TRUSTWORTHY
ONLINE CONTROLLED
EXPERIMENTS

A/B 테스트

신뢰도 높은 온라인 통제 실험

론 코하비 · 다이앤 탕 · 야 쉬 지음 이기홍 · 김기영 옮김

A/B 테스트

A/B 테스트

신뢰도 높은 온라인 통제 실험

론 코하비 · 다이앤 탕 · 야 쉬 지음 이기홍 · 김기영 옮김

에이콘

이 책에 쏟아진 찬사

"린lean 방법론의 핵심에는 가설 생성, 실험 실행, 데이터 수집, 통찰력과 가설의 검증과 수정과 같은 과학적 방법론이 있습니다. A/B 테스트는 검증 가능하고 반복 가능한 실험을 설계하는 표준이며, 이 책은 이를 위한 훌륭한 교과서입니다."

— 스티븐 블랭크(Steve Blank)/ 스탠포드 대학교의 겸임교수, 현대 기업가 정신의 아버지, 『기업 창업가 매뉴얼』과 『Four Steps to the Epiphany(깨달음을 향한 4단계)』의 저자

"이 책은 제품 기능, 프로젝트 효율성 또는 매출을 최적화하기 위해 온라인 종합 대조 실험을 사용하고자 하는 경영진, 리더, 연구원 또는 엔지니어에게 유용한 자료입니다. 코하비의 연구가 빙과 마이크로소프트에 미치는 영향을 직접 알고 있었기에 그의 배움이 이제 더 많은 청중에게 전달될 수 있게 돼 기쁩니다."

— 해리 셤(Harry Shum)/ 마이크로소프트 인공지능 및 리서치 그룹 전무이사

"엄밀하면서도 접근하기 쉬운 훌륭한 책입니다. 독자들은 인터넷 제품 개발에 혁명을 일으킨 신뢰도 높은 온라인 종합 대조 실험을 자신의 조직에 가져오는 방법을 배우게 될 것입니다."

— 애덤 단젤로(Adam D'Angelo)/ Quora의 공동 설립자이자 CEO, 페이스북의 전 CTO

"여러 회사가 온라인 실험과 A/B 테스트를 사용해 제품을 개선하는 방법을 한눈에 알아볼 수 있는 훌륭한 책입니다. 코하비와 탕, 쉬는 풍부한 경험과 훌륭한 조언을 하고자 수년 동안 배운 많은 실전 사례와 교훈을 이 책에 담았습니다."

— 제프 딘(Jeff Dean)/ 구글 선임연구원 및 구글 리서치 상무

"조직이 지속적으로 더 나은 결정을 내리기를 원하십니까? 이 책은 디지털 시대에서 데이터 기반으로 의사결정을 할 수 있도록 돕는 바이블입니다. 이 책을 읽는 것은 아마존, 구글, 링크드인, 마이크로소프트 내부의 회의에 참석하는 것과 같습니다. 저자들은 세계에서 가장 성공적인 기업들이 의사결정을 내리는 방법을 처음으로 공개합니다. 일반 비즈니스 서적에서 볼 수 있는 지침이나 일화 이외에도, 데이터에 기반한 의사결정을 위해 무엇을 어떻게 잘 해야 하는지를 보여줍니다. 이 책은 비즈니스 리더, 엔지니어, 데이터 분석가를 위한 디지털 세계의 의사 결정 매뉴얼입니다."

— 스콧 쿡(Scott Cook)/ Intuit 공동 설립자 겸 실행 위원회 회장

"온라인 종합 대조 실험은 강력한 도구입니다. 온라인 종합 대조 실험이 어떻게 작동하고, 강점이 무엇이고, 어떻게 최적화될 수 있는지를 이해하는 것은 전문가와 더 많은 청중 모두에게 도움이 될 것입니다. 이 책은 기술적으로 권위 있으면서도, 읽기 쉽고, 매우 중요한 문제를 다루는 세상에 많지 않은 책입니다."

— 존 P.A.(John P.A.)/ 이오아니디스 의학과 보건연구정책학 교수,
스탠포드 대학교의 생물의학 데이터 과학 및 통계학과

"어떤 온라인 옵션이 더 나을까요? 우리는 이에 대해 자주 선택하고 실수 또한 자주 해야 합니다. 무엇이 실제로 더 잘 작동할지 결정하려면 엄격한 종합 대조 실험, 즉 A/B 테스트가 필요합니다. 마이크로소프트, 구글 및 링크드인의 전문가가 저술한 이 우수하고 생생한 책은 A/B 테스트의 이론과 모범 사례를 보여줍니다. 온라인상에서 무언가를 한다면 반드시 이 책을 읽어야 합니다!"

— 그레고리 피아테스키-샤피로(Gregory Piatetsky-Shapiro) 박사/ KD Nuggets 사장,
SIGKDD 공동 설립자 및 링크드인 탑보이스의 공동 설립자

"저자들은 세계 최고의 온라인 실험 전문가입니다. 저는 몇 년 동안 그들의 제품을 사용해왔는데, 그들이 팀을 이뤄 이 책을 쓰게 돼 기쁩니다. 이 책을 제 학생들뿐만 아니라 온라인 제품 및 서비스에 관련된 모든 사람에게 추천합니다."

— 에릭 브린욜프슨(Erik Brynjolfsson)/ MIT의 교수이자 『제2의 기계 시대』 공동 저자

"현대 소프트웨어 기반 비즈니스는 온라인 종합 대조 실험 없이는 성공적으로 경쟁할 수 없습니다. 이 분야에서 가장 경험이 많은 지도자 세 명이 쓴 이 책은 기본 원리를 제시하고, 설득력 있는 예를 보여주고, 풍부한 실용적인 조언을 제시하기 위해 더 깊이 파고듭니다. 이 책은 그야말로 필독서입니다!"

— 포스터 프로보스트(Foster Provost)/ 뉴욕대학 경영대학원 교수이자
『비즈니스를 위한 데이터 과학』의 공동 저자

"과거 20년 동안 기술 산업은 과학자들이 수세기 동안 알고 있었던 것을 학습해왔는데, 그것은 바로 종합 대조 실험이 복잡한 현상을 이해하고 매우 어려운 문제를 해결하기 위한 최고의 도구 중 하나라는 것입니다. 종합 대조 실험을 설계하고, 규모에 맞게 실행하고, 그 결과를 해석할 수 있는 능력은 현대 첨단 기술 기업이 어떻게 운영되는지 보여주는 토대입니다. 저자들은 세계에서 가장 강력한 실험 플랫폼들을 설계하고 구현했습니다. 이 책은 이러한 도구와 기술을 사용하는 방법에 대한 경험을 통해 배울 수 있는 좋은 기회입니다."

— 케빈 스콧(Kevin Scott)/ 마이크로소프트의 전무 및 CTO

"온라인 실험은 아마존, 마이크로소프트, 링크드인 및 주요 디지털 기업의 성공을 촉진했습니다. 이 실용적인 책을 읽는다면 이러한 회사에서 수십 년 동안 쌓은 실험 경험을 접할 수 있는 드문 기회를 얻는 것과 같으며, 이 책은 모든 데이터 과학자, 소프트웨어 엔지니어, 제품 관리자의 책장에 있어야 합니다."

— 스테판 톰케(Stefan Thomke)/ 하버드 비즈니스 스쿨의 윌리엄 바클레이 하딩 교수,
『Experimentation Works(실험 연구)』의 저자

"성공적인 온라인 비즈니스의 비결은 실험입니다. 이는 더 이상 비밀이 아닙니다. 이 책은 온라인 종합 대조 실험의 대가들이 A/B 테스트의 모든 것을 설명해서 온라인 서비스를 지속적으로 개선할 수 있도록 도와줍니다."

— 할 바리안(Hal Variian)/ 구글 수석 이코노미스트,
『Intermediate Microeconomics with Calculus(중간 미시경제학)』의 저자

"실험은 온라인 제품 및 서비스를 위한 최고의 도구입니다. 이 책은 마이크로소프트, 구글, 링크드인에서 수년간 성공적으로 테스트해 얻은 현장 속의 지식으로 가득합니다. 실제 사례, 흔한 실수와 실수의 의미와 솔루션을 통해 모범 사례와 통찰력을 보여줍니다. 저는 이 책을 강력히 추천합니다!"

— 프레스턴 맥아피(Preston McAfee)/ 마이크로소프트의 전 수석 이코노미스트이자 부사장

"실험은 디지털 전략의 미래이며, 이 책은 바이블이 될 것입니다. 저자들은 오늘날 실험 분야에서 가장 주목할 만한 전문가 중 세 명이며, 이들의 책은 디지털 실험을 위해 진정으로 실용적인 로드맵을 제공합니다. 마이크로소프트, 아마존, 구글 및 링크드인에서 수십 년에 걸쳐 실시한 사례 연구는 심도 있는 깊이와 명확성으로 실질적인 교훈을 주며 쉽게 이해할 수 있도록 구성돼 있습니다. 디지털 비즈니스의 모든 관리자가 이 책을 읽어야 합니다."

— 시난 아랄(Sinan Aral)/ MIT의 데이비드 오스틴 경영학 교수, 『Hype Machine(하이프 머신)』의 저자

"진지한 실험의 전문가를 위한 필독서로 매우 실용적이며 이 정도로 깊이 있게 다루는 책은 보지 못했습니다. 이 책은 매우 유용해서 마치 초능력을 얻는 것처럼 느껴집니다. 통계적 뉘앙스부터 결과 평가, 장기적인 영향 측정에 이르기까지 이 책을 통해 모든 것을 확인할 수 있습니다."

— 핍 라자(Peep Laja)/ 전환율 전문가이자 CXL의 설립자 및 대표

"온라인 실험은 마이크로소프트의 문화를 바꾸는 데 중요한 역할을 했습니다. 사티아Satya, 마이크로소프트 CEO는 성장 마인드에 대해 말하며 실험은 새로운 아이디어를 시도하고 그로부터 배우는 가장 좋은 방법이라고 합니다. 종합 대조 실험을 신속하게 반복하는 방법을 통해 빙은 수익성을 향상시켰고 마이크로소프트는 오피스, 윈도우와 애저 제품을 확산시켰습니다."

— 에릭 보이드(Eric Boyd)/ 마이크로소프트 AI 플랫폼의 기업 부사장

"기업가, 과학자, 경영자로서 소량의 데이터도 수십 배의 직감에 맞먹는 가치를 지닌다는 것을 어렵게 배웠습니다. 하지만 어떻게 하면 좋은 데이터를 얻을 수 있을까요? 이 책은 아마존, 구글, 링크드인, 마이크로소프트에서 수십 년 동안 쌓은 경험을 보기 쉽게 잘 구성돼 있습니다. 온라인 실험의 바이블입니다."

— 오렌 에치오니(Oren Etzioni)/ AI 연구소 CEO 겸 워싱턴 대학교 컴퓨터 과학 교수

"인터넷 회사들은 전례 없는 규모, 속도, 정교함으로 실험을 해왔습니다. 저자들은 이러한 발전에 핵심적인 역할을 해왔고, 독자들은 그들의 경험으로부터 배울 수 있어서 다행입니다."

— 딘 에클레스(Dean Eckles)/ MIT의 커뮤니케이션과 기술 분야의 KDD 경력 개발 교수, 페이스북의 전직 과학자

"이 책은 중요하지만 아직 충분한 가치를 인정받지 못하는 분야를 살펴볼 수 있는 매우 풍부한 자원입니다. 모든 장의 실제 사례는 성공적인 비즈니스의 내부 운영 방식을 보여줍니다. OEC전체 평가 기준를 개발하고 최적화하는 데 중점을 두는 것이 특히 중요한 교훈입니다."

— 제리미 하워드(Jeremy Howard)/ 싱귤래리티 대학교, fast.ai 설립자, 캐글의 전 대표 및 수석 과학자

"A/B 테스트에 대한 가이드는 많지만, 신뢰할 수 있는 온라인 종합 대조 실험 분야의 가이드는 적습니다. 저는 18년 동안 코하비를 따라다녔습니다. 그의 조언은 실무에 몰두하며 쌓여갔고 경험에 의해 연마되며, 실제 환경에서 연구를 하며 달궈졌다는 것을 알게 됐습니다. 다이앤 탕과 야 쉬와 함께라면 이해의 폭은 무엇과도 비교할 수 없을 것입니다. 다른 모든 책과 비교해보면 더 확실히 알게 될 것입니다."

— 짐 스턴(Jim Sterne)/ 마케팅 분석 서밋 창립자이자 디지털 분석 협회 명예 이사

"분석적 정교함, 명확한 설명, 어렵게 얻은 실제 경험의 교훈을 결합한 온라인 실험 실행을 위한 매우 유용한 방법이 담긴 책입니다."

— 짐 만지(Jim Manzi)/ Foundry.ai의 설립자, Applied Predictive Technologies의 전 CEO,
『Uncontrolled(비통제)』의 저자

"실험 디자인은 농업, 화학, 의학, 이제는 온라인 전자상거래와 같은 새로운 영역에 적용될 때마다 발전합니다. 최고 전문가들이 집필한 이 책은 온라인 실험을 하는 이유와 방법 그리고 실패하지 않는 방법 모두 다루는 실용적인 조언으로 가득합니다. 실험에 많은 비용이 들 수도 있지만, 무슨 일이 일어날지 모를 때는 더 큰 비용이 들 수 있습니다."

— 아트 오웬(Art Owen)/ 스탠포드 대학교 통계학 교수

"경영 간부와 운영 관리자가 꼭 읽어야 할 책입니다. 운영, 재무, 회계, 전략이 비즈니스의 기본 구성 요소를 형성하는 것처럼 오늘날 AI 시대에 온라인 종합 대조 실험을 이해하고 실행하는 것은 필수적입니다. 저자들은 실질적으로 접근할 수 있는 새롭고 중요한 지식 영역의 필수 요소를 제시했습니다."

— 카림 라카니(Karim Lakhani)/ 하버드 대학교의 혁신과학 연구소 책임자이자 교수,
모질라 코퍼레이션 이사

"실력 있는 '데이터 기반' 조직은 분석만으로는 충분하지 않다는 것을 알고 있습니다. 그들은 실험에 전념해야 합니다. 이 책은 어느 책보다 접근하기 쉽고 명확하며, 영향력이 큰 실험 설계를 할 수 있도록 돕는 매뉴얼이자 선언문입니다. 이 책에서 실용적인 영감을 많이 받았는데, 무엇보다도 문화가 기술 역량과 경쟁하는 방식을 중요한 성공 요인으로서 명확하게 알려줬습니다."

— 마이클 슈레이즈(Michael Schrage)/ MIT 디지털 경제 이니셔티브의 연구원이자 『Innovator's Hypothesis(혁신자의 가설)』의 저자

"실험에 관한 이 중요한 책은 세계에서 가장 큰 기술 기업 중 일부에서 일한 세 명의 저명한 리더의 지혜를 담았습니다. 조직 내에서 데이터 중심 문화를 구현하려는 소프트웨어 엔지니어, 데이터 과학자, 제품 관리자에게 훌륭하고 실용적인 책으로 추천합니다."

— 다니엘 툰켈랑(Daniel Tunkelang)/ Endeca의 수석 과학자이자 링크드인의 전 데이터 과학 및 엔지니어링 이사

"모든 산업이 디지털화되고 데이터 중심이 되면서 통제된 온라인 실험을 수행하고 그 혜택을 누리는 것이 필수가 됐습니다. 코하비와 탕, 쉬는 데이터 실무자와 경영진 모두에게 필요한 자료가 될 수 있는, 완벽하고 잘 연구된 가이드를 제공합니다."

— 에반젤로스 시무디스(Evangelos Simoudis)/ Synapse Partners 공동 설립자 겸 전무 이사, 『Big Data Opportunity in Our Driverless Future(이끄는 이가 없는 미래의 빅데이터 기회)』의 저자

"저자들은 10년 이상 동안 어렵게 얻은 실험 교훈을 가장 전략적으로 제공합니다."

— 콜린 맥팔랜드(Colin McFarland)/ Netflix의 실험 플랫폼 디렉터

"A/B 테스트에 대한 이 실용적인 가이드는 실험 실습의 최고 전문가 세 명의 경험을 이해하기 쉽게 추출했습니다. 올바른 지표 선택부터 기관 기억의 이점처럼 실험에서 가장 중요한 몇 가지 고려 사항을 알려줍니다. 과학과 실용성의 균형을 이루는 실험 코치를 찾고 있다면 이 책은 당신을 위한 것입니다."

— 딜런 루이스(Dylan Lewis)/ Intuit 실험 리더

"실험을 하지 않는 것보다 더 나쁜 것은 오해의 소지가 있는 실험을 하는 것입니다. 왜냐하면 그것은 여러분에게 잘못된 확신을 주기 때문입니다! 이 책은 세계 최대 규모의 실험 플랫폼들에서 나온 통찰력을 바탕으로 온라인 종합 대조 실험의 기술적 측면을 자세히 설명합니다. 어떤 규모라도 온라인 실험에 참여한다면 지금 이 책을 읽고 실수를 방지하고 결과에 대한 신뢰를 얻길 바란다."

— 크리스 고워드(Chris Goward)/ 『You Should Test That!(꼭 그 테스트를 해야 해요!)』의 저자, Widerfunnel의 설립자이자 CEO

"경이로운 책입니다. 저자들은 풍부한 경험을 바탕으로 읽기 쉬운 참고 자료를 만들었고 이는 포괄적이면서도 상세하게 설명합니다. 디지털 실험에 진지하게 임하는 분께 이 책을 적극 추천합니다."

— 피트 쿠멘(Pete Koomen)/ Optimely의 공동창업자

"저자들은 온라인 실험의 선구자입니다. 그들이 만든 플랫폼과 그들이 가능하게 만든 실험은 몇몇 가장 큰 인터넷 브랜드를 변화시켰습니다. 이들의 연구와 강연은 업계 전반의 팀들이 실험을 채택할 수 있도록 영감을 줬습니다. 이 책은 업계에서 기다려온 권위적이면서도 실용적인 텍스트입니다."

— 아이디얼 아이자즈(Adil Aijaz)/ Split Software의 공동 설립자이자 CEO

한국어판 추천의 글

IT 서비스에서 프로덕트를 성장시키고, 사용자를 더 잘 이해하고자 한다면 실험이 필수적이고, 신뢰성 있는 실험을 통해 올바른 결론을 얻는 것이 중요합니다. 하지만 실제 현업에서 신뢰성 있는 실험을 하기란 쉽지 않습니다. 실험에 대한 조직 구성원들의 이해, 실험 진행을 기술적으로 뒷받침할 수 있는 실험 플랫폼, 데이터와 지표의 신뢰성 등 많은 요소를 갖추고 있어야 신뢰성 있는 실험을 진행할 수 있습니다.

이 책은 신뢰성 있는 온라인 종합 대조 실험을 진행하기 위해 알아야 할 항목들을 많은 사례와 함께 설명합니다. 주로 실험의 진행 과정이나 통계 및 기술적인 측면을 다루지만, 실험을 잘 진행하기 위한 조직과 문화적인 부분도 함께 다룹니다. 개인적으로 이 책에 수록된 마이크로소프트 빙과 링크드인 등의 실제 사례에서 많은 도움을 받았는데, 이 사례들을 바탕으로 현업에서의 시행착오를 많이 줄일 수 있었습니다. 실험을 더 잘하고자 하거나 실험을 통해 서비스를 성장시키고 싶은 모든 분께 이 책을 추천드립니다.

— 허성연/ 소프트웨어 엔지니어, 당근마켓 데이터 가치화 팀

지은이 소개

론 코하비 Ron Kohavi

에어비앤비의 상무 이사이자 기술 펠로우이다. 이 책은 마이크로소프트의 기술 펠로우이자 본사 상무 이사로 있을 때 쓴 것으로, 그 이전에는 아마존의 데이터 마이닝 및 개인화 책임자였다. 스탠포드 대학교에서 컴퓨터 과학 박사 학위를 받았다. 그의 논문은 40,000개 이상의 인용됐으며, 그중 3개가 컴퓨터 과학에서 가장 많이 인용된 상위 1,000개의 논문에 속한다.

다이앤 탕 Diane Tang

대규모 데이터 분석 및 인프라, 온라인 종합 대조 실험 및 광고 시스템에 대한 전문 지식을 보유한 구글 연구원이다. 하버드 대학교에서 학사를, 스탠포드 대학교에서 석사 및 박사 학위를 취득했으며 모바일 네트워킹, 정보 시각화, 실험 방법론, 데이터 인프라, 데이터 마이닝, 대용량 데이터에 대한 특허를 받았으며, 관련 저서를 썼다.

야 쉬 Ya Xu

링크드인에서 데이터 과학 및 실험을 주관한다. 실험에 관한 여러 논문을 발표했으며, 탑티어 콘퍼런스와 대학에서 종종 발표를 한다. 이전에 마이크로소프트에서 일했으며 스탠포드 대학교에서 통계학 박사 학위를 받았다.

감사의 글

수년간 우리와 함께 일해준 동료들에게 감사를 표한다. 감사한 분들이 너무 많아서 일일이 이름을 댈 수 없지만, 이 책은 업계 전반의 사람들뿐만 아니라 온라인 종합 대조 실험을 연구하고 수행하는 것 이상의 합동 작업에 바탕을 두고 있다. 여러분 모두에게 많은 것을 배웠다. 다시 한 번 감사하다.

책을 쓰는 내내 편집자인 Lauren Cowles와 연락했다. Cherie Woodward는 훌륭한 편집으로 우리 셋의 목소리가 어우러질 수 있도록 도움을 줬다. Stephanie Grey는 우리와 함께 모든 다이어그램과 그림을 개선했다. Kim Vernon은 최종본을 편집 및 감수했다.

무엇보다 출간 작업으로 가족들과 함께할 시간을 갖지 못한 것을 이해해준 우리 가족에게 깊은 감사를 전한다. Ronny의 가족 Yael, Oren, Ittai, Noga, Diane의 가족 Ben, Emma, Leah, Ya의 가족 Thomas, Leray, and Tavis. 여러분의 성원이 없었다면 이 책을 쓸 수 없었을 것이다!

구글 팀: Hal Varian, Dan Russell, Carrie Grimes, Niall Cardin, Deirdre O'Brien, Henning Hohnhold, Mukund Sundararajan, Amir Najmi, Patrick Riley, Eric Tassone, Jen Gennai, Shannon Vallor, Eric Miraglia, David Price, Crystal Dahlen, Tammy Jih Murray, Lanah Donnelly과 구글에서 실험을 수행하는 모든 분들

링크드인 팀: Stephen Lynch, Yav Bojinov, Jiada Liu, Weitao Duan, Nanyu Chen, Guillaume Saint-Jacques, Elaine Call, Min Li, Arun Swami, Kiran Prasad, Igor Perisic 및 전체 실험 팀

마이크로소프트 팀: Omar Alonso, Benjamin Arai, Jordan Atlas, Richa Bahayan, Eric Boyd, Johnny Chan, Alex Deng, Andy Drake, Aleksander

Fabijan, Brian Frasca, Scott Gudit Gupta, Adam Gustson, Tommy Gufson, Tommy Guy, Randy Henne, Edward Jezierski, Jing Jin, Dongwoo Kim, Waldo Kuipers, Jonathan Litz, Sophia Liu, Jiannan Lu, Qi Lu, Daniel Miller, Carl Mitchell, Nils Pohlmann, Wen Qin, Thomas Schreiter, Harry Shum, Dan Sommerfield, Garnet Vaz, Toby Walker, Michele Zunker 및 분석 실험 팀.

책 전체를 살펴봐준 Maria Stone, Marcus Persson과 윤리에 관한 장에 대한 전문가 피드백을 준 Michelle N. Mey에게 특별히 감사드린다.

피드백을 준 다른 사람들은 다음과 같다. Adil Aijaz, Jonas Alves, Alon Amit, Kevin Anderson, Joel Barajas, Houman Bedayat, Beau Bender, Bahador Biglari, Stuart Buck, Jike Chong, Jed Chou, Pavel Dmitriev, Yurong Fan, Georgi Georgiev, Ilias Gerostathopoulos, Matt Gershoff, William Grosso, Aditya Gupta, Rajesh Gupta, Shilpa Gupta, Kris Jack, Jacob Jarnvall, Dave Karow, Slawek Kierner, Pete Koomen, Dylan Lewis, Bryan Liu, David Manheim, Colin McFarland, Tanapol Nearunchron, Dheeraj Ravindranath, Aaditya Ramdas, Andre Richter, Jianhong Shen, Gang Su, Anthony Tang, Lukas Vermeer, Rowel Willems, Yu Yang, Yufeng Wang

이름을 밝히지 않고 도와준 많은 분께도 감사드린다.

한국어판 감사의 글

보다 정확하고 바르게 전달될 수 있도록 한국어판 출간에 도움을 주신 김진구 박사님, 이명지 박사님, 이민용 박사님, 이민형 님, 황민하 박사님, 허성연 님께 특별히 감사드립니다.

— 편집 팀 일동

옮긴이 소개

이기홍

(keerhee@gmail.com)

카네기멜론대학교에서 석사 학위를 받았고, 피츠버그대학교 Finance Ph.D, CFA, FRM이자 금융, 투자, 경제 분석 전문가다. 삼성생명, HSBC, 새마을금고 중앙회, 한국투자공사 등과 같은 국내 유수의 금융기관, 금융 공기업에서 자산운용 포트폴리오 매니저로 근무했으며, 현재 딥러닝과 강화학습을 금융에 접목시켜 이를 전파하고 저변을 확대하는 것을 보람으로 삼고 있다. 저서(공저)로는 『엑셀 VBA로 쉽게 배우는 금융공학 프로그래밍』(한빛미디어, 2009)이 있으며, 번역서로는 『포트폴리오 성공 운용』(미래에셋투자교육연구소, 2010), 『딥러닝 부트캠프 with 케라스』(길벗, 2017), 『프로그래머를 위한 기초 해석학』(길벗, 2018)과 에이콘출판사에서 펴낸 『실용 최적화 알고리즘』(2020), 『초과 수익을 찾아서 2/e』(2020), 『자산운용을 위한 금융 머신러닝』(2021), 『존 헐의 비즈니스 금융 머신러닝 2/e』(2021), 『퀀트 투자를 위한 머신러닝 • 딥러닝 알고리듬 트레이딩 2/e』(2021), 『자동머신러닝』(2021), 『금융 머신러닝』(2022), 『퇴직 연금 전략』(2022) 등이 있다. 누구나 자유롭게 머신러닝과 딥러닝을 자신의 연구나 업무에 적용해 활용하는 그날이 오기를 바라며 매진하고 있다.

김기영
(kky416@gmail.com)

서울대학교 기계항공공학부를 졸업한 뒤 동대학원에서 유체역학, 응용수학 분야 연구로 박사 학위를 받았다. 이후 디지털 마케팅 플랫폼에서 인공지능 기반 웹데이터 분석 솔루션을 만들었으며 금융 데이터 분석 전문 회사에서 인공지능 연구소장으로 다양한 기업 및 결제 데이터를 분석하고 활용하는 일을 했다. 현재는 인공지능 기술을 개발 및 서비스하는 아티피셜 소사이어티Artificial Society의 대표로 메타버스와 헬스케어를 결합한 서비스를 운영하고 있다.

옮긴이의 말

이 책은 빙Bing에서의 광고 페이지 속 한 줄의 변화로 전체 매출의 10%를 향상시킨 놀라운 사례로 시작한다. 단순한 A/B 테스트의 기술서가 아닌 마이크로소프트, 구글, 링크드인에서 수년간 온라인 종합 대조 실험을 주도했던 저자들의 경험과 교훈을 공유하는 책이다.

숫자를 얻는 것은 쉽다. 하지만 믿을 수 있는 숫자를 얻는 것은 어렵다. 실험 결과를 단순한 숫자로 얻는 것이 아니라 믿을 수 있는 숫자를 얻을 수 있도록 신뢰도 높은 온라인 종합 대조 실험을 설계하고, A/B 테스트를 사용해 혁신을 가속화하는 방법에 대해 이야기하고 있다. 저자들은 현재 연간 20,000건 이상의 종합 대조 실험을 실행하는 각 기업에서의 경험을 바탕으로 학생과 업계 전문가가 실험을 시작할 수 있도록 예시, 빠지기 쉬운 실수 및 조언을 공유하고 있을 뿐만 아니라 데이터 기반 의사결정 방식을 개선하고자 하는 숙련된 실무자들에게도 도움될 심화 주제에 대해서도 깊이 있게 논의한다.

온라인 종합 대조 실험인 A/B 테스트는 2000년 중반부터 시작된 테크기업들의 문화적 혁신, 예를 들면 에릭 리스의 린 스타트업Lean Startup 및 MVP(최소 기능 제품)의 개념과 그 맥을 같이 한다. 기능 또는 서비스의 작은 변화를 지속적으로 온라인에서 테스트하면서 조금씩 기능 또는 서비스를 개선한다. A/B 테스트는 과거의 설계–개발–테스트–배포의 상품개발이 3년의 주기에 걸쳐 진행되던 행태를 해방시켜 이제는 빠르게 기능 또는 서비스를 출시하고 이를 지속적으로 테스트해 고객과 실시간으로 소통할 수 있도록 한다. 또한 이러한 점에서 A/B 테스트는 개인화된 서비스를 중시하는 여러 분야에서 전통적인 설문조사나 시장조사를 탈피해 실제 상황에서 고객의 취향을 가장 잘 반

영하는 서비스와 추천을 가장 효율적으로 달성시킬 수 있는 방법이기도 하다.

A/B 테스트의 기능 중 가장 중요한 것은 한꺼번에 모든 것을 테스트하는 것이 아니라 조금씩 테스트하는 것이다. 통제된 상황에서 이를 실행하는데, 이는 온라인상의 통제이므로 실제 상황을 반영하는 통제이다. 따라서 연관분석(예를 들어 장바구니 분석)에서와 같이 상관관계를 발견하는 것을 넘어서 원인과 결과를 밝히는 설명력이 가능한 인과성을 발견하고자 하는 것이 주된 목적이다. (그래서 그 테스트 대상은 작은 무엇인가가 될 것이다.) 이러한 발견은 보다 설명력이 있기 때문에 테스트의 승자가 상품 또는 서비스의 개선을 위해 큰 신뢰도로 강건하게 도입될 수 있는 것이다. 이에 대한 많은 예제가 이 책에 담겨있으니 독자들은 이를 즐기길 바란다. 1부는 통제실험의 설계, 구축, 과정 및 평가에 대한 기본적 설명이 포함돼 있으며, 2부부터는 데이터 기반의 문화를 도입하고, 신뢰도 있는 온라인 종합 대조 실험이 실제 상황에서 상시적으로 실행되기 위한 전반적인 지침과 교훈, 앞으로의 방향을 담고 있어 데이터 과학자들 및 관련 실무자들뿐 아니라 기업 내의 리더들 및 임원들에게도 크게 유용하리라 본다. 마지막으로 HiPPO(최고의 보수를 받는 최고 경영자)의 의견이 주도하던 시대를 넘어서 고객들의 의견을 중시하고 데이터를 기반으로 의사결정하는 문화를 구축하는 데 있어 획기적인 전기가 되기를 바란다.

차례

들어가며

데이터를 갖고 있다면 데이터를 살펴봐라.
만약 갖고 있는 것이 의견뿐이라면 자신의 의견대로 가라.
전 넷스케이프 CEO인 짐 박스데일

아마존과 마이크로소프트(Ron), 구글(Diane), 마이크로소프트와 링크드인(Ya)에서 수십 년 동안 온라인 종합 대조 실험online controlled experiment을 여러 규모로 실행해온 경험에서 얻은 실질적인 교훈을 공유하는 것이 이 책의 목표다. 구글, 링크드인 또는 마이크로소프트의 담당자가 아닌 한 개인으로써 이 책을 집필하는 동안, 수년간에 걸쳐 마주했던 주요 교훈과 함정을 선정해, HiPPOHighest Paid Person's Opinion: 최고 보수를 받는 자의 의견, 즉 최고 경영자의 의견에 의존하는 것이 아니라 그들에게 정보를 제공하는 데이터 기반 문화를 구축하기 위해 소프트웨어 플랫폼과 기업 문화 측면 모두에 대한 지침을 제공하고자 한다(Kohavi, HiPPO FAQ 2019). 이러한 많은 교훈이 온라인 환경, 대기업 또는 중소기업, 심지어 회사 내의 팀과 조직에도 적용된다고 믿으며, 우리 모두는 실험 결과의 신뢰도를 평가할 필요가 있다고 생각한다. 또한 트위먼Twyman의 법칙이 암시하는 회의론을 믿는다. **흥미로워 보이거나 다르게 보이는 어떤 수치는 대체로 틀린다.** 독자들이 다시 한 번 결과를 확인하고, 특히 획기적이고 긍정적인 결과를 위해 유효성 검사를 실행할 것을 권한다. 숫자를 얻는 것은 쉽지만, 믿을 수 있는 숫자를 얻는 것은 어렵다!

이 책에서는 다음 방법들을 배운다.

- 과학적 방법을 사용한 종합 대조 실험을 통해 가설 평가
- 주요 지표[1] 및 전반적인 평가 기준 정의
- 결과의 신뢰도를 검증하고, 실험자들에게 위배된 가정에 대한 경고 제공
- 실험 결과의 빠른 해석과 반복 실험
- 주요 사업목표를 보호하기 위한 가드레일 구축
- 실험의 한계 비용을 0에 가깝게 낮추는 확장 가능한 플랫폼 구축
- 이월 효과, 트위먼의 법칙, 심슨의 역설, 네트워크 상호작용과 같은 흔한 실수 피하기
- 일반적인 가정 위반을 포함해 통계 문제가 실제로 어떻게 처리되는지 이해

1부는 배경에 상관없이 모든 사람이 읽을 수 있도록 설계됐으며, 4장으로 구성돼 있다.

- 1장에서는 온라인 종합 대조 실험 실행의 이점을 간략히 설명하고 실험 용어를 소개한다.
- 2장에서는 예를 들어 엔드-투-엔드로 실험을 실행하는 과정을 살펴본다.
- 3장에서는 일반적인 오류 및 실험 신뢰도 구축 방법을 설명한다.
- 4장에서는 실험 플랫폼을 구축하고 온라인 실험을 확장하기 위해 필요한 사항을 개략적으로 설명한다.

2부에서 5부까지는 필요에 따라 모든 사람이 이용할 수 있지만, 각 장은 특정 청중에 초점을 두고 작성됐다. 2부에는 조직 지표와 같은 펀더멘털에 관한 5개의 장들이 포함돼 있다. 2부의 주제는 특히 리더와 임원 모두에게 권장된다. 3부의 2개 장은 리더, 데이터 과학자, 엔지니어, 분석가, 제품 관리자들이 온라인 종합 대조 실험을 보완하는 기법을 소개한다. 이 기법은 자

1 지표는 metrics로 메트릭, 지표 또는 성과지표와 혼용해서 사용되나 이 책에서는 특별히 구별되지 않는 한 지표로 통일한다. – 옮긴이

원 및 시간을 투자할 때, 유용한 지침으로 사용될 수 있을 것이다. 4부는 실험 플랫폼 구축에 중점을 두고 엔지니어를 대상으로 설명한다. 마지막으로, 5부는 고급 분석 주제를 파고들어 데이터 과학자를 대상으로 한다.

웹사이트 https://experimentguide.com은 이 책의 동반자이다. 여기에는 추가 자료와 오타 정보가 포함돼 있으며, 공개 토론의 영역을 제공한다. 저자들은 이 책의 모든 수익금을 자선단체에 기부할 계획이다.

1부

모두를 위한 소개

01

소개와 동기

하나의 정확한 측정이 수천 개의 전문가 의견보다 가치 있다.

그레이스 호퍼 제독Admiral Grace Hopper

2012년 마이크로소프트의 검색 엔진인 빙Bing에서 일하는 한 직원이 광고 헤드라인 표시법을 바꾸자고 제안했다(Kohavi, Thomke 2017). 아이디어는 그림 1.1과 같이 타이틀 라인을 타이틀 바로 밑 첫째 줄의 문장과 합쳐서 광고의 타이틀 줄을 길게 만드는 것이었다.

아무도 이 간단한 변화가 수많은 아이디어 중 빙 역사상 최고의 매출 창출 아이디어가 될 줄 몰랐다!

아이디어의 우선순위는 낮았고 코드를 바꾸는 것이 쉽다는 전제하에 소프트웨어 개발자가 변화를 한 번 시도하자고 하기 전까지 6개월 이상 밀려 있었다. 개발자는 아이디어를 구현하고, 실제 사용자 일부에게 새로운 타이틀 배포를, 또 다른 일부에게는 이전의 것을 무작위로 보여주면서 평가하기 시작했다. 웹사이트에 대한 사용자 반응은 광고 클릭과 이를 통한 매출을 포함해 기록됐다. 이것은 A와 B 또는 대조군control과 실험군treatment의 2개 종류를 비교하는 가장 간단한 형태의 종합 대조 실험controlled experiment인 A/B 테스트의 예시다.

테스트를 시작한 몇 시간 후 매출이 너무 많다는 경고가 발생했으며, 이는 보통 무엇인가 실험에 잘못된 것이 있다는 의미다. 새로운 타이틀이 배치된 실험군이 광고 수입을 너무 많이 증가시켰던 것이다. "사실이기에 너무 좋다"

고 하는 경고는 로그를 두 번 카운트하거나(더블 빌링) 다른 웹 광고가 뜨지 않고 단지 해당 광고만 표시되는 것과 같은 심각한 버그를 발견할 수 있기 때문에 매우 유용하다.

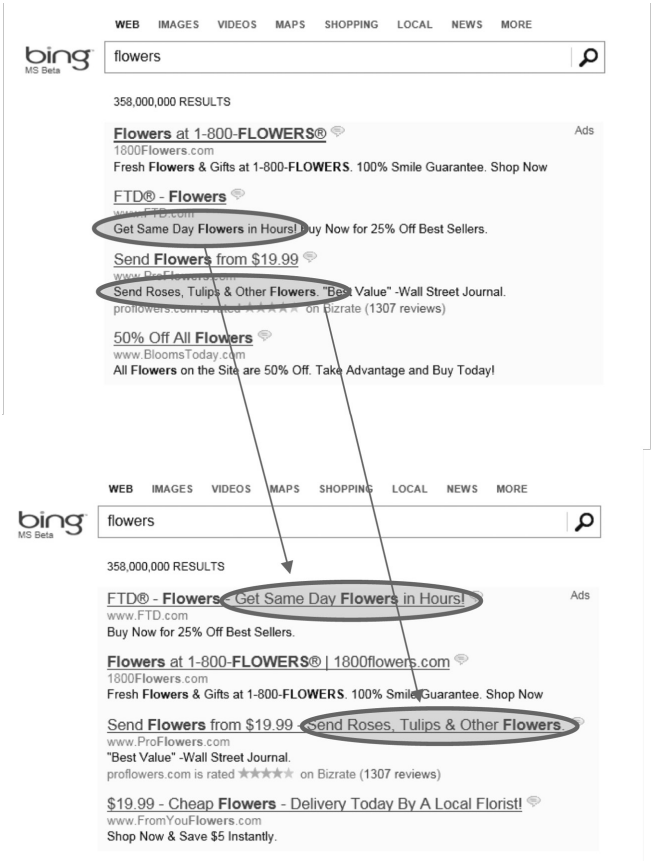


그림 1.1 빙에서의 광고 배포 변화 실험

그러나 이 실험에서의 매출 증가는 유효한 것이었다. 빙의 매출은 12%나 증가했으며 당시 미국에서만 연간 1억불에 해당하는 것이었다. 다른 주요 사용자 경험 지표를 저해하지도 않았다. 이 실험은 오랜 기간 동안 여러 번 되풀이됐다.

이 예시를 통해 온라인 종합 대조 실험의 핵심 주제를 알 수 있다.

- 아이디어의 가치는 평가하기 힘들다. 연 1억불 이상의 가치를 가진 간단한 변화가 여러 달 연기됐다.
- 작은 변화도 큰 영향을 줄 수 있다. 한 개발자가 며칠 작업하고, 연 1억불의 수익을 낸다면 최고의 투자수익률(ROI)을 올린 것 아닌가?
- 큰 영향을 발생시키는 실험은 드물다. 빙은 1년에 10,000건의 실험을 수행하지만, 큰 개선을 주는 간단한 아이디어는 단지 수년에 한 번 발생한다.
- 실험을 실행하는 오버헤드는 적어야 한다. 빙의 개발자들은 마이크로소프트의 실험 시스템인 ExP에 접근해, 과학적으로 아이디어를 쉽게 평가할 수 있었다.
- 전체 평가 기준(OEC)이 분명해야 한다. 앞의 예시에서는 매출(revenue)이 OEC의 핵심 요소였지만 매출만으로는 OEC가 되기는 부족하다. 매출을 위해 사용자 경험을 저해하는 것으로 알려진 광고를 웹사이트에 도배할 수 있지만, 빙은 사용자당 세션(사용자들이 떠나는지 아니면 더 사용하는지)과 여러 다른 요소를 포함해 사용자 경험 지표를 비교해 매출을 판단하는 OEC를 사용한다. 여기서의 핵심은 수익이 극적으로 늘었어도 사용자 경험 지표가 크게 저하되지 않았다는 점이다.

이제 종합 대조 실험의 용어를 알아본다.

온라인 종합 대조 실험 용어

종합 대조 실험controlled experiment은 길고도 매혹적인 역사를 갖고 있으며 이에 대한 우리의 연구는 온라인에서 공유하고 있다(Kohavi, Tang, Xu 2019). 종합 대조 실험은 때때로 A/B 테스트, A/B/n 테스트(다중 변형을 강조하기 위함), 현장 실험, 무작위 종합 대조 실험, 분할 테스트, 버킷 테스트, 및 플라이트라고 불린다. 이 책에서는 이런 여러 변형에 관계없이 종합 대조 실험과 A/B 테스트라는 용어를 섞어 사용할 것이다.

온라인 종합 대조 실험은 에어비앤비, 아마존, 부킹닷컴, 이베이, 페이스북, 구글, 링크드인, 리프트, 마이크로소프트, 넷플릭스, 트위터, 우버, 야후/

오쓰, 얀덱스(Gupta et al. 2019)와 같은 기업에서 많이 사용된다. 이러한 회사는 매년 수천에서 수만 개의 실험을 실행하며, 때로는 수백만 명의 사용자와 연관돼 모든 것을 테스트한다. 사용자 인터페이스(UI), 관련 알고리즘(검색, 광고, 개인 정보 확인, 추천 상품 등), 지연 시간/성능, 콘텐츠 관리 시스템, 고객 지원 시스템 등을 실험하며, 웹사이트, 데스크톱 애플리케이션, 모바일 애플리케이션, 이메일 등 여러 채널에서 실행된다.

가장 일반적인 온라인 종합 대조 실험에서 사용자는 실험군과 대조군에 무작위로 분할되며, 한 번 지정된 분할은 바뀌지 않는다(따라서 사용자는 여러 번의 사이트 방문에서 동일한 경험을 하게 된다). 앞서 빙의 예에서 대조군Control은 광고의 원래 표시였고, 실험군Treatment은 더 긴 제목을 가진 광고의 표시였다. 사용자의 빙 웹사이트에서의 사용내역을 관찰하고 기록해, 기록된 데이터로부터 지표를 계산해 각 지표에서 대한 광고 간의 차이를 평가할 수 있다.

가장 간단한 종합 대조 실험에서는 그림 1.2와 같이 대조군(A)과 실험군(B)의 두 가지 변형군이 있다.

이 책에서는 Kohavi, Longbottom(2017)과 Kohavi, Longbottom et al.(2009)의 용어를 따르며, 다른 분야와 관련한 용어에 관해서는 다음 페이지에서 설명할 것이다. 이 장의 끝에 있는 참고문헌에서 실험과 A/B 테스트에 관한 다른 자료를 확인할 수 있다.

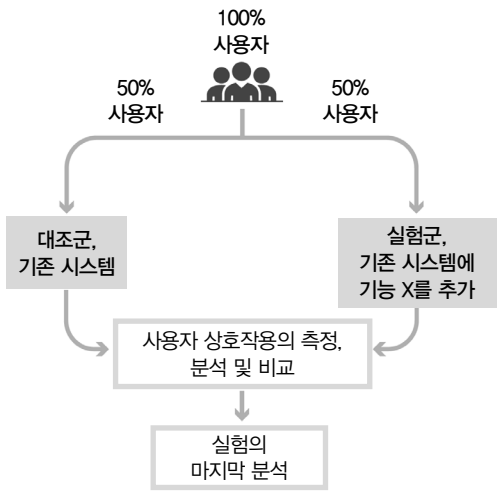


그림 1.2 간단한 종합 대조 실험: A/B 테스트

전체 평가 기준OEC, Overall Evaluation Criterion: 실험 목적의 계량적 지표. 예를 들어 OEC를 사용자별 활동일 수active days per user로 정할 수 있는데, 이는 실험 중 사용자가 활동한 일 수(사용자가 방문해서 어떤 행동을 취한 일 수)를 가리킨다. 이 OEC를 늘린다는 것은 사용자들이 해당 사이트를 더 자주 방문한다는 것을 의미하며, 이는 좋은 결과물이다. OEC는 단기적으로(실험 기간 동안) 측정할 수 있어야 하면서도 결과적으로는 장기적으로 봤을 때 전략적인 목표를 추진하는 것이 돼야 한다(이 장의 뒷부분과 7장의 전략, 전술 및 실험과의 관계 참조). 검색 엔진의 경우 OEC는 사용량usage(예: 사용자당 세션 수), 관련성relevance(예: 성공적인 세션, 성공까지의 시간), 광고 수익의 조합이다(모든 검색 엔진이 이러한 모든 지표 또는 단지 이런 지표만을 사용하는 것은 아니다).

이는 통계에서 흔히 반응Response 또는 종속Dependent 변수라고 불리며(Mason, Gunst, Hess 1989, Box, Hunter, Hunter 2005), 결과Outcome, 평가Evaluation, 적합도 함수Fitness Function가 동의어로 사용될 수 있다(Quarto-vonTivadar 2006). 단일 지표를 선택하는 것이 매우 바람직하고 권장되지만(Roy 2001, 50, 405-429), 실험은 여러 목표를 가질 수 있으며 밸런스 스코어 카드 접근 방식(Kaplan, Norton 1996)을 사용할 수 있다.

실험을 위한 OEC를 결정하는 방법은 7장에서 자세히 알아본다.

파라미터: OEC 또는 기타 관심 지표에 영향을 미치는 것으로 간주되는 통제 가능한 실험 변수. 파라미터는 때때로 요인factors 또는 변수variables라고도 불린다. 파라미터에는 값이 할당되는데, 이를 수준level이라고 한다. 단순 A/B 테스트에서는 일반적으로 두 개의 값을 갖는 단일 파라미터가 있다. 온라인 세계에서는 여러 값(예를 들면, A/B/C/D)을 가진 단일변수 설계를 사용하는 것이 일반적이다. 다변수 테스트MVTs, Multivariate Tests는 글꼴 색상과 글꼴 크기와 같은 다중 파라미터(변수)를 함께 평가해 실험자가 파라미터가 상호작용할 때의 전역적 최적값을 발견할 수 있도록 한다(4장 참조).

변형군: 일반적으로 파라미터에 값을 할당해서 테스트하는 사용자 경험. 간단한 A/B 테스트에서 A와 B는 보통 대조군과 실험군이라고 불리는 두 가

지 변형군이 있다. 일부 문헌에서 변형군은 오직 실험군만을 의미하며, 대조군을 특별한 변형군, 즉 비교를 할 대상이 되는 기존 버전이라고 간주한다. 예를 들어 실험에서 버그가 발견된 경우 실험을 중단하고 모든 사용자를 대조군에 할당해야 할 것이다.

무작위 추출 단위: 준무작위 추출(예: 해싱) 과정을 각 단위(예: 사용자 또는 페이지)에 적용해 변형군에 할당한다. 높은 확률로 인과관계를 판별할 수 있도록 상이한 변형군에 할당된 사용자들이 통계적으로 비슷해야 하므로, 적절한 무작위 추출은 중요하다. 단위를 지속적이고 독립적인 방식으로 변형군에 할당해야 한다(즉, 사용자가 무작위 추출 단위인 경우 사용자는 동일한 경험을 일관되게 가져야 하며, 한 사용자의 한 변형군에의 할당은 다른 사용자의 동일 변형군에의 할당과 아무런 관련이 없어야 한다). 온라인 종합 대조 실험을 실행할 때, 사용자를 무작위 추출 단위로 사용하는 것은 매우 일반적이며, 적극 권장한다. 일부 실험 설계에서는 페이지, 세션 또는 사용일별로 무작위 추출을 선택하는데, 사용자 경험이 서버에 의해 정의된 매 24시간 동안 일관성을 유지하도록 주어져야 한다. 자세한 내용은 14장을 참조하자.

적절한 무작위 추출이 중요하다. 실험 설계가 각 변형군에 동일한 비율의 사용자를 할당하는 경우, 각 사용자는 각 변형군에 할당될 확률이 같아야 한다. 무작위 추출을 가볍게 생각하지 말자. 다음 소개하는 예시는 적절한 무작위 추출의 어려운 점과 중요성을 보여준다.

- 랜드 법인은 1940년대 몬테카를로 방식에 난수가 필요했기 때문에 맥박 기계를 이용해 생성된 백만 개의 난수를 책으로 만들었다. 그러나 하드웨어의 편향(bias)으로 인해 원래 표에는 현저한 편향이 있었고, 새로 펴낸 책에서 숫자를 다시 무작위 추출해야 했다(RAND 1955).
- 종합 대조 실험은 처음에 의료 분야에서 사용됐다. 미국 재향군인청(VA)이 결핵에 대한 스트렙토마이신(streptomycin) 실험(제약 실험)을 실시했으나 의사들이 내과적으로 편견을 갖고 선발 과정에 개입했기 때문에 실험은 실패했다(Marks 1997년). 영국에서 유사한 실험이 블라인드 프로토콜로 행해졌는데 이는 성공했으며, 종합 대조 실험의 역사에서 분수령이 되는 순간(watershed moment)을 만들었다(Doll 1998).

어떠한 요인도 변형군 배정에 영향을 주도록 허용해서는 안 된다. 사용자(단위)는 "어느 오래된 어느 방식"으로 분배될 수 없다(Weiss 1997). 랜덤성은 "마구잡이식이거나 계획되지 않은" 것이 아니라, "확률에 기초한 의도적인 선택"을 의미한다는 점에 유의해야 한다(Mosteller, Gilbert, McPeek 1983). Senn(2012)은 무작위 추출에 대한 몇 가지 잘못된 생각에 대해 논의한다.

실험의 이유? 상관관계, 인과관계, 신뢰성

매달 사용자의 X%가 이탈(가입 종료)하는 넷플릭스와 같은 구독 사업에서 일하고 있다고 가정하자. 당신은 새로운 기능을 도입하기로 결정했고 그 기능을 사용하는 사용자의 이탈률이 X%/2, 즉, 반이라는 것을 보게 된다. 당신은 이에 대해 인과관계를 주장하고 싶을지도 모른다. 기능은 이탈을 절반으로 줄이고 있다. 이 기능을 더 쉽게 검색하고 더 자주 사용하게 할 수 있으면, 가입자가 급증할 것이라는 결론을 얻을 것이다. 그러나 이는 잘못된 논리다! 주어진 데이터에서 이 기능이 사용자 이탈을 감소시키는지 증가시키는지에 대한 결론을 내릴 수 없으며, 그 기능으로 인해 사용자 이탈을 감소시키는 것과 증가시키는 것 두 상황 모두 가능하다.

이러한 오류를 보여주는 예는 또 다른 구독 사업인 마이크로소프트 오피스 365에서 발견할 수 있다. 오류 메시지가 뜨고 충돌을 경험한 오피스 365 사용자가 이탈률이 낮다고 하더라도, 오피스 365에 오류 메시지를 더 많이 표시하거나 마이크로소프트에서 굳이 코드 품질을 낮춰 더 많은 충돌을 유발해야 하는 것은 아니다. 세 가지 이벤트 모두 사용률usage이라는 한 가지 요인에 의해 발생한다. 제품을 많이 사용하는 사용자는 오류 메시지가 더 많이 표시되고, 충돌이 더 많이 발생하며, 이탈률이 더 낮다. 상관관계는 인과관계를 의미하지 않으며 이러한 관찰에 지나치게 의존하면 잘못된 결정을 내리는 것으로 이어질 수 있다.

1995년에 가이야트와 연구진(Guyatt et al., 1995)은 의학 문헌에서의 추

천을 평가하는 방법으로 증거 계층hierarchy of evidence을 도입했으며, 그린할 Greenhalgh은 증거 기반 의학의 실행에 대한 논의(1997, 2014)에서 이를 확장했다. 그림 1.3은 바일라(Bailar, 1983, 1)에 근거하는 간단한 증거 계층을 보여준다. 무작위 추출 종합 대조 실험은 인과관계를 확립하는 최고의 기준이다. 종합 대조 실험의 체계적 검토, 즉 메타분석은 더 많은 증거와 일반화 가능성을 제공한다.

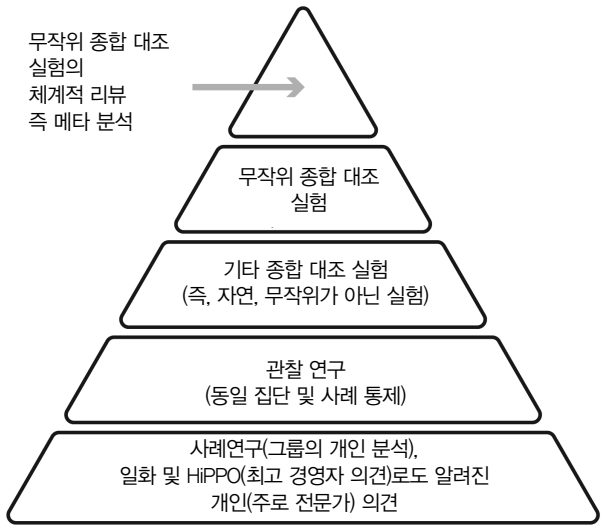


그림 1.3 임상시험 설계의 품질을 평가하기 위한 간단한 증거 계층(Greenhalgh 2014)

옥스포드 증거 기반 의학 센터Oxford Centre for Evidence-based Medicine의 증거 수준Levels of Evidence과 같은 좀 더 복잡한 모델도 이용할 수 있다(2009).

우리 회사들에서 사용하는 실험 플랫폼은 구글, 링크드인, 마이크로소프트의 실험자들이 그 결과에 대한 높은 신뢰도를 갖고 연간 수만 건의 온라인 실험들을 실행할 수 있게 해준다. 우리는 온라인 종합 대조 실험이 다음과 같은 것이라고 믿는다.

- 높은 확률로 인과관계를 확립할 수 있는 최선의 과학적인 방법이다.

- 시간에 따른 변화와 같이 다른 기법으로 탐지하기 어려운 작은 변화 감지가 가능하다(민감도).
- 예상치 못한 변화 감지가 가능하다. 종종 과소평가되지만 성능 저하, 충돌/오류 증가, 또는 다른 기능으로부터의 클릭 수 감소 등 많은 실험에서 여러 지표에 대한 놀라운 영향을 발견한다.

이 책의 핵심은 실험에서의 잠재적 오류를 알아보고 결과에 대한 신뢰도를 높이는 방법을 알려주는 것이다. 온라인 종합 대조 실험은 신뢰할 수 있는 데이터를 대규모로 전자적으로 수집하고, 적절하게 무작위 추출하며, 함정을 피하거나 탐지하는 데 있어 최고의 능력을 제공한다(11장 참조). 온라인 종합 대조 실험이 불가능한 경우에만 신뢰도가 떨어지는 다른 방법(관찰 연구 등)을 사용하길 권장한다.

유용한 종합 대조 실험 실행을 위한 필수 재료

종합 대조 실험의 과학적인 엄격함으로 모든 결정을 내릴 수 있는 것은 아니다. 예를 들어 인수합병(M&A)과 이에 대한 반대되는 일(인수합병이 없는 사건)을 동시에 진행할 수 없기 때문에 인수합병에 대해 종합 대조 실험을 실행할 수 없다. 이제 유용한 대조 실험을 실행하는 데 필요한 기술적 요소들을 검토하고(Kohavi, Crook, Longbotham 2009), 그 다음에 조직의 원칙을 살펴본다. 4장에서 실험 성숙도 모델을 다룬다.

1. 서로 간의 간섭 효과 없이 여러 변형군에 할당될 수 있는 실험 단위(예: 사용자). 예를 들어, 실험군의 사용자들은 대조군의 사용자들에게 영향을 미치지 않는다(22장 참조).
2. 종합 대조 실험에 충분한 실험 단위(사용자 수). 통제된 경험이 유용하려면 수천 개의 실험 단위를 권장한다. 즉 숫자가 클수록 더 작은 효과도 탐지할 수 있다. 좋은 소식은 작은 소프트웨어 스타트업들조차도 일반적으로 초반에 실험에 필요한 충분한 수의 사용자들을 빨리 얻으므로, 큰 효과를 찾기 위한 종합 대조 실험을 시작할 수 있다는 것이다. 비즈니스가 성장함에 따라 작은 변화(예: 대규모 웹사이트는 사용자 경험과 수익률

변화율에 영향을 미치는 주요 지표에 대한 작은 변화를 감지할 수 있어야 함)를 탐지하는 것이 더욱 중요해지고, 민감도는 사용자 기반이 증가함에 따라 개선된다.

3. 실제로 평가 가능하면서도 다른 사람도 동의하는 핵심지표(이상적인 OEC). 목표를 측정하기가 너무 어렵다면 대체 지표에 대해 합의하는 것이 중요하다(7장 참조). 신뢰할 수 있는 데이터를 이상적이면서도 광범위하고 저렴하게 수집할 수 있다. 소프트웨어 분야에서는 일반적으로 시스템 이벤트와 사용자 액션을 기록하는 것이 쉽다(13장 참조).

4. 용이한 변경. 소프트웨어는 일반적으로 하드웨어보다 변경하기 쉽다. 그러나 소프트웨어에서도 일부 영역은 일정 수준의 품질 보증을 요구한다. 추천 알고리즘은 변경하기도 쉽고 변경한 것을 평가하기도 쉽다. 항공기 비행 통제 시스템의 소프트웨어를 변경할 때는 연방항공관리국(FAA)의 전혀 다른 승인 과정이 필요하다. 서버 측 소프트웨어는 클라이언트 측보다 훨씬 쉽게 변경할 수 있기 때문이다(12장 참조). 클라이언트 소프트웨어에서 서비스를 호출하는 것이 일반화되고 있고, 업그레이드와 서비스 변경은 보다 신속하게 수행되고 종합 대조 실험을 사용할 수 있게 된다.

대부분의 중요한 온라인 서비스는 종합 대조 실험에 기반한 민첩한 개발 프로세스를 실행하는 데 필요한 구성 요소를 충족하거나 충족시킬 수 있다. 소프트웨어+서비스의 많은 구현도 비교적 쉽게 요구사항을 충족할 수 있다. 톰케Thomke는 조직이 "혁신 시스템"과 함께 사용될 때 실험에서 최대한으로 이익을 얻을 것이라고 한다(Thomke 2003). 신속한 소프트웨어 개발은 그러한 혁신 시스템이다.

종합 대조 실험이 불가능할 경우 모델링을 수행할 수 있으며 다른 실험 기법을 사용할 수 있다(10장 참조). 핵심은 종합 대조 실험이 실행될 수 있다면 그것들은 변화를 평가하기 위한 가장 신뢰할 수 있고 민감한 메커니즘을 제공한다는 것이다.

원칙

온라인 제어 실험을 실행하려는 조직에게 도움될 세 가지 핵심 원칙이 있다(Kohavi et al. 2013).

1. 조직은 데이터 중심 결정을 내리고 OEC를 공식화한다.
2. 조직은 종합 대조 실험을 실행하고 그 결과가 신뢰할 수 있는지 확인하기 위해 인프라와 실험에 기꺼이 투자한다.
3. 조직은 아이디어의 가치를 평가하는 데 서툴다는 것을 인지한다.

원칙 1. 조직은 데이터 중심 결정을 내리고 OEC를 공식화한다.

어떤 조직의 수장이 데이터 중심적으로 생각하고 싶지 않다고 하는 것은 거의 볼 수 없다(켄 세갈이 "우리는 단 한 건의 광고도 테스트하지 않았다. 인쇄물, TV, 광고판, 웹, 리테일 또는 그 어떤 것에도 대해서도."(Segall 2012, 42)라고 주장한 스티브 잡스 집권하의 애플이 눈에 띄는 예외다). 그러나 새로운 기능을 도입했을 때 이에 대한 사용자의 추가적 편익을 측정하는 것은 비용이 들고, 객관적인 측정에 따르면 일반적으로 성과(지표의 개선 정도)가 처음에 계획한 것만큼 낙관적이지 않은 경우가 많다. 많은 조직에서 성과를 정의하고 측정하는 데 드는 자원을 투자하지 않을 것이다. 많은 경우, 새로운 기능이 핵심 지표에 긍정적인 영향을 미치는지 여부를 무시하고, 그저 계획을 짜고 실행해 "실행된 계획의 비율"로 성공을 측정하는 것이 쉽다.

데이터 중심적이 되려면 조직은 비교적 짧은 기간(예: 1~2주)에 걸쳐 쉽게 측정할 수 있는 OEC를 정의해야 한다. 대규모 조직에는 여러 개의 OEC 또는 주요 지표가 있을 수 있으며, 이들은 분야 간에 공유되며 분야별로 적절히 수정돼 사용된다. 어려운 부분은 단기간에 측정할 수 있고, 차이를 보일 정도로 민감하며, 장기 목표를 예측할 수 있는 지표를 찾아내는 것이다. 예를 들어

단기적인 수단(예: 가격 인상)은 단기 이익을 증가시킬 수 있지만 장기적으로는 오히려 이를 해칠 수 있기 때문에 "이익"은 좋은 OEC가 아니다. 고객생애가치는 전략적으로 강력한 OEC이다(Kohavi, Long-bottom et al. 2009). 조직이 전사적으로 이해관계가 일치하는 좋은 OEC에 동의하는 것의 중요성은 아무리 강조해도 지나치지 않다. 6장을 참조하라.

"데이터 정보에 기반data-informed" 또는 "데이터 인지data-aware"라는 용어는 의사결정이 단일 데이터 출처(예: 종합 대조 실험)에 의해 내려지지 않았음을 의미하기 위해 가끔 사용된다(King, Churchill, Tan 2017, Knapp et al. 2006). 이 책에서 데이터 주도data driven와 데이터 정보 기반data-informed을 동의어로 사용한다. 궁극적으로 종합 대조 실험, 조사, 새로운 코드의 유지보수 비용 추정 등을 포함한 다양한 데이터를 기반으로 의사결정이 이뤄져야 한다. 데이터 주도 조직이나 데이터에 정통한 조직은 직관에 의존하지 않고 관련 데이터를 수집해 의사결정을 추진하고 최고 경영자HiPPO, Highest Paid Person's Opinion에게 정보를 제공한다(Kohavi 2019).

원칙 2. 조직은 종합 대조 실험을 실행하고 그 결과가 신뢰할 수 있는지 확인하기 위해 인프라와 테스트에 기꺼이 투자할 용의가 있다.

온라인 소프트웨어 영역(웹사이트, 모바일, 데스크톱 애플리케이션 및 서비스)에서는 소프트 엔지니어링을 통해 종합 대조 실험에 필요한 조건을 충족할 수 있다(유용한 종합 대조 실험 실행에 필요한 필수 요소 참조). 즉 사용자를 안정적으로 무작위화할 수 있으며 원격 측정도 가능하고 새로운 기능과 같은 소프트웨어 변경사항을 도입하기가 매우 쉽다(4장 참조). 비교적 작은 웹사이트라도 필요한 통계적 테스트를 실행할 수 있는 충분한 사용자는 있다(Kohavi, Crook, Longbotham 2009).

종합 대조 실험은『린 스타트업』(Ries 2011)의 에릭 리스Eric Ries가 널리 알린

애자일[1] 소프트웨어 개발(Martin 2008, K. S. Rubin 2012), 고객 개발 프로세스(Blank 2005), MVP(Minimum Viable Products)와 결합할 때 특히 유용하다.

다른 분야에서는 종합 대조 실험을 신뢰성 있게 실행하는 것이 어렵거나 불가능할 수 있다. 의료기관에서의 종합 대조 실험에서 필요한 일부 개입은 비윤리적이거나 불법적일 수 있다. 하드웨어 장치는 제조에 대한 리드 타임(개발 시간)이 길거나 수정이 어려울 수 있으므로 사용자에 대한 종합 대조 실험은 거의 새로운 하드웨어 장치(예: 새로운 휴대전화)에서 실행되지 않는다. 이러한 상황에서는 종합 대조 실험을 실행할 수 없을 때 보완 기술(10장 참조)과 같은 다른 기법이 필요할 수 있다.

여러분이 종합 대조 실험을 할 수 있다고 가정한다면 실험에 대한 신뢰성을 확보하는 것이 중요하다. 온라인 실험을 할 때 숫자를 얻는 것은 쉽다. 하지만 신뢰할 수 있는 숫자를 얻는 것은 어렵다. 신뢰할 수 있는 결과에 대해서는 3장에서 다룬다.

원칙 3. 조직은 아이디어의 가치를 평가하는 데 서툴다는 것을 인지한다.

팀에서는 각 기능이 유용하다고 생각해 개발하지만 많은 곳에서 대부분의 아이디어는 핵심 지표를 개선하는 데 실패한다. 마이크로소프트에서 시험한 아이디어의 3분의 1만이 개선을 위해 고안된 지표를 개선했다(Kohavi, Crook 및 Longbotham 2009). 빙이나 구글과 같이 최적화된 도메인에서는 성공률이 더 낮다. 일부 지표의 성공률은 약 10 – 20%이다(Manzi 2012).

슬랙의 제품 및 라이프사이클 담당 이사인 파리드 모사밧(Fareed Mosavat)은 그의 트위터에서 슬랙의 경험으로부터 수익화 실험의 약 30%만이 긍정적인 결과를 보여준다고 밝혔다. “실험을 주도하는 팀에 있다면 최소한 70% 이상의

1 애자일(agile)은 신속함을 의미하며, 이는 린(lean) 방법론 및 MVP(최소 기능 제품)과 일맥상통한다. – 옮긴이

작업이 버려지는 것에 익숙해져야 한다. 이에 따라 프로세스를 구축해야 한다."(Mosavat 2019)

아비나쉬 카우쉭Avinash Kaushik은 실험 및 테스트 입문서(Kaushik 2006)에서 "고객의 요구가 무엇인지에 대해 우리의 80%가 잘못 알고 있다"고 썼다. 마이크 모란(Mike Moran 2007, 240)은 넷플릭스는 그들이 시도하는 것의 90%를 틀린 것으로 간주한다고 한다. Quicken Loss의 레지스 하디아리스Regis Hadiaris는 "수년 동안 테스트를 수행해왔지만, 결과의 추측은 메이저리그 야구 선수가 공을 치는 것을 추측하는 것만큼밖에 정확하지 않다"고 기술한다. 그렇다. 5년 동안 이 일을 해왔는데, 테스트의 결과를 33% 정도만 '알아맞힐 수 있다!'(Moran 2008) Etsy의 댄 매킨리Dan McKinley(McKinley 2013)는 "거의 모든 것이 실패한다"고 썼고, 기능에 대해서는 "첫 번째 시도에서 성공하는 것이 얼마나 드문 일인지 겸허하게 받아들여야 한다. 이런 경험이 보편적이라고 굳게 믿지만, 보편적으로 잘 인식되고 있지 않은 것 같다"고 기술하고 있다. 마침내 콜린 맥팔랜드는 『Experiment!(A/B 테스트를 통한 웹사이트 전환율 최적화)』이라는 책에서 이렇게 썼다!(2012) "아무리 쉽다고 생각하든 깊이 연구하든 경쟁자가 얼마나 많이 하고 있든 실험 아이디어는 생각보다 더 자주 실패한다."

모든 도메인이 그런 열악한 통계를 갖고 있는 것은 아니지만, 고객 대면 웹사이트와 애플리케이션에서 종합 대조 실험을 실행한 대부분의 사람들은 이런 겸손한 현실을 경험했다. 즉 우리는 아이디어의 가치를 평가하는 데 서툴다.

시간에 따른 개선

실제로 주요 지표의 개선은 0.1%~2%의 수많은 작은 변화로 달성된다. 많은 실험은 사용자들 일부에게만 영향을 미치기 때문에 10%의 사용자에게 5%의 개선을 했다면 그 영향은 희석되며(모집단이 나머지 사용자와 유사한 경

우, 10% x 5% = 0.5%), 따라서 전체 사용자 관점에서 그 영향은 훨씬 적다. 3장을 참조하라. 알 파치노가 영화 〈Any Given Sunday〉에서 말한 것처럼 "...승리는 조금씩 이루어진다."

구글 광고 사례

2011년, 구글은 1년 이상의 점진적 개발 실험development and incremental experiment을 거쳐 향상된 광고 순위 메커니즘을 출시했다(Google 2011). 엔지니어들은 광고 경매 자체에 대한 변화뿐만 아니라 기존 광고 순위 매커니즘 내에서 광고의 품질 점수를 측정할 수 있는 새롭고 개선된 모델을 개발하고 실험했다. 그들은 광고주들에게 미치는 영향을 더 깊이 이해하기 위해 모든 시장에 걸쳐 수백 개의 대조 실험과 여러 번 반복 시행을 실행했다. 이러한 대규모 백엔드 변화(와 종합 대조 실험)는 궁극적으로 어떻게 다양한 변화에 대한 계획과 계층화가 고품질의 광고를 제공함으로써 사용자의 경험을 개선하고, 동시에 이러한 고품질 광고을 제공하는 평균 가격을 낮춰 광고주의 경험도 개선하는지를 보여준다.

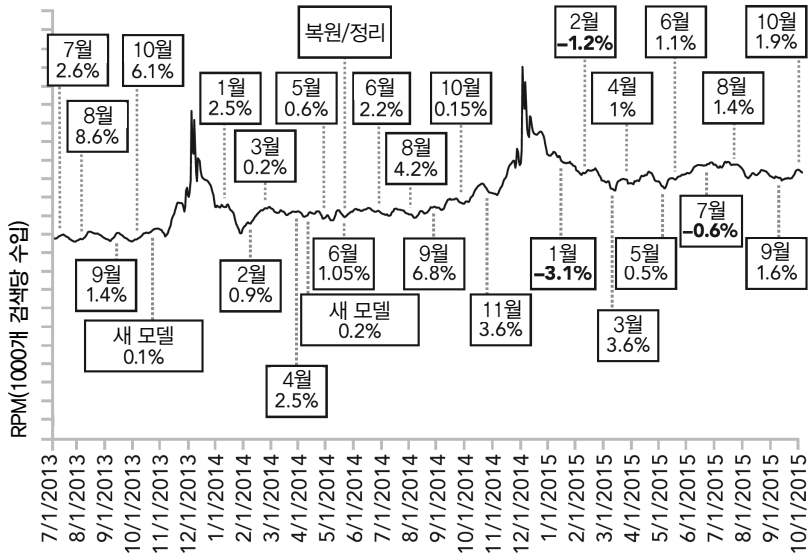
빙의 관련성 팀 사례

빙의 관련성 팀Relevance Team은 단일 OEC 측정지표를 매년 2%씩 개선하는 임무를 맡은 수백 명의 사람들로 구성돼 있다. 2%는 연간 사용자에게 실행되는 모든 대조 실험에서의 실험 효과의 합계(즉, OEC의 델타)이다. 그 팀은 수천 개의 실험을 진행하며, 그중 일부 실험은 우연히 긍정적으로 보일 수 있기 때문에(Lee, Shen 2018) 2%에 대한 공헌도는 반복 실험(뒤에 나오는 인증 실험)을 기반으로 부여된다. 여러 번의 반복 시행과 조정 후에 아이디어가 성공적으로 구현되면 인증 시험certification experiment이 단일 실험군에 대해 실행된다. 이 인증 실험의 실험군 효과가 2% 목표에 대한 공헌도를 결정한다. 최근에는

더 정밀한 공헌도 결정을 위해 베이지안 축소법bayesian shrinkage를 사용해 실험군 효과를 추정하는 법도 제안되고 있다(Coey, Cunningham 2019).

빙 광고 팀 사례

빙의 광고 팀Ads team은 꾸준히 연간 15%의 매출(eMarketer 2016)을 성장시켰지만 대부분의 개선이 조금씩 이뤄졌다. 매달 그림 1.4와 같이 많은 실험의 결과물인 "패키지"가 발송됐다. 대부분의 개선 사항은 작았고, 심지어 일부 월별 패키지는 공간 제약이나 법적 요건으로 인해 부정적으로 알려져 있었다.



(*) 숫자는 명백한 이유로 교란된다.

그림 1.4 시간 경과에 따른 빙 광고 수입
(y축은 연간 약 20% 성장률을 나타냄) 구체적인 숫자는 중요하지 않다.

사용자들의 구매 의향이 크게 상승하는 12월을 전후해 계절성이 급상승하는 것을 보면 광고 공간이 늘어나고, 1,000건 검색당 매출이 증가하는 것을 알 수 있다.

흥미로운 온라인 종합 대조 실험 사례

기대 결과와 실제 결과의 절대적 차이가 큰 실험은 흥미롭다. 만약 무슨 일이 일어날 거라고 생각했을 때 그런 일이 일어난다면, 그때는 새로 배울 점이 많지 않다. 만약 어떤 일이 일어날 것이라고 생각했지만 일어나지 않았다면, 여러분은 중요한 것을 배운 것이다. 그리고 사소한 일이 일어날 것이라고 생각했고 그 결과가 중대한 놀라움이고 돌파구로 이어진다면, 여러분은 매우 가치 있는 것을 배운 것이다.

이 장의 앞부분과 이 절의 빙의 예시들은 놀랍고 매우 긍정적인 결과가 얻어진 흔치 않은 성공 사례이다. 빙이 페이스북이나 트위터와 같은 소셜 네트워크와 통합하려는 시도는 강력한 결과를 결과가 기대됐으면서도 실현되지 못한 사례로, 2년 동안 많은 실험에서 아무런 가치도 보이지 않자 그 시도는 포기됐다.

지속적인 발전은 지속적인 실험과 많은 작은 개선과 관련된 문제지만, 빙 광고 사례에서 볼 수 있듯이, 여기 우리가 아이디어의 가치를 얼마나 형편없이 평가하는지를 강조하는 놀라운 몇 가지 예가 있다.

사용자 인터페이스 예: 41개 색조의 파란색

구글과 마이크로소프트가 모두 보여주듯이 작은 결정은 상당한 영향을 미칠 수 있다. 구글은 구글 검색 결과 페이지에 대해 41개 색조의 파란색을 시험했으며(Holson 2009), 당시 시각 디자이너들의 심기를 불편하게 만들었다. 그러나 구글의 색상 수정은 사용자 참여에 상당히 긍정적인 결과를 가져왔고 (구글은 각 개인의 변화의 결과를 보고하지 않는다는 점에 유의하라) 결국 디자인과 실험의 강력한 파트너십을 이끌어냈다. 마이크로소프트의 빙 색상 조정도 이와 유사하게 사용자가 작업을 완료하는 데 더 성공적이었고, 성공 시간이 향상됐으며, 미국에서의 매출이 연간 천만달러 이상으로 향상됐다(Kohavi et al. 2014, Kohavi et al. 2014, Kohavi, Thomke 2017).

이러한 변화들이 엄청난 결과를 유발하는 아주 작은 변화들의 좋은 사례지만 광범위한 색상이 테스트됐다는 점을 감안할 때, 추가 실험에서도 색상에 대해 실험한다면 더 뚜렷한 개선을 만들어낼 것 같지는 않다.

올바른 시점에 제안하기

2004년에 아마존은 신용카드 제안을 홈페이지에 올렸다. 그것은 매우 수익성이 좋았지만, 클릭률이 매우 낮았다. 연구 팀은 그림 1.5와 같이 아이템을 추가한 후 사용자가 볼 수 있는 쇼핑 카트 페이지로 제안을 이동하는 실험을 진행해 사용자가 받을 수 있는 절감액을 단순한 수학 공식으로 부각시켰다(Kohavi et al. 2014).

장바구니에 아이템을 추가한 사용자들은 명확한 구매 의도를 갖고 있기 때문에 이 제안은 적시에 표시된다. 대조 실험은 이러한 단순한 변화가 아마존의 연간 수입을 수천만 달러 증가시킬 수 있다는 것을 증명했다.

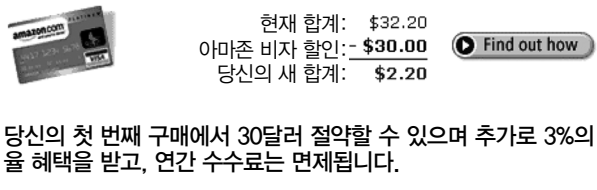


그림 1.5 아마존의 카트 총액에 대한 절감을 나타내는 신용카드 제안

개인화 추천

아마존의 그렉 린덴Greg Linden은 사용자의 쇼핑 카트에 있는 아이템을 기반으로 개인화된 추천을 보여주는 프로토타입을 만들었다(Linden 2006, Kohavi, Longbottom et al. 2009). 사용자가 어떤 항목을 추가하면 그에 따른 추천 사항이 나타나고 또 다른 항목을 추가하면 새로운 추천 사항이 나타난다. 그렉은 이 시제품이 유망해 보였지만 "마케팅 수석 부사장이 이 시제품이 사람들의 체크아웃을 방해할 것이라 주장하면서 이에 대해 완전히 반대했다"고 지적했다. 그렉은 더 이상 이 일을 진행시키는 것이 금지됐다. 그럼에도 그는 대조 실험을 했고, 새로운 기능의 성과는 너무 훌륭해서, 새로운 기능을 사용하지 않음으로써 아마존은 큰 비용을 초래하고 있음을 보여줬다. 결국 긴급히 쇼핑카트 아이템 기반 추천 기능이 도입됐다. 이제 아마존뿐 아니라 여러 서비스에서 쇼핑카트 아이템 기반 추천을 사용하고 있다.

속도는 "매우" 중요하다.

2012년 마이크로소프트 빙의 엔지니어가 자바스크립트 생성 방식을 변경해 클라이언트로 전송되는 HTML의 길이를 크게 줄여 성능을 향상시켰다. 종합 대조 실험은 놀랄 만큼 많은 개선된 지표를 보여주었다. 그들은 서버 성능에 미치는 영향을 추정하기 위해 후속 실험을 실시했다. 그 결과 성능 개선은 성공률과 성공 시간과 같은 주요 사용자 지표를 상당히 개선하며, 매 10밀리초의 성능 개선(눈 깜박임 속도의 1/30)은 엔지니어의 1년 연봉에 상응하는 수익을 초과창출했다(Kohavi et al. 2013).

2015년까지 빙의 성능이 향상됨에 따라 서버가 95번째 백분위에서(즉 쿼리의 95%에 대해) 1초 미만으로 결과를 반환할 때 성능 향상에 여전히 가치가 있는지에 대한 의문이 제기됐다. 빙의 팀은 후속 연구를 실시했고 주요 사용자 지표는 여전히 상당히 개선됐다. 수익에 대한 상대적인 영향은 다소 줄었지만, 빙의 수익은 그 기간 동안 너무 많이 향상돼 매 밀리초마다 개선된 성

능의 가치가 과거보다 더 높았다. 즉 매 4밀리초 감소마다 한 엔지니어를 고용할 수 있었다. 이 실험과 성능의 중요성에 대한 자세한 검토는 5장을 참조하라.

성능 실험은 여러 회사에서 수행됐으며, 성능이 얼마나 중요한지 알 수 있는 결과가 나왔다. 아마존에서는 100밀리초 느린 속도 실험으로 매출이 1% 감소했다(Linden 2006b, 10). 빙과 구글의 연사들이 공동으로 한 강연(Schurman, Brutlag 2009)에서 뚜렷한 쿼리, 수익, 클릭, 만족도, 클릭 시간 단축 등을 포함한 핵심 지표에 대한 성능의 현저한 영향을 보여주었다.

악성코드 감소

광고는 수익성이 좋은 사업이어서 사용자들이 설치한 '프리웨어'는 종종 광고로 페이지를 오염시키는 악성코드를 포함하고 있다. 그림 1.6은 악성코드malware를 가진 사용자에게 빙의 결과 페이지가 어떻게 나타나는지를 보여준다. 여러 개의 광고(빨간색 상자)가 페이지에 추가됐다는 점에 유의하라(Kohavi et al. 2014).

이로 인해 빙 광고가 제거됨으로써 마이크로소프트의 수입이 줄어 들었을 뿐만 아니라, 저품질 광고와 관련 없는 광고가 자주 게시돼 왜 그렇게 많은 광고를 보고 있는지 깨닫지 못했을 사용자들에게 나쁜 사용자 경험을 제공했다. 마이크로소프트는 380만 명의 사용자에게 종합 대조 실험을 실행했는데, 여기서 DOMDocument Object Mode 수정 기본 루틴은 신뢰할 수 있는 소스로부터의 제한된 수정만 허용되도록 재정의됐다(Kohavi et al. 2014). 그 결과 사용자당 세션을 포함해서 빙의 모든 핵심 지표가 개선돼 사용자들이 더 자주 방문하거나 혼란스러움을 덜 느끼는 것으로 나타났다.

게다가 사용자들은 검색에 더 자주 성공했고, 유용한 링크를 더 빨리 클릭했으며, 연간 수익은 수백만 달러 증가했다. 또한 이전에 논의한 주요 성능의 지표인 페이지 로드 시간이 영향을 받는 페이지에 대해 수백 밀리초 개선됐다.

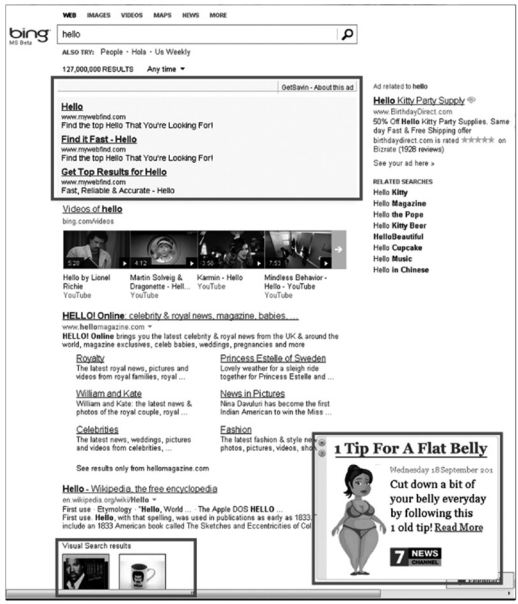


그림 1.6 사용자가 악성코드를 갖고 있는 경우 여러 광고를 표시하는 빙 페이지

백엔드 변화

백엔드 알고리즘 변화는 대조 실험을 사용하는 영역으로서 고려하지 않고 간과되는 경우가 많다(Kohavi, Longbottom et al. 2009). 그러나 이는 중요한 결과를 낼 수 있다. 앞에서 설명했듯이 구글, 링크드인, 마이크로소프트 팀이 많은 추가적인 소규모 변화에 대해 어떻게 작업하는지 아마존의 예로부터 알 수 있다.

2004년에 이미 두 가지 세트를 기반으로 하는 추천 알고리즘이 이미 존재했다. 아마존 추천의 대표 기능은 "X 품목을 산 사람은 Y 품목을 산다"였다. 그러나 이는 "X 품목을 본 사람은 Y 품목을 산다"와 "X 품목을 본 사람은 Y 품목을 본다"로 일반화됐다. "X를 검색한 사람은 Y 품목을 산다"에 대해 동일한 알고리즘을 사용할 것이 제안됐다. 알고리즘을 제안한 사람은 키퍼 서

덜랜드가 주연으로 나오는 TV 프로그램이 연관된 “24”와 같이 명확하게 지정되지 않은 검색의 예를 들었다. 24를 검색하면 기존의 아마존 검색은 24곡의 이태리 가곡 CD, 24개월 갓난아기 의류 및 24인치 타월 바와 같은 별로 좋지 못한 검색 결과를 보여주고 있었다(그림 1.7 좌측). 그러나 새로운 알고리즘은 아주 좋은 결과를 냈다(그림 1.7 우측). 아마존 검색은 “24”를 검색한 후 사람들이 실제로 구매한 품목을 기반으로 ‘24’ 드라마에 대한 DVD와 관련 책을 보여줬다. 이 알고리즘의 한 가지 약점은 검색 단계에서의 단어를 포함하지 않은 품목을 나타낸다는 것이다. 그러나 아마존은 종합 대조 실험을 실행했고, 이러한 약점에도 이 알고리즘은 아마존 전체 매출을 3% 증가시켰다. 이는 수억불에 해당하는 금액이다.

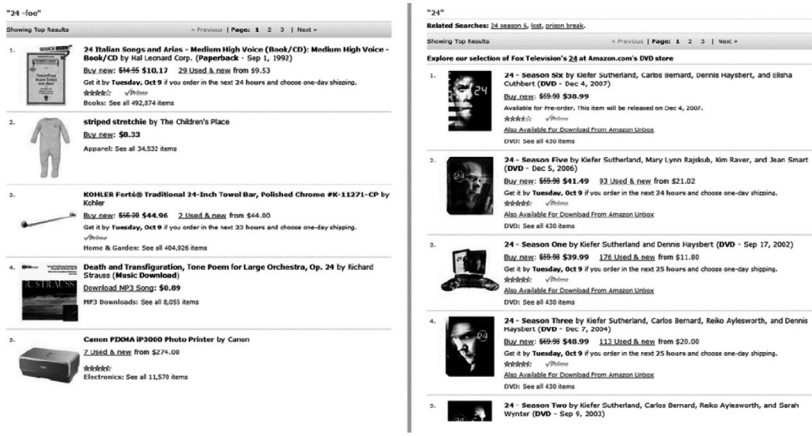


그림 1.7 BBS[2]가 있을 때와 없을 때의 ‘24’에 대한 아마존 검색

전략, 전술과 이들의 실험과의 관계

온라인 종합 대조 실험의 실행에 필요한 요소들이 충족되면 전략에서 전

2 Behavior Based Search의 약자로 사용자의 행동 이력 기반 검색 – 옮긴이

술에 이르기까지 모든 수준의 조직 결정에 정보를 주도록 실험이 수행돼야 한다. 전략(Porter 1996, 1998)과 종합 대조 실험은 시너지 효과를 낸다. 린 전략Lean Strategy의 데이빗 콜리스David Collis는 "효과적인 전략은 혁신과 실험이 이루어져야 하는 범위를 파악함으로써 기업가적 행동을 억제하기보다 이를 장려한다"고 기술한다(Collis 2016). 그는 경직된 계획과 제약 없는 실험의 양극단을 경계하는 린 전략 프로세스를 정의한다.

적절한 지표를 사용해 잘 실행된 실험은 비즈니스 전략, 제품 설계를 보완하고, 조직의 데이터 중심화를 통해 운영 효율성을 개선한다. 전략을 OEC에 요약하는 것으로 종합 대조 실험은 전략에 대한 훌륭한 피드백 루프를 제공할 수 있다. OEC를 개선하는 실험으로 아이디어를 평가했는가? 즉 실험의 놀라운 결과는 대안으로 실행할 수 있는 전략적 기회를 조명하게끔 만들 수 있으며, 그 방향으로 선회하게 만든다(Ries 2011). 제품의 설계 결정은 일관성coherency을 위해 중요하며 다중 설계 변형을 시도하는 것은 설계자들에게 유용한 피드백 루프를 제공한다. 그리고 전술적 변화는 운영 효율성을 향상시킬 수 있으며, 포터Porter에 의해 "경쟁사보다 유사한 활동을 더 잘 수행하는 것"으로 정의된다(Porter 1996).

이제 두 가지 주요 시나리오를 검토하겠다.

시나리오 1: 비즈니스 전략이 있고, 실험할 수 있는 충분한 사용자가 있는 상품이 있는 상황

이 시나리오에서 실험은 현재 전략과 제품에 기초해 국지적 최적화를 달성하는 데 도움이 될 수 있다.

- 실험은 ROI가 높은 분야, 즉 노력에 비해 OEC를 가장 많이 개선하는 분야를 식별하는 데 도움이 될 수 있다. MVP(최소 기능 제품)로 상이한 분야를 시도하면 큰 자원을 투입하기 전에 광범위한 분야를 더 빨리 탐색할 수 있다.
- 실험은 설계자에게 명백하지 않을 수 있지만 큰 차이를 낼 수 있는 최적화(예: 색상, 간격, 성능)에 도움이 될 수 있다.

- 실험은 팀이 사용자에게 초두 효과(primacy effect)[3](사용자는 이전 기능에 익숙하다. 즉, 이전 방식에 익숙하다)를 부여하는 완전한 사이트 재설계를 수행하도록 하는 것이 아니라 지속적으로 더 나은 사이트를 재설계하도록 돕는다. 완전한 재설계는 일반적으로 목표를 달성하는 데 실패할 뿐만 아니라 주요 지표에 대한 이전 사이트와의 동등성도 달성하지 못한다. (Goward 2015, slides 22 24, Rawat 2018, Wolf 2018, Laja 2019)
- 백엔드 알고리즘과 인프라 구조를 최적화하는 데 실험이 중요할 수 있다(예: 추천 알고리즘, 랭킹 알고리즘).

전략을 갖는 것은 실행 중인 실험에 매우 중요하다. 전략은 OEC의 선택을 주도하는 것이다. 일단 OEC가 정의되고 나면 종합 대조 실험은 팀이 OEC를 최적화하고 개선할 수 있도록 함으로써 혁신을 가속화하는 데 도움이 된다. 잘못 사용된 실험을 봤다면 그것은 OEC가 적절하게 선택되지 않은 것이다. 선택한 지표는 주요 기능을 충족해야 하며 지표만 좋게 나오도록 악용할 수 없어야 한다(7장 참조).

회사에는 어떻게 하면 실험을 제대로 실행할 수 있는가에 초점을 맞추는 팀도 있지만 지표 선택, 지표 검증 및 시간에 따른 지표 개선에 초점을 맞추는 팀도 있다. 지표 개선은 시간이 지남에 따라 진화하는 전략에 의해서 필요하며, 조작 가능하고도 개선이 필요한 CTR과 같은 전통적 지표에 한계가 발견했을 때에도 지표 개선이 필요하게 된다. 또한 성과 지표 팀은 실험이 일반적으로 짧은 기간 동안 실행되기 때문에 단기적으로 측정 가능한 지표를 결정하기 위해 노력한다. 하우저와 카츠(Hauser, Katz 1998)는 "현재 팀이 영향을 미칠 수 있지만 궁극적으로 해당 팀의 장기 목표에 영향을 미칠 지표를 파악해야 한다"고 기술한다(7장 참조).

이 전략을 OEC에 연결하면 전략 무결성strategy integrity(Sinofsky, Iansiti 2009)이란 개념이 생성된다. 이 개념을 말한 저자들은 "전략 무결성은 훌륭한 전략

3 먼저 제시된 정보가 나중에 들어온 정보보다 전반적인 인상 현상에 더욱 강력한 영향을 미치는 것을 의미하며, 변화 혐오 효과로도 알려져 있다. 최신 효과와 반대인 것으로 생각할 수 있다. - 옮긴이

을 짜거나 완벽한 조직을 갖추는 것이 아니다"라고 지적한다. 이해관계가 일치되고, 어떻게 수행하는지를 아는 조직에 의해 올바른 전략이 실행되는 것이다. 하향식 관점과 상향식 과제를 매칭하는 것이다." OEC는 전략을 명시적으로 만들고 기능을 전략과 일치시키는 완벽한 메커니즘이다.

궁극적으로 좋은 OEC가 없다면 자원을 낭비하고 있는 것이다. 예를 들어 침몰하는 유람선에서 음식이나 조명을 개선하기 위한 실험들을 생각해보라. 그러한 실험에 대한 OEC의 승객 안전 변수의 가중치는 매우 높아야 한다. 사실, 우리가 안전성을 떨어뜨릴 의사가 없을 정도로 안정성이 높아야 한다. 이는 OEC의 높은 가중치를 통해 또는 동등하게 승객 안전을 가드레일 지표guardrail metrics로 사용함으로써 반영할 수 있다(21장 참조). 소프트웨어에서 유람선 승객 안전과 유사한 것은 소프트웨어 충돌이다. 즉 기능이 제품의 충돌을 증가시키는 경우 그 경험은 너무 나쁜 것으로 간주되며, 다른 요소들은 이에 비교해서 전혀 중요하지 않다.

전략은 "트레이드오프인 무엇을 하지 않을지의 선택 역시 요구하기 때문에", 실험을 위한 가드레일 지표를 정의하는 것은 조직이 무엇을 변화시키려 하지 않는지를 식별하기 위해 중요하다(Porter 1996). 불운한 운명의 이스턴 항공기 401은 승무원이 불탄 착륙장치 표시등에 초점을 맞추다 실수로 자동 조종장치가 해제된 것을 알아차리지 못했기 때문에 추락했다. 즉 주요 가드레일 지표인 고도는 점차 낮아졌고, 1972년 플로리다 에버글레이즈에서 비행기가 추락해 101명의 사망자가 발생했다(위키디피아 기고자, Eastern Air Lines Flight 401, 2019).

포터Porter가 "Japanese Companies Rarely have Strategies"(1996) 섹션에서 바리안Varian은 Kaizen(2007)에 관한 기고문에서 언급한 바와 같이 운영 효율성 향상은 장기적으로 차별화된 이점을 제공할 수 있다.

시나리오 2: 제품과 전략을 갖고 있으나 결과는 방향 전환Pivot을 검토할 필요가 있다는 것을 제시하는 상황

시나리오 1에서 종합 대조 실험은 언덕 등반을 위한 훌륭한 도구다. OEC를 최적화하고 있는 "높이"로 하고, 아이디어의 다차원 공간을 생각한다면 당신은 정점을 향해 발걸음을 내딛고 있는 것일지도 모른다. 그러나 때때로 변화 속도에 대한 내부 데이터나 성장률 또는 기타 벤치마크에 대한 외부 데이터에 기초해 방향 전환pivot을 고려할 필요가 있다. 즉, 더 큰 언덕에 있을 수 있는 공간의 다른 위치로 점프하거나 전략과 OEC(따라서 지형의 모양)를 변경할 필요가 있다.

보통 항상 아이디어 포트폴리오를 만들 것을 권고한다. 즉 대부분은 현재 위치에 가까운 곳에서 최적화하는 시도에 대한 투자이어야 하지만, 그러한 점프가 더 큰 언덕으로 이어지는지를 보기 위해 몇 가지 급진적인 아이디어들이 시도돼야 한다. 우리의 경험에 의하면 대부분의 큰 점프는 실패하지만 (예: 대형 사이트 재설계) 위험/보상의 트레이드오프가 있다. 드물지만 성공이 다수의 실패을 상쇄하는 큰 보상을 제공하기도 한다.

급진적인 아이디어를 테스트할 때, 실험을 실행하고 평가하는 방법은 다소 바뀐다. 특히 다음 사항을 고려해야 한다.

- **실험 기간.** 예를 들어, 주요 UI 재설계를 테스트할 때 단기적으로 측정한 실험 변화는 초두 효과(primacy effect) 또는 변화 회피(change aversion)의 영향을 받을 수 있다. 실험군과 대조군의 직접적인 비교는 진정한 장기적 효과를 측정하지 못할 수 있다. 양면시장[4]에서, 변화가 충분히 크지 않으면 변화의 시험이 시장에 영향을 미치지 않을 수 있다. 좋은 비유는 매우 추운 방의 얼음 입방체인데, 실내 온도로 가는 작은 상승은 눈에 띄지 않을 수 있지만, 일단 녹는 점(예: 화씨 32도)을 넘어가면 얼음 입방체가 녹는다. 위의 구글 광고 품질 예제에 사용된 국가 수준의 실험과 같은 더 장기의 대규모 실험 또는

4 플랫폼을 통해 구매자와 판매자 간의 거래가 이뤄지는 시장을 '양면시장(two-sided market)'이라고 한다. 양면시장에서는 하나의 기업이 판매자와 구매자 간의 플랫폼 같은 연결고리 역할을 해서 거래가 이뤄지는 시장이다. 플랫폼 역할을 하는 기업에는 판매자와 구매자 모두가 고객이 된다. 이러한 역할을 하는 플랫폼의 예로, 앱스토어나 구글플레이, 삼성앱스 등의 앱 장터를 들 수 있다. - 옮긴이

대체 설계가 이러한 시나리오에서 필요할 수 있다(23장 참조).

- **테스트한 아이디어의 수.** 각 실험은 전체 전략의 구성원소인 특정 전술만 테스트하기 때문에 여러 가지 다른 실험이 필요할 수 있다. OEC를 개선하지 못하는 단일 실험은 특정 전술이 부실하기 때문일 수 있으며, 이는 전체 전략이 나쁘다는 것을 반드시 나타내는 것은 아니다. 실험은 설계에 의해 특정한 가설을 테스트하는 반면 전략은 훨씬 더 광범위하다. 즉, 종합 대조 실험은 전략을 수정하거나, 비효과적인 부분을 보여주고 방향 전환을 장려하는 것을 돕는다(Ries 2011). 종합 대조 실험을 통해 평가한 많은 전술이 실패한다면 '전략이 아무리 아름다워도 때로는 결과를 봐야 한다'는 윈스턴 처칠의 말을 생각해 볼 때가 올지도 모른다. 빙은 약 2년간 소셜미디어, 특히 페이스북과 트위터와 통합해 소셜 검색 결과가 나오는 제3의 창을 열겠다는 전략을 세웠다. 핵심 지표에 대해 눈에 띄는 영향을 미치지 않고 전략에 2,500만 달러 이상을 지출한 후 전략이 중단됐다(Kohavi, Thomke 2017). 큰 돈을 걸어둔 내기를 포기하기는 힘들겠지만, 경제 이론은 실패한 내기는 매몰 비용(sunk cost)이라는 것을 말해준다. 그리고 실험을 더 많이 할수록 쌓이는 사용가능한 데이터를 바탕으로 미래지향적인 결정을 내려야 한다.

에릭 리스Eric Ries는 완전히 실패한 것으로 판명된 계획을 성공적이고, 충실하게 그리고 엄격하게 집행하는 회사들에게 "달성된 실패"라는 용어를 사용한다. 대신 그는 다음과 같이 제안한다.

> 린 스타트업Lean Startup의 방법론은 스타트업의 노력을 어떤 부분이 훌륭하고 어떤 것이 말도 안되는지 그 전략을 테스트하는 실험으로 재인지하는 것이다. 진정한 실험은 과학적 방법을 따른다. 어떤 일이 일어날지 예측하는 명확한 가설에서 시작된다. 그러고 나서 실험은 그 예측들을 경험적으로 테스트한다.

전략을 평가하기 위해 실험을 실행해야 하는 시간과 도전 때문에 시노프스키와 이안시티(Sinofsky, Iansiti, 2009)는 다음과 같이 기술하고 있다.

> … 제품 개발 프로세스는 리스크와 불확실성으로 적재돼 있다. 이 두 개념은 매우 다르다. … 우리는 불확실성을 줄일 수 없다. 우리는 무엇을 모르는지 모른다.

하지만 우리는 이에 동의하지 않는다. 즉 종합 대조 실험을 실행할 수 있

는 능력은 당신이 최소 기능 제품MVP, Minimum Viable Product을 시도하고, 데이터를 얻고, 반복함으로써 불확실성을 줄일 수 있게 해준다(Ries 2011). 그렇기는 해도 모든 사람이 새로운 전략을 테스트하는 데 투자할 수 있는 시간으로 몇 년을 가질 수는 없고, 이러한 경우 불확실성하에서 결정을 내려야 할 수도 있다.

기억해야 할 유용한 개념 중 하나는 EVIExpected Value of Informaition이다. 즉 더글러스 허버드(Douglas Hubbard, 2014)가 제안한 정보의 기대값으로 이는 추가 정보가 의사결정에 어떻게 도움 될 수 있는지를 포착한다. 종합 대조 실험을 실행할 수 있는 능력은 최소 기능 제품(Ries 2011)의 시도, 데이터 수집 및 반복을 통해 불확실성을 현저하게 줄일 수 있다.

추가 참고문헌

온라인 실험 및 A/B 테스트와 직접 관련된 몇 권의 책이 있다(Siroker, Koomen 2013, Goward 2012, Schrage 2014, McFarland 2012, King et al. 2017). 대부분은 훌륭한 동기부여 스토리를 갖고 있지만 통계적으로는 부정확하다. 게오르기 게오르기예프Georgi Georgiev의 최근 저서에는 종합적인 통계적 설명이 포함돼 있다(Georgieves 2019).

종합 대조 실험과 관련된 문헌은 방대하다(Mason et al. 1989, Box et al. 2005, Keppel, Sauzey and Tokunaga 1992, Rossi, Lipsey, Freeman 2004, Imbens, Rubin 2015, Pearl 2009, Angist, Pischke 2014, Ger, Ger 2012).

웹상에서 종합 대조 실험을 실행하는 것에 관한 몇 가지 입문서가 있다(Peterson 2004, 76－78, Eisenberg 2005, 283－286, Chatham, Temkin, Amato 2004, Eisenberg 2005, Eisenberg 2004). (Peterson 2005, 248–253, Tyleryler, Redford 2006, 213–219, Sterne 2002, 116–119, Kaushik, Kaushik 2006).

멀티암드 밴딧multi-armed bandit은 실험이 진행됨에 따라 실험 트래픽 배분experiment traffic allocation이 동적으로 업데이트될 수 있는 실험의 한 유형이다(Li et

al. 2010, Scott 2010). 예를 들어, 매 시간마다 실험을 새롭게 관찰해 각 변형군이 어떤 성과를 냈는지를 볼 수 있고, 각 변형군이 받는 트래픽의 비율을 조정할 수 있다. 잘 하고 있다고 보이는 변형군은 더 많은 트래픽을 얻고, 성능이 떨어지는 변형군은 더 적게 얻는다.

멀티암드 밴딧에 기반한 실험은 대개 "일반적인" A/B 실험보다 더 효과적이다. 왜냐하면 그들은 실험 끝까지 기다리는 대신, 점차 승리하는 변형군 쪽으로 트래픽을 이동시키기 때문이다. 멀티암드 밴딧을 사용하는 것이 적절한 문제가 광범위하게 존재하지만(Bakshy, Balandal, Kashin 2019), 평가 목표는 단일 OEC가 돼야 하며(예: 여러 지표 간의 트레이드오프는 단순하게 공식화될 수 있다), OEC는 재할당 간에 합리적으로 잘 측정될 수 있어야 한다(예: 클릭율 대 세션)는 것이 주요 제한사항이다. 또한 사용자를 나쁜 변형군에 노출시키고, 다른 승자 변형군에 똑같지 않게 분산시킴으로써 잠재적인 편향이 생길 수 있다.

2018년 12월, 이 책의 공동저자들은 제1차 온라인 종합 대조 실험 콘퍼런스를 기획했다. 에어비앤비, 아마존, 부킹닷컴, 페이스북, 구글, 링크드인, 리프트, 마이크로소프트, 넷플릭스, 트위터, 우버, 얀덱스, 스탠퍼드대학교 등 13개 기관에서 총 34명의 전문가가 참가해 소집단회의breakout session에서 개요와 도전과제를 제시했다(Gupta et al. 2019). 도전과제에 관심이 있는 독자들에게는 이 자료가 도움이 될 것이다.