『통계품질관리』 워크샵

2004. 9. 17.

통계청 KNSO 통 계 청 통계청

목 차

『통계품질관리 워크샵』 1교시

조사오차 측정방법

이 석 훈

(충남대학교, shlee@stat.cnu.ac.kr)

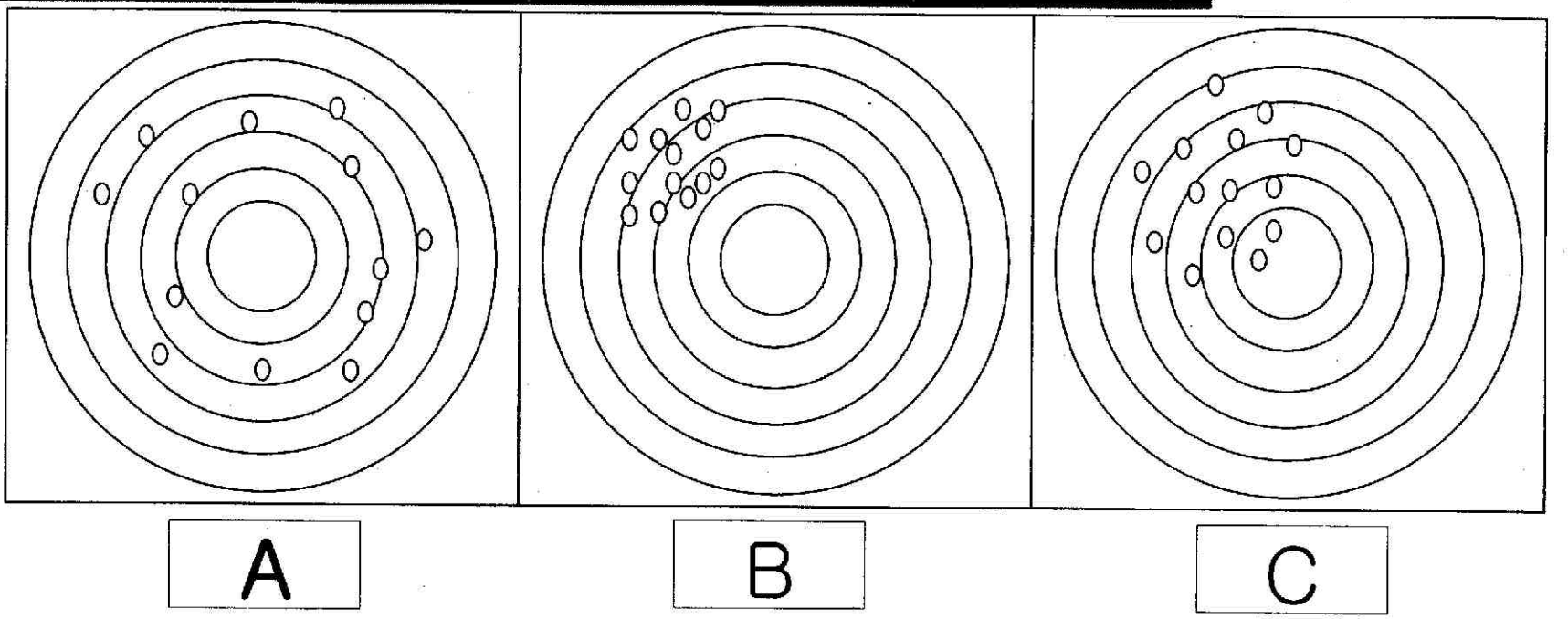
조사 오차 측정 방법

충남대학교 이 석 훈

e-mail : shlee@stat.cnu.ac.kr

차 례

(가)생각해보기



- ❖ 한발을 쏜다면 누구를 시킬 것인가?
- ❖ 세 사람 중 가장 신뢰할 만한 사람은?
- ❖ 세 사람 중 가장 오차가 많은 사람은?
- ❖ 만약 선택한 사람이 쏘게 된다면 과녁을 맞출 수 있을 것인가?
- ❖ 세 사람의 실력차이를 가능한 구체적으로 말해보시오

기술하시오.

생각해보기(1)

❖ 20세 이상 주민 10만 명의 정책 A에 대한 지지율 조사를 위하여 세 기관(가, 나, 다)에 서는 동일한 목표모집단, 동일한 연구모집단 동일한 추출틀, 동일한 표본추출법을 사용하여 조사한 결과가 다음과 같다고 한다.

기관	조사인원	지지인원	지지율
가	900명	720명	80%
나	625명	500명	80%
다	400명	320명	80%

❖ 세 기관의 결과를 비교하면서 전국민의 지지율에 대한 입장을 정리하시오.

기술하시오.

세 기관 모두 80%라고 나왔습니다.
다 같은 결과라고 해도 되겠습니까?

생각해보기(2)

특정지역에 대하여 부채경감정책수립을 위하여 정책 수립기관에서는 그 지역의 가구 당 평균 부채액을 추정하고자 한다. 조사방법론에서 제안하는 최적의 방법들을 이용하여 추출된 900가구를 대상으로 면접조사를 수행하였다. 이들 중 275 가구는 응답을 하지 않았다. 조사시점에서 응답한 625 가구로부터 얻은 결과는 평균 2300만원, 표준편차 500만원이 나왔다.

이 결과를 바탕으로 정책을 수립하려고 할 때, 이 결과의 해석에서 유의할 사항을 기술하시오.

이 때 사용한 질문은 다음과 같다.

"귀하의 가정이 갖고 있는 총 부채액수는 얼마 입니까?"

기술하시오.

2300만원을 평균 부채액으로 생각하고
정책을 입안해도 되겠습니까?

생각해보기(3)

- ❖ 모 이동통신사에서 자사의 고객DB를 추출틀로 하여 6개월 후의 국가경제에 대하여 어떻게 생각하는지 900명에게 전화 조사를 실시 하였다.
- ❖ 응답항목은 다음의 세 가지 중에 하나로 하였다.
  - ① 좋아질 것이다.
  - ② 현재와 같을 것이다.
  - ③ 나빠질 것이다.
- ❖ 조사결과 응답자 880명으로부터 30%가 나빠질 것이라고 하는 결과를 얻었다.
- ❖ 이 결과를 활용하려고 할 때 고려되어야 하는 사항을 논하시오.

기술하시오.

"우리 국민 30%는 경제가 나빠질 것이다"

라고 생각해도 되겠습니까?

생각해보기(4)

❖ 특정지역에 있는 3세 이하 자녀 1인을 갖고 있는 기혼 근로자들의 경제생활을 조사한 조사표에서 임의의 5명을 추출한 지난달 생활비 수치는 다음과 같다.

(예) 160, 100, 30, 120, 140 단위 (만원)

❖ 이 수치로부터 이 지역 안에 있는 결혼한 지 3년 이내의 가정의 생활비에 대하여 기술하시오.

기술하시오.

5명의 평균 110만원을 이 지역 안에 있는
결혼한지 3년 이하인 가정의 생활비라고
생각해도 되겠습니까?

생각해보기(5)

❖ 특정지역에서 정책 A에 대한 20세 이상 주민 10만 명의 통신비를 조사하고자 하여 세 기관으로부터 용역 제안서를 받았다

(법적 제약은 없다고 가정한다). 조사표는 동일하다고 하자.

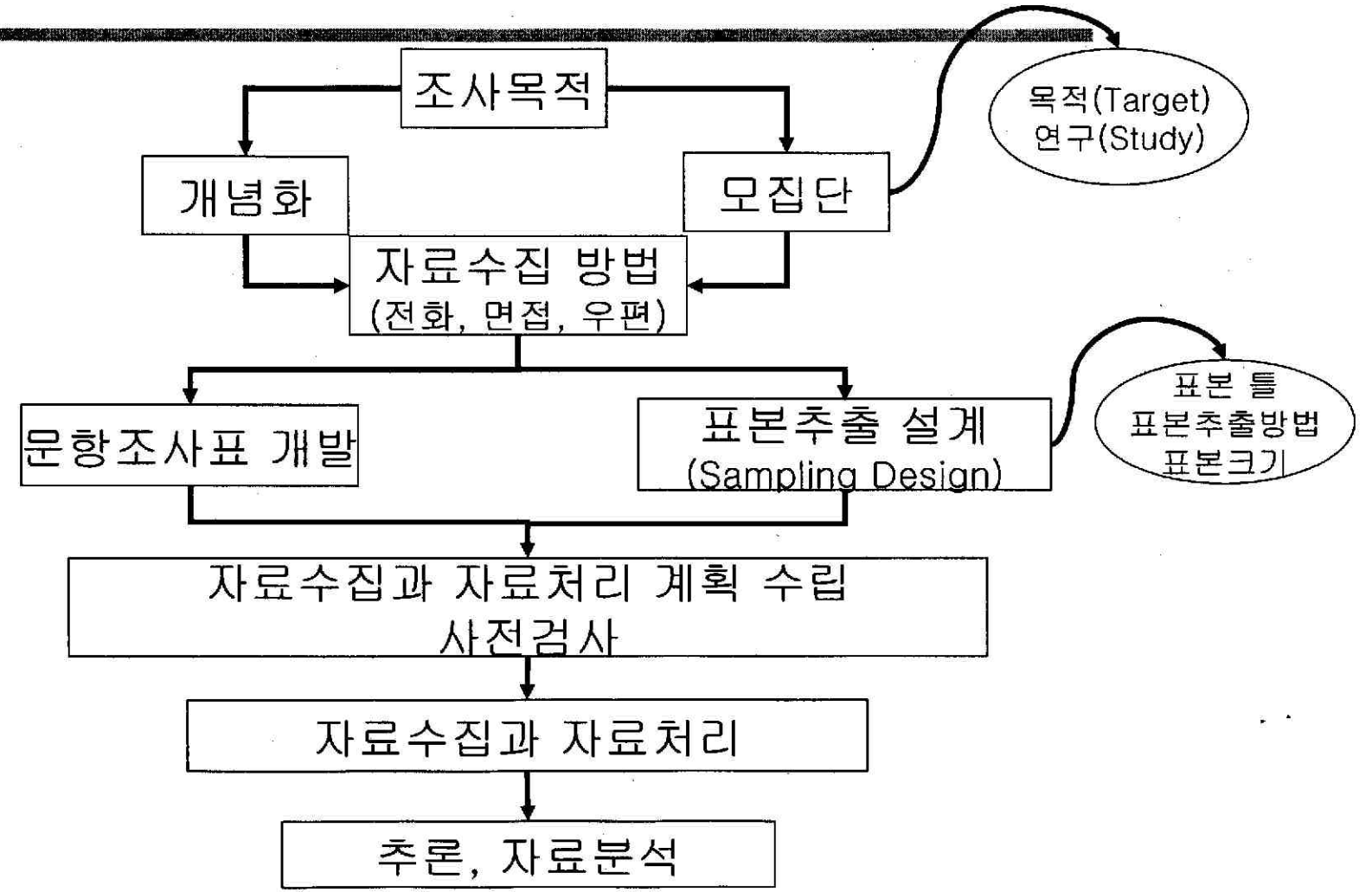
자료 수집방법	전 화(기관(가))	우 편(기관(나))	면 접 (기관(다))
표본 틀	7자리의 수 집합	전화회사 DB	GIS출력물
표본추출방법	단순임의 추출법		
표본 수	2,500	10,000	500
가격	동 일		

❖ 세 기관을 비교 심사한다고 할 때 고려하는 내용들을 구체적으로 기술 하시오.

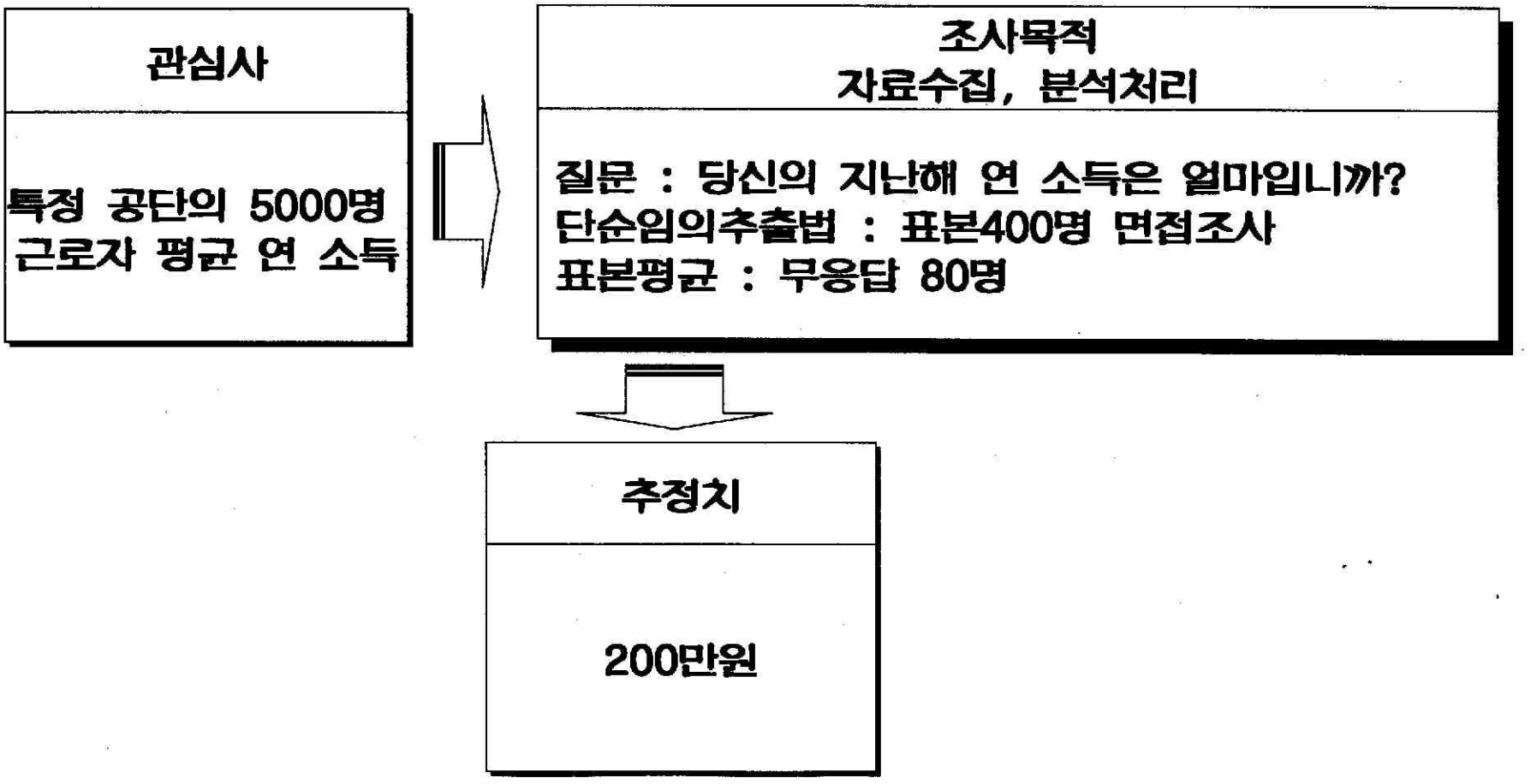
기술하시오.

예컨데 표본수가 제일 큰 기관 (나)가
제일 좋다고 해도 되겠습니까?

조사 연구 과정(Survey Process)



조사 연구 과정의 예
(Survey Process)




이 공장의 근로자들의
평균 연 소득은
200만원 입니다.

당신은 200만원이
실제 이곳 근로자들의
평균 연 소득(참값)이라고
생각하십니까?

예

아니오

기술하시오.

왜 아니라고 생각하는지를
기술하시오.

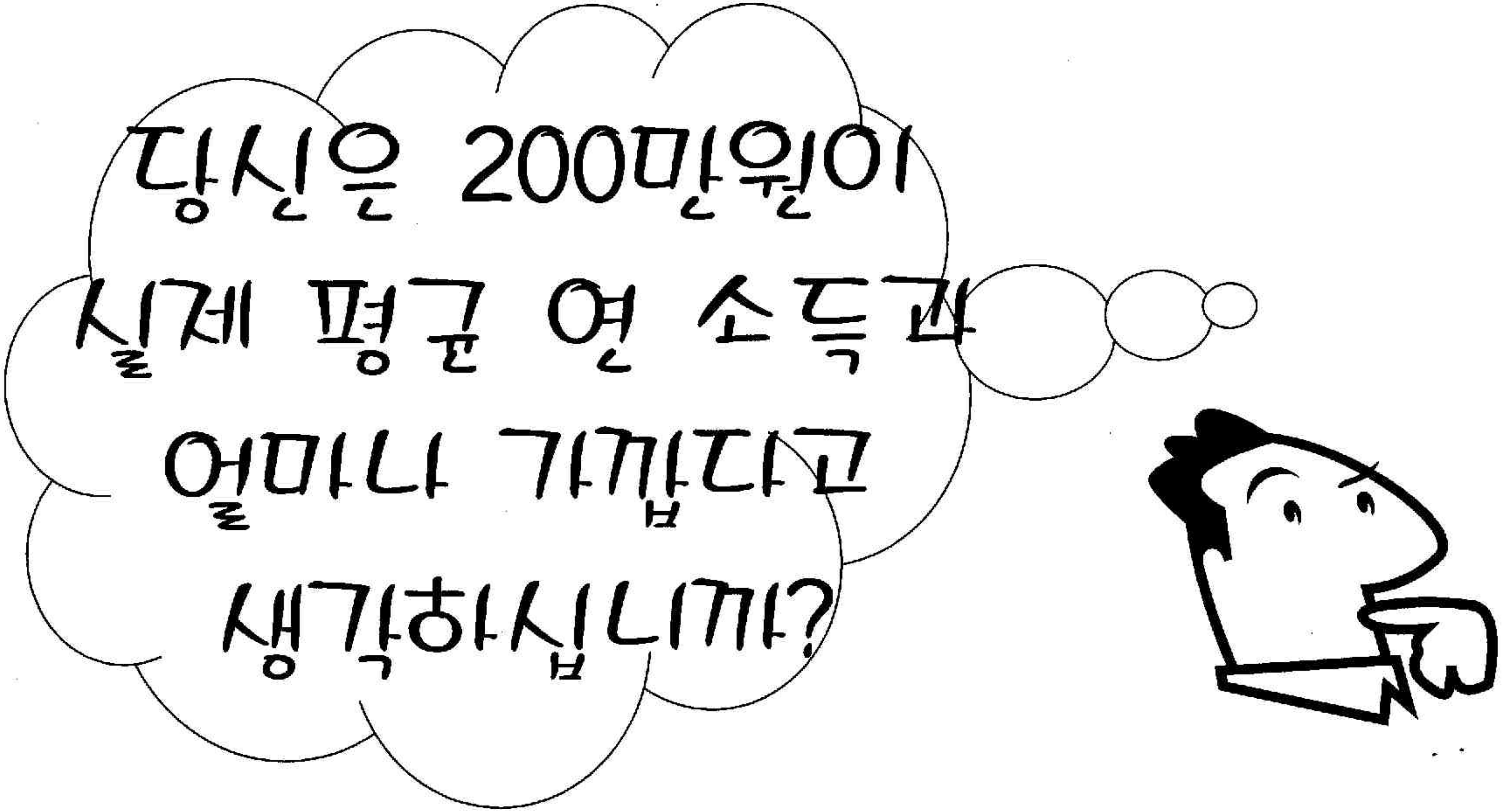
측정치가 참 값이라고 할 수 없는 이유들.

1. 조사목적이 불분명해서 응답자들이 응답을 기피하거나 불성실하게 했을 수 있음
2. 표본 400명을 추출한 명단에 비정규직이 포함되어 있지 않을 수 있었음
3. 800명의 무응답자들은 연 소득이 많은 사람들일지도 모른다.
4. 질문이 연 소득이라고 했는데 기타 소득의 포함여부가 불확실함

4.1 면접자가 50세 였기 때문에 응답자들이 실제 소득보다 낮게 응답했을 수 있음

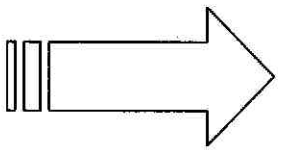
5. 조사가 늦어져서 2시간 만에 자료를 입력, 분석했다고 하는데...
6. 전체 5,000명을 다 조사하지 않고 일부 표본만 조사했음

6.1 단순임의추출을 하였기 때문에 초임자들이 많이 뽑힐 수도 있었음


당신은 200만원이
실제 평균 연 소득과
얼마나 가깝다고
생각하십니까?

총 조사오차 : 측정값 - 참값

참값은
모른다



총 조사오차도
모른다

(나). 조사오차의 발생 유형

A. 오차의 경향이 비계통적 오차 (Variable Error)

- ➢ 방향성이 없음.

B. 오차의 경향이 계통적 오차 (Systematic Error)

- ➢ 방향성이 있음

- ▪ 오차의 발생 이유에 따라 유형을 생각해 봅시다.

오차요인별 발생유형

발생이유	오 차 요 인	발생유형	
		A	B
1	명세사항 : 목적, 개념	약	강
2	추출 틀(표본추출 틀, 표본 틀) - 제거오류, 추가오류, 중복오류	약	강
3	무응답 : 항목, 개체	약	강
4	측정오차 : 자료수집방법, 면접자, 도구	강	강
5	자료처리오차 : 입력, 가중치	강	강
6	표본오차 : 표본크기, 확률 표본추출방법	강	약

(다).중간결론

이 지역 5000명의 근로자의 평균 연 소득을 200만원이라고 추정한 것에 대한 우리의 신뢰감은

총 조사오차 = 200만원 – 참값

을 생각하게 되는데 우리는 참 값을 알 수가 없다.

신뢰감은 조사목적 설정부터 자료입력 분석까지의 일체의 과정에 대한 종합적 느낌이다.

조사과정의 각 단계별 오차를 보는 두 가지 관점

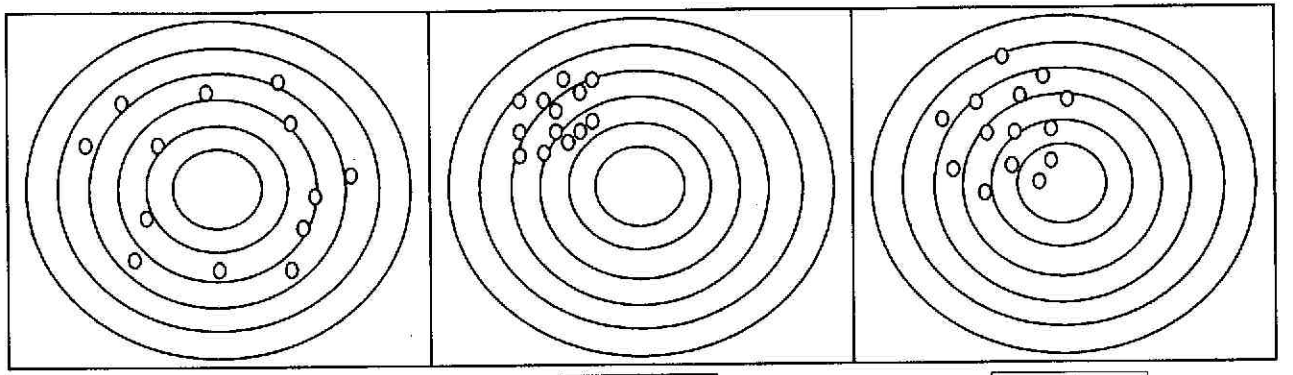
❖ 오차 발생요인에 따른 분류

- 표본오차
- 비표본 오차

❖ 오차 발생유형에 따른 분류

- 비계통적 오차
- 계통적 오차

(라)오차의 계량적 표현

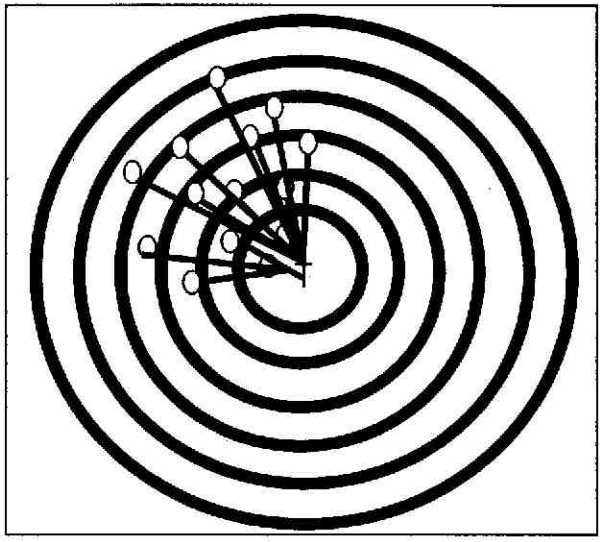


❖ 양궁의 세 선수에 대한 신뢰도, (오차정도)를 수치로 표현할 수 있습니까?

예 아니오

가능 : 과녁을 알고 있다. 여러 번 시행한 결과에서 오차의 유형이 나타났다.

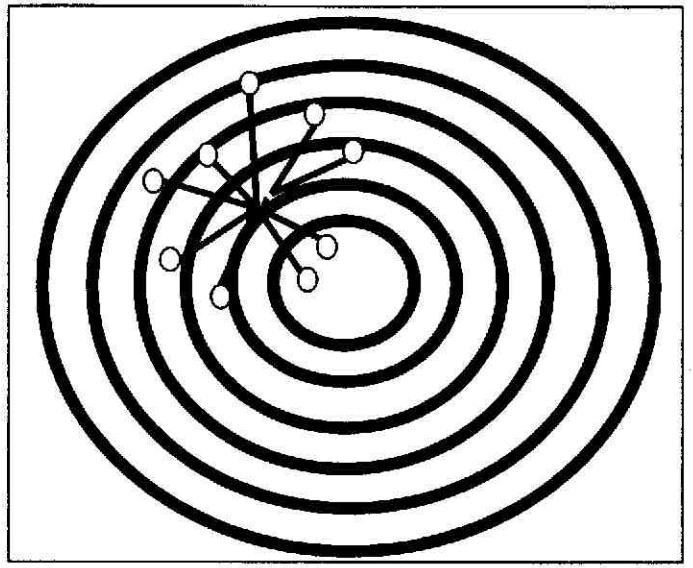
(1) 과녁을 기준으로...



- ❖ 직관 : 결과점들과 과녁(T)과의거리
- ❖ 계량화 (결과점들과 과녁과의 거리)의 제곱의 평균 =>평균 제곱합(MSE)

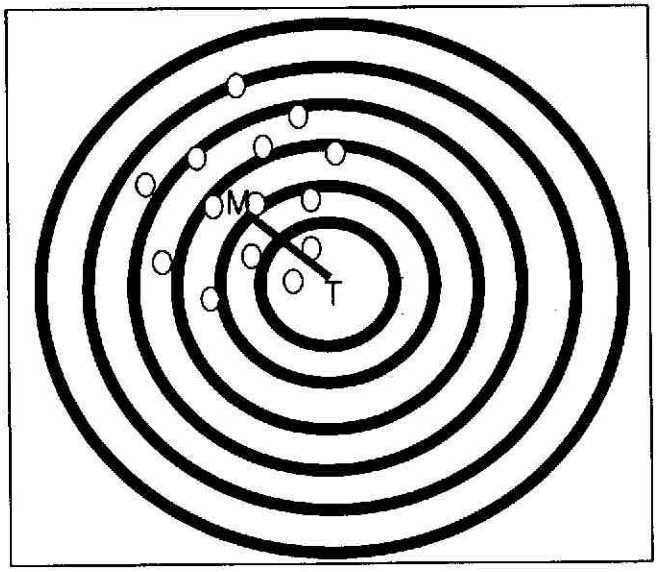
■흩어짐과 쏠림의 관점으로 본다면.....

(2)흩어짐:비계통적 오차의 척도



- 직관 : 결과 점들과의 그들의 중심(M)과의 거리
- 계량화 : 결과 점들과의 그들의 중심(M)과의 거리의 제곱의 평균(분산:variance)

(3)쏠림 : 계통적 오차의 척도



- ❖ 직관 : 결과점들의 중심(M)과 과녁(T)의 거리
- ❖ 계량화 : 편향(Bias)

(4)멋있는 수학적 결과

평균제곱오차=분산+편향의 제곱

MSE = Variance + (Bias)*(Bias)

양궁선수들 비교

오차 유형	A	B	C
비계통적 유형	강	약	중
계통적 유형	없음	강	중

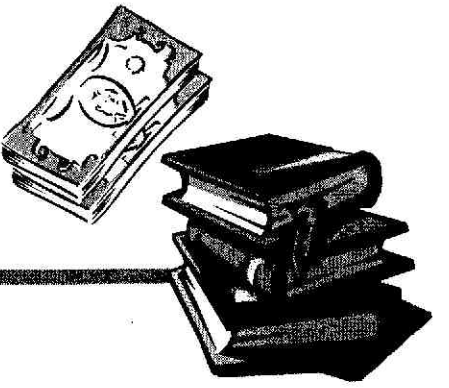
(5) 추정치에 대한 신뢰도

- "근로자의 소득 추정" 예에서 200만원 이라는 추정치에 대한 신뢰감을 수치화할 수 있겠는가?
- "근로자의 소득 추정"을 위한 조사과정에 대한 신뢰감
- 양궁의 예와 무엇이 다른가?

조사과정의 어려움

양궁	조사과정
과녁이 나타나 있다.	참 평균소득을 모른다.
한 선수가 여러 번 시행한 결과 점들이 있다.	1회의 조사 결과인 200만원 이라는 하나의 결과뿐이 없다.

돈과 이론이 필요하다



이론 ⇒ 표본오차

돈 + 이론 ⇒ 비표본 오차

표본오차 척도를 위한 이론

❖직관

- 전수조사에는 표본오차는 0 이다.
- 표본의 크기가 작으면 작을수록 표본오차는 커진다.

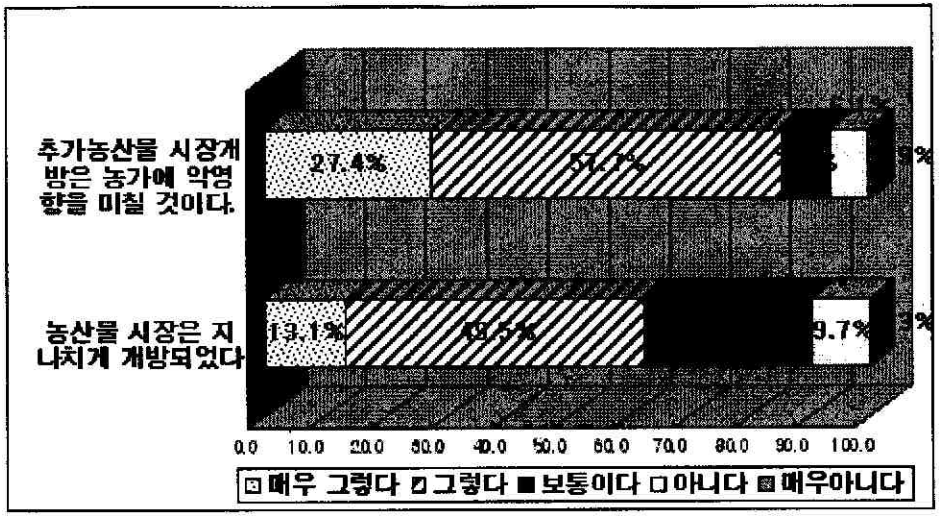
❖대표적인 예 : 표본 평균

표준오차 란?

- ❖ 정해진 조사방법으로 표본평균을 여러 번 구해본다면 비현실적인 가정이지만 가능하겠다.
- ❖ 이론에서는 이들 표본평균 값들은 편향은 0 이고, 분산은 (개인들의 소득 분산/ 표본크기) 가 된다.
  이때 분산의 제곱근 ; 표본평균의 표준오차라고 한다.
- ❖ 위의 조사방법에서 단순임의추출법이 아니고
  층화 추출법이 되면 이 결과는 조금 복잡해 진다.
- ❖ 위의 조사방법으로 표본평균이 아니라 표본 중위수를 구한다면 이론적인 표본 중위수의 편향과 표본 중위수의 표준오차는 상당히 복잡한 값이 된다.

=>표본평균 선호 이유

95 %, ±2.45% 포인트는???



ⓒ전농 전북도연맹

전농 전북도연맹은 이 조사가 전국 7개 특.광역시 거주 만 20세 이상 비농민 남·여를 인구비례에 의거 표본을 할당·추출해 2004년 8월 13일부터 20일까지 7일간 질문지를 통한 직접 면접조사로 진행되었고 유효 표본은 1,605명이며 95% 신뢰수준에서 표본오차는 ±2.45%포인트 라고 전했다.

김동현 기자

기술하시오.

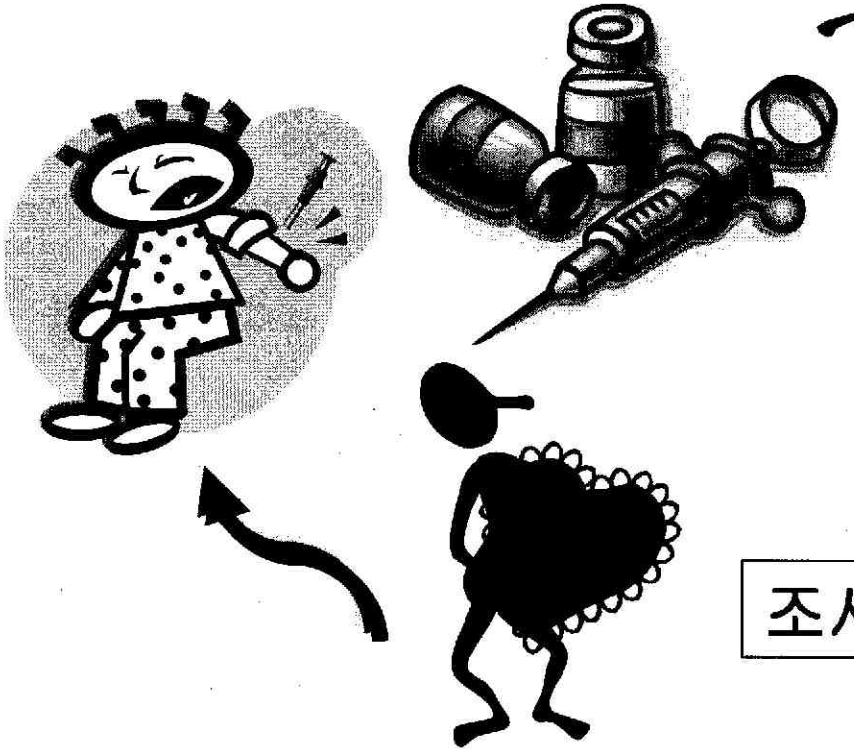
신뢰수준 95%의 의미는?

비표본 오차 척도를 위한 돈과 이론

❖ 비표본 오차는 분명히 존재하고 심각할 수 있다.
하지만 이론이 거의 없다. 따라서 많은 실험을 통하여
오차의 존재를 확인하고 심각성을 조사해야 한다.

- 예 : 추출틀 왜곡현상
  - 생각해 보기(3)에서 통신화사 DB와 국민전체와의 차이가 크다.
  - 실험
  - 통신회사 DB에 포함되어 있지 않은 사람들로부터 표본추출 조사하여 결과를 비교 => 결과는 일반적이지 않다.
- 예 : 무응답 현상
  - 생각해 보기 (2)에서 무응답들 중 다른 문항에 답한 사람의 경우 ⇒ 이론을 적용하여 추측해 본다.

돌 파 구



사전예방

조사자의 열정 요구

비표본 오차척도 : 실험결과, 경험에서 오는 추측, 이론적 결과

생각해보기(5)의 가상적 척도

	전화	우편	면접
추출 틀 편향제곱	1.50	2.10	0.00
무응답 편향제곱	2.60	3.70	0.85
측정 편향제곱	3.80	3.80	0.75
측정 분산	0.45	0.21	0.33
표본 평균의 분산	1.80	0.90	4.00
총 평균제곱오차 (TMSE)	10.15	9.81	5.93

❖ 비용이 다 같기 때문에

⇒ 면접방식을 제안한 (C)기관이 가장 작은 TMSE를 보이기 때문에 (C)기관을 선정한다.

결론

❖ 조사에 대한 신뢰감은

- 조사결과값이나 추정치로부터 얻어지는 것이 아니라
- 조사기획 단계부터 자료입력, 분석, 결과발표단계까지의 일체의 과정에 대한 종합적 느낌이다.

❖ 신뢰감에 대한 계량적 표현은 조사오차로 나타난다

- 조사오차는 오차발생 요인에 따라서 표본오차와 비표본오차 두가지로 측정된다.
- 조사오차는 오차발생 유형에 따라서 비계통적 오차(분산)와 계통적 오차(편향)으로 측정된다.

❖ 일반적인 발생요인별 발생유형은 다음과 같다

	오 차 요 인	발생유형	
		A	B
비 표 본 오 차	명세사항 : 목적, 개념	약	강
	추출 틀(표본추출 틀, 표본 틀) - 제거오류, 추가오류, 중복오류	약	강
	무응답 : 항목, 개체	약	강
	측정오차 : 자료수집방법, 면접자, 도구	강	강
	자료처리오차 : 입력, 가중치	강	강
	표본오차 : 표본크기, 확률 표본추출방법	강	약

유익한 시간이 되셨으면 좋겠습니다.

감사합니다.

조사 무응답 분석기법

김 규 성

(서울시립대학교 통계학과, kskim@uos.ac.kr)

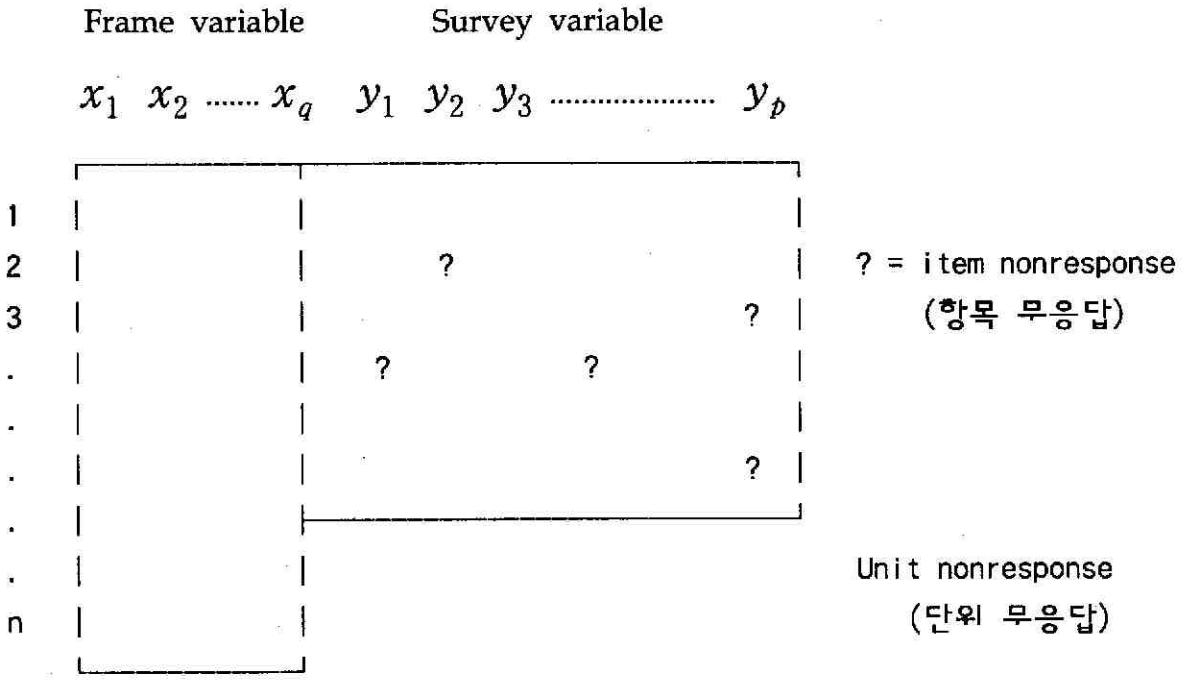
차 례

통계조사에서 무응답 데이터의 정의

- 무응답 데이터의 정의
  - 통상적인 정의 : 표본조사단위에서 비 고의적(unintentionally)으로 얻지 못하는 데이터.
  - 확장된 정의 : 고의적이든 비고의적이든 표본조사단위에서 얻지 못한 정보로서, 결과 데이터의 분석을 어렵게 함.

- 무응답 비율 : 미국의 CPS의 경우
  - 1994년 : 6.19% (Noncontact : 2.30%, Refusal : 2.41%)
  - 1995년 : 6.86% (Noncontact : 2.41%, Refusal : 3.89%)
  - 1996년 : 6.63% (Noncontact : 2.28%, Refusal : 4.09%)

무응답의 형태

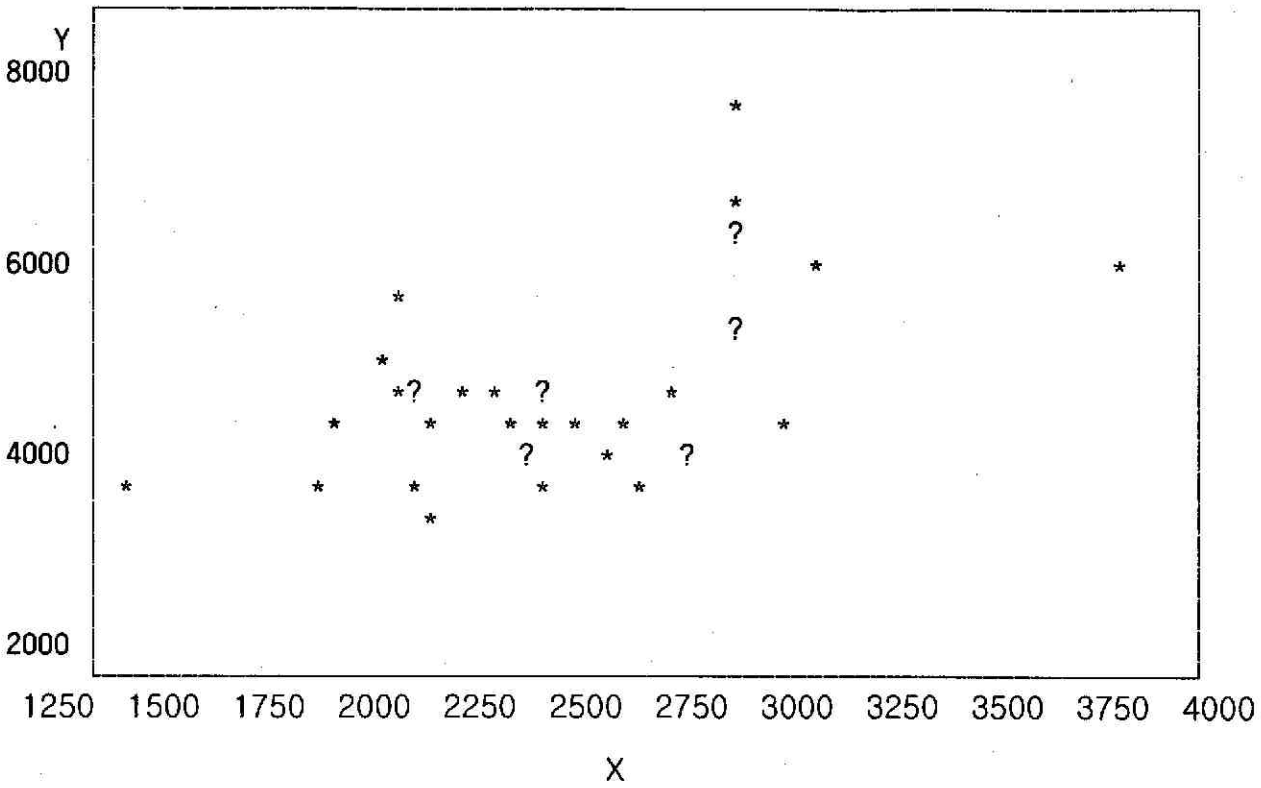

Frame variable
Survey variable
x_1 x_2 ..... x_q y_1 y_2 y_3 ..... y_p
1
2
3
.
.
.
.
.
n
?
?
?
?
? = item nonresponse
(항목 무응답)
Unit nonresponse
(단위 무응답)

예제 : 부산지역 제조업체의 연간매출액

<표 1> 조사변수 Y와 보조변수 X의 값 (단위 : 100,000)

번호	보조변수 X	주변수 Y	응답율		
			90%	80%	70%
1	2092.50	3450	3450	3450	3450
2	1293.24	3500	3500	3500	3500
3	2636.92	3598	3598	3598	3598
4	1791.74	3626	3626	3626	3626
5	2051.30	3676	3676	3676	3676
6	2357.96	3755	3755	3755	3755
7	2557.70	3890	3890	3890	?
8	2730.16	3898	3898	3898	3898
9	2330.70	4050	4050	?	?
.....	...	.....	.....	.....	.....
25	2017.98	5557	5557	5557	?
26	3077.87	5962	5962	5962	5962
27	3854.34	6136	6136	6136	6136
28	2873.63	6270	?	?	?
29	2861.77	6743	6743	6743	6743
30	2886.60	7600	7600	7600	7600

- 응답률 80%



<그림 1> 조사변수 Y와 보조변수 X의 산점도

무응답 편향

- 발생한 무응답을 무시했을 때 나타나는 현상 : 무응답 편향
- 예제 (가상 모집단) : 모집단을 응답층과 무응답층으로 구분한다고 가정

층	층 크기	층평균	추정량
응답층	$N_R = 800$	$\overline{Y}_{RU} = 180$	$\overline{y}_R$
무응답층	$N_M = 200$	$\overline{Y}_{MU} = 150$	
전체	$N = 1000$	$\overline{Y}_U$	$\overline{y} = ?$

- 모평균 : $\overline{Y}_U = \frac{800}{1000} \times 180 + \frac{200}{1000} \times 150 = 174$

- 응답평균 : $\bar{y}_R$
- 응답 평균의 편향 :

$$E\{\bar{y}_R\} - \bar{Y}_U = \frac{N_M}{N}\{\bar{Y}_{RU} - \bar{Y}_M\} = \frac{200}{1000}(180 - 150) = 6$$

⇒ (i) 응답층과 무응답층의 평균이 다르면 편향 발생

(ii) 무응답율이 커지면 편향은 커짐

⇒ 무응답율을 줄이는 것이 좋은 방법임

무응답을 다루는 방법

(1) 무응답 발생을 방지

- 무응답률이 최소가 되도록 조사 실시
- 최선의 방법임

(2) 무응답 단위 중에서 확률 재표집 & 재조사

- 무응답 층에 관한 추정을 위해

(3) 모형을 이용하여 무응답값 추정 : 모형기반 방법

- 정밀한 모형의 구축이 관건

(4) 무응답 무시

- 예전의 방법. 최악의 선택임

응답율에 영향을 주는 요인

[1] 조사 형태

- 인구변수 관련 조사
- 경제변수 관련 조사
- 사회/경제 관련 조사

[2] 조사자 특성

- 조사자 자질
- 조사자의 업무 동기
- 훈련
- 작업량
- 데이터 수집 방법

[3] 응답자 특성

- 응답자 응답 동기
- 가용성
- 응답자 부담

예제 : 무응답 발생 원인

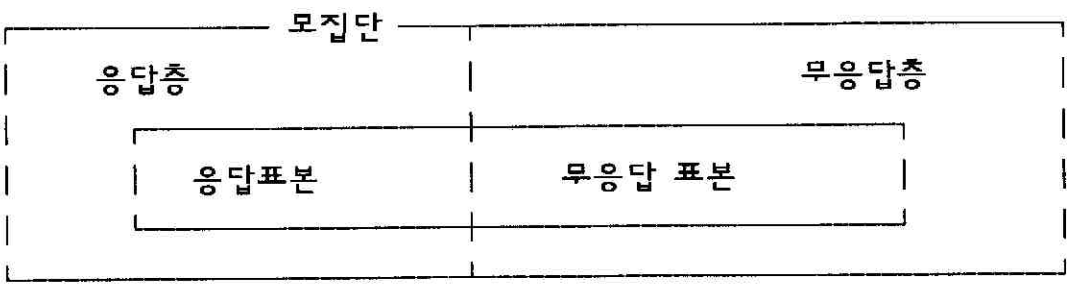
<표 2> 2003년 농가경제조사

교체사유	표본농가수
비농가	44 (28.0%)
전출	25 (15.9%)
단독가구	45 (28.7%)
조사불능	23 (14.6%)
조사불응	15 (9.6%)
장기출타	5 (3.2%)
합계	157 (100%)

통계조사에서 무응답을 어떻게 볼 것인가?

[1] 결정적인 관점

- 개개의 조사단위는 고정된 응답층 혹은 무응답층에 속한다는 관점
- 매우 제한적인 생각
- Hansen & Hurwitz (1946), Cochran(1977), Rao (1983).
- 재조사(follow up survey) 필요



[2] 2단계 확률화 관점

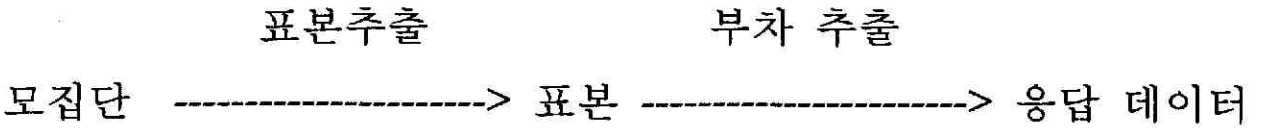
- 1단계 : 표본추출단계

$$I_k = \begin{cases} 1, & \text{표본에 포함} \\ 0, & \text{표본에 비포함} \end{cases}$$

- 2단계 : 개개의 조사단위는 응답을 확률적으로 한다는 생각

$$R_k = \begin{cases} 1, & \text{응답} \\ 0, & \text{무응답} \end{cases}$$

- 표본에 뽑히고($I_k=1$), 응답을 해야 ($R_k=1$) 데이터가 얻어짐.
- 단, 응답 확률을 미리 알 수 없음



[3] 모형 기반 관점

- 데이터와 응답과정은 결합 분포를 갖는다는 관점
  - $P(Y:\theta)$: 데이터의 분포
  - $P(R:\phi)$: 응답과정

  $\Rightarrow$ $P(Y,R|\theta,\phi)=P(Y|\theta)P(R|Y,\phi)$: 결합분포

- 우도 추론과 베이지안 추론에 유용

무응답 메카니즘

- 일반적으로 응답 메카니즘은 알 수 없음
- 무시가능한(Ignorable) 무응답 메카니즘
    - 응답 패턴이 조사변수와 무관

- 무시 불가능한 무응답 메카니즘 :
    - 응답 패턴이 조사변수와 관련이 있음 (무시 곤란)

무응답 처리 방법의 역사

- 1980년 이전 :
  - 무응답이 있는 단위를 제거 후 분석
  - 현재, 통상적인 통계 패키지에서도 이용

- 1980년 이후 : 여러 무응답 대처 기법이 개발
  - 설계 기반 추론 : 가중치 조정법, 대체 방법 (Imputation)
  - 모형 기반 추론 : EM 알고리즘, Data augumentation

- 현재는 가중치조정법과 대체법이 널리 쓰이고 있음

무응답을 위한 가중치 조정법

- 모집단 : $U=\{1,2,3,\ldots,N\}$
- 표본추출 : $s=\{i_1,\ldots,i_n\}$ - 표본
  - 포함확률 : $\pi_i>0$
  - 기본가중치 : $w_i=\frac{1}{\pi_i}$
- 응답 메커니즘 :
  - 응답확률 : $\phi_i>0$
  - 보정된 가중치 : $\tilde{w}_i=\frac{1}{\pi_i}\times\frac{1}{\phi_i}$

가중치 조정법의 고려사항

- 응답확률 ϕ_i을 어떻게 알 것인가?

  ⇒ 추정된 응답확률 $\hat{\phi}_i$을 사용하는 것이 현실적임

  ⇒ 응답확률 ϕ_i을 어떻게 추정할 것인가?

(1) 가중값 부여군 조정 (weighting-class adjustment) 방법

(2) 사후추정(post-stratification) 방법

가중값군 조정 (Weighting-class adjustment)

- 보조변수를 이용하여 응답률이 비슷한 단위들로 가중치 부여군을 형성
- c 군의 응답율 추정 :

$$\hat{\phi}_c = \frac{\sum_{i \in r_c} w_i}{\sum_{i \in s_c} w_i}$$

r_c : c 군에 속하는 응답자

s_c : c 군에 속하는 표본

$\Rightarrow$ 가중값 조정 : $\tilde{w}_i = w_i \times \frac{1}{\phi_c}, \quad i \in s_c$

예제 : 나이가 가중값군인 경우

	나이					합계
	15-24	25-34	35-44	45-64	65+	
표본수	202	220	180	195	203	1,000
응답자수	124	187	162	187	203	863
표본가중치 합계	30,322	33,013	27,046	29,272	30,451	150,104
응답가중치 합계	18,693	28,143	24,371	28,138	30,451	
ϕ_c	0.6165	0.8525	0.9011	0.9613	1	
응답가중값	1.622	1.173	1.110	1.040	1	

가중값 조정후 모총계추정

단순임의추출인 경우

(1) 표본추출 : 단순임의추출

(2) 가중값 조정군 형성 : c개의 군 형성. n_c : 표본수, n_{cR} : 응답자수

(3) 군별로 가중값 조정 : $\tilde{w}_i = w_i \frac{1}{\hat{\phi}_c} = w_i \frac{1}{n_{CR}/n_c}$

(4) 모총계추정치 :

$$\hat{t}_{wc} = \sum_{i \in r} \tilde{w}_i y_i = \sum_c \sum_{i \in r_c} \frac{N}{n} \frac{n_c}{n_{cR}} y_i = N \sum_c \frac{n_c}{n} \bar{y}_{cR}$$

예제 : 평균 가구소득 추정

C=1			C=2		
가구번호	가중치	소득액	가구번호	가중치	소득액
2	14	10.8	3	21	22.1
5	13	11.9	7	24	?
9	14	14.0	18	20	23.6
11	16	12.1	23	19	?
12	13	?	30	25	19.9
21	16	11.8			

- 평균 가구소득은?

예제 : 평균 가구소득 추정 (계속)

class	수	변수	N	합계
1	5	가중치	5	73.00
		소득액	5	884.30
2	3	가중치	3	66.00
		소득액	3	1433.60

- $\bar{y}_{wc} = \dfrac{884.3 \times 6/5 + 1433.6 \times 5/3}{73 \times 6/5 + 66 \times 5/3} = 17.4776$

사후층화(Post-stratification) 방법

- 가중값군 방법과 유사
- 차이점 : 모집단 비율이 층 가중치로 이용

$$\bar{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{hR}$$

- 참고 : 가중값군 방법은 가중치로 표본비율 사용

$$\bar{y}_{wc} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{hR}$$

예제 : 평균 가구소득 추정 (사후층화)

사후층 1 : $N_1 = 120$			사후층 2 : $N_2 = 90$		
가구번호	가중치	소득액	가구번호	가중치	소득액
2	14	10.8	3	21	22.1
5	13	11.9	7	24	?
9	14	14.0	18	20	23.6
11	16	12.1	23	19	?
12	13	?	30	25	19.9
21	16	11.8			

- $\bar{y}_{post} = \dfrac{884.3 \times 120/5 + 1433.6 \times 90/3}{73 \times 120/5 + 66 \times 90/3} = 15.9959$

사후층화(Post-stratification) 방법의 확장

- 사후층화법의 사용조건 :
  - 사후층화 변수가 1개
  - 층내 모집단 값을 알고 있을때

- 사후층화법의 확장 :
  - 사후층화 변수가 2개 이상인 경우
  - 사후층의 주변 모집단 값만 알고 있을때

  ⇒ 가중값 조정 방법은?

  ⇒ 갈퀴비(raking ratio) 조정법 이용

예제 : 사후층화 변수가 2개인 경우

• 표본의 가중치의 합

	흑인	백인	아시안	아메리칸	기타	가중치합
여성	300	1200	60	30	30	1620
남성	150	1080	90	30	30	1380
가중치합	450	2280	150	60	60	3000

• 예 : 흑인이면서 여성인 표본의 가중치의 합 = 300

• 모집단 값

	흑인	백인	아시안	아메리칸	기타	가중치합
여성	?	?	?	?	?	1510
남성	?	?	?	?	?	1490
가중치합	600	2120	150	100	30	3000

• 사후층화법을 사용하기 위해서는 셀안의 모집단 값을 알아야함

• 갈퀴비 방법을 이용하여 셀안의 모집단 값을 구함

• 단계 1 : 성별을 모집단 값으로 조정

여성 × 1510/1620, 남성 × 1490/1380

	흑인	백인	아시안	아메리칸	기타	가중치합
여성	279.63	1118.52	55.93	27.96	27.96	1510
남성	161.96	1166.09	97.17	32.39	32.39	1490
가중치합	441.59	2284.61	153.10	60.35	60.30	3000

• 단계 2 : 인종을 모집단 값으로 조정

흑인 × 600/441.59, 백인 × 2120/2284.61 등

	흑인	백인	아시안	아메리칸	기타	가중치합
여성	379.94	1037.93	54.79	46.33	13.90	1532.90
남성	220.06	1082.07	95.21	53.67	16.10	1467.10
가중치합	600	2120	150	100	30	3000

• 단계 3 : 위 과정을 반복적용

• 단계 4 : 최종결과

	흑인	백인	아시안	아메리칸	기타	가중치합
여성	375.59	1021.47	53.72	45.56	13.67	1510
남성	224.41	1098.53	96.28	54.44	16.33	1490
가중치합	600	2120	150	100	30	3000

무응답 대체법 (Imputation)

- 대체법 : 결측치 자리에 다른값을 채워 넣는 방법
- 대체법을 쓰는 이유 :
  - 완비 데이터 셀을 만듦 : 통상적인 통계분석 프로그램 이용 용이
  - 무응답으로 인한 편향 감소

- 대체의 원칙 : 분포 유지
  - 응답 이전의 데이터 분포를 유지하도록 대체
  - 변수간의 관계를 유지하도록 대체
  - 오차를 줄이는 방향으로 대체를 하면 안 됨

대체 방법

- 결정적 대체 방법
  - 논리적 대체법
  - 최근방 대체법
  - 평균 대체법
  - 회귀 대체법
  - 시기적 대체법
  - 축차핫덱 대체법
  - 비 대체법
  - 적합값 대체법
- 확률적 대체 방법
  - 핫덱 대체법
  - 랜덤 회귀 대체법
  - 랜덤 비 대체법

예제

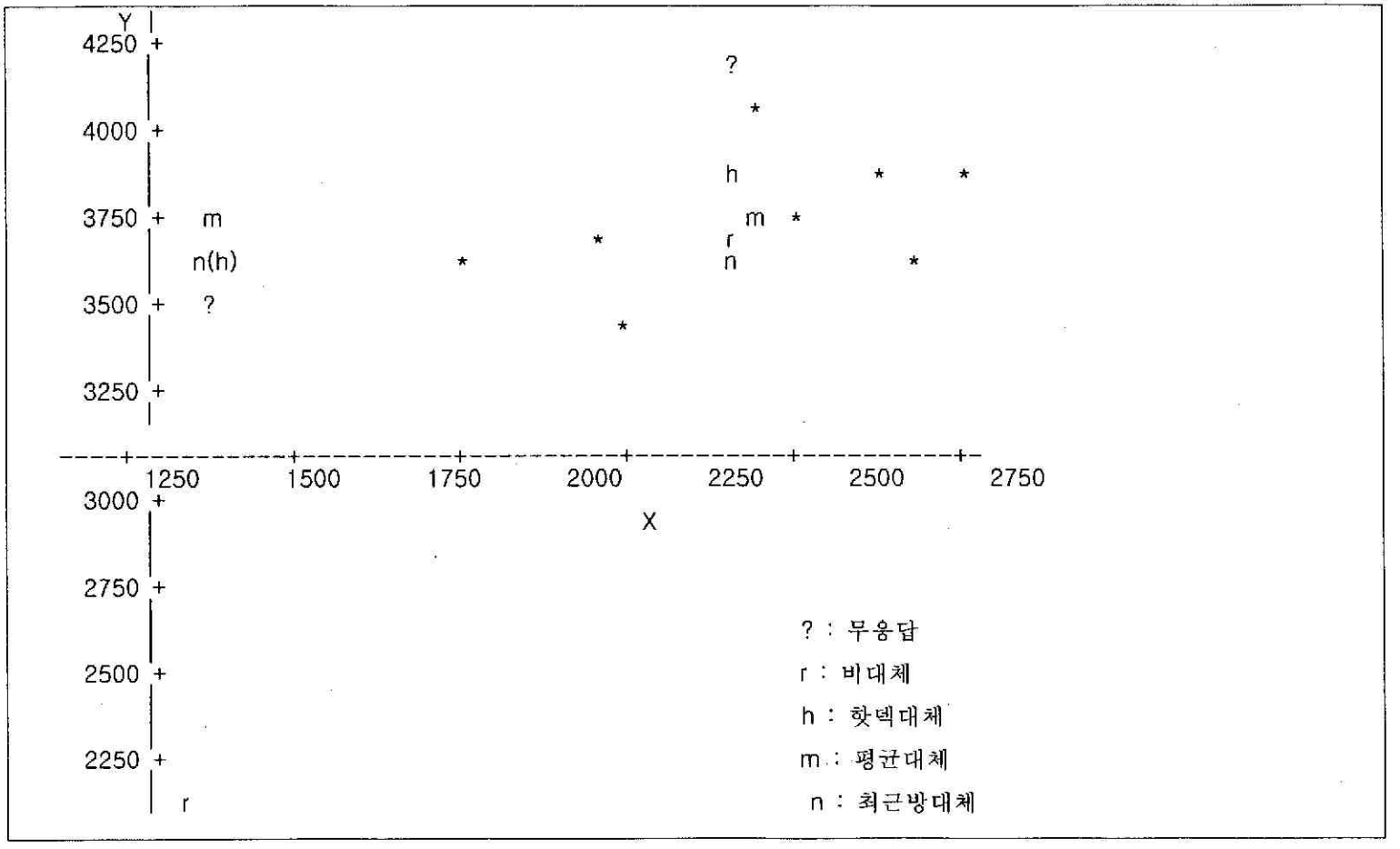
번호	X	Y	응답률	평균대체	비 대체	최근방대체	핫덱대체
			80%				
1	2092.50	3450	3450	3450	3450	3450	3450
2	1293.24	3500	?	* 3742.9	* 2087.6	* 3626	* 3626
3	2636.92	3598	3598	3598	3598	3598	3598
4	1791.74	3626	3626	3626	3626	3626	3626
5	2051.30	3676	3676	3676	3676	3676	3676
6	2357.96	3755	3755	3755	3755	3755	3755
7	2557.70	3890	3890	3890	3890	3890	3890
8	2730.16	3898	3898	3898	3898	3898	3898
9	2330.70	4050	4050	4050	4050	4050	4050
10	2298.95	4199	?	* 3742.9	* 3710.5	* 3676	* 3898

난수, 3, 7

대체 방법 비교

Variable	Mean	Std Dev
Y	3764.20	242.11
평균대체	3742.88	171.62
비 대체	3574.11	549.87
최근방대체	3724.50	176.33
핫덱 대체	3746.70	183.38

대체 방법 비교 (그림)


Y
4250
4000
3750
3500
3250
3000
2750
2500
2250
1250
1500
1750
2000
2250
2500
2750
X
?
h
m
r
n(h)
n
?
r
? : 무응답
r : 비대체
h : 핫덱대체
m : 평균대체
n : 최근방대체

무응답 대체후 추정치들의 비교

추정량	대체방법	응답률 70%	80%	90%	100%
표본평균	최근방대체	4633.97	4666.23	4604.00	4648.67
	평균대체	4579.38	4591.63	4593.67	4648.67
	비대체	4626.58	4667.34	4629.38	4648.67
	핫덱	4563.73	4601.93	4549.77	4648.67
상관계수	최근방대체	0.598	0.559	0.558	0.567
	평균대체	0.571	0.521	0.543	0.567
	비대체	0.651	0.585	0.580	0.567
	핫덱	0.566	0.504	0.500	0.567
회귀계수	최근방대체	1.900	1.910	1.880	1.896
	평균대체	1.892	1.903	1.888	1.896
	비대체	1.903	1.888	1.896	1.896
	핫덱	1.861	1.872	1.850	1.896

무응답 대체 효과

(1) 추정치의 구조적 편향

- 응답 메카니즘에 영향을 받음

⇒ 추정값의 신뢰도 저하

(2) 대체 후 표본평균의 분산은 증가

- 대체 분산이 포함됨

(3) 대체값을 응답값처럼 취급한 통상적인 분산추정량은 과소 추정

⇒ 통계 결과를 과대 평가함

⇒ 수정 필요 : 합리적 이용 방안

과소 분산 추정의 오류 (평균대체의 경우)

(1) 상대표본변이계수를 작게 평가 : $cv(\%) = \frac{s/\sqrt{n}}{\bar{x}} \times 100$

- 완전응답 : $cv = \frac{242.11/\sqrt{10}}{3764.20} \times 100 = 2.03\ \%$
- 대체 후 : $cv = \frac{171.62/\sqrt{10}}{3742.88} \times 100 = 1.48\ \%$

(2) 좁은 신뢰구간 제공 : $\bar{x} \pm t(n-1:\alpha/2) \times s/\sqrt{n}$

- 완전응답 : $3764.20 \pm 2.262 \times \frac{242.11}{\sqrt{10}} = [3591.0,\ 3937.4]$
- 대체 후 : $3742.88 \pm 2.262 \times \frac{171.62}{\sqrt{10}} = [3620.1,\ 3865.6]$: 71%

대체 후 표본평균의 성질

(1) 대체 후 표본평균의 오차 = 대체 오차 + 표본추출오차

$$\bar{y}_I - \bar{Y} = (\bar{y}_I - \bar{y}_s) + (\bar{y}_s - \bar{Y})$$

(2) 대체 후 표본평균의 분산 = 대체분산 +표본추출분산

$$Var\{\bar{y}_I - \bar{Y}\} \approx Var\{\bar{y}_I - \bar{y}_s\} + Var\{\bar{y}_s - \bar{Y}\}$$

(3) 통상적인 분산추정량 : 표본추출분산만 추정 $\Rightarrow$ 과수추정의 원인

모의실험 :

<표 2> 대체 후 추정량의 상대 편향 (B=10,000, n=20, R=70%)

응답패턴	모형	평균대체	비대체	회귀대체	핫덱대체
-0.7	1	-16.2239	-0.1858	-0.1818	-16.2713
	2	-12.6127	4.1577	0.0578	-12.4923
	3	-8.8885	4.4307	0.0006	-8.9590
0.0	1	0.0999	0.0429	0.0408	0.1712
	2	0.0167	0.2010	-0.0032	0.1498
	3	0.1027	0.5240	0.0335	0.0799
0.7	1	18.4046	0.0122	0.0128	18.5377
	2	12.7132	-2.7119	-0.0343	12.6876
	3	9.1951	-2.7188	0.0511	9.2899

대체 후 분산추정 방법

(1) 대체모형을 이용한 방법

- 표본추출분산과 대체분산을 분리해서 추정
- Sarndal(1992) : 비대체 모형 이용

(2) 반복표본을 이용한 방법

- 수정된 잭나이프 방법 : Rao&Shao(1992), Rao&Sitter(1995)
- 수정된 BHS 방법 : Shao, Chan & Chen (1998)
- 붓스트랩 방법 : Efron(1994), Shao & Sitter (1996)

(3) 다중대체방법 : Rubin (1987)

회귀대체모형을 이용한 분산 추정

(1) 회귀대체 모형

$$y_k = x_k'\beta + \varepsilon_k, \quad \varepsilon_k \sim (0, v_k\sigma^2), \quad k=1,\ldots,N$$

(2) 회귀 대체값

$$\hat{y}_k = x_k'(X_r' W_r^{-1} X_r)^{-1} X_r' W_r^{-1} y_r$$

(3) 회귀대체 후 표본평균의 비편향 분산추정량 :

$$v_G = \frac{s_I^2}{n} + \frac{1}{n(n-1)}\{(1_s'X_s + 1_r'X_r)(X_r' W_r^{-1} X_r)^{-1} X_{s-r}' 1_{s-r} - \sum_{k\in s-r} x_k'(X_r' W_r^{-1} X_r)^{-1} x_k\}\hat{\sigma}^2$$

수정된 잭나이프 방법

(1) Rao & Shao (1992)

(2) Idea : 잭나이프 추정치에 대체 효과를 반영

(3) 수정된 잭나이프 분산추정량 :

$$v_H = \frac{n-1}{n} \sum_{j=1}^{n} \{ \bar{y}_I^{\,a}(-j) - \bar{y}_I \}^2$$

여기서 $$\bar{y}_I^{\,a}(-j) = \begin{cases} \frac{1}{n-1} \{ n\,\bar{y}_I - y_j - \frac{n-r}{r-1}(y_j - \bar{y}_r) \} & (j \in r) \\ \frac{1}{n-1} \{ n\,\bar{y}_I - y_j^* \} & (j \in s-r) \end{cases}$$

모의실험

- **데이터 :** $y_k = \beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2} + \varepsilon_k, \quad k = 1, \ldots, n$

<표 3> 비수정 분산 추정량의 상대 편향(B=10,000, n=20, R=70%)

응답패턴	모형	평균대체	비대체	회귀대체	핫덱대체
-0.7	1	-50.2686	-32.9367	-30.7849	-42.7180
	2	-50.9822	-10.7623	-26.0959	-41.8826
	3	-50.7542	-2.3770	-17.8994	-41.4634
0.0	1	-51.5656	-23.6933	-22.1953	-42.7848
	2	-51.9143	-8.7327	-18.5003	-43.7404
	3	-51.7202	-0.1714	-13.1322	-43.3116
0.7	1	-52.4146	-13.4685	-12.7478	-43.8113
	2	-51.8376	-3.5693	-13.3433	-43.5980
	3	-50.8735	3.2900	-10.6101	-42.7631

<표 4 > 수정된 분산추정량의 상대 편향 (B=10,000, n=20, R=70%)

응답패턴	모형	v_C	v_R	v_{JR}	v_G	v_{JH}
-0.7	1	-24.445	2.389	0.549	1.342	3.671
	2	-26.952	30.609	3.911	0.150	4.715
	3	-28.122	79.687	7.519	1.983	5.849
0.0	1	0.260	1.332	0.528	1.169	3.441
	2	0.074	19.152	0.950	0.529	1.501
	3	-0.142	68.366	4.508	2.094	2.640
0.7	1	22.105	0.458	0.671	0.868	1.448
	2	20.526	11.981	2.675	0.715	1.962
	3	27.409	53.526	4.640	1.697	3.502

참고문헌

[1] 김규성 (2000). 무응답 대체 방법과 대체 효과. 조사연구, 1, 1-14.

[2] 김영원, 정선경 (1996). 표본조사에서 항목 무응답 대체 방법. 한국통계학회 논문집, 3, 145-159.

[3] 조사통계연구회 (2000). 무응답오차. 자유아카데미, 서울.

[4 Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. Journal of Official Statistics, 16, 113-131.

[5] Ford, B.L. (1983). An overview of hot-deck procedures. in W.G. Incomplete data in sample surveys, vol 2. New York, Acdemic Press, pp.185-207.

[6] Kim, Kyuseong. (2000). Variance estimation under regression imputation model. Proceeding of the Survey Research Methods Section, American Statistical Association.

[7] Kovar, J.G. and Whitridge, P.J. (1995). Imputation of business survey data. Business survey methods. Edited by Cox, B.G. et. al. pp.403-423.

[8] Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. Biometrika, 79, 811-822.

[9] Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581-592.

[10] Rubin, D.B. (1987). Multiple imputation for nonresponse in survey. Wiley, New York.

[11] Särndal, C.E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. Survey Methodology, 18, 241-252.

[12] Shao, J. and Sitter, R.R. (1996). Bootstrap for imputed survey data. Journal of American Statistical Association, 91, 1278-1288.

[13] Shao, J. Chen, Y., and Chen, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. Journal of American Statistical Association, 93, 819-831.

[14] Yung, W. and Rao, J.N.K. (2000). Jackknife variance estimation under imputation of estimators using poststratification information. Journal of the American Statistical Association, 95, 903-915.

통계에도 품질관리가 필요합니다.

박 성 현

(서울대학교 통계학과, parksh@plaza.snu.ac.kr)

통계에도 품질관리가 필요합니다

1. 통계품질관리란 무엇인가?

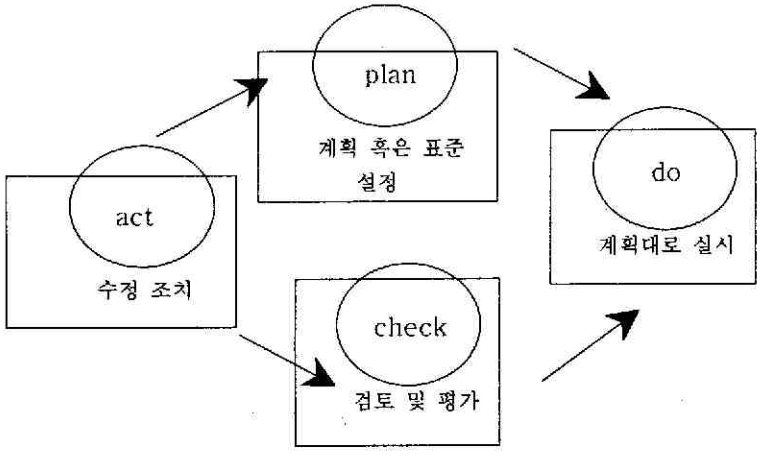
■ 품질관리와 통계품질관리

품질관리(Quality Control): 고객이 요구하는 품질을 확보·유지하기 위하여 조직이 품질목표를 세우고, 이것을 합리적이고도 경제적으로 달성할 수 있도록 PDCA 사이클에 따라 수행해 나가는 모든 활동.

통계 데이터 또한 통계작성기관의 입장에서 본다면 하나의 제품이며, 따라서 통계품질관리도 일반 상품의 품질관리와 같은 맥락에서 이루어져야 한다.

전통적 의미에서 품질 좋은 통계: 정확하고 신속한 통계.
새로운 통계품질의 의미: 통계가 이용자들에게 얼마나 사용 적합하도록 작성 및 제공되고 있는가의 사용품질
고객이 요구하는 품질을 확보·유지하기 위해서 새로운 개념의 통계품질관리가 요구되고 있음.

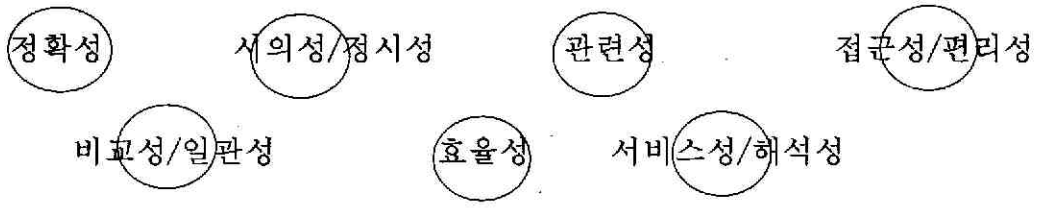
PDCA 사이클



■ **통계품질을 규정하는 요소들**

통계품질의 척도로 삼고 있는 요소들.

1) 정확성(accuracy) : 통계가 얼마나 정확한가를 나타내는 가장 기본적인 요소
2) 시의성/정시성(timeliness/punctuality) : 통계가 얼마나 신속하고 빈번하게 생산, 공표되는가 나타내는 요소
3) 관련성(relevance) : 자료 이용자에게 얼마나 의미 있고 유용한 통계를 작성하여 제공하고 있는가를 평가하는 요소
4) 접근성/편리성(accessibility/convenience) : 통계이용자들이 통계조사 결과를 얼마나 쉽게, 원하는 양식으로 이용할 수 있는가 하는 요소
5) 비교성/일관성(comparability/coherence) : 시간 또는 공간이 서로 다른 자료가 공통된 기준을 근거로 집계 또는 분석되어 서로 신뢰할 만한 비교가 가능한지를 평가하는 요소
6) 효율성(efficiency) : 통계작성기관이 통계자료를 얻는 데 사용된 비용에 비하여 통계 자료가 어느 정도 효율적으로 생산되었는가를 평가하는 요소
7) 서비스성/해석성(serviceability/interpretability) : 통계자료가 사용자들이 만족할 수 있도록 제공되며 서비스되고 있는가, 이용자가 자료를 쉽게 이해하고 활용하며 분석할 수 있는가를 평가하는 요소



2. 통계품질관리, 왜 필요한가?

■ **국가통계 생산 현황**

2003년 8월 현재 통계법 제8조(또는 제9조)에 의거 승인받은 통계는 총 446종으로서 지정통계 85종, 일반통계 361종이다. 작성방법별로는 조사통계 223종, 보고통계 173종, 가공통계는 50종이다. 작성기관을 정부기관과 지정기관으로 구분하여 볼 때 정부기관에 의하여 작성되고 있는 통계는 334종(통계청은 52종)이며, 지정기관에서 작성하고 있는 통계는 112종이다.

부문별로는 총 446종 중 보건사회복지(76종) 분야가 가장 높은 비중을 차지하고 있으며, 다음은 경기 • 기업경영(60종), 농림 • 수산(45종), 교통 • 정보통신(38종), 교육 • 문화 • 과학(25종), 고용 • 임금(26종), 인구(23종)의 순이다.

<표 1> 국가통계 작성 부문별 현황

분야	1986	1996	2003
인 구	10	21	23
고 용·임 금	24	18	26
물가·가계소비	22	17	15
보건·사회·복지		55	76
환 경		9	18
농 림·수 산	42	60	45
광공업·에너지	39	15	21
건설·주택·토지	45	22	25
교통·정보통신		37	38
도소매·서비스	15	5	9
경기·기업경영	20	40	60
국민계정·지역계정	9	7	11
재 정·금 융	16	21	14
무역·외환·국제수지		8	6
교육·문화·과학		16	35
기 타	99	21	24
총 계	341	372	446

■ **통계 오류의 사례(신문에서 발췌)**

보건복지부 의보예상적자 축소보고

감사원은 2001년 의약분업 특감 결과에서, 의약분업 실시에 따른 건강보험재정 파탄은 보건복지부가 정확한 예측이나 분석 없이 의약분업을 무리하게 추진하고 관련 통계를 축소 조작해 빚어진 것으로 결론지었다.

노동부의 취업자 집계 부풀리기

노동부 산하 기관인 전국의 고용안정센터가 취업자수 통계를 조직적으로 조작한 사실이 밝혀져 충격을 주었다. 고용안정센터의 취업자 통계는 통계청이 매월 발표하는 실업률과 취업률에 반영되는 것이어서 정보 통계의 신뢰성이 큰 훼손을 입었을 뿐 아니라 이를 기초로 각종 고용정책이 수립되었음을 생각하면 그 국가적 손실은 헤아리기 어려울 정도이다.

산업자원부 통계수치 왜곡

산업자원부는 2001년 8월 7일 전기요금 누진제로 민원이 폭주하자 "누진제를 적용받는 가구는 8.5%에 불과하며, 나머지의 모든 서민 가구는 누진제에서 제외된다"고 발표했다. 그러나 산업자원부 자료는 냉방기기 사용이 많은 여름 통계가 아니라, 연평균 전력사용량을 기준으로 한 것이었다. 이처럼 여론무마용으로 통계를 고의적으로 축소하여 잘못 사용하는 것은 국민을 기만하는 행위가 아닐 수 없다.

외교 통상부의 재외동포 인구통계 오류

외교통상부가 2년마다 집계해 발표하는 각국별 재외동포 인구통계가 탁상에서 주먹구구식으로 수치를 조정하는 등 오류가 심각한 것으로 드러났다. 이에 대해 외교통상부는 '현지실정 반영'이라는 주먹구구식 조정단계가 있었음을 인정했다.

부처마다 크게 차이 나는 자살통계

경찰청 통계에 따르면 자살 건수는 꾸준히 증가하는 추세이나, 통계청 자료에 따르면 3년간 거의 비슷한 수준이고 전체 자살건수도 경찰 집계치의 절반에 불과하다. 경찰청의 경우는 경찰관이 직접 나가서 자살, 타살 여부를 판단한 뒤 통계에 반영하고 있지만, 통계청은 의사가 자살로 진단한 사망진단서를 기준으로 하고 있기 때문에 생긴 통계상의 차이로, 집계상의 차이로 인한 통계오류의 예라고 할 수 있다.

(경찰청 : 99년, 2000년, 2001년, 2002년에 각각 11,713건, 11,794건, 12,277건, 13,055건으로 / 통계청 : 99년, 2000년, 2001년의 자살건수는 각각 7,075건, 6,460건, 6,933건-그림에 적용)

농림부 주요 채소작물 재배의향면적 표본조사결과 신뢰성 의문

농업관측정보센터와 농협에서 각각 주요 채소작물 재배의향면적 조사한 결과가 적지 않은 차이를 보였다. 이 같은 조사결과는 농림정책 수립에 활용되고 있는 점을 감안할 때 좀더 정확한 통계 수립을 위한 보완대책이 요구되고 있다.

(농협조사에서는 마늘은 18.9%증가할 것으로 예상됐으며, 조생양파와 당근, 양배추 등은 각각 22.9%, 7.3%, 14.1% 감소할 것으로 나타났다. 결국 당근을 제외한 나머지 작물에서는 양측 조사결과가 큰 차이를 보이고 있는 것이다.)

건설교통부 주택통계 오류

건설교통부가 매달 정기적으로 발표하는 주택건설 실적통계가 실제 건립된 주택 수와 큰 차이가 있어 정부통계의 신뢰성을 떨어뜨리고 있다. 특히 주택건설통계는 주택 수의 파악, 주택보급률의 산정, 그리고 이에 근거하여 주택수급 불균형 해소나 주택경기 활성화 정책 수립의 근거가 되기 때문에 통계가 잘못되면 정책적 혼선을 빚을 우려가 있다.

충북 수출통계 오류

2003년 충북지역 수출통계가 엉망이었던 것으로 나타났다. 충북 지역의 수출통계는 관세청 자료를 받아 청주세관이 발표하고 있으나 충북에 소재한 기업이 수출을 해놓고도 타지에서 수출한 것으로 잘못 계산되는 바람에 충북수출이 연속 4개월 두 자릿수 감소라는 발표가 나간 것으로 밝혀졌다.

■ 통계오류의 원인과 국가적 손실

통계를 조사, 작성하여 공표하는 데는 여러 가지 면에서 오류가 발생할 소지가 내재되어 있다. 통계는 국가기관에서 정책을 수립하고 집행하는 데 기초자료가 되며, 통계를 생산하는 데만도 많은 비용이 투여되는 것이기 때문에 통계의 오류는 곧 국가적 손실로 이어진다. 그러므로 오류의 원인이 어디에 있는지를 파악하여 수정, 보완하는 조치가 필요하다.

① 조사통계의 원천적인 문제점

조사통계에서는 모집단의 정의에서부터 조사결과의 분석에 이르기까지 모든 단계에서 오류가 발생할 소지가 있다. 따라서 각 단계에서 오류가 발생하지 않도록 시스템적인 장치를 마련하는 것이 중요하다. 통계품질관리 표준을 마련하려는 것도 이런 오류를 원천적으로 막아 보려는 것이다.

② 보고통계에 내재된 문제점

보고통계의 경우 보고자에게 유리하도록 통계가 조작되거나 의도적 제외, 허위응답 등으로 인한 오류가 발생할 소지가 있다. 통계정보가 투명하고 정직하게 제공되지 않으면 국제금융시장의 오판이나 혼란을 야기하는 것은 물론, 국가의 신인도에도 큰 영향을 미치게 된다. 지난 IMF 체제를 겪으면서 통계의 투명성 부족으로 인한 문제점을 인식할 수 있었다.

③ 기관의 업적지상주의 전시행정에서 비롯되는 오류

기관의 업적을 과대 포장해서 홍보용으로 통계를 이용하거나, 잘못된 행정을 감추거나 호도하는 방편으로 통계를 이용할 때 일어날 수 있다.

④ 개념 정의의 차이로 인한 오류

실업통계의 경우, 체감실업률과 국가통계의 수치가 차이가 나 종종 혼란을 초래한다. 이것은 실업이나 취업에 대한 ILO 기준과 OECD 기준이 서로 다른 데서 연유한 것으로, 이러한 개념의 차이가 오류를 낳게 되는 것이다.

⑤ 산업구조의 변화에 대처하지 못하는 데서 오는 오류

산업구조의 변화에 따라 서비스 산업의 비중이 커지고 디지털 경제로의 전환이 급속하게 이루어지고 있는 데 비해 그 변화에 적합한 통계생산이 이루어지지 않고 있어, 기존의 통계지표로는 이해가 안 되거나 오류의 범주에 속하는 것으로 판단되는 경우가 종종 일어난다.

■ **통계품질관리의 필요성**

① 정확한 통계는 국가경영에 필수적인 인프라이다
② 통계의 왜곡은 곧 정책의 왜곡이 된다
③ 정보화 사회는 정확한 통계에서 시작된다
④ 정확한 통계 데이터베이스를 구축할 필요성이 대두되고 있다
⑤ 통계는 이제 지방자치 시대의 필수 요소가 되고 있다

3. 통계품질 관리, 어떻게 하고 있나?

■ **외국의 통계품질관리**

IMF의 통계품질관리
IMF는 자료제공기준을 두어 통계품질을 관리하고 있다.

유럽국가의 통계품질관리
대부분의 유럽 국가들은 유럽 통계시스템의 개발 프로그램(European Statistical System: ESS)에 참여하고 있다. 1999년에 스웨덴 통계청에서 ESS의 통계품질을 높이기 위하여 품질 리더십그룹(Leadership Group on Quality: LEG)의 구성을 제안한 후에 대부분의 유럽 국가들이 이를 받아들였다.

OECD의 통계품질관리

2001년, OECD에서는 특별팀을 구성하여 'OECD 품질관리체계' 구축에 착수하였다.

Eurostat의 통계품질관리

유럽연합통계국(Eurostat)은 1994년 처음으로 품질평가제도를 도입하여 현재 Eurostat에는 통계품질을 관리하는 조직이 별도 구성되어 있으며, 품질에 대한 보고서 체계(일명 Quality Report)가 개발되어 있다.

캐나다의 통계품질관리

캐나다 통계청은 통계품질의 중요성을 일찍이 인식하고 1985년에 '캐나다 통계품질 가이드라인(Statistics Canada Quality Guidelines)'이란 책자를 발간하였다. 통계품질에 대한 공식적인 책자는 이 가이드라인이 최초의 책자이다. 이 외에도 캐나다 통계청은 인적자원의 효과적인 개발과 관리를 위한 프로그램을 운영하고 있으며, 통계품질을 보증하기 위한 기관 내의 절차들을 문서화하고 있다.

■ 우리나라의 통계품질관리

통계품질관리 연혁

1999년 4월, 통계기획국 기획과 내에 품질평가팀을 설치하고 통계품질관리 업무 시작

↓

2000년 12월, IMF와 공동으로 '통계품질평가 국제 세미나(Statistical Quality Seminar 2000)'를 개최하여 정부통계의 품질평가 방안에 관한 자료 수집

↓

2002년 7월, 품질평가팀을 청장 직속부서로 개편

↓

2002년 가을, 품질관리팀을 중심으로 12종의 통계(광공업동태조사, 경기종합지수 등)를 선정하여 자체 품질평가 실시

↓

2003년, 11종의 통계 자체 품질평가 실시

우리나라의 통계작성 8단계

조사 기획 → 모집단 및 표본 추출 → 조사표 작성 → 조사 직원 관리 → 조사 실시 → 자료처리 및 집계 → 자료공표 → 자료이용

4. 통계품질관리 표준 매뉴얼의 필요성 및 구성

통계는 국가 정책을 수립하고 집행하는 데 중요한 기초 자료가 될 뿐만 아니라 그 외의 여러 분야에서도 점점 더 많이 다양한 형태로 이용되고 있다. 그와 동시에 통계 수요는 점점 더 정확하고 신속하며 다양한 서비스를 요구하고 있다. 따라서 통계작성기관에서는 보다 나은 통계를 생산하기 위하여 좀더 철저한 통계품질관리를 할 필요가 있다. 그 한 방법으로 통계 작성 및 서비스에 대한 품질관리 방법을 표준화시켜 그 매뉴얼을 제공하고자 한다.

국가통계기관에서 작성하는 통계는 다음과 같이 크게 4가지 범주로 나누어 볼 수 있다.

① 표본조사통계
② 전수조사통계
③ 행정(보고)통계
④ 가공통계

표본조사통계 품질관리 매뉴얼

● 표본조사통계란 - 통계를 내고자 하는 전체 모집단 중 일부의 부분집단을 과학적인 방법으로 추출하여, 그 추출된 일부를 대상으로 조사를 하고, 그렇게 하여 얻어진 정보를 토대로 전체 모집단의 통계를 추정해 내는 것을 말한다.

● 전수조사통계란 - 조사대상이 되는 집단의 모든 개개의 단위를 조사하는 방법으로, 인구 및 주택 조사 센서스, 농업 총조사, 사업체 총조사 등이 그 대표적인 예이다.

● 표본조사통계 작성과정
조사기획→표본설계→조사표설계→조사원훈련 및 자료수집→조사결과 분석→이용자 서비스→사후관리→통계품질평가 및 관리 이다.

1장 조사 기획

표본조사통계에서 조사기획 단계는 이후 진행될 모든 작업의 뼈대를 잡는 가장 중요한 단계이다. 특히, 같은 조사라고 하더라도 조사목적이 무엇이냐에 따라 조사의 방향이나 규모, 방법 등이 달라질 수 있으므로 명확하고 구체적인 조사목적을 세우는 것이 매우 중요하다.

● 조사목적을 세울 때 유의점

* 주요 이용자 및 관련 분야 전문가들의 통계 수요를 분석하여 조사목적을 정한다.
* 관련 있는 유사통계들을 분석한다.
* 통계수요를 어느 정도 충족시킬 것인지에 대한 한계를 명확히 한다.
* 조사목적을 주기적으로 검토하여 개선한다.
* 명확한 품질목표가 있을 경우 이를 구체화하여 조사목적 명세서에 포함시킨다.
* 본 조사 데이터를 이용하여 생산될 가공통계가 있으면 함께 고려하여 목적을 세운다.

● 조사의 구조와 규모 결정

* 주간조사인지, 연간조사인지, 혹은 일회조사인지 등 조사의 구조를 결정한다.
* 법률규정, 조사에 필요한 인원의 확보, 소요예산 등 기본적인 제약조건들을 명확히 한다.
* 다음과 같은 전체적인 조사의 규모, 조사사항, 조사방법 등을 결정한다
  - 조사단위는 가구인가 아니면 개개인인가?
  - 응답자가 사실 그대로 쉽게 이해해서 객관적으로 응답할 수 있는 사항인가
  - 대상의 전부를 조사할 것인가 표본만 추출할 것인가
  - 타계신 조사인가 자계식 조사인가
    *타계식 조사- 면접조사, 전화조사
    *자계식 조사 -배포조사, 우편조사, 인터넷조사
  - 언제 시작해서 얼마나 오랫동안 실시할 것인가.

* 조사예산을 확보한다.
* 조사 실행을 위한 시스템을 구축한다.
* 조사기획을 완료하기 전에 관련 전문가의 자문을 받는다.
* 계속조사의 경우 주기적으로 조사를 재설계할 계획을 세운다.

2장 표본설계

표본설계란 모집단을 잘 대표할 수 있는 표본을 추출하고 추출된 표본에서 조사된 정보를 이용하여 모집단의 특성치를 추정하는 전 과정을 말한다. 주어진 여건 하에서 가장 경제적이고 정확성이 높으며 효율적인 표본을 설계하는 것이 표본설계의 목표이다.

● 모집단

* 정의 - 조사목적에 의하여 규정되는 모든 조사단위의 집단.
* 모집단의 구분 : 목표모집단, 조사모집단
  - 목표모집단(target population) : 조사목적에 의해 개념상 규정된 모집단
  - 조사모집단(sampled population) - 표본추출을 위해 규정된 모집단
    ==> 목표모집단과 조사모집단의 차이를 최소화한다.

* 모집단에 포함되는 조사단위는 조사원이 자의적으로 판단(정의)할 위험이 없도록 명확하게 정의되어야 한다.

예) 대도시에 거주하는 성인을 조사하려고 할 때 : 호적에 등재된 생년월일이 1983년 6월 31일 이전이고 제주도를 제외한 인구 10만명 이상의 대도시에 거주하는 성인.

● 추출틀 (sampling frame)

* 모집단 내의 모든 추출단위들의 리스트를 말한다.
* 모집단의 모든 추출단위를 누락 없이 그리고 중복 없이 포함해야 한다.
* 추출틀은 조사모집단의 구체적 표현이다.
* 추출틀이 불완전하면 왜곡된 통계가 작성된다.
* 다양한 추출틀을 비교한 후 가장 적합한 추출틀을 정한다.
* 일관성 유지를 위해 동일 모집단에 대한 조사는 가능한 동일한 추출틀을 사용한다.
* 추출틀의 포함범위를 주기적으로 평가하고 보정한다.
* 추출틀의 품질을 유지, 향상시키기 위한 시스템을 갖춘다.
* 조사 관련 보조정보를 잘 갖춘 추출틀을 마련한다.
* 필요에 따라서는 복수의 추출틀을 활용할 수 있다.

우리 나라 가구당 평균소득을 조사하고자 하는 경우에 모집단은 우리 나라 전체 가구이고, 조사의 기본단위는 가구가 된다. 만약 추출단위를 가구로 결정하였

다면 우리 나라 전체 가구에 대한 방대한 양의 리스트가 필요하지만 추출단위를 동(洞)으로 한다면 전국의 전체 동에 대한 목록을 만들면 되기 때문에 훨씬 수월한 작업이다.

● **표본설계시 고려해야 할 내용들**

* 적은 비용과 노력으로 모집단을 잘 반영할 수 있는 표본추출법을 정한다.
  → 추출단위에 대하여 사전에 정해진 추출확률에 따라 표본을 추출하는 **확률추출법** 사용
* 표본오차를 목표하는 수준이내로 유지하면서 비용을 최소화하는 표본 크기 설정.
* 변수가 여러 가지인 표본의 크기를 결정할 때는 가장 중요한 몇 개를 선정한 뒤 이를 만족시킬 수 있는 표본의 크기를 결정하는 것이 일반적이다.

● **모집단 층화**

* 층화- 모집단을 특성에 따라 서로 동질적인 몇 개의 부분집단으로 나누는 과정
* 층화를 통해 기대할 수 있는 효과
  ① 단순무작위추출법에 비해 추정의 정도를 높일 수 있다.
  ② 전체 모집단에 결과뿐만 아니라 각 층별로도 추정도 가능하다.
  ③ 조사관리가 보다 편리하고 조사비용도 절감할 수 있다.
* 효과적인 층화를 위해 고려해야 할 사항

① 적절한 층화변수를 결정해야 한다.

여론조사에서는 일반적으로 지역, 성별, 연령, 학력 등이 중요한 층화변수로 쓰이고 있으며, 서울시내 주택가격에 대한 조사라면 지역(강남, 강북), 주택유형(단독주택, 아파트, 연립주택 등의 구분), 주택면적 등이 중요한 층화변수가 될 것이다.

② 표본의 크기를 결정하고 효율적인 표본배분법에 따라 각 층에 표본을 배분하여야 한다.

③ 층화를 할 때 미리 층의 수를 제한하지 말고 가능한 모든 경우를 다 나눈 후 역으로 합쳐가면서 적절한 층의 개수를 정하는 것이 좋다.

● **표본추출법을 결정할 때 고려해야 할 사항들**

* 단순한 추출법을 사용하는 것이 좋다. 추출법이 복잡해질수록 추정식이나 관리가 까다로워진다.
* 조사비용과 시간, 용이성 등을 고려하여 표본추출이 이루어져야 한다.

* 자체가중설계(self-weighting design)가 되도록 표본을 추출하는 것이 좋다.
  자체가중설계-모집단에 속하는 최종추출단위(ultimate sampling unit)들의 추출확률을 동일하게 하는 방법
* 조사가 주기적으로 반복되는 계속조사일 때에는 표본크기의 변화, 재층화, 추출확률의 수정 등이 가능하도록 융통성있게 설계한다.
* 계속조사의 경우 표본의 품질을 지속적으로 모니터 할 수 있는 절차를 개발한다.

모집단을 제대로 대표하지 못하는 표본을 사용할 경우 잘못된 통계를 만들게 된다.

1936년 미국 대통령 선거에 대한 여론조사 결과를 참조 "

==> 랜돈 후보(상류층이 지지)와 루즈벨트 후보(서민층이 지지)의 대결

.....전화번호부와 자동차 등록대장을 사용하여 표본추출된 200만 명에 대하여 우편조사

==> 랜돈 후보의 압도적 승리 예상 (잘못된 예측)

실제 결과 : 루즈벨트 후보의 압도적 승리

※ 표본추출을 위하여 전화번호부와 자동차 등록대장을 사용하였기 때문에 표본이 모집단 내의 서민층을 반영하지 못하여 나타난 결과이다.

3장 조사표 설계

조사 설계에서 가장 중요한 측면 중 하나로 수집되어야 할 데이터가 무엇인지를 명확하게 규명하고 그 데이터를 얻을 수 있는 도구를 개발하는 과정이다.

● 유의 사항

* '어떤 데이터를 얻을 것인가' 하는 문제에 대해 통계 사용그룹과 생산그룹 사이의 충분한 협의가 있어야 한다.
* 객관적으로 조사문항을 평가할 수 있는 외부의 전문가를 확보하여 조사문항을 자문을 받는다.
* 누가 보아도 동일한 개념을 가질 수 있도록 용어를 명확하게 정의하고. 가급적 사용자들의 요구를 반영한 용어와 개념를 사용한다.
* 통계청의 표준분류와 그 밖에 UN 통계처, ILO, Eurostat, ISO 등 국제 기구에서 정한 국제표준분류들의 분류체계를 이용한다.
* 원 데이터를 코드화하고 가능한 한 가장 하위 분류의 자료까지 생산한다.

● **단계1> 조사질문 작성 및 구성**

● **단계2> 조사표 양식 (layout), 인쇄된 조사표를 만드는 단계 .**

● **단계3> 조사표 시험**

* 예비조사(pilot survey)를 한다.
  ① 예비조사 - 데이터의 수집, 처리, 분석을 망라하는 조사의 전 단계를 축소된 규모로 시행하는 것. 따라서 가능한 한 실제 조사환경과 같은 환경에서 실시되어야 함.
  ② 예비조사에서 확인해야 할 사항.
  조사표 상의 조사사항, 문항들의 배열 등이 타당한지 검토한다.
  설계된 조사표와 조사방법의 타당성을 검토한다.
  응답률, 응답거부율, 조사소요 시간 등을 파악한다.
  조사원 훈련방법의 적합성을 검토한다.

4장. 조사원 훈련

조사의 질을 담보하는데 있어서 핵심 요소는 바로 조사를 수행하는 인력의 훈련이라고 할 수 있다. 자질을 갖춘 조사인력은 하루아침에 생길 수 있는 것이 아니므로 지속적이고 장기적인 노력이 필요하다.

● **조사의 수행을 위해 필요한 인력**

* 조사전문가
* 조사감독원
* 조사원

● **조사감독원 훈련**

* 동일한 조사수준을 유지하기 위한 훈련프로그램을 중앙 차원에서 마련한다.
* 훈련과정에 이론과 실무를 결합시켜 추출틀 마련, 표본선정, 실제 조사업무, 자료입력 등의 전 과정을 이해하고 경험하게 하는 과정이 있어야 한다.
* 조사감독원에 대한 상시 훈련 프로그램을 마련한다. 특히 감독능력, 품질관리능력과 관련한 훈련이 필요하다.
* 새로운 조사감독원에 대해서는 특별한 관심을 가지고 다양한 정보들을 전수한다.

● **조사원 훈련**

* 전반적인 조사목적과 조사개요를 숙지시킨다.
* 조사표의 모든 조사문항별 조사지침을 상세히 가르쳐 준다.
* 일반적인 조사기술을 숙지시킨다.
* 조사원이 판단해서 처리해야 할 사항과 조사감독원의 판단을 의뢰해야 할 사항을 명확하게 구분한다.
* 조사원들이 조사품질 의식을 갖도록 돕는다.
* 조사를 위해 필요한 조사원수 외에 예비로 조사원을 확보한다.
* 계속조사일 경우 조사원들에 대해 정기적 평가와 재교육의 기회를 제공한다.

5장. 자료수집

좋은 데이터를 수집할 수 있어야만 효과적인 통계 정보의 생산이 가능하다. 자료수집 활동이 최종 통계품질에 미치는 영향을 감안할 때 자료수집의 각 단계마다 신뢰성을 확보할 수 있는 조치를 마련하는 것이 필요하다.

● **자료수집 일반 원칙**

* 조사원들에 대한 감독과 지원, 조사 과정에서 일어날 수 있는 다양한 상황들에 대한 대비 지침 등 체계적인 현장조사 관리 시스템을 마련한다.
* 조사원의 주관에 의해 조사대상이 정해지지 않도록 명확한 조사명부를 마련하고 조사대상의 선정규칙을 명확히 한다.

● **무응답 대책**

* 조사의 전 과정에서 응답률을 극대화시킬 수 있는 방안을 마련한다.
  일반적으로 응답률에 큰 영향을 미치는 요소로는 조사방법, 조사원의 능력, 조사원의 업무량, 조사주제, 응답부담, 조사표의 길이와 복잡성, 응답자 인센티브 등이 있다.
* 가능하다면 무응답에 대해 재조사(callback)를 실시한다.
* 무응답의 원인을 기록하고 모니터 한다.
* 무응답 데이터에 대해 가중값 조정 또는 대체(imputation) 등 적절한 조처를 취한다.
  대체를 하는 경우 데이터 세트에서 대체값 여부를 나타내는 표시(flag)를 반드시 해주어야 한다.
* 모든 조사단위를 응답과 무응답으로 분류하여 표시하고 각 조사의 응답률을 공표하여 조사가 지니는 한계를 밝힌다.
* 무응답에 관한 정보들을 축적하고 체계적인 연구를 한다.
  응답자와 무응답자 사이의 특성 차이 등이 밝혀지면 무응답으로 인한 편향 등을 추측하는데 큰 도움이 된다.

● **데이터 사전점검 (data preparation)**

조사에 의해 수집된 데이터를 분석 가능한 형태의 데이터베이스로 만드는 과정.

* 코딩 체계를 표준화한다.
* 입력오류를 최소화시킬 수 있는 입력시스템을 활용한다. 입력자는 데이터를 입력시킨 후 반드시 입력오류가 없는 지를 점검하여야 한다.
* 자료점검 과정을 위한 품질관리 기법을 도입한다.
* 수집된 자료와 자료전송 과정에 외부인이 접근할 수 없도록 데이터의 전송과 취급 과정에서의 보안성을 확보한다.

● **데이터처리 (data processing)**

입력된 데이터파일에 대해 데이터편집(data editing), 대체(imputation) 등을 실시하는 과정.

(1) 데이터편집(editing)

* 정의 - 조사된 데이터 중 오류가능성이 높은 응답값을 찾아내어 점검하는 과정.
  (예:연령이 13세인데 혼인상태가 기혼으로 응답되었다든지, 출생년도가 1980년인데 현재 상태가 중학생으로 응답되었다든지 하는 경우)
* 가능한 한 자동적으로 보완할 수 있도록 하는 것이 바람직하다.
* 현지와 중앙, 두 단계에 걸쳐 데이터편집을 한다. 일차적으로 데이터를 수집한 현장의 기관에서 기본적인 데이터편집을 실시하고 최종적으로 중앙에서 다시 데이터편집을 하는 것이 바람직하다.
* 가능한 한 수작업보다 컴퓨터 데이터편집 프로그램을 마련하여 작업한다.

*** 데이터편집을 위한 체크리스트.**

① 외양 편집 -조사변수에 대해 식별 가능한 코드 체계가 정확히 부여되어 있는지 여부.
② 구조 편집- 자료로 인정할지, 어떤 요소를 유효한 것으로 인정할지 확인.
③ 범위 점검 -유효한 응답값의 범위를 벗어나는 데이터가 있는지를 점검.
④ 누락 점검 -응답자가 답해야 하는 문항인데 건너 뛴 문항이 없는지 아니면 응답해서는 안되는 항목인데 응답한 것이 없는 지를 점검
⑤ 일관성 점검 - 응답값의 일관성을 점검하여 응답 데이터의 질을 체크.

*** 오류가 발견될 경우 대처 방법**

① 단위의 응답값 전체가 믿을 수 없는 경우에는 데이터를 제거한다.

② 다른 보조정보나 다른 항목의 응답값에 의해 오류가 명확하게 고쳐질 수 있을 경우에는 수정한다.

③ 수정이 적절치 않은 항목의 오류값은 무응답 처리한다.

④ 대체(imputation)가 적절하다고 판단될 때에는 대체 지침에 따라 대체한다.

(2) 대체 (imputation)

* 대체 - 무응답을 그럴싸한 값으로 대체시키는 작업.

*** 대체의 기본원칙**

① 특정 데이터를 대체할 경우 이 값이 대체된 값인지 관찰된 값인지를 명확히 밝혀 두어야 한다.

② 일선 조사현장에서 각각 대체를 하도록 한다면 방법이나 기준이 달라질 우려가 있으므로 조사자료를 수집한 후 중앙에서 무응답 대체를 하는 것이 바람직하다.

③ 무응답 변수와 상관이 높은 보조정보들을 이용하여 조사단위들을 몇 개의 대체층으로 구분한 후 각 대체층 내의 응답값들을 대체자료제공세트로 활용한다.

④ 가장 적절한 대체방법을 채택한 후 채택된 대체방법을 명확히 밝힌다.

⑤ 대체법을 사용했을 경우 이를 반영한 추정식을 마련한다.

6장. 조사결과 분석

대규모 조사인 경우 조사 데이터에 기초하여 생산되는 통계의 양은 매우 방대하다. 각각의 조사항목별로 전체 통계뿐 아니라 지역별, 층별, 관심영역별로 세분된 통계 또한 생산된다.

따라서 조사결과를 분석하는 단계에서는 각종 일치성의 점검 및 정확성 확인에 세심한 주의를 기울여야 한다.

- **단계1> 통계표 작성 (tabulation)**
- **단계2> 추정 (estimation)**
- **단계 3> 일치성(contingency) 점검**
- **단계 4> 이차적 통계분석**

7장 이용자 서비스

많은 비용을 들여 실시한 조사를 통해 얻은 정보를 최대한 효율적으로 활용한다는 측면에서 담당자들은 이용자들의 다양한 요구에 지속적으로 귀를 기울여야 하며, 통계정보를 효과적으로 활용할 수 있도록 다양한 수준의 자료를 제공하기 위해 끊임없이 노력해야 한다.

● 문서화 (documentation)

* 정의 - 통계의 생산을 위해 사용된 모든 통계적 활동을 기록하는 과정이다.
* **문서에 포함되어 되는 내용**
  - 조사개요 : 조사의 특성을 전반적으로 파악할 수 있는 내용, 주의사항 등.
  - 통계분석결과 : 다양한 통계분석을 통해 나타난 주요 결과와 유용한 정보.
  - 통계표 : 조사를 통해 생산하고자 하는 모든 통계를 망라하여 작성한 것.
  - 부록 : 본문에서 다루지 못한 참고 사항.
* **문서화를 할 때 유의점**
  -다양한 독자층을 고려하여 품질 좋고 읽기에 편한 문서를 작성할 것.
  -미리 공표한 일정에 맞출 것.
  -책자, 인터넷, CD-Rom 등 다양한 매체 형태의 문서를 제공할 것.

● 데이터 제공

* 가능한 한 원데이터가 담고 있는 정보를 그대로 제공한다.
* 데이터의 응답자의 식별이 불가능한 형태의 데이터를 제공한다.
* 응답자 신분공개의 위험을 방지하기 위한 방법들을 사용한다.
* 원자료를 포함하여 데이터를 제공할 때에는 사용방법을 자세히 알려준다.
* 실제 응답 데이터에 의해 계산된 통계와 제공된 데이터에 의해 계산된 통계는 반드시 일치해야 한다.
* 파일로 저장한 제공 데이터가 원래 의도대로 실행되는지 확인한다.
* 데이터를 제공할 때 데이터의 품질관련 자료도 함께 제공한다.
* 데이터를 제공할 때 담당자의 연락처를 함께 알린다.
* 제공되는 데이터와 관계되는 모든 사항에 대한 질의 응답 체계를 구축한다.
* 검색 엔진을 통해 이용자가 원하는 통계 정보를 쉽게 검색할 수 있도록 한다.

8장 사후관리

일회성 조사가 아닌 계속 조사인 경우 처음 조사를 기획했던 때의 개념이나 품질, 수준이 시간이 흘러도 계속 유지될 수 있도록 체계적으로 관리되어야 한다.

방대한 양의 데이터와 작성된 통계들을 어떻게, 어떤 수준으로 보관, 관리할 것인지를 고려해야 하는 데이터 베이스 관리부분도 사후관리의 측면에서 고려되어야 할 중요한 부분이다.

● **조사시스템 구축**

● **추출틀 관리 및 표본 관리**

● **데이터베이스 관리**

9장 통계품질평가 및 관리

● **통계의 품질 평가**

* 기획단계에서부터 통계 품질의 평가를 미리 염두에 두고 예산과 인력 배정을 한다.
* 통계를 공표하기 전에 과거 통계와 일치하는지, 논리적으로는 문제가 없는지 등을 평가한다.
* 포함률, 응당률, 표본오차 등의 품질 정보를 산출하여 제공한다.
* 품질평가를 적절한 주기 마다 실시하고 그 결과를 시의 적절하게 알린다.
* 객관적이고 계량화된 통계 품질의 평가가 불가능한 경우 질적, 주관적 품질 평가를 할 수 있다.
* 조사 기관 내외부의 이용자를 통계 품질평가 목표 설정, 평가 과정에 참여시킨다.

● **통계품질시스템의 구축**

* 통계품질관리가 일시적이 아니라 항구적으로 이루어지도록 하기 위해 품질 조직을 마련한다.
* 모든 구성원들의 적극적 참여를 유도한다.
* 조사의 각 단계별로 체계적인 품질관리를 할 수 있도록 품질매뉴얼을 마련한다.

● **이용자 만족도조사**

*** 통계 이용자 만족도 조사를 위한 고려사항**

- 이용자 만족도조사를 위한 조사목적을 세운다.
- 조사목적, 조사여건 등을 감안하여 조사대상을 정한다.
- 이용자 데이터베이스를 구축한다.
- 만족도조사 결과에 대해 적절한 피드백을 할 것.

*** 이용자조사 내용에 포함될 사항**

① 응답자 특성
② 통계 이용실태 : 용도, 입수경로, 선호매체, 이용 빈도, 이용 자료 등
③ 통계 및 이용자 서비스에 대한 전반적 만족도
④ 정확성, 시의성, 적합성, 접근 용이성 등 통계품질의 여러 측면에서의 만족도
⑤ 해당 만족도조사 대상 통계에 국한된 이슈에 대한 사항
⑥ 통계의 문제점이나 통계관련 제안 등을 기록할 수 있는 문항

● **품질평가 피드백**

* 통계 작성기관의 관련 담당자들을 위한 통계 품질평가 보고서를 작성한다.
* 축적된 품질관련 자료를 근거로 품질향상을 위한 연구를 한다.

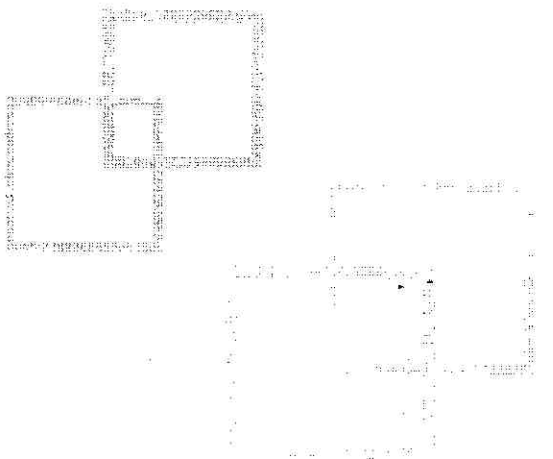
『통계품질관리 워크샵』 4교시

현장조사 관리 기법

이 미 경

(리서치플러스, mklee@researchplus.co.kr)

현장조사 관리 기법



현장조사 관리 기법

"현장조사" 한마디로 소비자의 소리를 듣는것이다

현장조사과정은 지금까지 여론조사의 전과정에서 가장 취약한 부분으로 알고 있다.
최근에는 설문조사에서 면접원의 역할이 점점 줄어들고 있는 경향에도 불구하고
조사과정에서 면접원이 차지하는 비율은 75%에 해당하는 중요한 역할을 담당하고 있다.
면접원들을 아무리 열심히 훈련시키고, 관리한다고 하더라도 실사의 통제를 벗어난 부분들이 있기 때문이다.
이를 조금이라도 보완하기 위해 현장에서 일어나는 사례로 알아보고저 한다

현장조사 관리 기법

프로젝트 진행과정

프로젝트 미팅 → 질문지 사전조사 → 실사준비 → 자료수집 → 자료입력

자료수집단계를 다시 세분화 하면,

면접원 모집 → 면접원 교육 → Role Playing → 사전조사(Pre-test) → 현장조사

설문지 회수 → Editing → 전화검증(통상30%) → 코딩 → 편칭

현장조사 관리 기법

자료수집단계를 다시 세분화 하면,

1. 면접원 모집

➢ 응답자와 직접 접촉하여 1차 자료를 수집하기 때문에, 상업적 마케팅조사에서 필수적이고 중요한 수단이 되며, 이들이 수행하는 면접의 효율성과 신뢰성에 따라 조사의 전체적인 구조와 조사결과의 정확성에 큰 영향을 미치기 때문에 면접원 선발에 신경을 쓴다. 90년대 이전만 하더라도 대부분 대학생들이 학업과 병행하여 아르바이트로 하였는데, 점차 조사만 전문으로 하는 주부들이 늘어나서 방학기간을 제외하고 95%이상이 주부들이 하고 있다.
(면접원의 종류는 일반면접원, 전속면접원, 프로젝트 단위 계약 면접원이 있다.)

2. 면접원 교육

➢ 연구자가 갖고 있는 의도를 공유하기 위해 교육, 훈련, 체계적인 관리는 아무리 강조해도 지나치지 않다

3. Role Playing

➢ 면접원의 미숙, 질문지 이해부족으로 인한 오차의 최소화를 위해 면접원끼리 역할을 바꿔해 보는 Role playing을 실시한다. (전속면접원은 제외되는 경우가 있다)

4. 사전조사

➢ 한 면접원이 2-3부씩 배부받아 주변의 사람에게 조사를 하여 한 면접원이 갖고 있는 바이어스를 배제시킨다.

5. 현장조사

➢ 현장조사를 조사방법별로 다시 세분화 하면,

개별면접(Personal Interviews) -> 우편조사 및 온라인 조사->전화면접->Gang Survey
CLT(Central Location Test) -> HUT(Home Use Test)->FGI(Focus Group Interview)

현장조사 관리 기법

5. 현장조사

➢ 현장조사를 조사방법별로 다시 세분화 하면,

개별면접(Personal Interviews) -> 우편조사 및 온라인 조사->전화면접->Gang Survey
CLT(Central Location Test) -> HUT(Home Use Test)->FGI(Focus Group Interview)

1) 개별면접
(Personal Interviews)

- 통반 즉 블록을 정해주고 그 블럭내의 가구원 모두에게 조사하는 방법
(정통부의 "정보화실태조사"나 노동부의 "청년패널")
현장에서의 문제점 및 장단점 2-3분정도 설명
-통반 즉 블록을 정해주고 그 블럭내의 가구원중 성인에 해당되는 사람중 생일이 가장 빠른 또는 늦은 사람과 조사하는 방법 (담배회사의 "담배흡연율"조사)
현장에서의 문제점 및 장단점 2-3분정도 설명
-동을 정해주고 그 동에서 연령을 정해주는 방법 (가장 일반적인 방법)
현장에서의 문제점 및 장단점 2-3분정도 설명
-면접원 주변에서 아는 사람을 찾아 조사하는 방법 (사용자를 찾아 하는 조사)
현장에서의 문제점 및 장단점 2-3분정도 설명
-Syndicated Research
: 담배소매점조사, 화장품 소매점조사, 식품 소매점 조사등
현장에서의 문제점 및 장단점 2-3분정도 설명
-Consumer Panel
: 일기장 패널이라고도 하며, 주간 또는 월간단위의 청취율, 시청률, 제품구입등을 알고저 할때 조사하는 방법
현장에서의 문제점 및 장단점 2-3분정도 설명

현장조사 관리 기법

5. 현장조사

➢ 현장조사를 조사방법별로 다시 세분화 하면,

개별면접(Personal Interviews) -> 우편조사 및 온라인 조사->전화면접->Gang Survey
CLT(Central Location Test) -> HUT(Home Use Test)->FGI(Focus Group Interview)

2) 우편조사 및 온라인 조사

설문지를 안내문, 회신용봉투등을 보내 우편 발송하는 조사와 응답자의 양해하에 이메일 주소로 보내주는 조사 방법 **현장에서의 문제점 및 장단점 2-3분정도 설명**

3) 전화면접

- 전화번호부를 이용한 랜덤조사 : 일반 여론조사
- 제공된 리스트를 이용하여 전화하는 방법
- CATI (Computer Aided Telephone Interviw)
: 면접원이 컴퓨터 모니터에 나타난 질문을 보고 면접하는 시스템으로, 조사가 진행되는 도중 누적된 데이터를 볼수있을뿐만 아니라, 전화번호를 무작위로 걸어주는것도 가능하다.

현장에서의 문제점 및 장단점 2-3분정도 설명

4) Gang Survey

개별면접의 문제점을 보완하는 조사 방법으로 응답자들을 정해진 시간에 일정한 장소로 모이게 한 후 조사 담당자가 응답자들로 부터 자료를 수집하는 기법이다.
조사 담당자가 직접 신제품 또는 광고카피 같은 보조물을 이용하여 조사 목적에 대한 상세한 설명을 하며 자료수집과정을 통제할수 있고, 조사 과정이 외부에 유출되는것을 방지 할수 있다 (냉장고, 에어컨, 자동차 조사등 신제품 테스트, 화장품, 담배등
(냉장고, 에어컨, 자동차 조사등 신제품 테스트, 화장품 담배등 사용 테스트에 이용)

현장에서의 문제점 및 장단점 2-3분정도 설명

현장조사 관리 기법

5. 현장조사

➢ 현장조사를 조사방법별로 다시 세분화 하면,

개별면접(Personal Interviews) -> 우편조사 및 온라인 조사->전화면접->Gang Survey
CLT(Central Location Test) -> HUT(Home Use Test)->FGI(Focus Group Interview)

5) CLT(Central Location Test)

특정 장소를 지정하여 그 곳을 지나가는 행인중에 조사 자격 요건에 해당되는 사람을 선정하여 면접하는 방법이다. (술, 음료,신제품의 맛테스트)
현장에서의 문제점 및 장단점 2-3분정도 설명

6) HUT (Home Use Test)

가정내 응답자가 실제 상황 하에서 제품을 7-10일 정도 사용기간을 갖고 소비자 반응을 조사 하는 조사 방법이다. (커피, 칫솔,고추장등 신제품을 경쟁사와 비교조사할때 사용)
현장에서의 문제점 및 장단점 2-3분정도 설명

7) FGI(Focus Group Interview)

가정내 응답자가 실제 상황 하에서 제품을 7-10일 정도 사용기간을 갖고 소비자 반응을 조사 하는 조사 방법이다. (커피, 칫솔,고추장등 신제품을 경쟁사와 비교조사할때 사용)
현장에서의 문제점 및 장단점 2-3분정도 설명