

켈러의 경영경제통계학

제4장

데이터 수집과 표본추출

Data Collection and Sampling

4.1 데이터 수집방법...

```
graph LR; A[데이터] --> B[통계학]; B --> C[정보];
```

데이터

통계학

정보

- 통계학은 데이터를 정보로 전환시키는 방법론이다.
- (데이터와 관련된 질문들) 데이터는 어디서 구해지나? 데이터는 어떻게 수집되나? 데이터가 정확하다는 것을 어떻게 확신하는가? 데이터는 신뢰할만한가? 데이터는 모집단을 대표하는가?

4.1 데이터 수집방법...

-통계분석을 위한 데이터를 수집하거나 또는 확보하기 위해 사용되는 많은 방법들이 존재한다.

-가장 널리 사용되는 3가지 방법은 다음과 같다.

- 직접 관측 (예. 한 은행으로 시간당 들어오는 고객의 수)
- 실험 (예. 비용을 최소화하는 새로운 제품 생산방법들)
- 서베이

서베이...

-서베이는 사람으로부터 정보를 수집하는 방법이다. 예. 갤럽 여론조사, 선거전 여론조사, 마케팅 서베이

-응답률(Response Rate)은 하나의 중요한 서베이 모수이다.
낮은 응답률은 통계분석의 결과를 무효화시킬 수 있다.

-서베이는 다양한 방법으로 수행될 수 있다.

예. 개인인터뷰, 전화인터뷰, 질문지서베이 등

질문지 설계 (Questionnaire Design)

1. 질문지는 가능한 한 짧아야 한다.
2. 질문은 짧고, 간단하고, 명료하게 서술되어야 한다.
3. 응답자가 편안하게 시작할 수 있도록 돕기 위해 인구통계학적 질문들로 시작한다.
4. 양자택일형 질문(예/아니오)과 선다식 질문을 사용한다.
5. 개방형 질문들을 세심하게 사용한다.
6. 유도하는 질문의 사용을 피한다.
7. 소수의 사람들을 대상으로 질문지를 사전에 테스트 한다.
8. 질문지를 준비할 때 수집된 데이터를 분석하고자 하는 방법을 생각한다.

4.2 표본추출(Sampling)...

- 통계적 추론은 표본에 기초하여 모집단에 관한 결론을 도출할 수 있도록 해준다.
- 모집단의 부분집합을 선택하는 표본추출(sampling)은 비용 cost (1억명의 TV시청자들을 인터뷰하는 것보다 표본으로 추출된 1000명의 시청자를 인터뷰하는 것이 비용이 훨씬 덜 든다)과 실용성 practicality (생산된 모든 자동차에 대하여 충돌실험을 하는 것은 실용적이지 못하다)의 이유때문에 이루어진다.

4.2 표본추출(Sampling)...

-목표모집단(target population)(통계적 추론을 하기 원하는 모집단)과 표본추출 모집단(sampled population)은 유사하여야 한다. ->이 경우에 표본통계량은 추정하고자 하는 모수에 가까울 수 있다. ->실제로 목표모집단과 표본추출모집단이 동일하지 않을 수 있다.

예/ 1936년 Literary Digest 의 대통령선거 여론조사

->Literary Digest 의 구독자명단과 전화번호부로부터 1000만명의 표본추출
->평균보다 부유한 사람들 중심의 표본추출

->230만명이 응답 (자기선택표본(self-selected sample) /주어진 이슈에 대하여 관심있는 대상만이 선택되는 표본 ->모집단을 대표하지 못하기 때문에 편의를 가짐) ->예측 실패 (->Franklin D. Roosevelt 승리)

4.2 표본추출방법 (Sampling Plans)

- 표본추출방법은 표본이 모집단으로부터 어떻게 추출될 것 인지를 규정하는 방법 또는 과정이다.
- 3가지의 표본추출방법에 초점을 맞추어 논의한다.
- 단순임의표본추출법 (simple random sampling)
- 층화임의표본추출법 (stratified random sampling)
- 군집표본추출법 (cluster sampling)

단순임의표본추출법 (Simple Random Sampling)

-단순임의표본(simple random sample)은 동일한 표본크기를 가지는 모든 가능한 표본이 선택될 가능성이 동일한 방식으로 선정된 표본이다.

예. 경영통계학과목을 수강하는 학생들의 모든 이름을 담은 모자로부터 3개의 이름을 추출하는 것->3개의 이름을 가지는 표본들 각각이 하나의 표본으로 선정될 가능성을 동일하다.

군집표본추출법(Cluster Sampling)

-군집표본(cluster sample)은 그룹들 또는 군집들의 단순임의표본이다.

-군집표본추출법은 모집단의 원소들을 모두 나열하기 어렵거나 또는 그렇게 하는 것이 비용이 많이 드는 경우 또는 모집단의 원들이 지리적으로 널리 흩어져 있는 경우에 유용하다.

예. 서울의 연간 평균가구소득을 추정하기 위한 표본추출

->서울의 각 블록을 하나의 군집으로 나타내고 단순임의추출방법을 사용하여 군집을 추출하고 추출된 군집들에 속한 가구들로부터 단순임의표본을 추출

표본크기(Sample Size)...

- 표본크기의 결정 (표본크기의 결정방법 사용)
- 표본크기가 클수록 통계적 추론의 정확도는 더 증가한다. (표본추정치가 추정하고자 하는 모수에 더 가까워진다)

4.4 표본오차와 비표본오차 (Sampling and Non-Sampling Errors)

...

-관측치들의 표본이 모집단으로부터 추출될 때 두가 종류의 오차, 즉 표본오차(sampling error)와 비표본오차(nonsampling error)가 발생될 수 있다.

-표본오차(sampling error)는 우연히도 표본에 선택되는 관측치들 때문에 존재하게 되는 표본과 모집단간의 차이를 의미한다.

->예.표본평균과 모평균의 차이가 표본오차이다.

-표본크기를 증가시키면 표본오차는 감소된다.

비표본오차(Nonsampling Error)

-비표본오차(nonsampling errors)는 데이터의 수집에서 발생하는 실수 또는 부적절하게 선택되는 표본관측치들 때문에 발생하는 오차를 의미한다.

-3가지 유형의 비표본오차:(1)데이터 수집상의 오차->데이터의 부정확한 기록, 부정확한 질문에 의한 부정확한 응답, 데이터의 기록 오류

(2)무응답 오차->표본의 일부로부터 응답을 받지 못하는 경우에 발생하는 오차 (표본이 모집단을 대표하지 못하는 가능성 존재)

(3)선택편의(selection bias)->표본추출방법이 모집단의 일부가 표본에 선택될 수 없게 만드는 경우에 발생하는 오차 (표본이 모집단을 대표하지 못하는 가능성 존재)

-표본크기를 증가시키는 것이 비표본오차를 감소시키지 못할 수 있다.