Hanbit eBook

Realtime 35

DATA SCIENCE

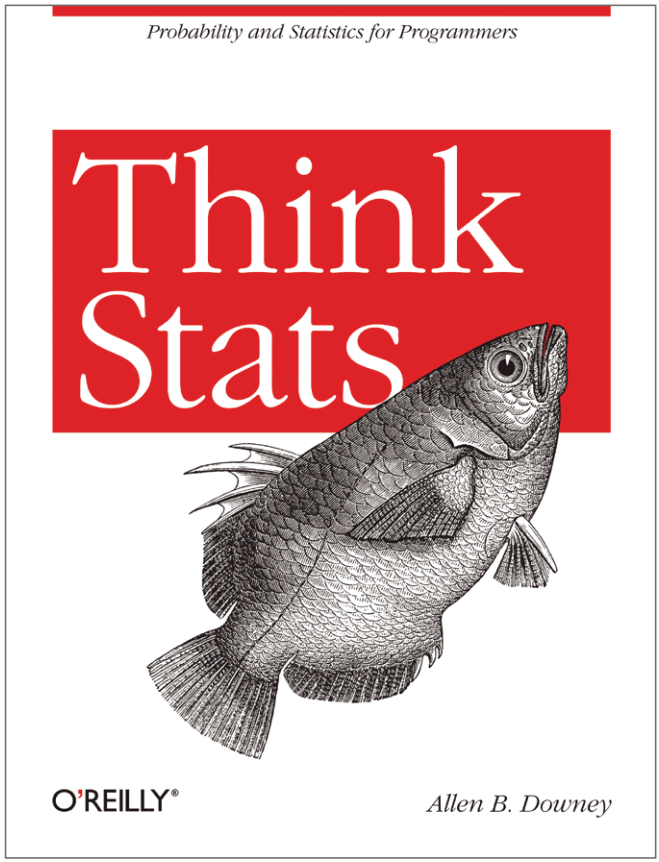
Think Stats

프로그래머를 위한 통계 및 데이터 분석 방법

Think Stats

앨런 B. 다우니 지음 / 김석우 옮김

O'REILLY® 한빛미디어 Hanbit Media, Inc.



이 도서는 O'REILLY의
Think Stats의
번역서입니다.

Data Science

Think Stats

프로그래머를 위한 통계 및 데이터 분석 방법

Data Science **Think Stats** 프로그래머를 위한 통계 및 데이터 분석 방법

초판발행 2013년 9월 3일

지은이 앨런 B. 다우니 / **옮긴이** 김석우 / **펴낸이** 김태헌
펴낸곳 한빛미디어(주) / **주소** 서울시 마포구 양화로 7길 83 한빛미디어(주) IT출판부
전화 02-325-5544 / **팩스** 02-336-7124
등록 1999년 6월 24일 제10-1779호
ISBN 978-89-6848-634-0 15000 / **정가** 9,900원

책임편집 배용석 / **기획** 이중민 / **편집** 김창수, 안선화
디자인 표지 여동일, 내지 스튜디오 [밈], 조판 북누리
마케팅 박상용, 박주훈

이 책에 대한 의견이나 오탈자 및 잘못된 내용에 대한 수정 정보는 한빛미디어(주)의 홈페이지나 아래 이메일로 알려주십시오.
한빛미디어 홈페이지 www.hanb.co.kr / **이메일** ask@hanb.co.kr

Published by HANBIT Media, Inc. Printed in Korea

지금 하지 않으면 할 수 없는 일이 있습니다.
책으로 펴내고 싶은 아이디어나 원고를 메일(ebookwriter@hanb.co.kr)로 보내주세요.
한빛미디어(주)는 여러분의 소중한 경험과 지식을 기다리고 있습니다.

저자 소개

지은이_ **앨런 B. 다우니**Allen B. Downey

MIT에서 학사와 석사 학위를 취득하였고, UC 버클리 대학U.C Berkeley에서 박사 학위를 받았다. 현재 올린 공과대학Olin College of Engineering 전산학과 부교수로 재직 중이며, 웰즐리 대학Wellesley College, 콜비 대학Colby College, UC 버클리 대학 전산학과에서도 강의를 하고 있다.

역자 소개

역자_ 김석우

데이터를 사랑하고 데이터 속에서 무엇을 발견할지 항상 고민하는 분석가 겸 개발자다. 학부 때는 수치 해석 및 수학적 최적화 방법에 매료되어 수학을 전공했고, 석사 때는 통계적 데이터 마이닝에 매료되어 통계학 석사를 취득하였다. 이후 Daum Communications 검색 본부 데이터 마이닝 팀을 거쳐 현재는 SK플래닛 데이터 기술 연구소의 Data Analytics 팀에서 근무하고 있다. Daum에 근무할 때부터 최근 화두가 되고 있는 빅데이터를 유용하게 분석하여 가치를 뽑아내는 것에 대해 고민해 왔으며, 단순히 분석뿐만 아니라 개발에 어떻게 활용할 수 있을지 연구하고 있다. 개발자들에게 분석 및 통계 이론을 전파하려고 노력하며, 분석가들에게는 분석을 위한 개발 방법과 최신 기술을 전파하기 위해 노력 중이다.

저자 서문

이 책을 집필하게 된 동기

이 책은 확률과 통계 입문 수업을 위한 새로운 종류의 교과서로, 크기가 큰 데이터 세트를 분석하는 데 통계를 어떻게 사용하는지에 초점을 맞췄다. 이 책은 또한 컴퓨터를 이용한 접근 방식을 취하는데, 여기에는 다음과 같은 장점이 있다.

- 프로그램을 작성해 봄으로써 학생들은 자신이 이해한 부분을 발전시켜 보고 검증해 볼 수 있다. 예를 들어, 최소제곱법least square fit, 잔차residuals, 그리고 결정계수coefficient of determination를 계산하는 함수를 작성할 수 있다. 코드를 작성하고 검증하기 위해서는 이와 관련된 개념을 이해해야 하며, 잘못 이해하고 있었던 부분은 무조건 바로 잡아야 한다.
- 학생들은 통계학적 거동statistical behavior을 검증하기 위해 실험을 해 볼 수 있다. 예를 들어, 몇몇 분포에서 샘플(표본)을 생성해 보면서 중심극한정리Central Limit Theorem, CLT를 탐구해 볼 수 있다. 파레토 분포에서 생성한 변수의 합이 정규로 수렴하지 않는 것을 보면서는 CLT의 기본 가정을 기억하게 된다.
- 시뮬레이션을 통해 수학적으로 이해하기 힘든 개념을 쉽게 이해할 수 있다. 예를 들어, 몬테카를로 시뮬레이션Monte Carlo simulation으로 p-value의 근사치를 계산함으로써 p-value의 의미를 더 잘 이해할 수 있다.
- 이산 분포와 컴퓨터를 이용한 계산을 통해 베이지안 추정Bayesian estimation 같은, 입문 수업에서 다루기 힘든 주제를 논할 수 있다. 예를 들어 '독일 탱크 문제German tank problem'[01]와 관련하여 학생들

01 (역자주) 유명한 통계학 기법의 응용 사례 중 하나다. 제2차 세계대전 당시 연합군은 독일의 탱크 제조 능력에 관한 정보가 부족하여 여러 가지의 정보 수집 방법을 동원했다. 전쟁 이전의 독일 기록을 참조하기도 하고, 독일 산업시설에 스파이를 침투시켜 관찰하기도 했다. 그러나 당시 독일 정부의 역정보와 선전 공작으로 인해 정확한 수치를 알아내기가 힘들었고, 또 전쟁 초기의 눈부신 독일의 전과에 매혹된 연합군의 정보장교들은 객관적인 정보에 대한 인식도 부족했다. 그때 이 문제에 도전한 것이 통계학자들이었다. 학자들은 파괴되거나 노획한 독일 탱크 부품의 일련번호들을 통계적으로 분석하여 탱크 제조 물량 정보를 알아냈고, 이런 분석 기법은 이후 V-2 로켓 등 다른 독일 무기들에도 적용되었다. 이 기법의 성공을 입증한 자료는 바로, 전쟁 후에 발견된 독일의 제조 기록이었다. 예를 들면, 연합군 정보장교들은 기존의 정보 수집 방법을 이용하여 독일이 1941년 6월 한 달 동안 탱크를 1,550대 제조했다고 주장했으나, 통계학자들은 244대라고 추산했다. 나중에 독일의 제조 기록과 비교해 본 결과 실제 수치는 271대였다. 통계학자들이 제임스 본드형 스파이들과 겨루어 대승을 거둔 셈이다. 이러한 통계적 분석은 후일 연합군의 독일 폭격 전략에 결정적인 수훈을 세웠다.

에게 사후 분포posterior distribution를 계산해 보라고 한다면? 이 문제는 수리적 또는 해석학적으로는 풀기 어렵지만 컴퓨터로 계산하면 놀라울 정도로 쉽게 답을 구할 수 있다.

- 학생들은 파이썬 같은 범용 프로그래밍 언어를 사용하기 때문에 어떤 종류의 데이터든 대부분 불러올 수 있다. 특정 통계 툴에 맞게 포맷 변경과 정제 작업을 거친 데이터도 아무 제약 없이 사용할 수 있다.

이 책은 프로젝트 중심으로 구성했다. 실제 강의에서 필자는 학생들에게 한 학기 동안 통계적 문제 해결 방식이 필요한 프로젝트를 주고, 그에 알맞은 데이터를 찾게 한다. 학생들은 여러 통계적 기술을 데이터에 직접 적용해 봄으로써 데이터에 대해 배울 수 있다.

이 같은 분석(방법)을 설명하기 위해 모든 장에 사례 연구를 실었으며, 다음 두 출처의 데이터를 사용했다.

- National Survey of Family Growth(NSFG)
  미국 질병통제예방센터Centers for Disease Control and Prevention, CDC에서 응답자의 가족 생활, 결혼, 이혼, 임신, 불임, 피임 여부, 남녀의 건강 상태에 대한 정보를 수집하고자 실행하는 조사다(http://cdc.gov/nchs/nsfg.htm 참조).
- Behavioral Risk Factor Surveillance System(BRFSS)
  미국 내 건강 조건과 위험 요소들을 추적하기 위해 '미국 만성질환예방 및 건강증진센터National Center for Chronic Disease Prevention and Health Promotion'에서 실시하는 조사다(http://cdc.gov/BRFSS/ 참조).

미국 국세청과 인구조사국, 마지막으로 보스톤 마라톤 대회의 데이터도 사용하였다.

이 책을 집필한 방법

새로운 교과서를 집필할 때, 사람들은 보통 기존 교과서를 읽고 참조한다. 그 결과, 교과서 대부분은 같은 내용 및 전개 방식을 가지고 있다. 그리고 종종 같은 구문과 오류 등이 여러 교과서에서 발견되기도 한다. 스티븐 제이 굴드Stephen Jay Gould도 에세이 『The case of the creeping fox terrier』에서 이를 지적했다. 필자는 이러한 방식에서 벗어나, 이 책을 집필하는 동안 인쇄물을 거의 보지 않았다.

- 필자의 목표는 새로운 방식으로 책을 집필하는 것이다. 기존 접근 방식에서 벗어나고 싶었다.
- 무료 라이센스로 이용할 수 있도록, 이 책의 어떠한 부분도 저작권에 의해 제한받길 원하지 않았다.
- 필자가 쓴 책의 많은 독자들은 도서관에 갈 수가 없기 때문에 인터넷에서 자유롭게 볼 수 있는 책을 만들고자 했다.
- 전통 미디어 지지자들은 전자적 형태로만 이용할 수 있다는 것에 대해, 성의가 부족해 보이고 신뢰할 수 없다고 생각한다. 성의가 부족해 보인다는 것은 맞는 말일 수도 있겠지만, 신뢰할 수 없다는 데 대해서는 그들의 말이 틀렸음을 증명해 보고 싶었다.

집필할 때 가장 많이 참조한 것이 위키피디아Wikipedia다. 필자는 위키피디아에서 정말 좋은 통계 관련 자료 및 주제들을 접했고, (수정을 조금 하긴 했지만) 이 책에서도 그 내용을 다루었다. 참조한 부분에 대해서는 위키피디아 참조 표시를 했으며, 이 책에서 사용한 단어나 표기법도 특별히 문제가 없는 한 위키피디아에 있는 것을 사용하였다.

위키피디아 이외에도 수학 관련 사이트인 'Wolfram MathWorld(http://MathWorld.Wolfram.com)'와 구글에서도 유용한 자료들을 얻을 수 있었다. 그 밖에, 데이비드 맥케이David Mackay의 저서 『Information Theory, Inference and Learning Algorithms』에서 베이지안 통계에 관한 정보를 얻었고, 『Numerical Recipes in C』에서도 역시 많은 정보를 얻을 수 있었다. 두 책 모두 온라인 버전이 있어서, 책의 내용을 참고하는 것에 대해서는 마음이 불편하지 않았다.

역자 서문

몇 년 전부터, 여러 곳에서 미래의 유망 직종으로 '데이터 사이언티스트'라는 신종 직업을 언급하고 있다. 데이터 사이언티스트란 분석과 개발이 명확히 구분되던 과거와 달리, 데이터 분석과 개발을 동시에 할 수 있는 고급 인력을 뜻한다. 특히 요즘처럼 빅데이터가 주목받는 시대에는 빅데이터를 자유자재로 다루는 것은 물론, 그 데이터를 분석하여 데이터 속의 숨은 가치를 찾아 주는 데이터 사이언티스트의 수요가 늘어날 수밖에 없다. 하지만 기존 분석가들에게는 개발이라는 장벽이, 기존 개발자들에게는 통계학이라는 장벽이 존재한다. 역자는 개발 회사 및 연구소에 근무하면서 통계학이라는 장벽을 넘지 못하여 좌절하는 개발자들을 심심찮게 보아 왔다.

이 책은 이런 개발자들에게 데이터를 분석하는 데 필요한 통계적 이론을 개발자의 관점에서 쉽게 설명해 준다. 과거, 통계학 입문 서적은 복잡한 수식을 통한 이론 설명에 집중하여 개발자가 쉽게 접근할 수 없었다. 이에 반해, 이 책은 복잡한 수식을 배제하고 파이썬 코드를 이용해 개발자적인 관점에서 이론 부분을 설명해 줌으로써, 개발자들도 복잡한 통계 이론을 쉽게 이해할 수 있도록 돕는다. 물론 통계 전문 용어를 그대로 사용하기 때문에, 초반에는 개발자들이 조금 어렵고 낯설게 느낄 수도 있다. 하지만 매 장 부록마다 해당 장에서 다룬 통계 전문 용어를 설명해 주고 있어서, 큰 문제가 되지는 않을 것이라 확신한다.

데이터를 탐구하는 일은 참으로 아름답고 매력적인 일이다. 개발자들이 이 책을 통해 자신의 개발 능력을 분석에 활용함으로써, 과거에는 다룰 수 없었던 빅데이터에서, 통계적 분석을 통해 데이터의 가치를 발견하는 진정한 데이터 사이언티스트가 되기를 기대한다.

끝으로 이 책의 출판을 결정하고 기회를 준 김창수 팀장님과 이중민 대리님, 번역을 허락해 주고 응원해 준 Data Analytics 팀의 김수진 팀장님, 데이터에서 가치를 발견하기 위해 항상 애쓰는 우리 팀원 분들에게 고마움을 전한다. 사랑하는 아내와 어머니, 아버지께도 감사의 마음을 전한다.

2013년 5월 김석우

예제 파일

이 책에서 사용된 예제 파일은 http://greenteapress.com/thinkstats/thinkstats.code.zip에서 받을 수 있습니다.

한빛 eBook 리얼타임

한빛 eBook 리얼타임은 IT 개발자를 위한 eBook입니다.

요즘 IT 업계에는 하루가 멀다 하고 수많은 기술이 나타나고 사라져 갑니다. 인터넷을 아무리 뒤져도 조금이나마 정리된 정보를 찾는 것도 쉽지 않습니다. 또한 잘 정리되어 책으로 나오기까지는 오랜 시간이 걸립니다. 어떻게 하면 조금이라도 더 유용한 정보를 빠르게 얻을 수 있을까요? 어떻게 하면 남보다 조금 더 빨리 경험하고 습득한 지식을 공유하고 발전시켜 나갈 수 있을까요? 세상에는 수많은 종이책이 있습니다. 그리고 그 종이책을 그대로 옮긴 전자책도 많습니다. 전자책에는 전자책에 적합한 콘텐츠와 전자책의 특성을 살린 형식이 있다고 생각합니다.

한빛이 지금 생각하고 추구하는, 개발자를 위한 리얼타임 전자책은 이렇습니다.

1. **eBook Only: 빠르게 변화하는 IT 기술에 대해 핵심적인 정보를 신속하게 제공합니다.**

   500페이지 가까운 분량의 잘 정리된 도서(종이책)가 아니라, 핵심적인 내용을 빠르게 전달하기 위해 조금은 거칠지만 100페이지 내외의 전자책 전용으로 개발한 서비스입니다. 독자에게는 새로운 정보를 빨리 얻을 수 있는 기회가 되고, 자신이 먼저 경험한 지식과 정보를 책으로 펴내고 싶지만 너무 바빠서 엄두를 못 내는 선배, 전문가, 고수분에게는 보다 쉽게 집필하실 기회가 되리라 생각합니다. 또한 새로운 정보와 지식을 빠르게 전달하기 위해 O'Reilly의 전자책 번역 서비스도 하고 있습니다.

2. **무료로 업데이트되는, 전자책 전용 서비스입니다.**

   종이책으로는 기술의 변화 속도를 따라잡기가 쉽지 않습니다. 책이 일정한 분량 이상으로 집필되고 정리되어 나오는 동안 기술은 이미 변해 있습니다. 전자책으로 출간된 이후에도 버전 업을 통해 중요한 기술적 변화가 있거나, 저자(역자)와 독자가 소통하면서 보완되고 발전된 노하우가 정리되면 구매하신 분께 무료로 업데이트해 드립니다.

3. 독자의 편의를 위하여, DRM-Free로 제공합니다.

구매한 전자책을 다양한 IT기기에서 자유롭게 활용하실 수 있도록 DRM-Free PDF 포맷으로 제공합니다. 이는 독자 여러분과 한빛이 생각하고 추구하는 전자책을 만들어 나가기 위해, 독자 여러분이 언제 어디서 어떤 기기를 사용하더라도 편리하게 전자책을 볼 수 있도록 하기 위함입니다.

4. 전자책 환경을 고려한 최적의 형태와 디자인에 담고자 노력했습니다.

종이책을 그대로 옮겨 놓아 가독성이 떨어지고 읽기 힘든 전자책이 아니라, 전자책의 환경에 가능한 최적화하여 쾌적한 경험을 드리고자 합니다. 링크 등의 기능을 적극적으로 이용할 수 있음은 물론이고 글자 크기나 행간, 여백 등을 전자책에 가장 최적화된 형태로 새롭게 디자인하였습니다.

앞으로도 독자 여러분의 충고에 귀 기울이며 지속해서 발전시켜 나가도록 하겠습니다.

차례

1 | 프로그래머를 위한 통계적 사고

이 책에서는 데이터를 지식으로 바꿔 주는 연금술에 대해 논할 것이다. 상대적이긴 하지만 데이터는 얻기 쉽다. 하지만 지식은 그에 비해 얻기가 어렵다.

우선 아래 세 가지에 대해 설명해 보겠다.

확률Probability

무작위 사상(임의 사건)을 표현하는 방법에 대한 연구다. 사람들은 대부분 '아마도 ~할 것 같다'나 '아마도 ~할 것 같지 않다'라는 말을 특별한 훈련 없이도 사용하는 것처럼, 확률의 개념에 대해서도 직관적으로 이해하고 있다. 하지만 우리는 앞으로 이 개념을 어떻게 '수치적'으로 표현할 수 있는지에 대해 논할 것이다.

통계Statistics

통계는 모집단에 관한 주장을 뒷받침하기 위해 표본 데이터를 사용하는 분야다. 통계적 분석의 대부분은 확률에 기초하고 있기 때문에, 확률과 통계는 보통 같이 쓰인다.

계산Computation

정량적 분석에 적합한 도구다. 통계를 처리하는 데 컴퓨터가 유용하게 사용되는 것처럼, 계산(전산) 실험은 확률이나 통계의 개념을 이해하는 데 유용하다.

필자가 말하고자 하는 이 책의 요지는, 프로그래밍을 할 줄 안다면 여러분은 이 능력을 확률과 통계를 이해하는 데 사용할 수 있다는 것이다. 확률과 통계는 수학적 관점에서 설명되는 것이 일반적인데, 이러한 접근 방식은 일부 사람에게는 잘 맞는다. 그러나 이 분야의 중요한 일부 개념은 수학적 접근보다는 컴퓨터를 사용하는 계산적 접근이 상대적으로 이해하기 쉽다.

먼저, 1장에서는 아내와 내가 첫아이를 가지려고 할 때 들었던 질문인 '첫아이는 예정일보다 늦게 태어나는 경향이 있는가?'에 대한 사례 연구를 논하겠다.

1.1 첫아이는 예정일보다 늦게 태어날까?

구글에서 이 질문을 검색해 보면, 이와 관련하여 많은 논의가 이루어지고 있다는 걸 알 수 있다. 몇몇 사람들은 사실이라고 주장하고, 또 어떤 사람들은 미신에 지나지 않는다고 주장한다. 그리고 또 다른 사람들은 오히려 첫아이일수록 빨리 출산한다며 반대되는 의견을 내기도 한다.

많은 사람들이 제각각 자기의 의견을 증명하기 위해 아래와 같이 데이터를 제시하고 있다.

> "최근에 첫아이를 출산한 친구들 두 명 다 예정일보다 2주나 지난 후에 자연 분만과 유도 분만으로 출산했다."
>
> "첫아이가 예정일보다 2주 늦게 태어났으니, 둘째 아이는 예정일보다 2주 빨리 태어날 거다."
>
> "우리 누나는 첫째지만 예정일보다 빨리 태어났고 내 사촌들도 마찬가지였으므로, 사실이 아니라고 생각한다."

공개되지 않고, 일반적으로 개인적 데이터에 바탕을 둔 이러한 보고를 가리켜 '일화적 증거anecdotal evidence'라고 한다. 일상적인 대화에서는 이처럼 개개인의 경험을 바탕으로 데이터를 제공하는 것이 잘못되었다고는 할 수 없으므로, 위에서 언급한 사람들의 말 역시 틀렸다고 볼 수 없다.

하지만 우리는 일화적 증거보다는 좀 더 설득력 있고, 신뢰할 수 있는 증거를 원한다. 왜냐하면, 자신 또는 타인의 경험을 바탕으로 한 일화적 증거에는 보통 아래와 같은 문제점이 있기 때문이다.

적은 관측 수 Small number of observations

첫아이의 임신 기간이 길 경우, 자연변이에 의한 초과 임신 기간과의 차이는 그다지 나지 않을 것이다. 이 경우, 확실한 차이가 존재하는지 확인하기 위해서는 많은 수의 임신 기간을 비교할 필요가 있다.

선택 편향 Selection bias

이 토론에 참여한 사람들은 그들의 첫아이를 예정일보다 늦게 출산했기 때문에 관심을 가졌을 것이다. 이런 경우, 데이터를 선택하는 과정이 편향되었다고 볼 수 있다.

확증 편향 Confirmation bias

논제를 믿는 사람들은 논제와 일치하는 예시를 드는 경향이 있고, 논제에 대해 의심하는 사람들은 반대되는 예시를 드는 경향이 있다.

부정확성Inaccuracy

일화적 증거들은 대부분 개인적인 경험에 바탕을 둔 증거들이다. 그러므로 이 경험들은 잘못 전해지거나 틀리게 기억될 수도 있고, 따라서 증거들은 부정확할 수 있다.

그렇다면 어떤 식으로 접근해야 더 잘할 수 있을까?

1.2 통계적 접근

일화적 증거의 한계점을 해결하기 위해 다음의 통계 도구를 사용할 것이다.

데이터 수집 Data collection

미국 인구에 대해 통계적으로 유효한 추론 생성을 목표로 설계된 대규모 전국 조사 데이터를 사용할 것이다.

기술 통계 Descriptive statistics

간결하게 데이터를 요약하는 통계값들을 산출하고, 데이터를 시각화하는 여러 방법들을 평가할 것이다.

탐색적 자료 분석 Exploratory data analysis

논제를 해결하기 위해 데이터의 패턴, 차이, 특징 등을 찾아낼 것이다. 이와 동시에, 데이터에 한계나 모순점 등이 있는지 확인해 볼 것이다.

가설 검정 Hypothesis testing

두 그룹이 차이가 날 경우에는 그 차이가 어떤 사건에 의해 발생한 명백한 효과인지, 아니면 단순히 우연히 발생한 것인지 평가해 볼 것이다.

추정Estimation

샘플 데이터를 사용하여 전체 모집단의 특징을 추정해 볼 것이다.

실수하지 않도록 주의하면서 이 같은 단계를 수행하면, 타당하고 정확한 결론에 도달할 가능성이 높아진다.

1.3 전미 가족 성장 조사

1973년부터 미국 질병통제예방센터U.S. Centers for Disease Control and Prevention, CDC는 결혼, 이혼, 임신, 불임, 피임 여부, 남녀의 건강 상태 같은 가족생활 관련 정보를 얻기 위해 '전미 가족 성장 조사National Survey of Family Growth, NSFG(이하 NSFG)'를 실시해 왔다.[01]

01 http://cdc.gov/nchs/nsfg.htm

우리는 이 조사에서 얻은 데이터를 사용하여 '첫아이는 예정일보다 늦게 태어나는가' 여부를 조사할 것이다. 데이터를 효과적으로 사용하기 위해서는, 우선 이 조사가 어떤 식으로 설계되었는지 올바르게 이해해야 한다.

NSFG는 '횡단면 연구cross-sectional study'에 속한다. 횡단면 연구란, 연구 대상에 관한 자료 측정을 어느 시점(동일한 시점)에서 시행하는 경우를 말한다. 이와 반대로 대상을 반복적으로 측정하는 경우는 '종단면 연구longitudinal study'라고 한다.

NSFG는 총 일곱 번 시행되었다. 시행된 각각의 조사를 사이클cycle이라 하는데, 이 책에서 우리는 여섯 번째 사이클(2002년 1월 ~ 2003년 3월에 시행)의 데이터를 사용할 것이다.

이 조사의 목적은 모집단population(미국 인구 전체)에서 여러 결론을 끄집어내는 것이며, 표적 집단target population은 15세에서 44세의 미국인이다.

조사에 참여한 사람들을 응답자respondents라고 하며 응답자 그룹을 코호트cohort라 한다. 일반적으로 횡단면 연구의 데이터는 대표성을 띤다. 표적 집단에 속한 모든 사람들이 동일한 기회를 가지고 연구에 참여한다고 전제하기 때문이다. 물론 이 가정은 이상적인 것일 뿐, 현실적으로는 가정을 만족시키기가 쉽지 않다. 그러나 조사를 주관하는 사람은 가능한 한, 최대한 가정을 만족시키려고 한다.

이에 반해 NSFG의 데이터는 대표성을 띤다고 할 수 없으며, 오히려 오버샘플되어 있다. NSFG 설계자는 히스패닉, 흑인, 10대 등 세 종류의 실험 대상 집단을 모집했는데, 이 집단의 비율은 미국 전체 인구에서 각 집단이 차지하는 비율보다 높았다. NSFG 설계자가 이처럼 오버샘플된 데이터를 사용한 이유는, 통계적 추론을 올바로 하기 위해 충분히 큰 표본 집단이 필요했기 때문이었다.

물론 이렇게 오버샘플된 데이터에서 도출한 통계값을 바탕으로 얻은 결론을 일반화하여 적용하는 것은 그리 쉽지 않다. 이 부분에 대해서는 나중에 다시 논의해 보도록 하겠다.

Exercise 1-1

NSFG는 총 일곱 번 시행되었지만 종단면 연구longitunital study라고는 할 수 없다. 위키피디아의 문서(http://wikipedia.org/wiki/Cross-sectional_study, http://wikipedia.org/wiki/Longitudinal_study)를 확인해 보면 그 이유를 알 수 있을 것이다.

Exercise 1-2

이번 연습 문제를 풀기 위해서는 NSFG 데이터를 내려받아야 하는데, 이는 이 책 전반에서 사용할 데이터다.

1. 이 책의 웹페이지 http://thinkstats.com/nsfg.html를 방문하여 내용을 확인한 후 'I accept these terms'을 클릭한다.
2. 2002FemResp.dat.gz 파일과 2002FemPreg.dat.gz 파일을 내려받는다. 2002FemResp.dat.gz 파일에는 여성 응답자 7,643명의 정보가 한 줄로 처리되어 있으며, 2002FemPreg.dat.gz 파일에는 임신 관련 데이터가 들어 있다.
3. 조사에 관한 온라인 문서가 궁금하다면 웹사이트 http://www.icpsr.umich.edu/nsfg6를 방문해 보기 바란다. 왼쪽 세로 영역의 각 섹션들을 클릭해 보면 어떤 데이터가 포함되었는지 볼 수 있다. 조사에 관련된 내용은 http://cdc.gov/nchs/data/nsfg/nsfg_2002_questionnaires.htm에서 찾아볼 수 있다.
4. 이 책의 웹페이지에서는 NSFG 데이터를 처리하는 코드를 제공한다. http://thinkstats.com/survey.py 파일을 내려받은 다음 이 파일을 데이터 파일이 위치하는 디렉토리에서 실행하면, 각 데이터 파일의 라인 개수를 아래와 같이 반환해 준다.

   ```
   Number of respondents 7643
   Number of pregnancies 13593
   ```

5. 파일의 각 코드들이 무엇을 의미하는지는 다음 섹션에서 다루도록 하겠다.

1.4 테이블과 레코드

시인이자 철학자인 스티브 마틴Steve Martin은 다음과 같이 말했다.

> "Oeuf는 계란을 의미하고, chapeau는 모자를 의미한다. 이와 같이 프랑스어는 모든 것에 대응하는 다른 단어를 가지고 있다."(프랑스어는 영어와 다른 단어를 가지고 있음을 의미)

프랑스에서 영어와 다른 단어를 사용하듯, 데이터베이스 프로그래머는 통계 용어와는 다른 용어를 사용한다. 또한 실무에서 통계를 사용할 때는, 데이터베이스 사용을 피할 수 없으므로 데이터베이스 프로그래머가 사용하는 용어를 알고 있어야 한다.

응답자 파일(2002FemResp.dat.gz, 2002FemPreg.dat.gz)의 각 라인은 응답자 한 명의 정보를 포함하고 있다. 이 정보를 '레코드record'라고 한다. 또한 DB에서는 이 레코드를 이루는 변수variable를 '필드field', 레코드들의 컬렉션을 '테이블table'이라고 한다.

파이썬 파일인 'survey.py'을 보면, 'Record' 클래스가 정의된 것을 볼 수 있는데 이는 레코드를 나타내는 객체이며, 'Table' 클래스는 테이블을 나타내는 객체다.

Record 클래스는 'Respondent'와 'Pregnancy'라는 두 개의 하위 클래스를 가지고 있는데, 이 두 하위 클래스는 각각 응답자respondent와 임신pregnancy 테이블의 레코드를 가지고 있다. 이 클래스들은 잠시 비어 있을 것이다. 자세히 말하자면, 이 두 클래스의 속성을 초기화시키는 init 메소드가 없기 때문이다. 대신에 'Table.MakeRecord'를 가지고 텍스트 라인을 Record 객체로 변환시킬 것이다.

Table 클래스 또한 'Respondents'와 'Preganacies' 두 개의 하위 클래스를 가지고 있다. 각 클래스의 init 메소드는 데이터 파일의 초기 이름과 생성할 레코드의 형식(타입)을 지정해 준다. Table 객체는 Record 객체들의 리스트인 'records'라는 속성attirbute을 가지고 있다.

각 Table에 대해서 'GetFileds' 메소드는 각 Record 객체에 속성으로 저장될 레코드의 필드를 지정하는 튜플tuple의 리스트를 반환한다.

Pregnancies.GetFields를 예로 들어 보자.

```
def GetFields(self):
  return [
      ('caseid', 1, 12, int),
      ('prglength', 275, 276, int),
      ('outcome', 277, 277, int),
      ('birthord', 278, 279, int),
      ('finalwgt', 423, 440, float),
     ]
```

첫 번째 튜플은 'caseid' 필드가 1행column부터 12행에 걸쳐 있으며 정수형integer임을 말해 준다. 각 튜플은 다음의 정보를 가지고 있다.

field

필드 값들이 저장될 속성명이다. 대부분 NSFG codebook에 있는 이름을 사용하였고 소문자로만 이루어져 있다.

start

필드의 첫 번째 행을 나타내는 인덱스 값이다. 예를 들어 caseid의 첫 번째 행의 인덱스 값은 '1'이다. 이런 인덱스 값들에 관해서는 NSFG codebook에서 찾아 볼 수 있다(http://www.icpsr.umich.edu/webdocs/).

end

필드의 마지막 행을 나타내는 인덱스 값이다. 예를 들어, caseid의 마지막 행의 인덱스 값은 '12'이다.

Conversion function

문자열을 적절한 다른 형태로 변환하는 함수다. int와 float 함수를 사용할 수 있고, 사용자가 직접 정의할 수도 있다. 만약 변환에 실패하면 문자열의 값은 'NA'가 된다. 필드(문자열)를 변환하고 싶지 않다면(변환 함수conversion function를 사용하고 싶지 않다면), str 함수를 사용하거나 identity 함수를 사용하면 된다.

임신pregnancy 레코드에 관한 내용은 다음과 같다.

caseid

정수로 표현된 응답자의 ID다.

prglength

정수로 표현된 임신 기간(주 단위)이다.

outcome

정수로 표현된 임신 후 출산 여부다. 정상 출산은 '1'이다.

birthord

정상 출산된 아이의 순서를 정수로 표현한 변수다. 예를 들어, 첫아이의 경우 '1'이다. 유산된 경우에는 '0'으로 나타낸다.

finalwgt

응답자와 관계된 통계적 가중치다. 실수형으로 이루어진 변수이며 응답자가 대표하는 미국 인구의 수를 의미한다.

Codebook을 자세히 읽다 보면, 변수는 대부분 실제 조사의 원시 데이터를 사용하지 않고 재조정된 것(이를 리코드recode라 한다)을 알 수 있는데, 이 재조정된 값들도 원시 데이터를 바탕으로 계산되었다.

예를 들어, 정상 출산한 아이의 prelength 변수 값은 원시 데이터의 wksgest(주 단위의 임신 기간) 값과 동일하다. 하지만 원시 데이터에 wksgest 값이 존재하지 않으면, '4.33 * mostgest'(월 단위의 임신 기간 * 한 달의 평균 주의 수)로 계산하여 prelength 변수 값을 얻는다.

리코드recode는 일관성 있고 정확한 데이터에 기반을 두어야 한다. 일반적으로, 원시 데이터를 재조정하지 말아야 할 큰 이유가 없다면, 리코드를 사용하는 것은 좋은 방법이다.

Pregnancies도 Recode 메소드를 가지고 있다는 것을 확인할 수 있다.

Exercise 1-3

이번 연습 문제에서는 Pregnancies 테이블 데이터를 탐색하는 프로그램을 작성해 볼 것이다.

1. Survey.py 파일과 데이터 파일이 존재하는 디렉터리에, 'first.py'라는 파일을 생성하고 아래 코드를 복사해 보자.

```
first.py and type or paste in the following code:
import survey
table = survey.Pregnancies()
table.ReadRecords()
print 'Number of pregnancies', len(table.records)
```

결과값은 'Number of pregnancies 13,593'이다.

2. Table을 반복하여 돌면서, 정상 출산된 아이가 얼마나 되는지 계산할 수 있는 loop 함수를 작성해 보자. 출산 관련 문서에서 실제 정상 출산된 아이의 수를 확인해 보고, 위에서 계산한 결과가 실제 데이터와 일치하는지 비교해 보라.

3. 정상 출산한 아이들의 레코드를 첫아이 그룹과 그 외 아이 그룹 두 가지로 나누어서 아이 수를 확인할 수 있도록, loop 함수를 수정하라. 그런 다음, 다시 출산 관련 문서를 이용해 위에서 계산한 값이 맞는지 확인해 보라.
   새로운 데이터로 작업할 때, 위와 같은 방식을 통해 에러 및 버그를 찾고 실제 데이터와 일치하는지 확인해 볼 수 있다.

4. 첫아이를 임신했을 때의 평균 임신 기간(주 단위)과, 그 외 아이를 임신했을 때 임신 주기의 평균(주 단위)을 계산해 보라. 이 두 값에 차이가 존재하는가? 존재한다면, 그 차이가 얼마나 큰가?

위 연습 문제들의 해답은 http://thinkstats.com/first.py에서 내려받을 수 있다.

1.5 유의성

우리는 앞선 연습 문제에서 첫아이와 그 외 아이의 임신 기간을 비교해 보았다. 그 결과, 첫아이들이 평균 13시간 정도 더 늦게 태어난다는 것을 확인했다.

이 차이를 '겉보기 효과apparent effect'라고 한다. 겉보기 효과란 두 그룹의 차이가 있는 것처럼 보이지만, 확신할 수는 없는 경우를 말한다. 이와 관련하여 아래와 같은 의문이 생길 것이다.

- 두 그룹의 평균값이 다르다면, 중앙값이나 분산 같은 다른 요약 통계값들에도 차이가 있는가? 두 그룹의 차이가 얼마 정도인지 좀 더 명확히 알 수 있는 방법은 없는가?
- 비교한 두 그룹이 실제로는 같은데, 이러한 차이는 우연히 발생할 수 있는가? 만약 그렇다면, 이 차이는 통계적으로 유의하지 않다고 결론 내릴 수 있다.
- 겉보기 효과가 선택 편향이나 잘못된 실험 설계에 의해 발생할 수 있는가? 만약 그렇다면, 겉보기 효과는 실험자에 의해 (의도치 않게) 생긴 인위적인 것이라고 할 수 있다.

이 질문의 답은 이 책의 나머지 부분에서 다룰 것이다.

Exercise 1-4

통계를 배우는 가장 좋은 방법은 관심 있는 분야의 프로젝트를 진행해 보는 것이다. 앞에서 언급한 '첫아이가 더 늦게 태어나는가'와 같은, 조사해 보거나 연구해 보고 싶은 주제가 있는가?

개인적으로 관심 있는 주제나 사회적 통념과 논쟁, 또는 정치적 이슈 등에 관한 질문을 던져 보고, 통계적 조사가 가능하도록 그 질문을 좀 더 명확하게 만들 수 있는지 알아보자.

질문을 제기하기 위해서는 먼저 데이터를 찾아야 한다. 공공 연구에 관련된 자료는 정부 관련 사이트에 가면 찾을 수 있다. 좋은 품질의 다른 데이터를 찾고 싶다면, Wolfram Alpha(http://wolframalpha.com) 웹사이트를 방문해 보라. 다만, Wolfram Alpha의 데이터에는 저작권이 있으니, 저작권과 관련된 조건 사항들을 먼저 확인하도록 한다.

구글이나 다른 검색 사이트에서 데이터를 직접 찾아볼 수도 있지만, 그러한 데이터들의 품질은 보장할 수 없다.

조사해 보고 싶었던 주제에 대한 분석 결과를 인터넷에서 찾을 수 있다면, 조사 및 분석이 올바르게 진행되었는지를 먼저 확인해 보라. 만약 데이터나 분석 방법에 문제가 있다면 다른 분석 방법을 사용하여 조사해 보기 바란다.

조사해 보고 싶은 주제가 이미 논문으로 발표되었다면, 조사와 관련된 원시 데이터를 구할 수 있을 것이다. 논문의 저자가 사용한 데이터는 보통 저자의 홈페이지에 함께 공개하기 때문이다. 단, 공개하기 민감한 데이터에 한해서는 공개하지 않는 경우도 있다. 이 경우에는 저자에게 데이터를 직접 부탁해야 한다. 데이터를 부탁할 때는 그것을 사용하는 목적과 계획을 분명하게 밝히기 바란다.

1.6 용어 정리

겉보기 효과apparent effect

의미 있거나 흥미 있는 결과값을 보여 주는 측정 또는 요약 통계값.

대표성을 띠는representative

모집단의 모든 값들의 확률이 표본의 확률과 같은 경우, '표본이 대표성을 띤다'고 한다.

레코드record

데이터베이스에서의 사람 한 명, 또는 연구에서 대상의 정보들.

리코드recode

원시 데이터에 로직이나 계산 등을 거쳐 변형된 값.

모집단population

연구 대상이 되는 집단. 보통은 인간 집단을 대상으로 많이 사용하지만, 동식물이나 미네랄 등도 사용한다.[02]

오버샘플링oversampling

샘플 크기가 작아서 나타나는 오류를 피하기 위해, 부분 모집단을 조정해 주는 방법.

요약 통계summary statistic

데이터의 특징을 숫자 하나로 함축하여 나타내는 계산 값.

원천 데이터raw data

가공(검증, 계산, 해석)되지 않은 채로 수집되고 기록된 값들.

응답자respondent

설문 조사에 대답하는 사람.

02 설명이 잘 이해되지 않는다면 'http://wikipedia.org/wiki/Twenty_Questions'를 참고하라.

인위 결과artifact

측정 오차 또는 다른 오차들로 인해 일어나는 겉보기 효과.

일화적 증거anecdotal evidence

정식적인 방법으로 수집된 증거가 아닌, 개인적인 경험을 바탕으로 구한 증거.

종단면 연구(경시적 자료 연구, 종단 연구)longitudinal study

같은 그룹에서 시간에 따라 반복적으로 데이터를 수집한 모집단을 분석하는 연구.

코호트Cohort

응답자들 그룹.

테이블table

데이터베이스에서 레코드 값들의 모음.

통계적 유의성statistically significant

겉보기 효과가 우연히 발생하기 힘든 경우, '통계적으로 유의미하다'고 한다.

표본(샘플)sample

수집된 데이터의 모집단의 부분 집합.

필드field

데이터베이스에서 레코드를 구성하는 이름을 가진 변수들.

횡단면 연구cross-sectional study

특정 시간의 모집단에서 추출된 데이터에 관한 연구.