The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 425-433

A Note on Statistical Reports on the Korean Anthropometric Survey

Jinwoo Park[1] and Eun-kyung Lee[2]

Abstract

Most of national-wide surveys are summarized by some statistical tables and graphs. In spite of high costs to get statistical results from surveys, we often find some statistical problems in the statistical reports. In this paper, we point out some statistical problems for the Korean Anthropometric Survey report. Also, we suggest some alternatives which may avoid the illustrated problems.

Keywords : anthropometric survey, self-weighting design, kernel smoothing method

1. 서 론

인체치수조사는 각 나라에서 자국민들의 성별, 연령별, 신체부위별 치수를 측정하는 조사이다. 의류, 신발, 자동차, 가구, 전자 업계 등 산업계의 다양한 인체치수 통계의 요구를 수용하고 반영하기 위해 각 나라에서는 인체치수조사를 실시하는데 우리나라에서는 1979년 1차 조사를 시작으로 하여 2004년까지 총 6차에 걸쳐 인체치수조사를 실시한 바 있다. 미국, 영국 등 구미 국가들과 일본, 대만 등의 나라에서도 각 나라 실정에 맞는 인체치수조사를 실시하고 있다(산업자원부 기술표준원, 2004).

ISO에서는 각국에서 측정, 구축되는 인체치수자료의 표준화를 위해 「인체치수 데이터베이스를 구축하는데 필요한 일반 규격(General requirements for establishing anthropometric database)」(ISO 15535, 2001)을 제안하였고 각 나라에서는 이를 준수하여 인체치수조사 사업을 실시하고 있다. 우리나라 최근의 조사인 제6차 조사는 2003년부터 2004년에 걸쳐 약 20,000 명에 이르는 인원을 대상으로 하여 Size Korea 사업이라는 이름 아래 실시되었다. 제6차 조사는 ISO 15535의 규격을 그대로 적용시켜 실시되었으며 조사 결과는 각각 보고서와 인터넷 홈페이지에 수록되어 있다.

많은 시간과 비용이 소요되는 인체치수조사 사업의 최종적인 결과는 각종 통계의 형태로 표현되어 나타나진다. 따라서 통계 발표양식에 있어서도 가능한 한 정확하면서도 효과적인 것이 되도록 작성하는 것이 중요하다. 그러나 현재의 보고서나 웹 상에 발표된 통계표를 보면 통계적인 측

1) Department of Applied Statistics, University of Suwon, Kyunggi-Do, 445-743.
E-mail : jwpark@suwon.ac.kr
2) Department of Statistics, Ewha Womans University, Seoul, 120-750

면에서 몇 가지 문제점이 있는데 이러한 문제점은 비단 인체치수조사에 국한된 것이 아니라 다른 유사한 표본통계에서도 빈번하게 발생하는 문제이므로 그것을 규명하고 올바른 대책을 제시하는 것은 절실하게 필요한 일이다.

본 연구의 목적은 현재 작성되는 인체치수 통계 양식이 지니는 통계적 문제점들을 지적하고 아울러 이에 대한 개선책을 제안하는 것이다. 본 연구에서는 인체치수 통계라는 특정한 통계를 다루고 있지만 이 연구의 결과는 다른 유사한 조사통계들의 발표양식에서도 일반적으로 널리 적용될 수 있다. 2절에서는 인체치수 통계 양식의 문제점을 지적한다. 그리고 3절에서는 새로운 양식을 제안한다. 마지막으로 4절에서는 결론을 내린다.

2. 인체치수 통계표 양식의 문제점

ISO 15535에서는 인체치수 데이터를 이용하여 각 인체부위 통계를 작성할 때 연령그룹의 구분과 각 연령그룹별 작성 통계량의 종류에 대해 기준을 제시하였는데 그 내용을 나타낸 것이 <표 2-1>이다. 성장속도가 빠른 5세에서 19세까지는 한 살 단위, 20세 이후의 성인에 대해서는 형편에 따라 5세, 10세, 20세 간격 중 어느 하나를 택하도록 하였다.

<표 2-1> ISO에서 제시하는 연령그룹 및 기술통계량

연령그룹의 구분				기술통계량 종류
연령그룹(세)	설명	연령그룹(세)	설명	
5	4.50- 5.49	20-24	19.50-24.49	표본수 최소값 최대값 산술평균 평균의 표준오차 표준편차 5, 95분위수의 표준오차 변동계수 도수분포 왜도 첨도
6	5.50- 6.49	25-29	24.50-29.49	
7	6.50- 7.49	30-34	29.50-34.49	
8	7.50- 8.49	35-39	34.50-39.49	
9	8.50- 9.49	40-49	39.50-49.49	
10	9.50-10.49	50-59	49.50-59.49	
11	10.50-11.49	60-69	59.50-69.49	
12	11.50-12.49	70 이상	69.50 이상	
13	12.50-13.49			
14	13.50-14.49			
15	14.50-15.49			
16	15.50-16.49			
17	16.50-17.49			사분위수 (1, 5, 25, 50, 75, 95, 99%) : 권장사항
18	17.50-18.49			
19	18.50-19.49			

한편 <표 2-2>는 2004년에 실시된 우리나라의 인체치수조사를 토대로 하여 작성한 통계표 중 키에 관한 표를 나타낸 것이다. 이 표를 보면 ISO에서 제시한 규격을 잘 따르고 있음을 알 수 있다. 뿐만 아니라 더 나아가서 ISO의 규격에 포함되지 않은 4세 미만의 영·유아들에 대해서도 조사를 실시하여 통계를 작성하였다. 2세 미만의 영아들의 경우에는 1세 간격이 아닌 3개월 내지 6개월 간격의 통계까지도 생산하여 발표하고 있다. ISO에서 제안한 모든 종류의 통계들을 다 생산하고 있는데 표준오차, 왜도, 첨도 등의 통계량은 부록으로 처리하였기 때문에 <표 2-2>에는 나와 있지 않다.

<표 2-2> 남자 키의 나이그룹별 기초통계표

나이 (Age)	측정수 (N)	평균 (Mean)	표준편차 (S.D.)	1th	5th	10th	25th	50th	75th	90th	95th	99th
0~3개월	51	591	39.2	669	645	630	619	605	555	530	522	505
3~6개월	50	667	32.5	742	720	715	687	668	649	618	613	600
6~9개월	51	714	34.8	817	776	746	730	712	691	675	663	630
9~12개월	50	750	29.7	828	816	782.5	765	748.5	725	716.5	712	685
12~18개월	50	801	34.8	897	850	840	825	805	770	753.5	745	737
2세	207	878	44.3	972	950	935	908	876	847	822	804	764
3세	210	952	39.9	1049	1020	1004	978	948	927	900	891	875
4세	209	1023	41.2	1113	1095	1080	1049	1023	994	973	956	931
5세	260	1090	46.6	1200	1165	1154.5	1122.5	1087.5	1058	1034	1016	981
6세	260	1155	46.0	1256	1236	1218	1186.5	1152.5	1122.5	1100	1087	1051
7세	251	1220	48.3	1334	1304	1278	1249	1224	1185	1153	1142	1120
8세	252	1278	53.5	1400	1365	1345	1309	1278.5	1243.5	1214	1186	1150
9세	257	1333	58.1	1460	1430	1410	1371	1330	1295	1257	1245	1197
10세	259	1380	57.2	1534	1493	1450	1410	1375	1347	1314	1295	1240
11세	257	1449	68.0	1635	1557	1530	1488	1447	1410	1363	1335	1287
12세	263	1507	80.3	1694	1647	1614	1556	1508	1450	1400	1375	1335
13세	269	1582	75.6	1773	1705	1670	1627	1589	1534	1480	1453	1410
14세	253	1647	74.4	1789	1754	1730	1695	1652	1610	1547	1495	1452
15세	276	1692	58.8	1805	1790	1770	1732	1694.5	1648	1619	1594	1550
16세	254	1703	57.8	1824	1806	1785	1740	1703.5	1665	1635	1613	1558
17세	269	1725	55.2	1864	1814	1792	1762	1725	1685	1655	1646	1598
18세	266	1729	54.7	1854	1814	1796	1770	1726	1688	1660	1643	1600
19세	277	1734	56.4	1878	1825	1804	1770	1732	1700	1665	1636	1587
20~24세	344	1738	58.3	1870	1838	1809	1775	1740	1699.5	1666	1642	1587
25~29세	336	1725	53.0	1846	1813	1795	1764.5	1725	1691.5	1655	1640	1610
30~34세	353	1713	54.0	1845	1800	1785	1748	1710	1677	1643	1630	1606
35~39세	358	1707	56.9	1835	1800	1780	1747	1707	1670	1636	1613	1570
40~49세	409	1686	54.9	1822	1775	1756	1722	1687	1649	1615	1595	1556
50~59세	382	1661	54.6	1780	1750	1734	1698	1660.5	1624	1595	1577	1528
60~69세	400	1644	52.6	1753.5	1727	1713	1677	1646.5	1614	1569.5	1547	1516
70세이상	341	1624	59.1	1762	1725	1702	1660	1620	1590	1546	1522	1488

위와 같은 통계 발표 양식은 우리나라 뿐 아니라 다른 나라에서도 공통적으로 널리 사용하는 표준적인 양식이다 (Pheasant, 1996). 그런데 이 양식 자체는 통계적으로 아무런 문제가 없지만 이를 이용하는 과정에서 몇 가지 문제를 야기할 수 있다. 실제 통계를 이용하는 이용자 입장에서는 위에서 제시한 표를 이용하여 연령 그룹들을 통합하여 새로운 통계를 생산하고자 할 수도 있는데 가령, 18세 이상 24세 사이 연령그룹들의 평균키를 계산하는 예를 생각할 수 있다. 일반적으로는 단순히 해당 연령그룹 통계량들의 산술평균으로 새로운 연령 그룹의 통계를 작성하게 된다. 만일 모집단에서 각 연령대별 인구수가 동일하다면 아무 문제가 없다. 하지만 연령그룹별 인구수가 서로 다른 경우에는 편향이 생긴다는 문제를 지적할 수 있다. 인체치수조사와 같이 모든 조사단위들

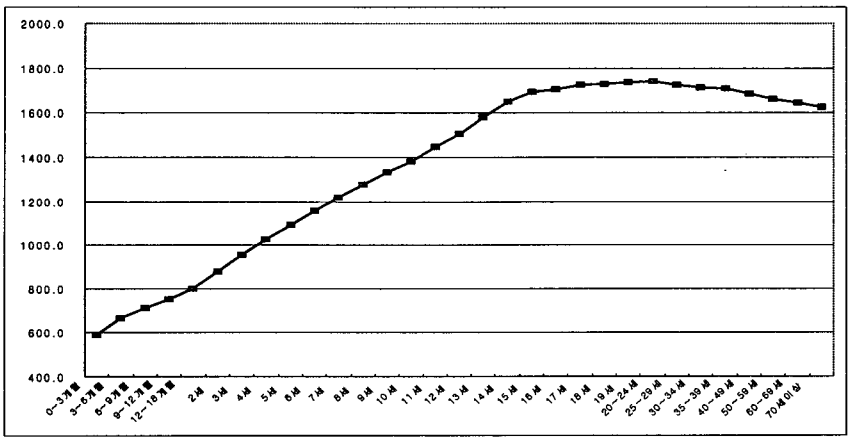
의 추출률이 동일하지 않은 경우(Park 등, 2003), 다시 말해 자체가중표본설계(self-weighting sampling design)가 되지 못하는 복합조사의 경우 설계단계에서의 추출확률을 무시하고 추정하면 생기는 편향의 문제는 널리 알려진 문제이다 (Skinner 등, 1989).

<표 2-2>의 통계값들을 이용하여 18세 이상 24세 남자들의 평균키를 추정해보자. 18세, 19세 그리고 20 ~ 24세 세 연령그룹들의 키와 표준편차를 단순평균하면 각각 1733.7mm와 56.5mm로 계산된다. 그러나 인구주택총조사(통계청, 2001) 결과를 토대로 각 연령그룹별 인구수를 나타낸 것이 <표 2-3>에 나와 있는데 각 연령 그룹별 모집단 인구수는 서로 다르다. 이럴 경우 인구수를 가중값으로 하는 가중평균을 구해야 한다. 18세 이상 24세 연령그룹의 평균키와 표준편차를 가중평균에 의해 계산한 결과는 각각 1736.1mm와 57.5mm로 단순평균과 약간의 차이가 난다. 동일한 통계를 기초로 하여 계산된 값 사이에 이런 불일치가 생기는 것은 바람직하지 않다.

<표 2-3> 남자 18 ~ 24세 키 자료

나이 (Age)	인구수	측정수	평균 (Mean)	표준편차 (S.D.)
18세	411,643	266	1729	54.7
19세	436,551	277	1734	56.4
20~24세	2,028,206	344	1738	58.3
합계	2,876,400	887		

다음으로 여러 연령대별 통계의 비교 양식의 문제점을 지적할 수 있다. <그림 2-1>은 기술표준원의 Size Korea 홈페이지에 실린 <표 2-2>의 연령대별 평균키를 나타낸 그림이다. <표 2-2>를 보면 관심 연령그룹에 대해 각각 평균, 표준편차 이외에도 분위수 통계들이 나와 있다. 그러나 <그림 2-1>에는 <표 2-2>에 나온 통계들 중 분위수 통계는 전혀 나타내지 못한 채 단지 평균값만 보여주고 있다. 다시 말해 그림이 연령그룹별로 생산된 통계정보들을 제대로 전달하지 못하고 매우 제한적인 정보만 나타내는 빈약한 것임을 지적할 수 있다. 뿐만 아니라 각 연령대별 가중값이 서로 다름에도 불구하고 이것은 전혀 반영하지 못하고 있다는 점도 문제점으로 지적되어야 한다.



<그림 2-1> 남자 연령대별 평균키 그림

3. 새로운 통계 발표양식

앞 장에서는 현재 사용하고 있는 통계 발표양식의 두 가지 문제점을 지적하였는데 이 장에서는 각각의 문제점에 대한 해결책으로서 바람직한 통계 발표양식을 제안한다. 인체치수조사와 같이 모든 조사단위들의 추출률이 동일하지 않은 비자체가중설계에서 통계표를 작성할 때에는 <표 3-1>와 같이 서로 각 연령그룹별 기술통계량 이외에 해당그룹의 가중값을 함께 나타내는 것이 바람직하다. 그럴 경우 서로 다른 연령그룹들을 서로 병합하여 통계를 생산하는 경우라고 해도 가중값을 이용한 가중평균을 사용하여 편향을 제거할 수 있기 때문이다.

<표 3-1> 새로운 통계표 제안 양식

나이 (Age)	인구수	측정수	평균 (Mean)	표준편차 (S.D.)
6세	369301	260	1155	46.0
7세	371804	251	1220	48.3
8세	373690	252	1278	53.5
9세	352388	257	1333	58.1
10세	338272	259	1380	57.2
11세	322993	257	1449	68.0
12세	317863	263	1507	80.3
13세	312843	269	1582	75.6
14세	323042	253	1647	74.4
15세	327213	276	1692	58.8
16세	354377	254	1703	57.8
17세	384101	269	1725	55.2
18세	411643	266	1729	54.7
19세	436551	277	1734	56.4
20~24세	2028206	344	1738	58.3
25~29세	2057321	336	1725	53.0
30~34세	2068202	353	1713	54.0
35~39세	2117492	358	1707	56.9
40~49세	3525517	409	1686	54.9
50~59세	2144919	382	1661	54.6
60~69세	1430439	400	1644	52.6
70세이상	693423	341	1624	59.1

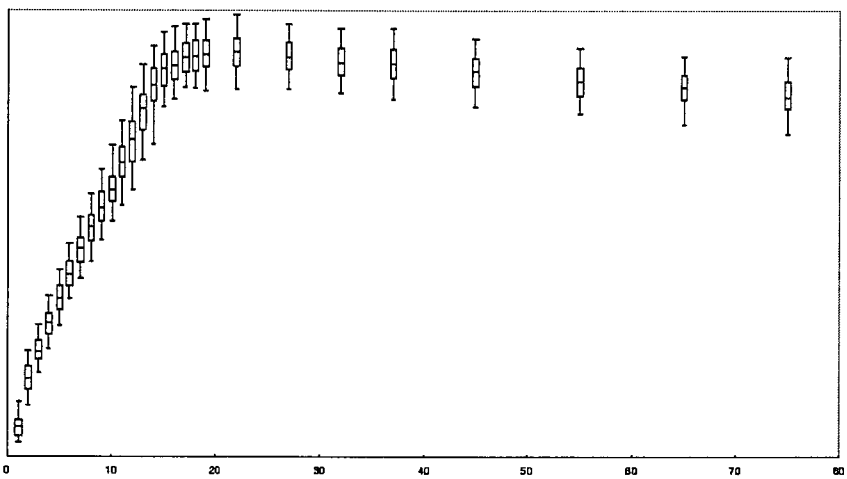
서로 다른 가중값을 갖는 H 개의 연속된 그룹을 서로 병합할 경우 병합된 그룹의 평균과 표본평균의 식은 식 (3.1)과 같이 표현된다.

$$\bar{y}_H = \frac{\sum_{i=1}^{H} w_i \bar{y}_i}{\sum_{i=1}^{H} w_i}, \quad s_H^2 = \frac{\sum_{i=1}^{H} w_i s_i^2}{\sum_{i=1}^{H} w_i} \tag{3.1}$$

여기서 w_i는 설계가중값이고 $\overline{y_i}$, s_i^2는 각각 i번째 연령그룹의 평균과 표본분산이다.

식 (3.1)을 이용하여 18세 이상 24세 남자들의 평균키와 표준편차를 계산한 결과는 각각 1735.8mm 와 57.4mm로 이 조사에서는 인구수를 가중값으로 한 경우와 거의 비슷하다.

두 번째로 <그림 2-1>은 <표 2-2>의 통계정보들을 제대로 반영하지 못하는 빈약한 그림이라는 문제점을 지적하였다. 이를 해결하기 위해 쉽게 생각할 수 있는 방법은 각 연령그룹별 상자그림을 한꺼번에 표현하여 <그림 3-1>과 같이 나타내는 방법이다. 이와 같이 상자그림으로 표현하면 각 연령그룹별 평균키의 추세뿐 아니라 연령그룹별 분포에 대해서도 일목요연하게 파악할 수 있으므로 더욱 유용한 그림이라고 할 수 있다. 이 그림을 그리기 위한 모든 정보들이 <표 2-2>에 모두 들어있기 때문에 이 그림을 그리는데 아무런 어려움이 없다.



<그림 3-1> 남자 나이그룹별 키의 상자그림

원래 조사대상이 되는 개인들의 나이는 연속형으로 주어지는데 분석과정에서 이를 특정한 몇 개의 나이범주로 구분하여 이산형 데이터로 변환시켜 <표 2-2>를 만들었다. <그림 3-1>은 <표 2-2>에 나타난 통계 정보를 이용하여 작성한 것이다. 만일 원 데이터를 활용한다면 <그림 3-1>보다 한 단계 더 발전된 분포함수를 추정한 그림으로 나타낼 수 있다. Korn과 Barry (1998)은 표본조사 데이터를 이용한 커널함수를 이용한 평활법(smooting method)을 소개하였는데 여기서도 그 방법을 그대로 적용할 수 있다. 다음과 같은 삼각 커널함수를 고려하자.

$$K(u) = (1 - |u|) I(|u| \leq 1)$$

이 때 나이 x가 주어질 때의 조건부 평균의 추정량은 다음의 식 (3.2)와 같이 나타낼 수 있다.

$$mean(y \mid x) = \sum_{i=1}^{n} w_i^F y_i \tag{3.2}$$

여기서 커널 가중값을 나타내는 w_i^F는 표본 가중값을 포함하는 것으로서 다음의 식으로 표현된다.

$$w_i^{KS} = w_i K\left(\frac{x-x_i}{h_x}\right) \Big/ \sum_{j=1}^{n} w_j K\left(\frac{x-x_j}{h_x}\right), \tag{3.3}$$

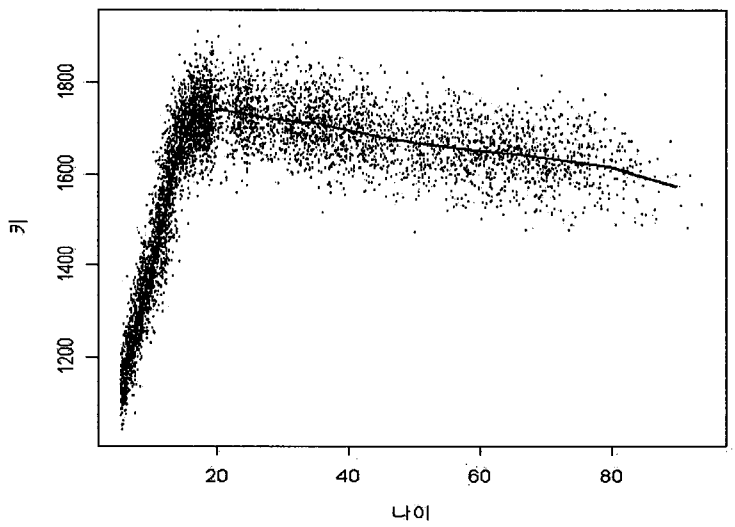
$$w_i^F = w_i^{KS}\left[1+\frac{(x_i-\bar{x}^{KS})(x-\bar{x}^{KS})}{\sum_{j=1}^{n} w_j^{KS}\left(x_j-\bar{x}^{KS}\right)^2}\right], \tag{3.4}$$

이 때 $\bar{x}^{KS} = \sum w_j^{KS} x_j$ 이며, w_i는 설계가중값을 나타낸다.

이 방법을 확장시켜 w_i^F를 이용하여 조건부 백분위수의 추정량을 구할 수 있다. 나이 x가 주어져 있을 때 키 y의 조건부 중앙값 $\mathrm{med}(y|x)$는 w_i^F를 가중값으로 이용한 y_i 들의 가중 경험적 누적분포함수를 이용하여 구한다. 중앙값보다 더 큰 백분위수를 구하기 위해서는 $z_i = y_i - \mathrm{med}(y|x_i)$를 이용한다.

$$\begin{aligned} \text{조건부 90분위수} &= \mathrm{med}(y|x) + h80(z|x) \\ \text{조건부 75분위수} &= \mathrm{med}(y|x) + h50(z|x), \end{aligned} \tag{3.5}$$

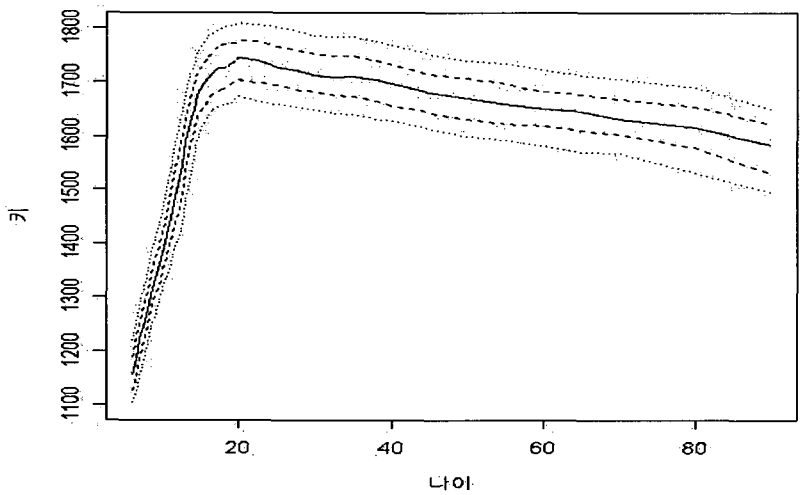
이때 $h80(z|x)$는 z_i중 양수인 자료만을 이용하여 조건부 80분위수를 구한 것이고 $h50(z|x)$는 조건부 50분위수를 구한 것이다. 조건부 10분위수와 25분위수는 z_i중 음수인 자료만을 이용하여 같은 방식으로 구한다.



<그림 3-2> 커널 평활법을 이용한 조건부 가중 평균의 추정함수

<그림 3-2>는 인체치수조사의 나이와 키에 대한 산점도 위에 조건부 평균의 추정함수(식 3.2)

를 선으로 나타낸 그림이다. 그림 <3-3>은 조건부 중앙값과 10, 25, 75, 90분위수들의 추정함수를 나타낸 그림이다. 이 그림들은 각 데이터의 가중값까지 고려하여 작성된 것이며 그림의 추정함수를 사용할 경우 어떤 나이에 대해서도 거기에 해당되는 평균값 또는 다섯 가지 수치요약 값을 추정할 수 있다는 면에서 <그림 3-1>에 비해 한 단계 발전된 형태의 그림이라고 할 수 있다. 띠너비 h_x 는 $[x-h_x, x]$ 또는 $[x, x+h_x]$ 의 구간에 포함되는 관측값의 수가 350개를 넘지 않도록 잡아주었다.



<그림 3-3> 커널 평활법을 이용한 조건부 가중 분위수의 추정함수

4. 맺음말

많은 비용과 인력이 소요되는 대규모 조사의 최종적인 결과는 통계표나 그림으로 표현된다. 그럼에도 불구하고 많은 조사기관에서는 통계표나 그림을 그리는 일에 대해 충분한 관심을 기울이지 않는 경향이 있는데 이것은 매우 안타까운 일이다. 본 연구에서는, 대표적인 예로서 인체치수통계의 양식에서 나타나는 두 가지 문제점을 지적하였고, 아울러 그 해결책을 모색해보았다.

인체치수통계의 경우 보고서나 홈페이지에 보고된 통계는 나이그룹별로 작성된 통계로서 그 자체로는 편향의 문제가 발생하지 않는다. 그러나 보고서에서 제공하는 나이구분과는 달리 나이그룹의 구분을 다르게 하는 경우 각 나이그룹별 추출률이 서로 다른 까닭에 편향의 문제가 생길 수 있음을 지적하였다. 한편 이에 대한 해결책으로 각 나이그룹별 가중값에 해당되는 인구수를 통계표에 함께 나타낼 것을 제안하였다.

다음으로 나이별 통계를 손쉽게 비교하기 위한 그래프의 작성에 대해서도 고찰하였다. 현재는 많은 유용한 정보들이 있음에도 불구하고 단지 평균의 그래프만 그리고 있는 실정임을 지적하였다. 아울러 이에 대한 개선책으로 나이그룹별 상자그림을 그려 비교하는 것과 커널 평활법을 사용한 추정함수 그림으로 나타내는 방법을 제안하였다.

대규모 표본조사에서는 인체치수조사와 같이 표본설계 단계에서 각 조사 단위들의 추출확률을

서로 다르게 추출하게 되는 것이 일반적이다. 이런 경우 인체치수통계에서 지적한 것과 동일한 문제들이 흔히 발생하므로 본 연구의 결과는 하나의 특정한 조사에만 국한된다고 하기 어렵다. 많은 비용과 노력을 들여 수행한 조사인데 분석 단계에서 사소한 부주의로 통계의 신뢰를 떨어뜨린다면 이것은 매우 어리석은 일이 될 것이다. 그러므로 통계 생산자 입장에서는 늘 예상되는 다양한 문제들을 미리 생각하여 통계의 품질에 손상이 가지 않도록 대비하는 것이 필요하다.

참고문헌

[1] 산업자원부 기술표준원 (2004). 「제5차 한국인 인체치수조사 사업보고서」.

[2] 통계청 (2001). 2000 인구주택총조사 cd.

[3] ISO 15535 (2003). *General requirements for establishing anthropometric databases.*

[4] Jinwoo Park, Jinho Kim, Inkeuk Hwang (2003). "A Sampling Design of the Korean Anthropometric Survey", *The Korean Communications in Statistics*, Vol. 10, 3, pp 707-718.

[5] Korn E. L. , Graubard, B. I. (1998), "Scatterplots With Survey Data", *The American Statistician*, Vol. 52, 1, pp 58-69.

[6] Pheasant, Stephen (1996), *Body Space*, 2nd ed., Taylor & Francis Books Ltd.

[7] Skinner, C. J., Holt, D., and Smith, T. M. F. (1989), *Analysis of Complex Surveys*, John Wiley & Sons.

[8] http://sizekorea.ats.go.kr/ (2005. 2. 14 접속)

[2005년 3월 접수, 2005년 6월 채택]