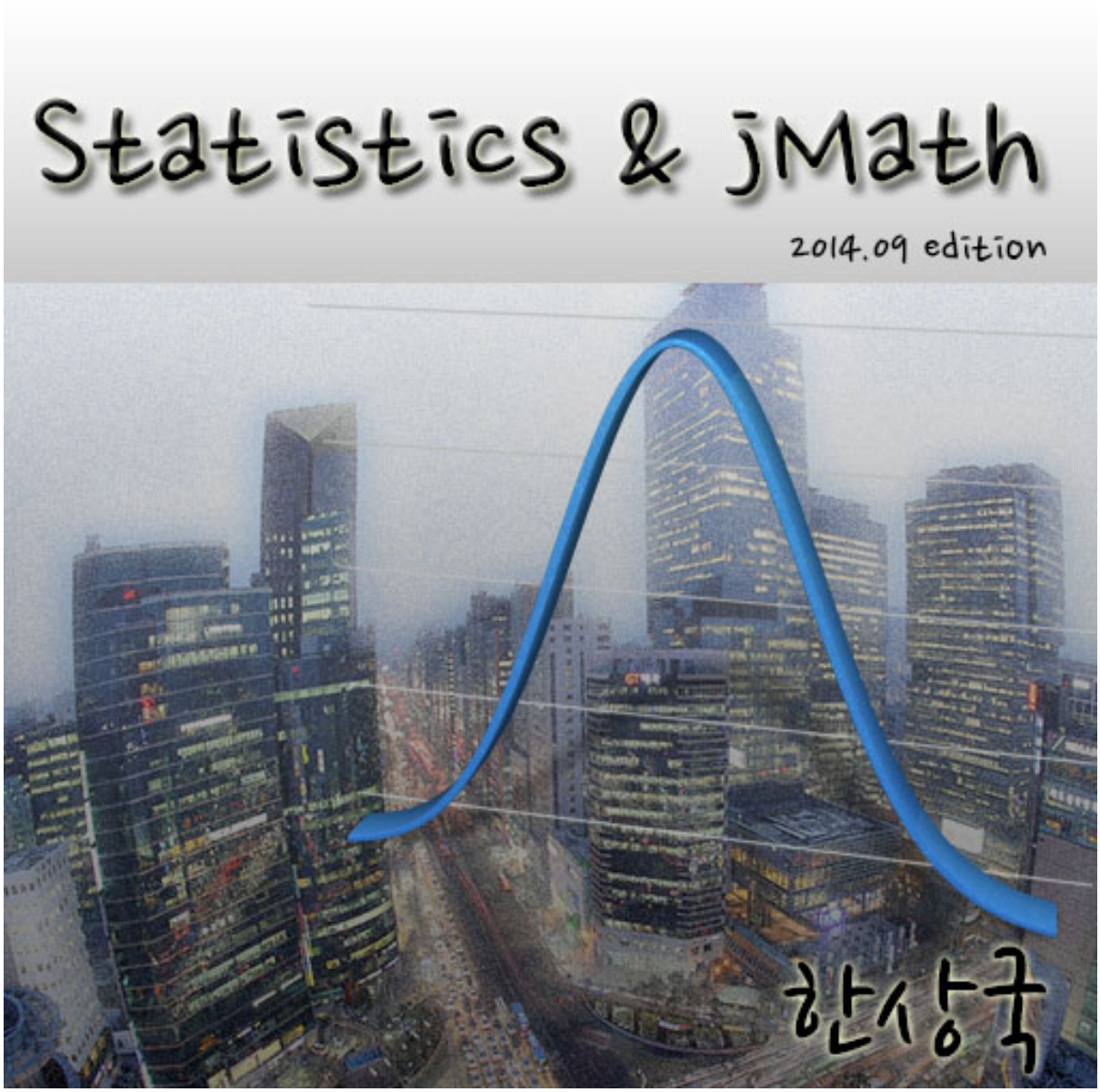


저자: 한상국

이메일: handucklive@gmail.com

Chapter 1 Statistics 개론

韓, 커피 연 242억잔 '홀짝'…시장 4조6000억원 '훌쩍'

작년 1인당 484잔…하루평균 1.3잔
시장규모 6년새 세배 이상 성장…관련 시장은 6조2000억대 육박
전체 소비량 62.6% 커피믹스 1위
인스턴트 27% 베트남서 수입…원두 한잔당 순수커피값 1420원

업계의 조사에 따르면 작년(2013) 한국인은 총 242억잔의 커피를 마셨다. 1인당 연평균 484잔에 해당하는 양으로 일평균 1.3잔은 마셨다는 의미다. 그렇다 보니 커피와 관련해 많은 돈이 오간다. 커피업계가 바라보는 국내 순수 커피시장 규모는 지난해 4조5680억원이었다. 6년새 세배 이상 성장했다.

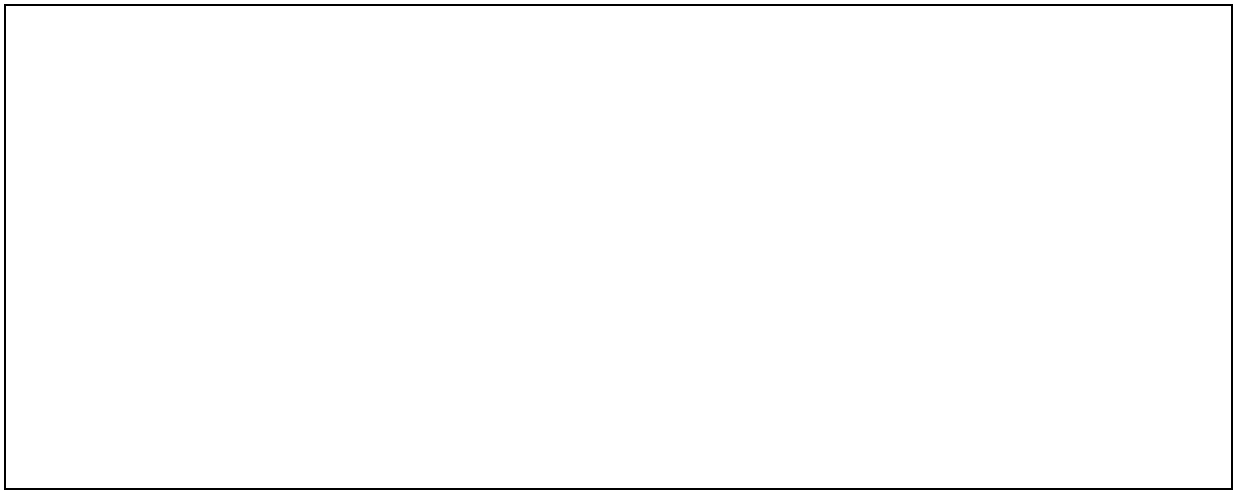
소비자가 지불하는 총 지출액 기준으로 보자면 작년 커피 관련 시장은 6조1560억원에 달했다. 소프트웨어 시장(6조5000억원)이나 아웃도어 시장(6조9000억원)에 육박한다.

한국인의 깊어지는 커피사랑은 길거리에 있는 커피전문점의 갯수로도 쉽게 체감할 수 있다. 2009년 전국 5200여개에 불과하던 커피전문점은 지난해 1만8000개까지 성장했다. 모처럼 읍내를 찾 은 시골 어르신이나, 최전방 부대에서 외출나온 군인들 조차도 느긋하게 맑은 아메리카노 커피 한잔을 사먹을 수 있을 정도로 커피전문점 갯수는 늘었다.

그렇지만 '한국인이 가장 사랑하는 커피'는 따로 있다. '커피계의 왕'이자 '전가의 보도'인 커피믹스다.

커피업계에 따르면 전체 커피 소비량의 62.6%는 커피믹스다. '커피 둘, 설탕 둘, 프림 둘'로 대변되는 직접 타먹는 가루 커피인 솔루블 커피도 13.2%나 된다. 음식료 제조사들이 내놓는 '카페○○'류의 커피도 11.6%다. 원두커피는 10% 수준에 불과하다.

(주)헤럴드 2014/06/30



예전에는 커피를 마시려면 다방이나 커피숍과 같이 모든 연령대가 즐기는 장소가 아니였지만 요즘 거리에 있는 수많은 커피 전문점들은 나이 연령에 상관없이 모두가 즐기는 장소가 되었습니다. 그런데 신기한건 커피 전문점이 거리에 너무 많아 모든 상점들이 장사가 잘 되나 의심될 정도 입니다. 이렇게 많은 상점 주인들이 커피전문점을 열려고 하는 이유는 커피시장의 6 조 2000 억원이라는 소비자 지출 규모 때문이라고 생각이 됩니다. 이 규모는 기사 내용과 같이 소프트웨어와 아웃도어 시장과의 비교했는데 거의 비슷합니다. 하지만 기사에서 보면 이 많은 커피 매출액을 주도하는 것은 커피 전문점이 아니라 일명 봉지커피라 불리우는 키피믹스라는 사실을 숫자를 통해서 알려 줍니다.

우리는 이와 같이 수 많은 자료들로 둘러 싸여 있습니다. 이러한 자료들은 잡지 기사, 뉴스, 스프츠와 같은 한번 보면 잊어버리는 내용들이 있고, 반면에 잊어버려서는 않되는 중요한 내용들이 있습니다. 전자의 경우는 단순히 보고 듣는것으로 끝나기 때문에 얻은 자료를 토대로 더 이상 의미가 없다면 머리속에서 사라지지만, 후자의 자료들은 이해를 돕기 다양한 분석 방법들을 적용을 합니다. 예를 들어 앞서 보았던 신문기사와 같이 자료를 숫자로 표현하여 눈으로 보이지 않는 현상을 알 수 있도록 합니다. 이렇게 숫자를 이용하여 알지 못했던 사실이나 구체적이지 못한 내용들을 명확하게 알 수 있도록 하여 현상을 이해하고 예측을 할 수 있도록 하는 것이 통계 입니다. 즉, 통계는 숫자로 표기되는 자료를 모아서 정보를 제공하는 것 뿐만 아니라 조사 대상 전체에 대한 추론을 하는 기능을 제공하는 것으로, 병원, 약품 개발, 과학에서 뿐만 아니라 기업과 자영업자들에게도 판단을 위해 중요한 수단입니다.

하지만 인류가 이러한 통계적 사고는 확률의 도움없이는 되지 않았습니다. 통계에 가장 기본인 물건을 세는 것인데 이를 이해하기 위해서는 확률이 반드시 필요합니다. 그런데

역사를 보면 확률의 발달은 피타고라스와 같은 수학자들이 있었던 고대 그리스 시대가 아니라 1500 년대에 와서야 도박에 의해서 수학으로 발달이 되었습니다. 이러한 이유는 고대 그리스인들에게는 확률은 신들이 다루는 즉 신의 뜻에 의해 결정이 되는 것이지 사람 다룰수 있는 것이 아니라고 생각했기 때문에 이 분야에 대한 수학이 발달 하지 않았습니다. 그래서 고대 그리스 시대의 수학을 살펴보면 삼각함수와 같은 정확하게 계산이 가능한 즉 진실을 찾을 수 있는 내용들을 발달을 했습니다. 이러한 사실을 기반으로 통계의 역사를 보면 BC 5 세기경 부터 통계적 방법을 사용 했다라는 증거가 있지만, 1663 년이 되어서야 학문적인 접근이 시작이 되기 시작했습니다. 이 시대에 통계가 적용된 곳은 인구 자료와 경제 자료 분석이였고, 통계는 관찰된 자료로 부터 현재 인구나 경제가 어떠한 상태인가를 연구하기 위해서 사용되었습니다. 그래서 상태를 의미하는 state 와 학문을 나타내는 접미어 ics 를 합쳐서 statistics 라는 용어가 만들어 지게 되었습니다. 결과적으로 확률과 통계의 발달로 인해서 통계를 본격적 활용하기 시작한 것은 19 세기 때 부터입니다.

통계의 발달이 다른 수학에 비해 늦게 시작된 다른 이유는 심리학적 관점에서도 볼 수 있습니다. Daniel Kahneman 의 저서 Thinking, Fast and Slow 를 보면 인간의 사고는 크게 두가지 사고로 구분을 합니다. 하나는 직감적인 사고로 예를 들어 사진에 얼굴을 찡그리고 있는 얼굴을 보면 고통을 느끼고 있다 순식간에 판단하는 것과 같이 매우 빠른 속도로 인지하게 됩니다. 이러한 것은 정상적인 사람들이라면 기본적으로 발달한 사고 능력입니다. 다른 하나는 시간을 두고 생각을 해야만 알 수 있는 사고로 예를 들어 1+1 은 2 이라는 것은 초등학교를 졸업한 사람이라면 바로 생각하지도 않고 답을 알 수 있지만 32+13+43-23 와 같이 약간만 복잡하게 되면 답을 얻기 위해서 시간을 두고 생각을 해야 합니다.
이 두가지 사고력을 각각 부지런한 system 1 과 게으른 system 2 라고 하는데, system1 의 직관적이고 빠른 처리 능력 덕분에 여러 가지를 동시에 처리할 수 있지만 오류가 발생하기가 쉽습니다. 예를 들어 편견이 있는 상태에서 system 1 에 의존하는 경우는 이러한 문제를 많이 발생시킵니다. 하지만 system1 은 부지런하게 계속 주변에서 정보를 인지를하고 판단을 장점을 갖고 있습니다. 반면 게으른 system 2 를 system 1 이 포기하고 넘겨 줄 때까지 동작을 하지 않습니다. 즉 대충 보고 복잡해 보이면 바로 포기하기도 하고 중요한 것이면 system 2 에게 기회를 줍니다. 통계는 system 1 의 직감에 의존하지 않고 바로 system 2 영역에서 자료를 보고 생각을 하여 숨어있는 정보를 얻어야 하는것이기에 일반적으로 사람들이 삶을 위해서 크게 의존하는 system 1 보다 의존성이 다소 떨어져 당장의 필요성을 느끼지 못합니다. 이러한 경향은 비록 통계를 훈련 받은 통계학

전공자에게도 나타나는 현상으로 사람의 판단이 얼마나 system 1 에 크게 의존하고 있는가를 알 수 있습니다.

통계를 위해 필요한 절차중 하나는 자료를 수집하는 과정과 수집된 자료를 정리하고 테이블이나 도표와 같은 방법으로 표현하는 단계를 거치는 것입니다. 다행이도 컴퓨터 기술의 발달은 전자 장비를 이용한 자료 수집에서 부터 결과를 보기 쉽게 정리할 수 있도록 해줍니다. 그런데 자료 수집과정은 아직 사람이 직접 해야 하는 경우가 많이 있습니다. 예를 들어 환자의 연령대별 질병의 종류를 알고 싶을 경우 환자를 진료하여 자료를 생성하는 과정은 현재 기술로는 의사 대신 기계가 할 수 없습니다. 이는 컴퓨터가 수집된 데이타를 근간으로 다양한 정보를 생산하는 deduction 처리 능력이 뛰어나지만 데이타를 수집하여 처리하는 induction 처리 능력이 인간에 비해 뒤쳐지기 때문입니다. 예를 들어 23+17 이라는 것을 계산을 위해서 induction 과정에서 두개의 숫자와 덧셈 연산자를 인식하는 과정을 거쳐야하는데 이는 음성인식으로 알수도 있고 그림을 통한 인식이 가능합니다. 이러한 과정은 사람이 더욱 빠르고 정확합니다. 하지만 자료를 컴퓨터가 알 수 있도록 기입만 된다면 즉 23, 17, +을 컴퓨터가 바로 이해 할 수 있도록 하게 되면 사람과 비교가 안될 정도로 매우 빠른 속도로 계산을 합니다.

이처럼 사람이 갖고 있는 특징과 컴퓨터가 갖고 있는 특징을 잘 조합하을 하여 통계처리를 한다면 효울적입니다. 그래서 통계 학습에는 통계 관련된 수학만 다룰 수 있으면 되는 것이 아니라 통계 프로그램 혹은 컴퓨터 언어를 병행하여 학습을 하지 않으면 안됩니다.

일반적으로 많이 사용되는 통계 프로그램으로 SPSS, SAS, R 과 같은 프로그램들이 있지만 이 교재는 회사는 소상공들이 회사를 운영하는 필요한 기업통계의 기초 내용을 JavaScript 언어를 통해서 웹 브라우저에서 통계처리가 바로 가능하도록 설계된 jMath library 와 도표를 그리는 jqplot 를 이용하는 방법 설명 할것입니다. 이렇게 JavaScript 을 이용한 이유는 요즘과 같이 웹이 중요한 시대에서 데이타를 처리하고 결과를 만들어 보여주는 모든 과정을 Microsoft 의 Internet explorer 나 Google chrome 에서 바로 처리 할 수 있도록 하여 빠른 정보 전달을 하기 위해서 입니다.

기업 통계 (Business Statistics)

큰 기업에서 부터 자영업자까지 통계는 자료로 부터 숨어있는 정보를 얻을 수 있게 해주므로 이윤 창출을 위해서 없어서는 않되는 중요한 것입니다. 예를 들어

1) 온라인 쇼핑몰을 통해 옷을 판매하는 운영자는 매 월마다 판매되는 옷의 수량을 통해서 판매되는 물건들의 차이점이 있는지를 알고 소비 유형을 미리 알아내어 앞으로 판매 수량을 예측하고 물건을 미리 준비하려고 합니다.
2) 컴퓨터와 노트북 수리점에서 서비스 의뢰로 소프트웨어와 하드웨어의 문의에서 소프트웨어에 문제가 더 많이 있다는 것을 알고 이 분야에 더욱 집중된 서비스를 준비하려고 할 때 통계를 이용할 수 있습니다.
3) 새 음료를 만들고 고객들에 시음을 하도록 하여 고객들의 만족에 대한 설문조사를 하여 연령별, 성별 차이점을 알고 이 후에 마케팅을 위한 자료로 활용하려 할 수 있습니다.

기업 통계는 기업 활동영역인 마케팅, 광고, 관리, 금융등에 자료를 수집하여 직관적으로 잘못된 판단을 피하도록 통계적인 근거로 접근을 통해 알고 싶은 정보를 얻는 것입니다. 하지만, 인터넷의 발달로 인해서 개인 혼자서도 기존에 큰 업체만이 누릴 수 있었던 업종들을 쉽게 할 수 있지만 자세한 통계적 분석은 아직 개인들이 하기에는 어려운 점이 많이 있습니다. 예를 들어 마케팅 전략을 모색을 하려고 할 경우 Google analytics 와 같은 사용자 방문에 대한 연령, 성별, 지역별로 구분하여 간단한 정보를 얻을 수 있지만, 비교 분석, 예측, 가설 검증과 같은 세부적인 내용들은 개인들이 통계를 별도로 알아 자신에게 맞는 시스템을 갖기 전에는 어려움이 있습니다.

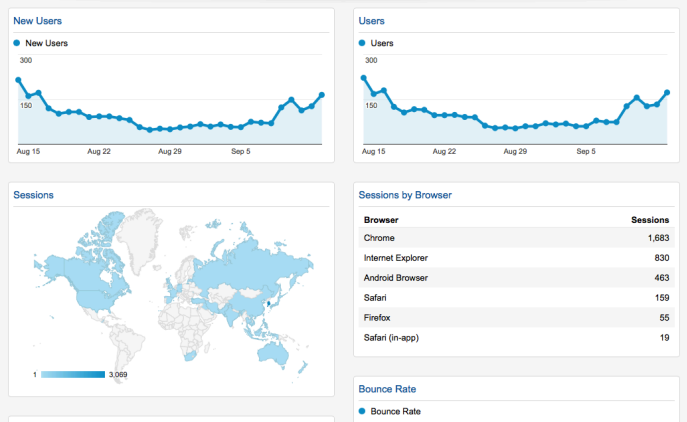


그림 1 Google Analytics

이러한 통계로 정보를 얻는 과정을 비유를 하자면, 봉합된 상자안에 물건이 있고 이를 흔들면서 상자와 물건이 붙이치는 소리 만으로 상자 안에 물건이 무엇인가를 아는 과정과 같다고 생각하시면 됩니다. 즉 관찰을 통해서 얻은 자료를 통합하여 정보를 유출하여 상자안 물건을 알아내는 것입니다. 장사나 기업을 운영하여 이윤 창출을 위해 소비 유형과 같은 정보는 상자안 물건이고 이를 여러 가지 방법으로 관찰하여 얻은 정보를 통해서 소비 유형을 발견하려 할 것입니다. 하지만 이렇게 얻은 결과는 정확한 값이 아니므로 통계에서는 항시 값의 범위 즉, 오차 범위를 함께 설명합니다.

관찰을 통해 얻은 결과가 관찰할 수 있는 전체를 통해서 얻은 것인가 아니면 일 부분만 관찰했는가에 따라 차이가 있습니다. 예를 들어 신상품을 만들기 위해 소비자 선호조사를 하려고 할 경우 모든 사람들에게 물어 보는 것은 불가능 합니다. 그래서 일부 사람들에게 선호도 조사를 하고 이 대상자들이 전체를 대표하는 것처럼 하여 통계결과를 만드는 것이 일반적입니다. 하지만 회사 직원에 100 명인 곳에서 직원들의 지각률과 같은 정보를 얻는 것은 전체 직원들을 관찰하는것은 문제가 되지 않기 때문에 대표로 몇 명만 추출하여 통계 결과를 만들 필요가 없습니다. 이에 관련된 통계 용어들로 population 은 관찰하려는 전체 대상을 말하고, sample 은 population 으로 부터 일부를 대표로 선발한 것입니다.

통계를 하는 주된 일은 내가 주장하려는 내용이 타당성이 있는가를 알아내는 것입니다. 예를 들어 영화 시사화를 통해서 남녀 성별 선호도를 조사하여 여성의 선호도가 남성 것보다 높다는 가설(hypothesis)을 만들고 조사를 하게 됩니다. 이를 위해서 필요한 것은 시사회를 참석한 남녀별 영화 관람후 평가 점수들을 하나로 표현할 수 있는 숫자값으로 전환을 해줘서 이 값을 통해 관람자의 특성을 알 수 있게 해줘야 합니다.

이러한 특성을 나타내기 위해 표현된 숫자값으로 대표적인것이 평균입니다. 만일 평균값이 전체 조사 대상 즉 population 으로 부터 얻은 값이라면 parameter 라고 하고, 조사 대상 중 대표들인 sample 로 부터 얻은 값이라면 statistic 이라고 합니다.

번호	1	2	3	4	5	6	7	8	9
남자	8	7	3	4	2	4	6	9	2
여자	5	5	6	5	4	5	6	5	4

남녀 둘다 평균 5 점입니다.

이렇게 얻은 statistic 으로 부터 population 의 parameter 값을 추론(inference) 하는 것을 parameter estimation 이라고 합니다. 하지만, 이 추론 값은 정확성을 보장 못하고 대신 parameter 가 있을 신뢰 구간과 확신률을 같이 알려 줘야 합니다.

이러한 관점에서 통계는 descriptive statistics 와 inferential statistics 로 구분이 됩니다. Descriptive statistics 은 자료를 정리할 수 있도록 전체 자료값을 대표하는 값들을 계산하는 통계방식이고, inferential statistics 은 부분을 통해서 전체를 이해하려는 통계 방식으로, 다시 말해 population 의 공통된 특징을 sample 을 통해서 알아내는 것입니다.

앞의 영화 평점의 예를 보시면 비록 두개의 sample 에 대한 평균은 같지만 남자의 점수의 분포가 여자의 분포보다 더 다양함을 알 수 있습니다. 이를 통해서 남자와 여자의 실제 평점이 같다고 판단하기는 어렵습니다. 하지만 만일 남성의 경우 영화에 대한 점수가 5 점보다 멀리 떨어진 9 점 2 점과 같은 점수가 몇명 안되지만 이 조사에서 이러한 경우들만 sample 로 모집된 것일 수 있습니다. 이러한 경우가 매우 심각하면 문제 이지만 sample 내에서도 많이 나타나지 않는다면 통계는 이러한 자료를 근거로 추론을 할 수 있도록 도와주는 도구라고 생각하시면 됩니다.

자료 구분

다음은 자료값(data)를 4 칙 연산인 덧셈, 뺄셈, 곱셈, 나눗셈이 가능한 유무에 따라 구분하는 양적 자료와 질적 자료에 대한 설명입니다.

1) 양적(Quantitative) 자료

숫자들로 4 칙 연산이 가능한 정보 입니다. 예를 들어 온라인에 음식점들의 평가 점수는 고객들이 식당 음식을 먹고 난 후에 음식의 맛있는 정도, 서비스 정도에 따른 점수로 10 점 만점으로 점수를 주었을 때 이렇게 모아진 자료를 갖고 평균 점수를 만들 수 있습니다. 다른 예로 자동차 연비, 제품의 온라인 쇼핑몰 별 판매 가격등이 있습니다.

2) 질적(Qualitative) 자료

교육의 정도를 나타태는 고졸, 대졸, 석사, 박사는 평균을 나타낼 수 없습니다. 또한 지역을 나타내는 우편번호는 비록 숫자이지만 우편번호를 갖고 평균을 만드는 것은 무의미한 정보를 얻기만 합니다.

자료(data)를 구분하는 다른 방법으로 다음의 4 가지로 구분이 될 수 있습니다.

1) **nominal** 자료: 인종, 지역, 성별, 우편번호와 것이 대표적인데 이름으로 의미가 있을 뿐 순위(rank)와 같은 정보는 없습니다.
2) **ordinal** 자료: 순위(rank)가 있는 것이 nominal 자료와 차이로 고졸, 대졸, 대학원 졸과 같은 최종 학력, 상품 만족도등이 대표적인 예입니다. 순위는 있지만 category 별로 실제 거리는 없습니다.
3) **interval** 자료: 온도나 년도와 같이 2000, 2001 과 같은 것으로 순서가 있지만 비율로 나타날 때 의미가 없고 또한 절대값 0 이 없습니다. 예를 들어 2001/2013 으로 나타낸 비율은 의미가 없고, 15 도는 30 도에 비해 두배로 춥다고 말을 하지 못합니다. 절대값 0 에 대해서 0 년은 모든 연도의 시작되는 연도가 아닙니다.
4) **ratio** 자료는 interval 자료와 같이 순위가 있고 비율에 의미가 있고 절대적인 0 이 있습니다. 예를 들어 나이, 무게, 가격, 연봉등이 있습니다. 연봉 1000 만원은 연봉 2000 만원의 2 배입니다.

여기서 nominal 과 ordinal 자료는 질적자료(qualitative)에 해당하고 interval 과 ratio 는 양적자료(quantitative)에 해당합니다.

통계 주의 사항

Sample 로 부터 값을 얻어 Population 이 어떠할 것이라는 것은 선거의 예를 보면 쉽게 이해가 될시 겁니다. 하지만 여기서 주의할 점은 조사 대상인 sample 을 만드는 방법과 해석입니다. 선거에서 보면 대통령 선거 기간 동안 지지율을 조사하기 위해서 sample 로 선택된 사람들 대부분 보수 경향이 강한 사람들로 구성되어 있다면 보수파 대통령 후보자의 지지도가 매우 높게 나타날 것입니다.

다시 말해, 만일 sample 에 대한 이해를 확실하게 하지 않고 통계를 만들다 보면 잘못된 결과가 나타나는데, 이것이 만일 회사에 자신의 제품을 홍보하기 위해서 악의적으로 사용했다면 도덕적으로 범죄와 같은 행위라 볼 수 있습니다. 따라서 통계를 하기 위해서 sample 을 만드는 과정에 매우 신중해야 하며 어느 한 쪽으로 편향된 대상들만 선출이 되었는가 확인을 잘 해야 합니다. 다음 기사는 2014 년 충청북도 교육감 선거에서 잘못된 통계자료로 여론조사 기관에 벌금을 부과한 내용입니다.

'여론조사 결과 오류' 조사기관 과태료 1500 만원

충북도선거관리위원회는 12 일 충북교육감선거의 여론조사 결과 후보자별 지지도를 사실과 다르게 분석해 언론사에 제공한 A 여론조사기관에 과태료 1500 만원을 부과했다고 밝혔다.
선관위에 따르면 A 여론조사기관은 지난달 9 일 충북교육감 여론조사를 실시해 그 결과를 충북지역 모 일간지에 제공했다. 이 언론사는 이를 근거로 같은 달 14 일자 신문에 결과를 보도했다.
그러나 여론조사결과에 대한 예비후보의 이의신청으로 충북선거여론조사공정심의위원회에서 심의한 결과 A 기관이 자료를 잘못 분석해 특정 후보자들의 지지율이 실제 결과와 다르게 취합된 것으로 드러났다.
이후 이 조사기관과 해당 언론사는 공식 사과문을 냈다.
선관위 관계자는 “잘못된 여론조사 결과는 선거인의 자유로운 의사결정을 방해하고 선거에 미치는 영향이 크기 때문에 1500 만원의 과태료 처분을 내렸다”며 “선거일이 20 여일 남은 시점에서 선거에 영향을 미칠 수 있는 행위에 대해 예의주시 하겠다”고 말했다.

뉴스 1 2014/05/12

만일 이러한 통계의 문제가 사업과 연관되어 신 제품을 출시하고 제품을 좋아하는 사람들만 선별하여 조사를 하고 결과 보고를 했다면 업체에게는 치명적인 손해를 보게 됩니다.