des}$]와 단순임의표본에 의한 통계량의 표준오차 [$SE(p)_{srs}$]의 비를 말한다 (Kish[9]).

이들 표의 퍼센트와 평균값의 결과를 보면 몇 가지 차이가 있는 것을 알 수 있다. 우선 설계효과 (DEFT)의 값은 1.15~1.74의 범위로 모두 1보다 큰 값을 나타내고 있다. Tabale 1에서 당뇨병 환자의 평균 백분율을 보면 일반 통계패키지인 SAS를 사용한 경우 7.59%이고 표준오차는 0.39%이었다. 그러나 설계효과를 고려한 SUDAAN 또는 SAS의 복합표본설계 프로시져인 Proc Surveymean을 사용한 경우 백분율의 평균은 6.37%이고 표준오차는 0.45%이었다. 그러므로 설계효과는 0.45/0.39 = 1.15이다. 당뇨병 뿐 아니라 비만, 고혈압 환자의 백분율도 설계효과를 고려한 SUDAAN을 사용하였을 때의 표준오차 값이 크게 나오므로 설계효과 값이 모두 1보다 크게 나왔다. 설계효과를 고려한 통계패키지인 SUDAAN과 SAS의 복합표본설계 프로시져는 같은 값으로 나타났다. Table 2의 영양소섭취량, 식품군 섭취량의 계산 결과도 평균섭취량의 값이 SAS의 단순임의표본설계 프로시져를 사용한 경우보다 층, 집락, 가중치를 고려한 SAS의 복합표본설계 프로시져인 proc surveymeans를 사용하거나 SUDAAN을 사용했을 때의 평균값, 표준오차 값이 크게 나타났다.

Chi-square 검정 및 *t*-검정에 대한 표본설계효과

2005년도 한국 국민건강영양조사 자료를 이용하여 당뇨병에 걸린 사람과 건강한 사람들의 나이, 교육정도, 결혼상태, 흡연, 영양소 섭취량, 식품섭취량 등에 대하여 *t*-검정, chi-squre 검정을 단순임의표본설계 SAS 프로시져를 이용한 결과, SAS의 Proc Surveyfreq, Proc Surveymeans 등 복합표본설계 SAS 프로시져를 이용한 결과, SUDAAN을 이용한 분석 결과를 비교해 보고저 한다.

Chi-square 검정

성별, 연령군 (10대, 20대 등), 교육정도 (중졸, 고졸, 대졸), 결

Table 1. Design effects for survey estimates of percentage of diseased population aged over 20, data from the KNHANES 2005

	SAS		SAS		SUDAAN		DEFT (p)[1]
	Proc Means		Proc Surveymeans		Proc Descript		
	p (%)	SE (p)	p (%)	SE (p)	p (%)	SE (p)	
Diabetes	7.59	0.39	6.37	0.45	6.37	0.45	1.15
Obesity	31.97	0.68	31.05	0.89	31.05	0.90	1.32
Hypertension	25.60	0.63	22.86	0.81	22.86	0.83	1.32

1) $SE(p)_{des}/SE(p)_{srs}$

Table 2. Design effects for survey estimates of means of nutrients and foods consumption aged over 20, data from the KNHANES 2005

	SAS		SAS		SUDAAN		DEFT (Mean)[1]
	Proc Means		Proc Surveymeans		Proc Descript		
	Mean	SE	Mean	SE	Mean	SE	
Energy (kcal)	1995.33	12.13	2075.35	20.35	2075.35	20.48	1.69
Protein (g)	76.64	0.60	79.83	0.94	79.83	0.96	1.60
Fat (g)	41.10	0.47	44.73	0.80	44.73	0.80	1.70
Carbohydrate (g)	311.53	1.74	317.13	3.01	317.13	3.02	1.74
Sugars	10.63	0.23	10.87	0.32	10.87	0.32	1.39
Fruits	247.55	5.89	252.53	9.50	252.53	9.76	1.66
Fats and oils	10.47	0.15	10.86	0.25	10.86	0.25	1.67

1) SE_{des}/SE_{srs}

혼상태 (미혼, 결혼, 이혼), 흡연여부 (흡연, 비흡연), 당뇨병유무 (당뇨군, 비당뇨군) 등의 이산자료 (discrete data)들은 범주로 분류되며 그 범주에 해당하는 수의 결과는 빈도수 또는 도수자료 (frequency data)이다. 종속변수인 질병유무와 여러 일반사항의 독립변수들과 분할표 (contingency table)를 작성하고 이들 자료에 있어서 두 변수사이에 어떠한 관계가 있는지 알아보기 위하여 각각 chi-square 검정을 이용하여 독립성 검정을 하였다. 두 변수 간의 독립성 검정에서의 귀무가설은 '두 개의 변수는 서로 독립이다 또는 관계가 없다'이다. 주어진 현상에 대하여 chi-square 검정은 각각 범주의 기대도수와 관찰도수를 비교한다. 관찰도수와 기대도수의 차이를 편차라고 하며 각 범주의 편차의 제곱을 기대도수로 나누어 모두 더한 값이 검정통계량 chi-square (χ^2) 값이 되며 이것은 chi-square 분포를 하게 된다.[10-12]

Table 3에는 한국 국민건강영양조사 자료를 이용하여 당뇨병유무와 일반사항들과의 chi-square 검정 결과를 요약하여 검정통계량 값 (χ^2)과 p value를 제시하였다. 이 경우 단순임의표본설계 SAS 프로시져를 이용한 경우 당뇨병 유무와 연령군, 성별, 교육정도, 결혼상태, 흡연상태가 모두 유의적인 것으로 나타났다. 반면, 표본설계효과를 고려한 SAS의 Proc Surveyfreq이나 SUDAAN을 이용하면 당뇨병 유무와 성별, 흡연상태와는 유의적인 관계가 없는 것으로 나타났다. 즉 설계효과를 고려하면 분석결과가 다르게 나오게 된 경우이다.

Table 3에서 복합표본설계 SAS 프로시져의 χ^2값과 SUDAAN의 χ^2값이 다른 것은 두 프로그램 모두 복합표본설계추출을 고려하여 층화, 집락, 가중치를 고려하여 분산과 검정통계량을 계산 하였지만 주장한 학자들의 검정통계량 계산과정이 약간씩 다르기 때문인 것으로 사려되며 유의성 검정결과에 영향을 미치지는 않았다.

t-검정

서로 다른 두 모집단의 모수에 대한 추론을 할 경우, 즉 서로 관련이 없는 두 개의 독립적인 집단에서 얻어진 표본이 있고 종속변수는 연속형 변수인 경우, 두 모평균의 차이에 대한 추정, 검정문제를 다룰 때 *t*-test를 이용한다. Table 4는 당뇨환자와 건강한 성인의 영양소 섭취량과 식품 섭취량의 *t*-test 결과를 요약한 표이다. 유의수준 5%에서 검정하였을 때 SAS의 단순임의표본설계 프로시져로 검정한 경우, 단백질, 지질, 당

Table 3. Design effects for survey estimates of test statistics and p values from Chi-square tests between diabetes and socioeconomic variables aged over 20, data from the KNHANES 2005

	SAS		SAS		SUDAAN	
	Proc freq		Proc surveyfreq		Proc crosstab	
	χ^2	p value	$\chi^{2\dagger}$	p value	χ^2	p value
Age group	202.2	<0.0001*	117.80	<0.0001*	38.32	0.0000*
Gender	15.8	<0.0001*	1.77	0.1832NS	1.76	0.1863NS
Education	113.6	<0.0001*	85.60	<0.0001*	23.43	0.0000*
Marriage	60.4	<0.0001*	43.20	<0.0001*	33.1	0.0000*
Smoke	4.65	0.0311*	1.10	0.2952NS	1.01	0.3174NS

*: Significantly different between diabetes and each variable at α = 0.05, †: Rao-Scott Chi-square. NS: Not significantly different between diabetes and each variable at α = 0.05

Table 4. Effect of weighting and sample design on test statistics (*t*) and p values by *t*-test for Diabetes and Non-Diabetes over 20 year, data from the KNHANES 2005

Dependent variables	SAS				SUDAAN				DEFT (mean)
	Difference[1]		Test statistic (*t*)	p value	Difference[1]		Test statistic (*t*)	p value	
	Mean	SE			Mean	SE			
Energy	88.67	46.22	1.92	0.0551NS	93.03	59.71	1.56	0.1211NS	1.29
Protein	5.80	2.29	2.53	0.0113*	5.43	2.82	1.92	0.0562NS	1.23
Fat	8.21	1.81	4.52	<0.001*	8.86	2.27	3.9	0.0001*	1.25
Carbohydrate	7.09	6.63	1.07	0.2847NS	6.32	8.88	0.71	0.4779NS	1.34
Sugars	2.06	0.94	2.2	0.0278*	2.57	0.81	3.17	0.0018*	0.86
Fruits	45.57	25.08	1.82	0.0270*	59.58	21.59	2.76	0.0064*	0.86
Fat and oils	1.68	0.61	2.75	0.0061*	1.44	0.79	1.83	0.0693NS	1.29

1) nonDM-DM
*: Significantly different between diabetes and non-diabetes at α = 0.05. NS: Not significantly different between diabetes and non-diabetes at α = 0.05

류 과실류, 유지류의 섭취량이 차이가 있는 것으로 나타났으나, 표본설계효과를 고려한 SUDAAN으로 검정한 경우에는 지질, 당류, 과실류의 섭취량만 유의적인 차이가 나타났다. 당뇨환자군과 건강한 사람들 간에 SAS 결과와 다르게 단백질과 유지류의 섭취량은 차이가 없는 것으로 나타났다. 즉 국민건강영양조사와 같은 층화집락계통추출표본 data를 일반 단순임의표본설계 SAS 프로시져로 검정하게 되면 이와 같이 틀린 결과를 발표하는 오류를 범하게 된다.

다변량의 모수추정 및 가설검정에 대한 표본설계효과

분산분석 (Analysis of Variance)과 사후검정 (Multiple Comparison)

3집단 이상의 독립된 집단들로 구성된 자료가 있을 때 여러 집단 간의 평균을 비교할 경우 분산분석 (analysis of variance: ANOVA)을 이용하게 된다. 여러 집단 (k 집단)의 평균의 차이가 있는지를 분석할 경우 귀무가설은 'k개의 모든 군의 평균은 같다. 또는 차이가 없다' 이고 수식적으로 표현하면 H_0: $\mu_1 = \mu_2 = \mu_3 - - - - = \mu_k$와 같다. 대립가설은 '$k$개의 평균 중 적어도 1개 이상이 다른 평균들과 같지 않다'이다. 분산분석이란 집단 간 변량과 집단 내 변량으로 분리하여 분산의 원인이 어디에 있는가를 알아보는 통계적 방법이다. 만일 처치효과가 있다면 k개 집단은 서로 다른 평균을 가지므로 집단 간 분산이 클 것이고 처치효과가 없다면 집단 간 분산을 적을 것이다. 검정통계량은 집단간분산/집단내분산이고 이 값은 F 분포를 하게 된다. 만일 귀무가설이 틀렸다면 즉, k개 집단의 평균이 같지 않다면 어느 집단 간의 평균이 다른지를 알아보기 위해 사후검정을 하게 된다. 사후 검정은 여러 집단의 평균을 여러 개의 짝으로 나누어 비교하게 되므로 다중비교 (multiple comparison)라고도 한다. 결혼상태 (미혼, 결혼, 이혼)에 따른 비타민 C 섭취량에 대한 분산분석 결과 결혼상태에 따라 차이가 있는 것으로 나타났으므로, 사후검정으로 비타민 C 섭취량을 2 그룹씩 비교한 결과를 Table 5에 제시하였다. 사후검정법은 Bonferroni t-test 외에 Schffee방법, Turkey방법, Duncan 방법, LSD (pair-wise t-test)방법, Sidak방법 등 여러 가지 방법이 있다.

Table 5의 사후검정 결과를 보면 SAS의 단순임의표본설계 프로시져 결과는 1군과 2군, 1군과 3군, 2군과 3군 모두 유의적인 차이가 있는 것으로 나타났으나 SAS의 복합표본설계 프로시져인 Proc Surveyreg의 결과와 SUDAAN의 결과에서는 1군과 2군의 비교는 유의적인 차이가 나타나지 않았다. 즉 미혼자와 결혼자의 비타민 C 섭취량은 SAS의 단순임의표본설계 프로시져를 사용한 경우 유의적인 차이가 있는 것으로 나타났으나 design effect를 고려한 SAS의 복합표본설계 procedure와 SUDAAN을 사용한 결과는 유의적인 차이가 없는 것으로 나타나 통계패키지 또는 프로시져를 어느 것을 사용하느냐에 따라 결론이 다르게 도출되었다.

회귀분석

회귀분석이란 관심있는 두 변수 간에 서로 관련성이 있는지 변수들 사이의 관계를 모형화 시키는 기법이다. Data는 (x_i, y_i)로 주어지며 독립변수 X로 부터 종속변수 Y를 예측하는데 사용할 수 있는 수리적인 관계식을 구하는 것이 회귀 분석이다.

x가 변할 때 y도 변하게 되고 이런 경우에 두 변수 간의 선형적인 관계를 잘 설명하는 직선을 찾으면 이것이 선형회귀직선이다. 하나의 독립변수만을 고려한 분석인 경우는 단순회귀 (simple regression)라고 하고, 두 개 이상의 독립변수를 고려한 경우는 다중회귀 (multiple regression)라 한다. 다중회귀 직선의 모형은 다음과 같다.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 - - - - + \beta_k x_k + \varepsilon_i$$

각각의 $H_0 : \beta_i = 0$을 검정하게 되며 직선의 기울기가 '0'이면 y와 χ_i 간에 선형적인 관계가 없다는 의미이다.[31]

Table 6은 종속변수인 공복혈당의 농도와 여러 가지 독립변수들 즉, 에너지섭취량, 성별, 연령군, 교육정도, 결혼상태 흡연상태 등과의 관계식을 알아보기 위한 다중회귀 직선의 유의성을 검증하여 $H_0 : \beta_i = 0$에 대한 t 검정 검정 결과를 요약해 놓은 것이다. SUDAAN의 경우 모형의 적합도 검정을 위한 검정통계량은 Wald chi-square값으로 제시하였다. SAS의 단순임

Table 5. Effect of sample design on Bonferroni's multiple comparison of Vitamin C consumption by marital status aged over 20[1], data from the KNHANES 2005

Multiple comparison[1]	SAS		SAS		SUDAAN	
	Proc ANOVA		Proc Surveyreg		Proc Descript	
	Difference	p value	Difference	p value	Difference	p value
1 vs. 2	9.87	$<0.05^*$	4.46	0.5385^{NS}	4.46	0.5400^{NS}
1 vs. 3	20.34	$<0.05^*$	27.38	0.0004^*	27.38	0.0004^*
2 vs. 3	30.23	$<0.05^*$	31.85	0.0000^*	31.85	0.0000^*

Dependent variable: Consumption of Vitamin C.
Independent variable: [1]marital status. 1: not-married, 2: married, 3: divorced
*: Significantly different between two groups at $\alpha = 0.05$. NS: Not significantly different between two groups at $\alpha = 0.05$

Table 6. Effect of weighting and sample design on test statistics and P values from multiple regression model of blood glucose aged over 20, data from the KNHANES 2005

Independent variable	SAS Proc Reg		SAS Proc Surveyreg		SUDAAN	
	t value	p value[1)]	*t* value	p value[1)]	Wald Chi-square	p value[1)]
X1: energy	−0.74	0.4598[NS]	0.43	0.6669[NS]	0.0466	0.8291[NS]
X2: gender	−7.52	0.0001*	−5.95	<0.0001*	32.5809	0.0000*
X3: age group	7.04	0.0001*	7.39	<0.0001*	28.0367	0.0000*
X4: education	−4.3	0.0001*	−2.37	0.1019[NS]	1.9226	0.1278[NS]
X5: marriage status	2.63	0.0086*	2.69	0.0078*	4.6407	0.0109*
X6: smoke	0.09	0.9307[NS]	−0.33	0.7419[NS]	0.1301	0.7187[NS]

Dependent variable: fasting blood glucose. Model: $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon_i$
1) p values from testing $H_0 : \beta_i = 0$
*: Significant at $\alpha = 0.05$

의표본설계 프로시져 결과를 보면 공복시 혈당은 성별, 연령군, 교육정도, 결혼상태가 유의적인 변수로 나왔고, SAS의 복합표본설계 프로시져와 복합표본설계에 의한 SUDAAN의 결과는 성별 연령군, 결혼상태는 유의적이었으나 교육정도는 유의적이 아닌 변수로 나타났다. 그러므로 사용하는 통계 패키지에 따라 결과가 다르게 나타나므로 도출하는 결론이 달라지게 된다.

로지스틱 회귀 분석

로지스틱 회귀모형은 종속변수가 1 또는 2, 네 또는 아니오, 등 이분형으로만 나타나는 이항분포 (binomial distribution, y~B (n, p)를 하는 경우이다. 그러므로 로지스틱 모형에서는 이항분포의 모수 (parameter)인 p [y의 기대값 (expected value)]과 관심있는 독립변수들과의 관계를 알아보게 된다.[13-15]

다음은 로지스틱 회귀직선의 모형이다.

$$\log (p/1\text{-}p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \varepsilon_i$$

표본설계효과가 복합표본조사 자료 분석에 어떠한 영향을 주는지 알아보기 위하여 2005년도 한국 국민건강영양조사 자료를 이용하여 20~74세 성인의 건강, 영양, 보건의식행태의 가중치를 적용하여 로지스틱회귀분석을 실시해보았다. 종속변수는 당뇨유무 (1: 유, 0: 무)이고, 독립변수는 범주형변수 또는 연속형변수 모두 가능하므로 성별 (남, 여), 연령군 (20~29, 30~49, 50~64, 65~74), 교육정도 (중졸, 고졸, 대졸), 결혼상태 (미혼, 결혼, 이혼), 흡연 (흡연, 비흡연), 그리고 에너지, 단백질, 지방, 탄수화물의 섭취량은 연속변수를 그대로 사용하여 분석하였다.

Table 7에는 로지스틱 화귀분석의 계수인 $\hat{\beta}$값과 SE($\hat{\beta}$)의 값을 알아보기 위해 층화, 집락, 가중치의 표본설계효과를 고려한 SUDAAN의 결과와 고려하지 않은 SAS의 단순임의표본설계 프로시져의 결과를 제시하였다. 표본설계효과를 고려한 SUDAAN 결과의 SE($\hat{\beta}$)의 값이 대부분 크게 나오고 DEFT ($\hat{\beta}$)의 값은 대부분 1.0 이상으로 나타났다. 예를 들면 성별에 대한 회귀계수의 표준오차를 비교해 보면 DEFT (β_{gender}) = 0.1745/0.1550 = 1.13이다. β_{gender}의 신뢰구간은 표본설계효과를 고려한 경우 고려하지 않은 경우보다 13% 넓게 된다. 그러므로 유의성 검정을 하면 종속변수에 영향을 미치는 독립변수의 결과가 다르게 나타나게 된다.

당뇨병에 영향을 미치는 유의적인 변수를 알아보고저 SAS의 단순임의표본설계 프로시져와 SUDAAN으로 분석한 결과를 로지스틱 모형의 검정통계량인 Wald chi-square 값과 p value를 비교하였다 (Table 8). 대부분의 경우 model요인에 대한 유의성 검정은 Wald chi-square test를 사용한다. 설계효과를 고려한 SUDAAN으로 분석한 Wald chi-square 값이 SAS의 단순임의표본설계 프로시져로 분석한 Wald chi-square 값보다 훨씬 작게 나타났고 p값은 크게 나타났다. 유의수준 5%에서 유의성 분석을 하였을 때 SAS의 단순임의표본설계 프로시져 결과는 에너지, 지방, 탄수화물, 성별, 연령군, 교육정도, 결혼상태가 유의적인 영향을 미치는 변수로 나타났다. 그러나 SUDAAN의 결과는 에너지와 연령군 만이 유의적인 변수로 나타났다. 당뇨병에 영향을 미치는 변수로 지방, 탄수화물, 성별, 교육정도, 결혼상태는 SAS의 단순임의표본설계 프로시져 분석결과 유의적인 변수로 나타났으나 SUDAAN의 분석결과는 유의적인 변수가 아닌 것으로 나타났다. 그러므로 국민건강영양조사 자료를 표본설계효과를 고려하지 않은 일반 SAS의 단순임의표본설계 프로시져를 이용하여 분석하여 결론을 내린다면 잘못된 결론을 도출하게 되므로 오류를 범하게 된다.

고 찰

모든 data의 표본추출은 단순임의추출, 층화집락계통추출

Table 7. Effect of weighting and sample design on model estimation, logistic model of Diabetes of 20−74 year, data from the KNHANES 2005

Independent variables	SAS V 9.13		SUDAAN V 10.01		DEFT ($\hat{\beta}$)
	Proc Logistic		Proc Rlogistic		
	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)	
Intercept	−2.2325	0.3135	−2.2654	0.3528	1.13
Energy	0.0006	0.0002	0.0005	0.0002	1.03
Protein	−0.0009	0.0028	−0.0001	0.0030	1.06
Fat	−0.0077	0.0037	−0.0069	0.0041	1.09
Carbohydrate	−0.0027	0.0010	−0.0022	0.0012	1.17
Gender	0.4953	0.1550	0.2427	0.1745	1.13
Age group					
20−29	−2.6658	0.6705	−2.6067	0.6270	0.94
30−49	−1.3110	0.2078	−1.2839	0.2535	1.22
50−64	−0.0293	0.1567	−0.0894	0.1984	1.27
65−74	0		0		
Education					
Elementary	0.5273	0.2282	0.4009	0.2638	1.16
Middle school	0.1928	0.2478	0.2089	0.3131	1.26
High school	0.3241	0.2039	0.2459	0.2323	1.14
College	0		0		
Marital status					
Non-married	−0.8770	0.4940	−0.7257	0.4440	0.90
Married	−0.2210	0.1683	−0.0664	0.2104	1.25
Devorced	0		0		
Smoke	0.0847	0.1586	0.1770	0.1829	1.15

$DEFT\ (\hat{\beta}) = SE(\hat{\beta})_{des} / SE(\hat{\beta})_{srs}$

Table 8. Effect of weighting and sample design on test statistics and p values. Logistic model of diabetes aged 20−74, data from the KNHANES 2005

Independent variable	SAS		SUDAAN	
	Proc Logistic		Proc Rlogist	
	Wald chi-square	p value[1)]	Wald chi-square	p value[1)]
X_1: energy	6.5892	0.0103*	4.6079	0.0333*
X_2: protein	0.0162	0.8988^{NS}	0.0008	0.9771^{NS}
X_3: fat	5.3090	0.0212*	2.9069	0.0901^{NS}
X_4: carbohydrate	6.1957	0.0128*	3.1889	0.0759^{NS}
X_5: gender	12.3740	0.0004*	1.9343	0.1661^{NS}
X_6: age group	50.5442	<0.0001*	14.9992	0.0000*
X_7: education	8.2554	0.0041*	0.8960	0.4445^{NS}
X_8: marriage status	4.7172	0.0299*	1.4304	0.2421^{NS}
X_9: smoke	0.2934	0.5881^{NS}	0.9357	0.3347^{NS}

Model: $\log (p/1-p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \varepsilon_i$
*: Significant at $\alpha = 0.05$. NS: Not Significant at $\alpha = 0.05$

등 여러 가지가 있고 표본추출방법에 따라 data의 분산 및 검정통계량을 계산하는 법이 달라지게 되며 이에 따라 통계적 유의성 검정의 결과가 달라지게 된다.[16-21)]

한국 국민건강영양조사의 표본추출방법은 제1기~제3기(1998년, 2001년, 2005년)에는 우선 지역군, 지역층, 행정구역, 주택유형의 조합으로 층화하였고, 2단계 추출로 1차 조사구를 선정하고 2차 가구를 선정하는 층화집락계통추출법을 사용하고 있다.[1)] 제4기 (2007~2009년)에는 전국 11개 지역군을 성별, 연령대별 인구구성비 기준으로 29개 층으로 층화 하였고, 1차 동읍면 추출, 2차 조사구 추출, 3차 가구 추출을 하였다. 이러한 표본 추출방법은 전국적인 대표성이 있어야 하므로 전체적인 추정이나 통계분석을 할 경우 '복합 표본설계 (com-

plex designs)', 즉 층화 (stratification), 집락 (clustering), 가중치 (weight)를 고려해서 통계처리를 해야 한다. 그런데 일반적으로 많이 쓰이는 일반 통계패키지인 SAS, SPSS 등의 프로시져들은 이러한 것을 고려하지 않았고 모든 data를 단순임의추출 (simple random sampling: SRS)로 가정한 분석방법이다. 단순임의추출은 모든 data를 각각 독립적 (independent and identically distributed: IID)이라고 가정하므로 이러한 일반 통계패키지에 의해 산출된 추정치의 표준오차나 검정통계량 값은 표본 data가 IID일 경우에만 정확한 값이 될 수 있다. 그러나 표본이 독립적이라는 가정은 국민건강영양조사와 같은 복합표본설계에는 맞지 않는다. 복합표본설계 data를 SRS, IID라고 가정한다면 기술통계량이나 추정치의 표준오차를 과소평가하게 되며 또한 검정통계량 값은 편향되게 된다. Kish[9]는 복합표본설계에 의한 통계량의 표준오차와 단순임의표본에 의한 통계량의 표준오차의 비를 '설계효과 (design effect) 또는 표본설계효과'라고 하였다.

$$DEFT\ (t) = \frac{SE(t)_{des}}{SE(t)_{srs}}$$

t는 survey data의 추정치 즉 평균, 비율, 상관관계, 회귀식의 계수, 또는 검정통계량 등을 의미한다. 설계효과(design effect)는 복합표본설계와 단순임의표본설계에 의한 통계량의 상대효율(relative efficiency)를 나타내며 값이 클수록 복합표본설계와 단순임의표본설계의 값이 차이가 많은 것을 의미한다. 표본설계에서 survey data는 표본추출 전 설계과정에서 모집단을 층화하고, 표본추출과정에서 clustering 또는 grouping하며, 자료 분석과정에서 가중치를 적용하므로서 표본 설계효과가 나타나게 된다. 즉, 조사자료는 층화, 집락, 가중치를 적용함에 따라 설계효과가 나타난다.[2-4]

국민건강영양조사는 전국을 대표하는 표본을 추출하여 국가단위의 통계를 생산하는데 그 목적이 있으므로 조사별로 우리나라 국민에 대한 결과를 추정할 수 있도록 자료에 가중치를 부여하고 있다. 가중치는 추출률, 응답률을 고려하였고, 여기에 모집단의 성별, 연령별 인구구조를 반영하기 위해 사후 가중치 조정을 하였으며, 가중치는 개별조사 가중치와 연관성분석 가중치로 구분된다. 이와 같이 국민건강영양조사 자료와 같은 층화집락추출방법의 분산을 계산할 때 반드시 1차 추출단위, 층화변수, 가중치를 고려해야한다. 그 결과 분산추정치는 단순임의추출에 의한 자료를 분석하는 일반 통계패키지를 사용한 경우와 분산의 값이 다르므로 검정통계량 값이 달라지고, 신뢰구간, 유의확률 (p value), 통계적 유의성 검정의 결과가 다르게 나오게 된다.

표본으로부터 주어지는 정보를 이용하여 모수에 대해 추측, 또는 주장 등의 옳고 그름을 판정하는 통계적 가설검정 또는 유의성 검정 절차를 간단히 설명하면 다음과 같다. 연구자가 주장하고자 하는 것을 대립가설로 설정하고 대립가설과 반대되는 것을 귀무가설로 설정한다. 표본의 자료를 수집하여 data 특성에 맞은 통계처리 방법을 택하여 검정통계량, 유의확률 (p value), 임계점 (critical point), 신뢰구간 등을 계산하여 이미 알려진 확률분포의 값과 비교한다. 만일 검정통계량 값이 기각역에 속하거나, p값이 유의수준 α보다 작을 경우, 또는 신뢰구간이 가정한 값을 포함하지 않을 경우 귀무가설을 기각하게 된다. 그러므로 국민건강영양조사 자료를 이용하여 통계분석을 할 경우에는 반드시 층화집락계통추출에 의한 분산추정이 가능한 통계패키지를 이용해야 정확한 검정통계량, p값 등 올바른 결과를 얻을 수 있다.

Survey자료는 표본설계 또는 추정값 계산 시 다음과 같은 층화, 집락, 가중치의 3가지 특성에 의해서 표본설계효과가 나타나게 된다. 층화 (stratification)는 단순임의표본과 비교해 볼 때 효율성이 증가하게 되므로 효율적인 층화는 조사 자료를 분석할 때 추정치의 설계효과 (design effect)를 감소시키게 된다. 집락 (clustering 또는 grouping)을 하게 되면 data 수집 비용을 절약할 수 있다. 사람이 직접 방문하여 조사하는 경우 가족, 가족구성원의 집락을 추출하는 다단계 확률 표본설계 (multistage probability sample design)을 많이 사용한다. 그런데 집락으로 추출된 표본단위는 독립 표본 추출이 아니다. 집락 내에 있는 조사대상자들의 집락 내 상관관계 (intra-class correlations)가 있으면 추정치의 표본설계의 효율성이 감소하고 설계효과는 증가하게 된다 (Kish[22]). 가중치(weighing)는 조사 자료 분석 시 꼭 필요하며 그 이유는 여러 가지가 있다. 표본 추출할 때의 확률과 실제로 조사할 때의 확률은 다를 경우가 많은데 이것은 표본 추출할 때의 가중치와 불응답에 대한 가중치를 사용하므로서 감소시킬 수 있다. 만일 이런 가중치를 사용하지 않는다면 조사 자료의 추정치는 편향될 것이며 결과는 달라지게 된다. 특히 인구사회학적 변수 (나이, 성별, 지역 등)들에 대해 가중치를 적용해야한다. 이런 가중치의 사용은 이들 변수들의 상관관계에 따라서 설계효과 (DEFT)가 증가 또는 감소하게 된다 (Kish[22]).

복합효과 또는 설계효과에 영향을 미치는 층화, 집락, 가중치의 교호작용에 따라 수학적인 모델이 어렵게 된다. 복잡한 표본조사 자료의 분산을 정확하게 계산하기 위한 표본추출 이론이 매우 오래 전 부터 개발되어 왔으나 이를 컴퓨터 프로그램에 적용하는 것은 매우 제한적이었다. 모집단을 추정하기 위한 검정통계량, 표준오차, 신뢰구간 등의 값이 편향되지않도록 정확하게 계산되어야 한다. 이러한 편향을 방지하기 위해서는 조사 자료 분석 프로그램은 표본조사에서 층화, 집락, 가중치를 적

용하는 design effects를 반드시 도입하여야 한다.[2]

복합표본 통계량 (complex sample survey statistic)의 분산 또는 표준오차를 계산하는 방법에 대한 연구는 끊임없이 이어져왔다.[23-32] Goldstein[23]은 층화, 집락 표본 자료 분석을 할 경우 모형에 근거한 모수적 접근 (model-based parametric)방법을 제안했다. 그러나 많은 사람들은 대규모 조사 자료 분석 시 설계에 근거한 비모수적 접근 (design-based nonparametric) 방법을 주장하였으며,[24] Wolter[25]는 조사자료 분석을 위한 design based 방법과 software를 제안하였다. 많이 사용하는 예를 들면 Taylor의 계열선형화 방법 (Taylor series linearization method),[26] 균형반복추출법 (balanced repeated replication: BRR),[27,28] 잭나이프 반복추출법 (jack-knife repeated replication: JRR), 또는 Bootstrap 방법과 같은 반복추출분산추정 (resampling variance estimation)[29,30] 방법이있다.

집락의 크기가 다른 복합표본설계에 의한 조사 자료일 경우, 조사 자료들의 대부분의 통계량은 단순 선형함수가 아닌 것이 많다. 이런 비선형통계량을 Taylor의 계열선형화 방법 (Taylor series linearization method)에 의해 선형으로 만들고 이들의 분산, 공분산을 계산할 수 있게 되었다 (Woodruff[26]). 통계패키지인 SUDAAN은 이 방법을 사용하고 있다. SUDAAN[8]은 1966년에 미국 North Carolina주 Research Triangle Park에 있는 연구소에서 Shah 등에 의해 개발되었고, 그 후 계속 개발되어 2012년에는 SUDAAN version 11이 출시되었다. 실행프로시져로 평균을 계산하기 위한 Proc Descript; 비율 등의 계산을 위한 Proc Ratio; 분할표 분석을 위한 Proc Corsstab; 선형회귀 분석을 위한 Proc Regress; logistic regression 분석을 위한 Proc Rlogist; log-link (Poisson) models로 Proc Loglink; 생존분석을 위한 Proc Survival 등이 있고 그 외 Proc Kapmeier, Proc Wtadjust, Proc Hotdeck 등의 프로시져가 있다. 그 후 SUDAAN은 SAS내에서 같이 실행되는 프로그램이 출시되어 쉽게 사용할 수 있으며 http://www.rti.org에서 자세한 정보를 제공 받을 수 있다.

Stata[33]는 1997년에 개발되었으며 복합표본조사자료 (complex sample survey data)의 분석을 위한 표본가중치 (sampling weights), 다단계 설계 (multistage designs); 층화, 설계효과 (DEFF)에 의한 평균, 비, 비율, 합계 등 요약표, predictive margins; bootstrap, 잭나이프, 선형화에 의한 분산 추정: 회귀분석, 도구변수 (instrumental variables), 프로빗, Cox 회귀분석 등이 가능하다. 최근 Stata 12가 출시되었으며 http://www.stata.com에서 자세한 정보를 얻을 수 있다.

반복추출법 (resampling methods)은 복합표본 자료의 추정과 검정을 위하여 균형반복추출법 (BRR), 잭나이프 반복추출법 (JRR), bootstrap 등의 비모수 방법을 적용한다. 이 방법은 선형 또는 비선형 통계량의 분산을 측정하기 위하여 반복적으로 추출을 하는 방법이며 통계패키지로 Westat이 있다. WesVar PC[34]는 Westat Inc.에 의해 개발되었으며 BRR 또는 JRR의 계산방법을 사용한다. http://www.westat.com에서 자세한 정보를 제공받을 수 있다.

그 후 SAS에서도 복합표본설계 data를 분석하기위한 procedure로 Proc Surveyfreq, Proc Surveymeans, Proc Surveyreg, Proc Surveylogistic, Proc Surveyphreg, Proc Surveyselect 등이 개발되었고 SAS version 9.3이 2011년에 출시되었다. SPSS도 복합표본설계효과를 고려한 SPSS Complex Samples라는 프로그램으로 층화, 집락, 가중치를 고려한 survey data를 분석할 수 있는 프로그램이 포함되었고 SPSS version 14가 출시되었다.

그러므로 층화집락계통추출에 의한 복합표본설계 data를 분석할 경우에는 이를 고려한 위와 같은 컴퓨터 소프트웨어를 사용하여야 올바른 결과를 도출할 수 있다.

본 논문에서는 이를 증명하기 위하여 통계분석에 의한 모수추정과 가설검정 시 나타나는 복합표본의 설계효과를 비교해 보았다. 국민건강영양조사 data를 일반 단순임의표본설계 SAS를 사용하여 분석한 경우와 1차추출단위, 층화변수, 가중치를 지정한 SUDAAN 또는 SAS의 Proc Survey procedure를 사용하여 분석한 경우를 비교하였다. SUDAAN을 사용한 경우가 단순임의표본설계 SAS를 사용한 경우 보다 질병 유병률의 표준오차 값이 크게 나타나 설계효과 (DEFT)의 값이 1.15~1.32로 나타났고 (Table 1), 영양소 섭취량, 식품군 섭취량의 평균값과, 표준오차 값이 크게 나타났으며 설계효과 (DEFT)의 값이 1.39~1.74로 나타났다 (Table 2). 이는 대단위 survey data에서 평균, 비율, 백분율 등 일변량 통계량의 모수에 대한 추정을 할 경우 설계효과가 매우 크게 발생한다고 하였고, 또한 대부분의 복합표본조사에서 이들의 표본 설계효과 (DEFT)는 1.0보다 크게 나타나며 이들 일변량 통계량의 점추정치의 실제 신뢰구간은 일반 통계프로그램의 결과 보다 매우 넓게 나타난다고 한 Heeringa[2]의 결과와 일치하였다.

당뇨병과 여러 가지 변수들과의 관계를 알아보기 위하여 로지스틱 다중 회귀분석을 한 결과 로지스틱 회귀분석의 계수인 $\hat{\beta}$의 표준오차는 층화, 집락, 가중치의 표본설계효과를 고려한 SUDAAN의 결과가 SAS의 결과보다 크게 나오고 DEFT ($\hat{\beta}$)의 값은 대부분 1.0 이상으로 나타났다 (Table 7). 또한 검정통계량인 Wald chi-square 값도 SUDAAN으로 분석한 값이 SAS의 단순임의표본설계 프로시져의 값보다 훨씬 작게 나타났고 p값은 크게 나타났다 (Table 8). 즉, SAS의 단순임의표본설계 프로시져로 분석한 결과 유의적인 변수로 나타났으나 SUDAAN의 분석결과는 유의적인 변수가 아닌 것으로 나타

난 변수들이 많았다. 이는 미국인을 대상으로 조사한 정신건강 서비스 data를 SAS와 SUDAAN으로 분석한 결과와도 일치하였다.[2)]

이와 같이 국민건강영양조사 자료를 표본설계효과를 고려하지 않은 일반 SAS의 단순임의표본설계 프로시져를 이용하여 분석하여 결론을 내린다면 잘못된 결론을 도출하게 되고 오류를 범하게 된다.

요약 및 결론

오늘날 통계적 사고와 통계적 기법의 활용은 모든 학문분야에서 일상의 다양한 영역에 이르기까지 보편화되었으며, 사용되는 기법들의 종류가 다양해지며 그 수준도 높아지고 있는 추세이다. 그러나 통계 기법의 잘못된 선택과 잘못된 통계패키지, 또는 프로시져를 사용하여 많은 문제점과 오류 등이 발생하는 것을 알 수 있었다. 이것은 잘못된 결과와 결론을 도출하게 되고 이는 잘못된 정책 설정으로 이어질 수 있다. 이러한 문제와 그 심각성에 대한 평가는 매우 시급한 과제로 생각된다. 국민건강영양조사는 보건복지가족부에서 주관하는 국가를 대표하는 자료로서 이 data를 활용한 보고서 및 논문의 숫자가 증가하는 추세임에 반하여 그 자료의 특성과 표본추출법에 대한 정확한 이해 없이 올바른 통계패키지, 또는 올바른 실행프로시져를 사용하지 못하여 잘못된 결론을 도출하는 문제점이 제기되고 있다.

국민건강영양조사의 표본 추출은 층화집락계통추출법에 의해 추출된 자료이므로 추정값, 분산 계산과 검정통계량을 이용한 가설검정 시 층화, 집락, 가중치의 design effect 특성을 고려하여 통계처리를 해야한다. 단순임의표본을 가정한 일반 통계프로그램은 추정치의 분산 또는 표준오차, 검정통계량값이 편향되고, 유의성검정결과가 다르게 나타날 수 있다. 본 연구에서는 실제로 2005년도 국민건강영양조사 자료를 활용하여 SAS (version 9.13)의 단순임의표본설계 프로시져, design effect를 고려한 SAS의 복합표본설계 프로시져, design effect를 고려한 SUDAAN 10.1을 이용하여 모집단의 평균, 비율의 추정값 계산, *t* 검정, chi-square 검정, ANOVA와 사후검정 (multiple comparison), 회귀분석, 로지스틱회귀분석 등을 실시하여 결과를 비교하였다. 분석 결과 SAS의 단순임의표본설계 프로시져 결과는 매우 유의적인 것으로 나타났으나, design effect를 고려한 SAS의 복합표본설계 프로시져와 SUDAAN의 결과는 유의적이 아닌 것으로 나타난 경우가 많이 나타났다. 국민건강영양조사 자료와 같이 단순임의표본이 아닌 층화집락계통추출 표본 data를 분석할 경우 단순임의표본설계 프로시져를 사용하면 그릇된 결론을 도출하게 된다는 것이 판명되었다.

그러므로 국민건강영양조사 자료와 같이 단순임의표본이 아닌 층화집락계통추출 표본을 분석할 경우에는 반드시 층화집락계통추출 표본설계효과, 즉 층화, 집락, 가중치를 고려한 통계패키지로 SUDAAN, Stata, Westat, SAS의 복합표본설계 프로시져, SPSS의 Complex Samples 등의 프로시져를 이용하여 분석하여야 하며 논문작성 시 연구방법에서 이를 반드시 명시하여야 한다.

본 연구는 국민건강영양조사 자료와 같은 층화집락계통추출에 의한 복합표본 data를 분석할 경우 통계패키지 또는 통계프로시져 사용에 관한 문제점을 제시하였다. 앞으로 통계 이론적 측면에서의 더 많은 연구와 교육를 통하여 올바른 통계패키지, 올바른 통계 프로시져 사용법을 전파하여야 할 것으로 사려된다. 또한 올바른 결론 도출과 올바른 식품영양 정책수립으로 이어질 수 있도록 통계분석에 관한 지식과 통계패키지 또는 통계프로시져 사용법을 발전시켜 나가야 할 것으로 사려된다.

Literature cited

1) Ministry of Health and Welfare. Korea National Health and Nutrition Examination Survey. Seoul: Ministry of Health and Welfare. Available from: http://knhanes.cdc.go.kr
2) Heeringa SG, Liu J. Complex sample design effects and inference for mental health survey data. *Int J Methods Psychiatr Res* 1998; 7(1): 56-65
3) Muthén BO, Satorra A. Complex sample data in structural equation modeling. *Sociol Methodol* 1995; 25: 267-316
4) Koch GG, Lemeshow S. An application of multivariate analysis to complex sample survey data. In: Institute of Statistics Mimeo Series No. 802. Chapel Hill: University of North Carolina; 1972
5) Chung CE. Evaluation of statistical methodology in national journals related with food science, cooking, and food culture. Seoul: Youlchon Foundation; 2010. p.591-703
6) SAS version 9.3. Cary: SAS Institute Inc.; 2011. Available from: http://www.sas.com
7) SPSS. Armonk: IBM; 2012. Available from: http://www.spss.com
8) SUDAAN version 11. Research Triangle Park: RTI International; 2011
9) Kish L, Groves RM, Krotki KP. World fertility survey. Sampling errors for fertility surveys. In: Occasional Paper, No. 17. Voorburg: International Statistical Institute; 1975
10) Agresti A. Categorical data analysis, 2nd edition. New York: John Wiley & Sons; 2002
11) Agresti A. An introduction to categorical data analysis, 2nd edition. New York: John Wiley & Sons; 2007
12) Lee JH, Moon IO, Chung CE. Health statistics. Seoul: Power Book Co.; 2008
13) Roberts G, Rao JNK, Kumar S. Logistic regression analysis of sample survey data. *Biometrika* 1987; 74(1): 1-12
14) Morel JG. Logistic regression under complex survey designs. *Surv Methodol* 1989; 15: 203-223
15) Hosmer DW Jr, Lemeshow S. Applied logistic regression, 2nd edition. New York: John Wiley & Sons; 2000
16) Cochran WG. Sampling techniques. New York: John Wiley &

Sons; 1977
17) Skinner CJ, Holt D, Smith TMF. Analysis of complex surveys. New York: John Wiley & Sons; 1989
18) Särndal CE, Swensson B, Wretman J. Model assisted survey sampling. New York: Springer; 1992
19) Binder DA, Roberts GR. Design-based and model-based methods for estimating model parameters. In: Analysis of Survey Data. New York: John Wiley & Sons; 2003
20) Fuller WA. Sampling statistics. Hoboken: John Wiley & Sons; 2009
21) Lohr SL. Sampling: design and analysis, 2nd edition. Boston: Brooks/Cole; 2010
22) Kish L. Survey sampling. New York: John Wiley & Sons; 1965
23) Goldstein H. Multi-level models in educational and social research. London: Oxford University Press; 1987
24) Rust K. Variance estimation for complex estimators in sample surveys. *J Off Stat* 1985; 1(4): 381-397
25) Wolter KM. Introduction to variance estimation. New York: Springer; 1985
26) Woodruff RS. A simple method for approximating the variance of a complicated estimate. *J Am Stat Assoc* 1971: 66(334): 411-414
27) Kish L, Frankel MR. Balanced repeated replications for standard errors. *J Am Stat Assoc* 1970; 65(331): 1071-1094
28) Rao JNK, Shao J. Modified balanced repeated replication for complex survey data. *Biometrika* 1999; 86(2): 403-415
29) Rao JNK, Wu CFJ. Resampling inference with complex survey data. *J Am Stat Assoc* 1988; 83(401): 231-241
30) Rao JNK, Wu CFJ, Yue K. Some recent work on resampling methods for complex surveys. *Surv Methodol* 1992; 18: 209-217
31) Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics* 2000; 56(2): 645-646
32) Wolter KM. Introduction to variance estimation, 2nd edition. New York: Springer; 2007
33) Stata Corp. Stata statistical software: release 5. College Station: Stata Corp.; 1997
34) Brick JM, Broene P, James P, Severynse J. A user's guide to WesVar PC. Rockville: Westat Inc.; 1996

□ Appendix □

Analysis Program

Table 1 & Table 2 (Means)

```
DATA knhs05;set knhs05;
  if DM_g ne . then do; if DM_g=1 then DMp=100;
    else DMp=0; end; /* 당뇨 */
  if obe_g ne . then do; if obe_g=1 then obep=100;
    else obep=0; end; /* 비만 */
  if hp_g ne . then do; if hp_g=1 then hpp=100;
    else hpp=0; end; /* 고혈압 */

    /* SAS */
    proc means  N mean stderr std min max;
      var DMp obep hpp nf_en nf_prot nf_fat nf_cho
        sugars fruits fatoil; run;

    Proc surveymeans data=knhs05;
      strata kstrata;
      cluster psu;
      weight wt_all;
      var  DMp obep hpp nf_en nf_prot nf_fat nf_cho
        sugars fruits fatoil; run;

    /* SUDAAN */
    Proc descript data=knhs05  filetype=sas design=wr;
      nest kstrata psu;
      weight wt_all;
      var DMp obep hpp nf_en nf_prot nf_fat nf_cho
        sugars fruits fatoil;
      print nsum mean semean /style=nchs; run;
```

Table 3 (Frequency)

```
  /* SAS */
  proc freq data=knhs05;
    table (ageg sex edu marriage smoke) * DM_g/chisq ;
      run;

  proc surveyfreq data=knhs05;
    strata kstrata;
    cluster psu;
    weight wt_all;
    table (ageg sex edu marriage smoke) * DM_g/chisq ;
      run;

  /* SUDAAN */
  proc crosstab data=knhs05 filetype=sas design=wr;
    nest kstrata psu;
    weight wt_all;
    tables (ageg sex edu marriage smoke)* DM_g;
    subgroup ageg sex edu marriage smoke DM_g;
    levels      4    2   4      3          2      2
    print nsum rowper serow chisq chisqdf chisqp/
      style=nchs; run;
```

Table 4 (*t-test*)

```
  /* SAS */
  proc ttest data=knhs05;
    class DM_g;
    var nf_en nf_prot nf_fat nf_cho sugars fruits fatoil;
      run;

  proc surveyreg data=knhs05;
    strata kstrata;
    cluster  psu;  weight  wt_all;
    class  DM_g;
    model  nf_en=DM_od/solution;  run; quit;

  /* SUDAAN */
  proc Descript data=knhs05  filetype=sas design=wr;
    nest kstrata psu;
    weight wt_all;
    subgroup  DM_g;
    levels  2
    Diffvar DM_g=(1 2)/name= "DM vs non-DM" ;
    var nf_en nf_prot nf_fat nf_cho sugars fruits fatoil;
    print nsum mean semean t_mean p_mean/
      style=nchs; run;
```

Table 5 (ANOVA & Multiple comparison)

```
  /* SAS */
  proc anova data=knhs05;
    class marriage;
```

```
  model  nf_vitc=marriage;
  means marriage/Bon; run;

proc surveyreg data=knhs05;
  strata kstrata;
  cluster  psu;
  weight  wt_all;
  class marriage;
  model  nf_vitc=marriage/solution
  lsmeans marriage/pdiff adjust=Bon; run; quit ;

/* SUDAAN */
Proc Descript Data=knhs05 Filetype=sas Design=WR ;
  Weight WT_all;
  NEST kSTRATA PSU;
  SubGroup marriage;
  Levels 3;
  Var nf_vitc;
  Output nsum mean semean /Filename=mean
    replace; run;

Proc Descript Data=knhs05 Filetype=sas Design=WR ;
  Weight WT_all;
  NEST  kSTRATA  PSU;
  SubGroup marriage;
  Levels 3;
  Var nf_vitc;
  Diffvar marriage=(1 2)/name="unmarried vs married"
  Diffvar marriage=(1 3)/name="unmarried vs devorced"
  Diffvar marriage=(2 3)/name="married vs devorced"
  Output nsum mean semean t_mean p_mean/
    Filename=ttest replace;run;

  Data gt1(Keep=SYM1 P_MEAN1 T_MEAN1 DIFF1
    SE_DIFF1); Set ttest;
  format SYM1 $2.
  If p_mean in (.) Then SYM1=" . "
  Else If p_mean LE (0.01/3) Then SYM1=" ** "
  Else If p_mean LE (0.05/3) Then SYM1=" * "
  Else SYM1="  "
  T_MEAN1=T_MEAN;
  P_MEAN1=P_MEAN;
  DIFF1=MEAN;
  SE_DIFF1=SEMEAN;
  If contrast In (1) and _one_ In (0) Then Output; run;
```

Table 6 (Regression)

```
 /* SAS */
 proc reg data=knhs05;
 model HE_glu= nf_en sex  ageg  edu  marriage  smoke;
  run;

 proc surveyreg data=knhs05;
  strata kstrata;
  cluster  psu;
  weight  wt_all;
  model HE_glu = nf_en sex ageg edu marriage smoke;
    run;

 /* SUDAAN */
 proc regress data=knhs05  filetype=sas design=wr;
  nest kstrata psu;
  weight wt_all;
  subgroup sex ageg edu marriage smoke;
  Levels     2    4    4       3          2 ;
  model  nf_en sex  ageg  edu  marriage  smoke; run;
```

Table 7 & 8 (Logistic Regression)

```
 /* SAS */
 proc logistic data=knhs05;
  model DM_od= nf_en nf_prot nf_fat nf_cho sex1
    sex2  ageg1  ageg2  ageg3
  ageg4  edu1  edu2  edu3  edu4 marri1  marri2  marri3
    smoke1 smoke2; run;

 /* SUDAAN */
 proc RLogist  data=knhs05  filetype=sas design=wr;
  nest kstrata psu;
  weight wt_all;
  subgroup  sex ageg edu marriage smoke;
  Levels     2    4    4       3          2 ;
  model DM_g= nf_en nf_prot nf_fat nf_cho  sex ageg
    edu marriage smoke; run;
```